Marketing Mutual Funds[*]

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

November 30, 2017

Abstract

Marketing and distribution expenses constitute a large fraction of the cost of active management in the mutual fund industry. We investigate their impact on the allocation of capital to funds and on returns earned by mutual fund investors. We develop and estimate a structural model of costly investor search and fund competition with learning about fund skill and endogenous marketing expenditures. We find that marketing is nearly as important as performance and fees for determining fund size. Restricting the amount that funds can spend on marketing substantially improves investor welfare, as more capital is invested with passive index funds and price competition decreases fees on actively managed funds. Average alpha increases as active fund size is reduced, and the relationship between fund size and fund manager skill net of fees is closer to that implied by a frictionless model. Decreasing investor search costs would also imply a reduction in marketing expenses and management fees as well as a shift towards passive investing.

Keywords: mutual funds, distribution costs, broker commissions, performance evaluation, capital misallocation, investor welfare, financial regulation, structural estimation, search costs, information frictions, household finance
JEL codes: G11, G28, D14, M31

[*] We benefited from comments and suggestions by Eduardo Azevedo, Joao Gomes, Rich Evans, Juhani Linnainmaa, Gregor Matvos, David Musto, Stijn van Nieuwerburgh, Kent Smetters, Luke Taylor, and Jessica Wachter, as well as audiences at ESSFM Gerzensee, LBS, LSE, UT Dallas, and Wharton.

[†] The Wharton School, University of Pennsylvania and NBER

[‡] The Wharton School, University of Pennsylvania

[§] Marshall School of Business, University of Southern California

1 Introduction

In 2016, active mutual funds in the U.S. managed a total of 11.6 trillion dollars. This industry's annual revenue is on the order of $100 billion, with over one third of this amount representing expenditures on marketing, largely consisting of sales loads and distribution fees (known as 12b-1 fees). Although the relation between mutual fund size, performance, and fees has been actively debated in the academic literature (e.g., Berk and Green 2004, Chen et al. 2004, Pástor and Stambaugh 2012, Berk and van Binsbergen 2015, Pástor, Stambaugh and Taylor 2015), the role of marketing and distribution expenditures in steering investors into particular funds is not fully understood. Positive relationship between funds' marketing efforts and flows is well-documented (e.g., Sirri and Tufano 1998, Barber, Odean, and Zhang 2005, Gallaher, Kaniel, and Starks 2006, Bergstresser, Chalmers, and Tufano 2009, Christoffersen, Evans, and Musto 2013). Yet marketing expenses contribute substantially to total fund costs, thus reducing returns earned by investors for a given level of fund manager's skill. Is marketing a purely wasteful rat race, or does it help imperfectly informed investors find attractive investment opportunities more easily? Does it enable capital to flow towards more skilled managers or, instead, distort allocation of assets by channeling them towards underperforming funds?

We start with a benchmark model based on Berk and Green (2004), which describes allocation of assets to mutual funds by rational investors in a frictionless market. By estimating the model, we document substantial differences between such an efficient allocation and the observed distribution of fund size. The vast majority of funds are "too big" relative to the model and deliver substantially negative abnormal returns to investors, while the top decile of funds are actually smaller than is "efficient," and thus are able to outperform.[1] To explain these differences, we introduce information frictions by generalizing the framework developed by Hortaçsu and Syverson (2004) in the context of index funds. We allow funds' marketing activities, as well as exogenous characteristics, to affect their inclusion in the investors' information sets. In our setting, both the expense ratios (fees paid by investors) and the marketing/distribution costs (components of these fees related to broker compensation) are endogenous choices of each fund. By estimating the search model, we find that marketing expenses are nearly as important as price (i.e., expense ratio) or performance (i.e., the Bayesian estimate of manager skill based on historical returns) for explaining the observed variation in fund size.

We follow Hortaçsu and Syverson (2004) and model the impediments to investor's ability to allocate capital optimally to mutual funds as a search friction, whereby investors randomly sample and evaluate funds until deciding to invest in one of the funds drawn. This approach is intuitive at least when applied to retail investors: the task of choosing among thousands of funds can be daunting even for the most sophisticated individuals, and far more so for those lacking even basic financial literacy. Investors care about the fund's performance and the expense ratio charged by the fund. Mutual fund performance is determined by managerial skill as well as decreasing returns to scale. Hence, our model nests Berk and Green (2004) as a special case when search costs go to zero. Our key innovation is allowing mutual funds to influence the

[1] The bulk of the empirical evidence of performance persistence among mutual funds indicates that consistent *underperformance* is much more prevalent than outperformance - e.g. Hendricks, Patel and Zeckhauser (1993), Brown and Goetzmann (1995), Carhart (1997).

likelihood of being sampled by expending resources on marketing (e.g., via broker commissions). Thus, mutual funds choose their expense ratios and marketing expenses, which increase a fund's probability of being sampled but decrease its profit margin.

We estimate our structural model using data on well-diversified U.S. domestic equity mutual funds, which we think are representative of the industry as a whole. Our estimation results reveal sizable information frictions in the mutual fund market. The average investor implicitly incurs a cost of 39 basis points to "sample" an additional mutual fund. This friction's magnitude is about 2/3 of the mean annual gross alpha in our sample. The large magnitude of the estimated search cost is a manifestation of the asset misallocation problem that we document. The intuition is simple: high search costs prevent investors from sampling more funds. Less intensive search leads to an inferior allocation, as many investors are forced to "settle" for high-cost, low-skill funds since it is too costly for them to continue searching. In comparison, Hortaçsu and Syverson (2004) find the mean search cost for an average S&P 500 index fund investor is between 11 to 20 basis points. This should not be surprising since it is far easier for investors to evaluate index funds (which are essentially identical in terms of the returns they deliver, at least before fees) than actively managed funds. It is also possible that the index fund market is dominated by more sophisticated investors (i.e., those who know to look for them rather than, say, rely on a recommendation of a broker or financial advisor). Our higher estimated search cost indicates that asset misallocation problem is more severe in the mutual fund industry as a whole (including both active funds and passive index funds) than it is within the S&P 500 index funds sector.

Our estimates imply that marketing is relatively useful as a means of increasing fund size. On average, a one basis point increase in marketing expenses leads to a 1% increase in a fund's size. This effect is heterogeneous across funds. For high-skill funds, it amounts to a 1.15% increase in assets under management, while for low-skill funds it generates only a 0.97% increase. This result is intuitive: since, conditional on being included in an investor's information set, a high-skill fund is more likely to be chosen, such funds benefit more from a higher probability of being sampled than low-skill funds. We find that marketing expenses alone can explain 10% of the variation in mutual fund size; this explanatory power is comparable to both fund manager skill and fund price.

We use our model to quantify the importance of marketing expenses and search costs in shaping the equilibrium distribution of fund size as well as its impact on investor welfare via counterfactual experiments. First, we explore the consequences of tightening the regulatory constraint on marketing. We simulate the impact of preventing funds from doing any marketing by solving for the equilibrium size distribution and funds' fees choices using the estimated model parameters. We find that if the cap on marketing is set to zero, the mean expense ratio drops from 160 bps in the current equilibrium to 83 bps. Interestingly, funds lower their expense ratios by more than the original amount of marketing costs. The observed average marketing cost is 62 bps, but in the no-marketing equilibrium the average fund price drops by 77 bps. This indicates that restricting funds from competing on non-price attributes (such as marketing) could significantly intensify price competition. We also find the total share of active funds drops from 74% to 68%. This drop is accompanied by an increase in average fund performance as measured by mean gross alpha. The increase in alpha is due to the effect of decreasing returns to scale on fund performance. In the no-marketing equilibrium, the "index fund" takes up the market share lost by active funds.

Total investor welfare increases by 57% in the counterfactual equilibrium. Three factors contribute to this increase: in the no-marketing equilibrium, (i) active funds are cheaper, (ii) more investors invest in the passive index fund, (iii) active funds' alpha is on average higher due to their smaller sizes. In order to further understand the large increase in investor welfare, we examine the cross-section of investor search costs implied by our model. Naturally, high search cost investors search less and pay higher expense ratios than those with low search costs, while the funds they invest in have high marketing fees and lower alphas. Comparing investor welfare in the two equilibria, we show that the bulk of the welfare gain of eliminating marketing is driven by such high search cost investors. The intuition is simple: these are the investors who are "stuck" with the worst funds (unless they happen to be lucky to "find" the index fund or a high-skill active fund). In the no-marketing equilibrium, even the worst funds are much cheaper than in the current equilibrium. This leads to a significant welfare gain for the high search cost investors.[2]

In addition, we examine the impact of search costs on equilibrium market outcomes. With the advancement in information technology and development of services enabling more transparent comparison between funds, we would expect the search frictions to decrease over time. We conduct counterfactual experiments where we set the mean search cost to 35 bps and 20 bps respectively. Given a new search cost distribution, funds reoptimize their prices and marketing expenses. We find that as mean search costs decreases from 39 bps to 35 bps, mean marketing expenses drop from 61 bps to 44 bps. But when mean search cost further drops to 20 bps, the equilibrium marketing expenses fall to zero, even though we maintain the regulatory cap at 100 bps. Thus, low search costs render marketing unprofitable. In the model with a high mean search cost, a subset of funds specifically exploit the high search cost investors. Those funds invest aggressively in marketing so as to enter more of the high search cost investors' choice sets. Since high search cost investors will not search much, they will invest with those funds. But when mean search cost drops to a sufficiently low level, this strategy is no longer effective. This suggests that our model's mechanism is consistent with the observed decline in fees charged by active mutual funds along with the growth in passive index funds over the last two decades highlighted by Stambaugh (2014).

There is a growing literature examining the role of financial advice. Bergstresser, Chalmers and Tufano (2009) study broker-sold and direct-sold funds and find little tangible benefit of the former to fund investors. Del Guercio and Reuter (2014) show that the relationship between fund flows and past performance is muted among funds that are sold through brokers, presumably because such funds are targeting investors with higher search costs. Chalmers and Reuter (2012) show that broker recommendations steer retirement savers towards higher fees funds resulting in lower investor returns; Mullainathan, Noeth and Schoar (2012) provide similar evidence from an audit study of retail financial advisors. Christoffersen, Evans, and Musto (2013) find that the

[2] A potential caveat is that a drastic reduction in marketing expenses could reduce access to financial advice, especially for small investors. If the role of financial advisors is in establishing investors' trust, as argued by Gennaioli, Shleifer, and Vishny (2015), then investor welfare could be reduced, as would their allocation to the mutual fund sector and, potentially, the equity markets generally.

broker incentives impact investor flow to funds, especially for brokers not affiliated with the fund family. Egan, Matvos, and Seru (2016) exhibit the potentially severe conflict of interest between brokers/financial advisors and their retail investor clients, as exemplified by repeat incidence of misconduct in the industry (only about 5 percent of reported misconduct involves mutual funds, however). An alternative to the conflict of interest view is presented by Linnainmaa, Melzer and Previtero (2016), who show that financial advisors tend to commit common investment mistakes in their own portfolios.

More closely related to our work, Hastings, Hortaçsu and Syverson (2016) study the impact of sales force on observed market outcomes in the Mexico privatized retirement savings systems. In their model, a fund's sales force can both increase investors' awareness of the product and impact their price sensitivity. In our data we cannot distinguish between these two effects. We thus assume that the marketing expenses are purely informative (rather than persuasive). Egan (2017) uses a search-based structural framework similar to ours to study the conflict of interest between brokers and retail investors in the market for structured convertible bonds.

Our paper is also related to the literature that aims to understand the observed underperformance of active funds. Pástor and Stambaugh (2012) develop a tractable model of the active management industry. They explain the popularity of the active funds despite their poor past performance using two components: decreasing returns to scale and slow learning about the true skill level. In our model of the active management industry, we also include decreasing returns to scale and investor learning about unobserved skill (at the fund level). However, our model largely attributes the popularity of active funds to the information friction that prevents investors from easily finding out about index funds.[3]

This paper is related to those studying the role of advertising and media attention in the mutual fund industry. Gallaher, Kaniel and Starks (2006), Reuter and Zitzewitz (2006), and Kaniel and Parham (2016) study the impact of fund family-level advertising expenditures and the resulting media prominence of the funds on fund flows. In our model, we capture some of these effects parsimoniously by allowing fund family size to impact fund's probability of being included in investor's information set.[4]

The remainder of the paper is organized as follows. Section 2 develops our model. Section 3 describes the data used to estimate the model. Section 4 discusses the estimation methods. Section 5 presents the estimation results. Section 6 conducts the counterfactual analysis. Section 7 concludes the paper.

2 Model

Every period, heterogeneous investors conduct costly search to sample mutual funds to invest their (identical) endowments. Investors care about expected fund performance and expense ratio

[3] Huang, Wei, and Yan (2007) argue that differences in mutual fund prominence as well as heterogeneity in the degree of sophistication across investors help explain the observed asymmetry in the response of flows to fund performance. Garleanu and Pedersen (2016) incorporate search costs in their model of active management and market equilibrium, but assume that a passive index is freely available to all investors without the need to search.

[4] We follow this simple approach to incorporating advertising since the latter constitutes a very small fraction of fund expenditure, compared to the distribution costs that we focus on. Advertising can be potentially quite important for steering consumers into financial products - e.g., Honka, Hortaçsu and Vitorino (2016) and Gurun, Matvos and Seru (2016).

(i.e., its price). Mutual fund performance is determined by managerial skill as well as the impact of decreasing returns to scale. Mutual funds choose their expense ratios and marketing expenses to maximize profits. Marketing expenditures can increase a fund's probability of being sampled but decrease its profit margins.

We proceed by first describing how fund's performance is determined and then the investor's problem and lastly describe the funds' behavior.

2.1 Fund performance

In a time period t, the realized alpha $r_{j,t}$ for an active fund $j \in \{1, 2, ..., N\}$ 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 $D(M_t s_{j,t}; \eta)$ where M_t is the total size of the market and $s_{j,t}$ is the market share of the fund j, and $M_t s_{j,t}$ denoting fund size, η is a parameter measuring the degree of decreasing returns to scale, and (iii) an idiosyncratic shock $\varepsilon_{j,t} \sim \mathcal{N}(0, \delta^2)$.

$$r_{j,t} = a_{j,t} - D(M_t s_{j,t}; \eta) + \varepsilon_{j,t}, \quad j = 1, ..., N, \tag{1}$$

An important question in the mutual fund literature concerns the relative size of active funds vis-a-vis passive funds (e.g., Pástor and Stambaugh 2012). To be able to address this important extensive margin, we include a single index fund $j = 0$ into our model, and thus abstract from competition *between* index funds. We assume that the alpha of the index fund is zero, in that it neither has skill nor affected by the decreasing returns to scale. The total market size M_t includes both active funds and the index fund. We treat M_t as an exogenous variable in the model.

Our specification is very similar to Berk and Green (2004) with one exception: the manager's skill is allowed to vary over time. 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 $v_{j,t} \sim \mathcal{N}(0, \kappa^2)$. When a fund is born, its first period skill will be drawn from the stationary distribution $\mathcal{N}(\mu, \kappa^2)$. Parameter ρ captures the persistence of the skill level. As with other parameters, its value will be estimated from data. In the limiting case, when $\rho = 1$, skill is fixed over time, which is what Berk and Green (2004) assume.

Following Berk and Green, we assume the manager's skill is not observable to either the investor or fund manager herself: it is treated as a hidden state. Let $\widetilde{a}_{j,t}$ be investor's belief about the manager's skill in that period. Since equation (2) can be regarded as describing how the hidden state $a_{j,t}$ evolves over time, and equation (1) says that $r_{j,t} + D(M_t s_{j,t}; \eta)$ is a signal on the hidden state, one can apply Kalman filter to obtain the following recursive formulas for the belief on manager's skill and the variance of that belief:

$$\begin{aligned}\widetilde{a}_{j,t} &\equiv \boldsymbol{E}\left(a_{j,t}|r_{j,t-1}, s_{j,t-1}, r_{j,t-2}, s_{j,t-2}, ...\right) \\ &= \rho\left\{\widetilde{a}_{j,t-1} + \frac{\widetilde{\sigma}^2_{j,t-1}}{\widetilde{\sigma}^2_{j,t-1} + \delta^2}\left[r_{j,t-1} + D(M_{t-1}s_{j,t-1};\eta) - \widetilde{a}_{j,t-1}\right]\right\} + (1-\rho)\mu, \qquad (3) \\ \widetilde{\sigma}^2_{j,t} &\equiv \boldsymbol{Var}\left(a_{j,t}|r_{j,t-1}, s_{j,t-1}, r_{j,t-2}, s_{j,t-2}, ...\right) \\ &= \rho^2\left(1 - \frac{\widetilde{\sigma}^2_{j,t-1}}{\widetilde{\sigma}^2_{j,t-1} + \delta^2}\right)\widetilde{\sigma}^2_{j,t-1} + (1-\rho^2)\kappa^2. \qquad (4)\end{aligned}$$

and $\widetilde{a}_{j,t} = \mu$, $\widetilde{\sigma}^2_{j,t} = \kappa^2$ for the period t when j was born. When ρ is close to 1, these formulas reduce to what Berk and Green (2004) derived in their Proposition 1. The difference between our updating rule and theirs is that in Berk and Green model, all the historical signals receive the same weight in determining the investor's belief, whereas in our case, when ρ is smaller than 1, the signals in the more recent periods receive larger weights. This allows us to capture the fact that fund managers and/or their strategies change over time, and that investors might therefore *rationally* overweight the recent history.[5]

2.2 Investor search

Each investor allocates a unit of capital to a single mutual fund identified as a result of sequential search (conducted at the beginning of each period t). Investors are short lived, in the sense that they derive utility from their investment in the fund of their choice, and the capital they invest in the funds dissipates at the end of the period. A new population of investors enters in the subsequent period $t+1$ with new capital endowments that they allocate to the funds, and so on. Let $p_{j,t}$ be the expense ratio charged by fund j. An investor's utility derived from investing in fund j is given by

$$u_{j,t} = \gamma\widetilde{r}_{j,t} - p_{j,t}, \qquad (5)$$

where

$$\widetilde{r}_{j,t} = \widetilde{a}_{j,t} - \eta\log(M_t s_{j,t}).$$

Recall that $\widetilde{a}_{j,t}$ is the investors' belief on the manager's skill for fund j for this period t and $\widetilde{r}_{j,t}$ is the fund j's expected alpha in period t implied by these updated beliefs as well as the size of the fund, given the decreasing returns to scale function parameterized as $D(M_t s_{j,t};\eta) = \eta\log(M_t s_{j,t})$. The coefficient in front of the expense ratio is normalized to 1. If $\gamma = 1$ then investors simply care about the expected outperformance net of of fees (net alpha), as assumed by Berk and Green (2004). We allow the more general formulation to account for the potential difference in salience of fees vs. performance as well as the investors' imperfect ability to estimate manager skill. The utility derived from investing in the index fund is given by $u_{0,t} = -p_{0,t}$, where $p_{0,t}$ is the expense ratio charged by the index fund in period t; the alpha of the index fund is set to be zero.

[5] There is evidence that investors "chase" recent performance, potentially more actively than would be justified from a purely Bayesian perspective. Our framework could be used in quantitatively assessing the degree to which this behavior is driven by irrational extrapolation - e.g. Bailey, Kumar and Ng (2011), Greenwood and Shleifer (2014)

Fix a time period t. Investor i pays search cost c_i to sample one fund from the distribution of funds. The search costs in the population follow a continuous distribution G (which we will parameterize as the exponential distribution with mean λ.). As in Hortaçsu and Syverson (2004), we endow investors with one free search, so that every investor will invest in a fund (even if his search cost is very high). Let $\Psi_t(u)$ be the probability of sampling a fund that delivers the investor an indirect utility smaller or equal to u. Standard Bellman equation arguments imply that it is optimal for the investor to follow a cutoff strategy (see Appendix for details). Let u^* be the highest indirect utility among the funds sampled thus far. The investor continues searching iff $u^* \leq \bar{u}(c_i)$, where the threshold $\bar{u}$ is defined by

$$c_i = \int_{\bar{u}}^{+\infty} (u' - \bar{u}) d\Psi_t(u').$$

Since we have a finite number of funds, the above expression becomes

$$c_i = \sum_{j=0}^{N} \psi_{j,t}(u_j - \bar{u}) \cdot \mathbf{1}\{u_j > \bar{u}\},$$

where $\psi_{j,t}$ is the probability of sampling fund $j \in \{0, 1, ..., N\}$. Intuitively, the left hand side is the cost for an additional search, and the right hand side is the expected gain. Note that the right hand side is strictly decreasing in $\bar{u}$. So $\bar{u}(c_i)$ is *strictly* decreasing in c_i. Intuitively, the bigger c_i is, the smaller the cut-off $\bar{u}(c_i)$ becomes, and the less persistent the investor is in searching. Following Hortaçsu and Syverson (2004), we can solve for the market share of each fund, $s_{j,t}$, explicitly as a function of the utilities $\{u_{j,t}\}_{j=0}^{N}$, sampling probabilities $\{\psi_{j,t}\}_{j=1}^{N}$ and the distribution of search costs $G(c_i)$ (see detailed derivations in the Appendix).

2.3 Marketing and equilibrium market shares

Fund sampling probabilities depend on fund characteristics and, crucially, on funds' marketing efforts. Let $b_{j,t}$ denote marketing expenses paid by fund j, $\boldsymbol{x}_{j,t}$ denote a vector collecting the (observable) exogenous characteristics of the fund, and $\xi_{j,t}$ represent the unobservable shock that affects the sampling probability of this fund. Vector $\boldsymbol{x}_{j,t}$ includes year dummies, fund age, and the number of funds in the same family. Then the probability that an investor randomly draws fund j in year t is specified as

$$\psi_{j,t} = \frac{e^{\theta b_{j,t} + \boldsymbol{\beta}' \boldsymbol{x}_{j,t} + \xi_{j,t}}}{1 + \sum_{k=1}^{N} e^{\theta b_{k,t} + \boldsymbol{\beta}' \boldsymbol{x}_{k,t} + \xi_{k,t}}}, \tag{6}$$

$$\psi_{0,t} = 1 - \sum\nolimits_{k=1}^{N} \psi_{k,t}. \tag{7}$$

Thus, θ is a key parameter that characterizes the effectiveness of marketing expenditure as a means of attracting investors. As long as θ is positive, an increase in $b_{j,t}$ increases the probability that the fund is sampled by investors, all else equal. We assume that the index fund does not engage in any marketing activities; thus, increasing marketing by all of the active funds automatically reduces its sampling probability.

We use vector $\boldsymbol{p}_t$ to denote the vector that collects $p_{j,t}$ for $j = 1, ..., N$; the similar notation applies to other fund-specific variables in the model. With the specifications in (5), (6), and (7), the search model in Section 2.2 implies a mapping from $\boldsymbol{p}_t$, $\boldsymbol{b}_t$, $\widetilde{\boldsymbol{r}}_t$, $\boldsymbol{x}_t$, $\boldsymbol{\xi}_t$, and $p_{0,t}$ to a set of market shares. Let us write this mapping as

$$s_{j,t} = F_{j,t}\left(\boldsymbol{p}_t, \boldsymbol{b}_t, \widetilde{\boldsymbol{r}}_t, \boldsymbol{x}_t, \boldsymbol{\xi}_t, p_{0,t}; \Theta\right), \quad j = 1, ..., N, \tag{8}$$

where Θ collects the relevant parameters, which in this case include γ, $\boldsymbol{\beta}$, θ, and the parameter λ for G. The share for the index fund is given by $s_{0,t} = 1 - \sum_{j=1}^{N} s_{j,t}$. We use vector $\boldsymbol{s}_t$ to collect $s_{j,t}$ for $j = 1, ..., N$.

Decreasing returns to scale imply that $\widetilde{\boldsymbol{r}}_t$ depends on the funds' market shares:

$$\boldsymbol{s}_t = \boldsymbol{F}_t\left[\boldsymbol{p}_t, \boldsymbol{b}_t, \widetilde{\boldsymbol{a}}_t - \eta \log\left(M_t \boldsymbol{s}_t\right), \boldsymbol{x}_t, \boldsymbol{\xi}_t, p_{0,t}; \Theta\right]. \tag{9}$$

As in Berk and Green (2004), investors understand that their returns depend on the size of the fund they invest in, and therefore the equilibrium vector of fund market shares $\boldsymbol{s}_t$ is a fixed point of the above relation. We can write the fixed point as a function of the other inputs on the right hand side of (9),

$$s_{j,t} = H_{j,t}(\boldsymbol{p}_t, \boldsymbol{b}_t, \tilde{\boldsymbol{a}}_t, \boldsymbol{x}_t, \boldsymbol{\xi}_t, p_{0,t}; \Theta), \quad j = 1, ..., N, \tag{10}$$

with Θ now also including parameter η. In the appendix, we show that this fixed point is unique. Unlike $F_{j,t}$, we do not have a closed-form expression for $H_{j,t}$ and so it requires fixed point iteration to compute.

Profits for an active fund j in period t are given by

$$\pi_{j,t} := M_t \cdot H_{j,t}(\boldsymbol{p}_t, \boldsymbol{b}_t, \tilde{\boldsymbol{a}}_t, \boldsymbol{x}_t, \boldsymbol{\xi}_t, p_{0,t}; \Theta) \cdot (p_{j,t} - b_{j,t}). \tag{11}$$

We assume a Nash equilibrium, where each fund chooses $p_{j,t}$ and $b_{j,t}$ to maximize $\pi_{j,t}$, given other funds' choices $p_{-j,t}$ and $b_{-j,t}$ (as well as its own and other funds' estimated skills and exogenous characteristics).[6] Because the SEC currently imposes a one-percent upper bound on the 12b-1 fees, we restrict $b_{j,t} \leq \bar{b} \equiv 0.01$ in equilibrium.

In sum, in our model mutual funds choose their fees and marketing efforts to maximize profits each period while taking into the equilibrium distribution of fund size (and therefore expected outperformance of each fund).

3 Data

The data come from CRSP and Morningstar. Our sample contains 2,285 well-diversified actively managed domestic equity mutual funds from the United States between 1964 and 2015. Our dataset has 27,621 fund/year observations. In the data appendix, we provide the details about how we construct our sample. We closely follow data-cleaning procedures in Berk and van Binsbergen (2015) and Pastor, Stambaugh and Taylor (2015).

To compute the annual realized alpha $r_{j,t}$, we start with monthly return data. We first

[6] Our notion of profits most closely approximates management fees paid to the fund's investment advisor. We can think of this either as profits accruing to the fund family or as compensation paid to the fund manager, although in reality the latter is a much more complicated object, e.g., see Ibert, Kaniel, Van Nieuwerburgh and Vestman (2017)

augment each fund's monthly net return with the fund's monthly expense ratio to get the monthly gross return $r_{j,t}^{Gross}$. 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 4 different benchmark models: CAPM, Fama-French three-factor model, Fama-French and Carhart four-factor model and Fama-French five-factor model. For our main results, we use the Fama-French five-factor model as the benchmark, but our results are robust to other risk adjustments. In our sample, the average annual realized alpha for Fama-French five-factor model is 54 bps. This result is very close to Pastor, Stambaugh and Taylor (2015)'s estimates, where they find the monthly alpha is 5 bps, which translates to 60 bps of annual alpha.

Since our focus is on the efficient allocation of assets across active funds, we choose to minimize the details related to modeling index funds.[7] We aggregate all index funds from Vanguard to build a single index fund. We choose Vanguard because, as proposed in Berk and van Binsbergen (2015), index funds from Vanguard historically have been the most accessible index funds for retail investors. Specifically, we compute its assets under management (hereafter AUM) by summing AUM across all funds; we compute the combined fund's expense ratio by asset-weighting across index funds. We count the combined index fund's age from the inception year of Vanguard, which is 1975.

We define the total mutual fund market M_t as the sum of AUMs of all the active funds and the combined index fund in year t. We define market share $s_{j,t}$ as the ratio between fund j's AUM and the total fund market. $M_t s_{j,t}$ gives the fund j's AUM in millions of dollars in year t. We exclude fund/year observations with fund's AUM below \$15 million in 2015 dollars. A \$15 million minimum is also used by Elton, Gruber, and Blake (2001), Chen et al. (2004), Yan (2008), and Pastor, Stambaugh and Taylor (2015). In our dataset, there is a huge skewness in fund's AUM. From the summary statistics, we can see the mean of fund's AUM is much larger than the median. The funds at the 99 percentile is over 1,100 times larger than the funds at the 1 percentile. This skewness could potentially affect our estimates. Following Chen et al (2004) we use the logarithm of a fund's AUM as our measure of fund size.

In taking our model to the data, we use the reported distribution costs (sales loads and 12b-1 fees, which are typically used to compensate brokers for directing client investment to funds) as our combined proxies for marketing costs (rather than using, for example, advertising expenditures). The reason is that in the U.S., many investors purchase mutual funds through intermediaries such as brokers or financial advisors. Among all the expenses that mutual fund companies categorized as marketing, advertising expenses constitute only a tiny portion (according to the ICI). The bulk of the marketing costs is compensation paid to brokers and financial advisors, albeit we do not observe this compensation directly.

In the mutual fund industry, a single mutual fund may provide several share classes to investors that differ in their fees structures (typically, the difference is in the combination of front loads and 12b-1 fees). Following much of the literature (with some exceptions, e.g., Bergstresser,

[7] For a detailed study of search frictions *within* the index fund market, see Hortaçsu and Syverson (2004).

Chalmers, and Tufano 2009), we conduct our analysis at the fund level instead of the share class level. To be able to do so, we need to aggregate the share class level expense ratios, 12b-1 fees and front loads up to the fund level. We define the marketing expense $b_{j,t}$ as the "effective" 12b-1 fees that includes amortized loads (see appendix for details).

4 Estimation

Our estimation proceeds in two steps. We first estimate the set of parameters governing mutual fund investment performance: μ, κ, δ, ρ, and η, using the observed panel of fund returns and market shares: $\{r_{j,t}, s_{j,t} | j = 1, ..., N, t = 1, ..., T\}$ using maximum likelihood estimation (MLE). This also gives us the posterior beliefs on the funds' skills in every period. Then we estimate the other parameters (which are related to the search model) using generalized method of moments (GMM) by relating the observed $s_{j,t}$ to the fund characteristics, as well as making inferences from the equilibrium restrictions on the fee-setting behavior of funds, taking the Bayesian posterior beliefs about funds' skills as given.

4.1 Fund performance

From expression (1), we can write down the probability of observing $r_{j,t}$ conditional on observed market shares and realized outperformances up to t:

$$\Pr\left(r_{j,t} \middle| s_{j,t}, r_{j,t-1}, s_{j,t-1}, r_{j,t-2}, s_{j,t-2}, ...\right) \sim \mathcal{N}\left[\widetilde{a}_{j,t} - \eta \log(M_t s_{j,t}),\ \widetilde{\sigma}^2_{j,t} + \delta^2\right].$$

In writing down the above conditional likelihood, note that the current market share $s_{j,t}$ does not provide further information about the skill $a_{j,t}$ beyond $\{r_{j,t-1}, s_{j,t-1}, r_{j,t-2}, s_{j,t-2}, ...\}$, because it is a function of $\widetilde{a}_{j,t}$ but not $a_{j,t}$ directly. Neither does $s_{j,t}$ provide any information on $\varepsilon_{j,t}$ for the same reason.

We can use the above conditional probability to construct a partial log likelihood function (see Wooldridge, 2010, § 13.8):

$$\sum_{j=1}^{N} \sum_{t} \log \Pr\left(r_{j,t} \middle| s_{j,t}, r_{j,t-1}, s_{j,t-1}, r_{j,t-2}, s_{j,t-2}, ...\right).$$

The first summation is across all the funds. The second summation is across all the periods in which fund j existed. One maximizes this likelihood with respect to μ, κ, δ, ρ, and η to obtain their estimates.

4.2 Search model

The parameters in the search model are estimated using (i) a set of moment conditions constructed with $\xi_{j,t}$ and (ii) the optimality of the funds' behaviors. For (i), we first need to back out the $\xi_{j,t}$'s from the data given any set of parameter values. This amounts to finding the $\boldsymbol{\xi}_t$ that equates the model-predicted market shares $\boldsymbol{H}_t(\boldsymbol{p}_t, \boldsymbol{b}_t, \widetilde{\boldsymbol{a}}_t, \boldsymbol{x}_t, \boldsymbol{\xi}_t, p_{0,t}; \Theta)$ to the observed shares $\boldsymbol{s}_t$ for each period t. Since the fixed point of $\boldsymbol{H}_t$ is observed as $\boldsymbol{s}_t$ in the data we can achieve this by solving $\boldsymbol{F}_t\left[\boldsymbol{p}_t, \boldsymbol{b}_t, \widetilde{\boldsymbol{a}}_t - \eta \log\left(M_t \boldsymbol{s}_t\right), \boldsymbol{x}_t, \boldsymbol{\xi}_t, p_{0,t}; \Theta\right] = \boldsymbol{s}_t$ for $\boldsymbol{\xi}_t$ (given a set of parameter values and observed fund choices).

The definition of $\boldsymbol{\xi}_t$ gives us our first set of moment conditions: $\boldsymbol{E}(\xi_{j,t}|\boldsymbol{x}_t, \widetilde{a}_{j,t}) = 0$. This condition states that $\xi_{j,t}$ is mean independent of the $\boldsymbol{x}_{j,t}$, the exogenous variables that affect fund's sampling probability in addition to marketing expenses, and $\widetilde{a}_{j,t}$, the posterior belief about the fund skill at the beginning of period t. Let $j \in t$ denote any active fund j that is alive in period t. The sample version of the moment conditions is

$$\sum\nolimits_{t=1}^{T} \sum\nolimits_{j\in t} \xi_{j,t} \begin{pmatrix} \boldsymbol{x}_{j,t} \\ \widetilde{a}_{j,t} \end{pmatrix} = \boldsymbol{0}.$$

Following Hortaçsu and Syverson (2004) and Chen et al. (2004) we include in $\boldsymbol{x}_{j,t}$ both log age and the number of funds in the same fund family to capture the fund level social learning effects, as well as advertising that is conducted at the family level. Importantly, we do not include lagged fund size into $\boldsymbol{x}_{j,t}$. In the data, fund size is persistent over time, so including lagged fund size creates an over-fitting problem, where $s_{j,t}$ is almost mechanically explained by $s_{j,t-1}$. From the point of view of the moment conditions, such a problem arises due to the fact that lagged size depends on $\xi_{j,t-1}$, which is also likely persistent, and so lagged size $s_{j,t-1}$ is likely correlated with $\xi_{j,t}$.

In contrast to this first set of moment conditions, we do *not* require $\boldsymbol{E}(\xi_{j,t}|p_{j,t}, b_{j,t}) = 0$ because $p_{j,t}$ and $b_{j,t}$ are endogenous outcomes of the model and thus depend on $\xi_{j,t}$. One typical approach that the literature explores to deal with such endogeneity is using instruments for firms' pricing or marketing choices. Another common approach, which we follow here, is to rely on the optimality of the observed firm choices. Intuitively, the levels of fees and marketing expenses that are optimal for different funds will depend on the elasticities of demand. Therefore, as long as the observed choices are optimal, they help to identify the demand function.

The first order condition for the price for fund j at period t is

$$s_{j,t} + \partial H_{j,t}/\partial p_{j,t} \cdot (p_{j,t} - b_{j,t}) = 0.$$

In order to exactly align the behaviors predicted by a model with the observed behaviors of each individual fund in the data, one must either introduce unobserved heterogeneity in costs or allow for the first-order conditions to be satisfied with error.[8] In our estimation, we implicitly allow decision errors as each fund chooses its price and marketing expense. Specifically, we allow the first-order condition above to be satisfied up to an error:

$$s_{j,t} + \left(\frac{\partial H_{j,t}}{\partial p_{j,t}} \cdot e^{\zeta_{j,t}}\right)(p_{j,t} - b_{j,t}) = 0, \tag{12}$$

where $\zeta_{j,t}$ represents the fund's "error" in setting its price (e.g., due a to mis-assessment of the slope of the demand curve). We will assume that $\zeta_{j,t}$ has a mean of zero across all periods and funds. In other words, while discrepancies are allowed at the individual fund level, we still ask

[8] See Baye and Morgan (2004), which shows that allowing only a small amount of bounded rationality in players' optimization behaviors can be of great use in reconciling the Nash hypothesis with the commonly observed price patterns in the data.

the average behavior to be consistent with the model.

The first order condition for the marketing expenses is similar but sightly more involved because of the corner restrictions $0 \le b_{j,t} \le \bar{b}$. We let

$$- s_{j,t} + \left(\frac{\partial H_{j,t}}{\partial b_{j,t}} \cdot e^{\omega_{j,t}} \right) (p_{j,t} - b_{j,t}) \begin{cases} \le 0, & \text{if } b_{j,t} = 0; \\ \ge 0, & \text{if } b_{j,t} = \bar{b}; \\ = 0, & \text{otherwise.} \end{cases} \tag{13}$$

Here we again allow a mean zero error $\omega_{j,t}$. One interpretation of these decision errors is inertia: if it is costly for funds to change the fees that they charge, including the component that covers marketing costs, these will be sticky over time, typically deviating from the level that is optimal at a particular point in time (we abstract from modeling the dynamic fee-setting behavior here). Another source of errors would come from fund-family-related constraints (e.g., some families have financial advisory arms and might choose to cross-subsidize those by channeling marketing fees to the advisors they employ, even if it is suboptimal from the standpoint of maximizing profits on some of the funds they manage, while other families might eschew marketing altogether even if some of their funds might benefit from it).[9]

Thus, we assume that fund choices of prices and marketing expenses are optimal *on average*, so that the first order conditions are satisfied up to a fund-period-specific errors. Given (12) and (13), this amounts to $\boldsymbol{E}(\zeta_{j,t}) = 0$ and $\boldsymbol{E}(\omega_{j,t}) = 0$. Notice that these moments do not impose any distributional or correlational assumptions on $\zeta_{j,t}$ or $\omega_{j,t}$. In sample versions,

$$\sum\nolimits_{t=1}^{T} \sum\nolimits_{j \in t} \zeta_{j,t} = 0, \tag{14}$$

$$\sum\nolimits_{t=1}^{T} \sum\nolimits_{j \in t} \omega_{j,t} = 0. \tag{15}$$

The first error, $\zeta_{j,t}$, can be directly backed out from the first order condition given any set of parameter values:

$$\zeta_{j,t} = -\log \left(\frac{-\partial H_{j,t}/\partial p_{j,t}}{s_{j,t}} \right) - \log (p_{j,t} - b_{j,t}).$$

The second error, $\omega_{j,t}$, can be computed exactly for j with $0 < b_{j,t} < \bar{b}$, but unfortunately not for the boundary cases:

$$\omega_{j,t} \begin{cases} \le \bar{\omega}_{j,t}, & \text{if } b_{j,t} = 0; \\ \ge \bar{\omega}_{j,t}, & \text{if } b_{j,t} = \bar{b}; \\ = \bar{\omega}_{j,t}, & \text{otherwise,} \end{cases}$$

where

$$\bar{\omega}_{j,t} \equiv -\log \left(\frac{\partial H_{j,t}/\partial b_{j,t}}{s_{j,t}} \right) - \log (p_{j,t} - b_{j,t}).$$

[9] There is a connection between the decision errors that we introduce here and the notion of ϵ-equilibrium in game theory, first introduced by Radner (1980). A set of choices constitutes an ϵ-equilibrium if the difference between what a player achieves and what he could optimally achieve is less than ϵ. In other words, it only requires each player to behave near-optimally, which turns out to be the same as what we ask in (12) and (13). Specifically, there is a mapping from $\zeta_{j,t}$ and $\omega_{j,t}$ to the loss that firm j incurs relative to its optimal payoff. When both errors are zero, such loss is zero. More importantly, it can be shown that this mapping is insensitive, in the sense that fairly large errors only lead to a relatively small loss reduction in profits.

In principle, we cannot simply use the average of $\bar{\omega}_{j,t}$ as an estimate of $E(\omega_{j,t})$. A conventional way to deal with this kind of truncation problem is to make an additional distributional assumption and apply an MLE estimator. However, a key issue here is that the truncation interval is not fixed but varies across funds endogenously, and thus it may be correlated with $\omega_{j,t}$.

We take a simpler approach of comparing the estimates based on several subsample versions of (15):

$$\text{(i)} \sum_{0<b_{j,t}<\bar{b}} \omega_{j,t} = 0; \;\; \text{(ii)} \sum_{b_{j,t}=0} \bar{\omega}_{j,t} = 0; \;\; \text{(iii)} \sum_{b_{j,t}=\bar{b}} \bar{\omega}_{j,t} = 0; \;\; \text{(iv)} \sum_{\text{all } j,t} \bar{\omega}_{j,t} = 0.$$

The first version (i) assumes that on average, the funds that choose an interior level of marketing expenditure are right about the effect of marketing on market share. These are the funds for which we can exactly calculate the $\omega_{j,t}$. We acknowledge that these funds are a selected sub-sample of all funds; their average does not necessarily reflect the average across all funds. However, these are the funds that choose the less extreme marketing expenses. In addition, they make up a substantial portion (about 30%) of the funds in the data, so it is reasonable to believe that their average assessment is not far from the population average. The second version (ii) uses the truncated values (lower bounds) of the $\omega_{j,t}$ of the funds that choose zero broker marketing expenses. The third version (iii) uses the truncated values (upper bounds) of the $\omega_{j,t}$ of the funds that choose the highest possible marketing expenses, $\bar{b}$, which has been 1 percent imposed by the SEC. The last version (iv) uses all the values for $\omega_{j,t}$. We use these three latter cases as robust checks. If the estimates based on these four different assumptions are similar, then we can be confident that estimates based on the full sample moment (15) are not too severely impacted by the truncation.

Our GMM estimation is just identified, since there are five unknown parameters and five moment conditions. The parameters are the average search cost λ, the utility weight of outperformance γ, the sensitivity of sampling probability to marketing θ, and a two-dimensional vector of sensitivities β (for number of funds in the fund family and fund age). There are three moment conditions for the sampling probability residual ξ and two more moment conditions based on the first order necessary conditions for the optimality of funds' pricing and marketing behavior in equations (14) and (15), respectively. We conduct this second-stage estimation in one step using the identity weighting matrix.

5 Results

5.1 Fund performance

Table 1 reports estimates of the fund performance-related parameters using our full sample.[10] The magnitude of decreasing returns to scale parameter η is 0.0048, and it is statistically significant. Since one standard deviation of log fund size is 1.628, a one standard deviation *increase* in log fund size is associated with approximately 78 basis points *decrease* in mean annual alpha. This result is close to Chen et al. (2004). This magnitude is economically significant, in particular as compared to the mean gross alpha of 54 basis points. For robustness, we also estimate the model using linear rather than logarithmic specification similar to Pastor, Stambaugh and Taylor (2015) and obtain estimates broadly consistent with theirs.

Existence of stock-picking skill among mutual fund managers is one of the oldest queries in empirical finance. Early literature used persistence of fund-level performance as an indicator of skill in active fund management, an approach that is called into question by Berk and Green (2004) [11]. Here we take a different approach by estimating a version of the Berk and Green model directly. We find that the mean of the prior distribution of managerial skill is 3.05% (per annum). This number is positive and statistically significant, which means that an average active mutual fund manager is skilled (we plot the prior distribution of fund manager skill implied by the estimated parameters μ and κ in Figure A4 in the Appendix). Over 71% of the funds have fundamental skill levels that are higher than the mean expense ratio, at least when applied to the first dollar of assets under management (i.e., before any of the effects of decreasing returns to scale).

Another parameter of interest is ρ, the persistence of fund manager's skill. Our empirically estimated persistence is 0.94, which means past beliefs are quite useful in predicting future performance. Our skill persistence result is consistent with Berk and van Binsbergen (2015) who find that the cross sectional differences in value added persist for as long as 10 years. One of the reasons that skill persistence is not perfect as assumed in the Berk and Green model is managerial turnover. If we believe the skill of a mutual fund is partially due to the mutual fund manager, then a change of management team might affect the skill level of the fund. Fidelity Magellan fund manager Peter Lynch is a case in point: during his tenure from 1977 to 1990, according to our measure of performance, Magellan fund achieved 14 consecutive years of positive alpha. After Peter Lynch's departure, Magellan's performance becomes less impressive, reverting towards the mean.

[Insert Table 1 Here]

[10] In our dataset, the first period with non-missing data is the year 1964, so our full sample estimates use the data from 1964 to 2015. It is important to use all the available information to estimate the learning model that is the core of Berk and Green (2004). In the model, when a fund is born, it draws an initial skill level from the prior skill distribution. Then investors use the *entire* history of subsequent realized performance to update their beliefs about each fund's skill level. If we were to start the sample at a later date, for example, year 1995, we would lose the performance information for a lot of funds that were in operation well before 1995. One way to circumvent the above truncation problem is to pick a starting year and keep only the funds which are founded after this year. But this approach would bias the estimates toward newer funds.

[11] Berk and Green (2004) argue that the lack of performance persistence doesn't mean lack of skill if capital flows to outperforming funds and if fund size erodes fund performance.

5.2 Asset misallocation

Equipped with estimated parameters of fund skill distribution, we compute the investor beliefs about each fund's skill level at each point in time. Then we can derive implied fund size according to a benchmark frictionless model following Berk and Green (2004) (henceforth BG). By comparing BG-implied fund size with the data, we can assess the degree of asset misallocation in the mutual fund industry.

First, we compute the investor beliefs about each fund's skill level $\widetilde{a}_{j,t}$ using the recursive expression derived in section 2.1. As an example, consider a fund j that was born in period $t = 1$. At the fund's birth, we assign the fund an expected skill level of μ, then we use realized return $r_{j,1}$ and fund size $M_1 s_{j,1}$ to get the updated belief, $\widetilde{a}_{j,1}$. By iterating forward, we can generate the whole series of fund's expected skill levels. Next, we compute the BG-implied fund size. Berk and Green's model predicts that fund's size (i.e., total assets under management), which we denote by $s^{BG}_{j,t}$, should be such that the decreasing returns to scale exactly offsets the investor belief less fund expense ratio, denoted as "net skill": $D(s^{BG}_{j,t}; \eta) = \widetilde{a}_{j,t} - p_{j,t}$. So, with a log specification for $D(\cdot)$, we have

$$\log(s^{BG}_{j,t}) = \frac{\widetilde{a}_{j,t} - p_{j,t}}{\eta}. \tag{16}$$

This expression is intuitive: the higher the net skill of a fund, the larger is the efficient fund size; the stronger the effect of decreasing returns to scale, the smaller the fund's size will be.

To compare BG-implied fund size with data, we construct ten portfolios of mutual funds sorted on net skill. We then compute mean of log size in the data and in the BG model for each portfolio.[12] Figure 1 presents the result. First, we can see that in the data, the mean fund size monotonically increases with net skill. This result is consistent with the Berk and Green model's prediction. But we also witness a discrepancy between the data and the model. On the higher end, BG predicts the mean size of funds in portfolio 10 to be 7.3 billion. In the data, the mean fund size in portfolio 10 is 936 million. On the lower end, according to BG, the mean fund size in portfolio 1 is 0.7 million. And in the data, it is 134 million. These differences are statistically significant as indicated by the 95-percent confidence intervals. From this figure, we can draw the conclusion that asset misallocation exists in both bad funds and good funds in the data.

[Insert Figure 1 Here]

The key prediction of Berk and Green (2004) is that asset inflows into funds that are estimated to be skilled based on their past returns will erode their subsequent performance due to decreasing returns to scale. In addition, fund managers who have been revealed as skilled raise their fees. As a result the net alpha of these funds should be zero in the future. We can test this prediction of the (generalized) model by looking at abnormal returns on the portfolios of mutual funds formed on their net skill discussed above. Thus, using the updated belief about fund skill as well as its fees in year t, funds are placed into portfolios, and we track equal-weighted returns on these portfolios over the subsequent 12 months, until the portfolios are re-sorted based on

[12] We winsorize the belief $\tilde{a}$ at 1% and 99% level because there are some outliers in the estimated beliefs.

the updated information from $t+1$. Estimated five-factor alphas on these portfolios along with their 95% confidence intervals are displayed in Figure 2. Consistent with much of the mutual fund literature, the vast majority of alphas are negative (i.e., for all but the top two deciles of net skill). Perhaps more surprisingly, funds in the top decile of estimated net skill actually do display statistically significant outperformance. Overall, realized alpha is monotonically increasing with the estimated net skill, ranging from close to -3% per annum for the funds in the bottom decile, to about 0.7% for those in the top decile. This result indicates a stark rejection of the key prediction of the BG model. Not only don't assets seem to flow out of unskilled and/or expensive funds towards the relatively more skilled ones, as suggested by evidence in Figure 1 above, but even the "best" funds that are "too small" relative to the BG model don't raise their prices sufficiently to fully capture their outperformance. Thus the observed allocation of capital across equity mutual funds, combined with their price setting behavior, present a quantitative puzzle for the frictionless BG model.

[Insert Figure 2 Here]

5.3 Search model parameters

Our search model is meant to bridge the gap between the efficient capital allocation described by the Berk and Green model and the actual allocation of assets across mutual funds observed in the U.S. data. Table 2 reports the estimated parameters of the structural search model. With the view towards conducting counterfactual analysis, we rely on the more recent sample of the data for this part of the estimation, choosing 2001 as our starting point (our estimation results are robust to various starting points, however). As described in the estimation section, we estimate the model using four versions of the moment conditions in (15). All the parameters other than θ are quite stable across the four sets of estimates. This assures us that even though our identification of θ relies on one subsample, it does not affect other parameters drastically.

[Insert Table 2 Here]

Our estimate of λ, the mean of search cost, is 39 basis points. Hortaçsu and Syverson (2004) find that the mean search cost for the S&P 500 index fund market is from 11 bps to 20 bps across different specifications.[13] Our estimated search cost is somewhat higher than theirs since, presumably, investors in their sample have higher than average level of financial sophistication (implying a lower level of search costs). Indeed they focus on investment in S&P 500 index funds in the late 90s, when these funds were not as prominent as they are today. Alternatively, it may be the case that it is harder to evaluate actively managed mutual funds (compared to index funds, which are relatively simple products), and hence implied barriers to information acquisition that are implied by the observed distribution of fund size are greater.

The magnitude of the mean search cost is quite significant. For the average investor, the cost of drawing another sample fund is 39 basis points, which is comparable to the mean alpha

[13] Hortaçsu and Syverson (2004) estimated two variants of a search model. In the first type, the sampling probabilities across funds are different whereas in the second type, the sampling probabilities are the same. They estimate search costs for both types of models. We view our model as being closer to the first type. The estimation results for the first type of model is reported in Table III in their paper. The log mean search cost is around -6.17 to -6.78. So the mean search costs ranging from 11 bps to 20 bps in the S&P 500 index fund market.

in our sample. The large magnitude of estimated search cost is a reflection of the active fund under-performance puzzle. In the mutual fund literature, numerous papers documented the (persistent) underperformance of (at least a large subset of) active funds (e.g., Carhart 1997). Since many under-performing funds enjoy sizable market shares, our model requires a high search cost to rationalize those facts. In our model, high search cost investors will find it suboptimal to continue searching for a better fund than those drawn in the first couple of attempts. In the counterfactual case, if the search costs were low then index funds would be much larger than observed in the data, and underperforming active funds would be substantially smaller.

Our key parameter of interest is θ, the coefficient in front of marketing expenses in the sampling probability function. First, we notice that the estimated θ is the smallest when we use the moment conditions of the funds that choose to do no marketing, and the largest for the funds that choose the upper bound of 1%. For the funds that choose the interior levels, θ is in the middle. This is intuitive because θ measures the effectiveness of marketing. The funds that are at the upper bound are more likely to be constrained in their ability to increase their marketing in an effort to increase investors' awareness. Consequently, the first order condition (15) is likely to be satisfied with an inequality, and forcing it to be zero in estimation biases the estimate upward. Similarly, the funds at the lower bound are likely to find it optimal to receive a "rebate" on marketing in order to increase their profits, but since such rebates are not available their first order condition is also likely not to be satisfied, biasing the estimate of θ downwards. In what follows, we rely on the estimates obtained with funds in the interior of the marketing expenditures as our benchmark.

In the sampling probability function, besides marketing expenses, we include fund family size, log fund age and year fixed effect. The coefficient of family size is positive and significant, confirming the idea that larger fund families are better at informing investors about their products. The fund age coefficient is positive and significant, which is intuitive, as older funds also have more visibility than younger funds. This result is also consistent with Hortaçsu and Syverson (2004) evidence from the S&P 500 index fund market.

In order to put these estimates into perspective, we conduct the following experiments. We compute the percentage changes in fund size for various groups of funds when marketing expense increases by 1 bp. Table 3 provides the results. Each column corresponds to results computed using different values of the θ parameter - those obtained with the funds on the upper bound ($\theta = 133.18$), lower bound ($\theta = 111.22$), and in the interior ($\theta = 113.11$) of marketing expenditures, as described above. All the other parameters are fixed at the benchmark levels (estimates from the "interior" funds). When we change fund j 's marketing, we fix all the other funds' prices and marketing expenses and fund j's price (i.e., this is a comparative static, not counter-factual analysis). Thus, holding total fees fixed, a 1 bp increase in marketing implies an equivalent reduction in profit margins.

[Insert Table 3 Here]

Overall, a 1 bp increase in marketing expenses leads to a roughly 1% increase in fund's size, but there is substantial variation across different types of funds, and this elasticity naturally increases with θ. In panel A, we sort funds by their size. We find that as fund size decreases, the sensitivity of size to a 1 bp increase in marketing rises. Using the benchmark estimates

($\theta = 113.11$) it goes from approximately 0.87% for large funds to 0.9% for small funds. This is intuitive because as a prior, marketing investment should be much more effective for smaller funds because they have smaller probabilities of being known (e.g., typically, they are younger). Investing in marketing is a good way for small funds to attract greater investor attention. Interestingly, this sensitivity is higher both at the upper and at the lower estimates of θ.

In panel B, we sort funds by their skill level $\tilde{a}$. We find that marketing is much more useful for highly skilled funds. If high-skill funds can get into the consideration sets of more investors they will be picked by more investors. But for the low-skill funds, even if they are known to more investors, their size will not increase sufficiently to justify the extra expense. In fact, in Figure 3 we show that, for a fund of average age and belonging to a fund family of average size, with fund/year shock $\xi = 0$, the optimal level of marketing is increasing in the posterior belief about its skill. This result indicates that marketing is complementary to skill, yet it does not mean that it helps improve welfare in the presence of the search friction, since high-skill funds may be forced to spend "too much" on marketing, leading to a wasteful "arms race."

[Insert Figure 3 Here]

Lastly, in panel C, we sort funds by their original marketing expenses levels. Lower Bound funds are funds that originally choose zero marketing expenses. Upper Bound funds are funds that originally chose 1% marketing expenses. Non binding funds are the rest of funds, which choose interior marketing levels. We find that an additional 1 bp increase in marketing is not very useful to funds at the upper bound (suggesting that many of these funds are at suboptimally high levels of marketing, perhaps due to inertia). Similarly, for funds at the lower bound extra marketing appears more worthwhile. Some of these funds might belong to fund families that choose to sell their funds directly rather than through brokers, for example, and as a consequence do not charge any 12b-1 fees, even though it might be beneficial for some of their funds.

Next we analyze the impact of marketing on fund profits. Table 4 displays the results. In panel A, we sort funds by size; we find that for the small funds extra marketing increases profits, if all the other funds' strategies in pricing and marketing stay the same (since we are not recomputing their best responses in this exercise). In panel B, we show that when θ is at the higher level of estimates, it is profitable for high-skill funds to do more marketing. In panel C, we find that essentially all of the funds are worse off if they increase their marketing, which is not surprising given that the estimation procedure assumes that funds are at their optimal levels of marketing (on average).

[Insert Table 4 Here]

5.4 Sampling probabilities and fund size

In this section, we quantify the impact of various components of the sampling probability in explaining the size distribution of funds. Our method is as follows: we first set one of the components in sampling probability to be equal 0. Then we recompute the model-implied market shares of all funds. Notice that here we are not recomputing the whole equilibrium. We fix all other variables and parameters. Lastly we regress the log of market share of funds in the data onto model-implied log market shares and report the R-squared. In Table 5, we report the results. The lower the R-squared, the more important that component is in terms of explaining the size distribution. Among all of them, the unobserved characteristics of the fund ξ is the most important one (responsible for almost half of the R-squared). This is reasonable because we only include a limited number of variables in our estimation; any other variables that could potentially affect fund size would be subsumed by ξ. The second most important variable is age. After controlling for fund's age and other variables, the family size doesn't add much explanatory power.

What about the key features of mutual funds that have been the main focus of the literature - skill and costs? Removing either variation in posterior skill or in the fund price (expense ratio) reduces the R-squared to about 90% in each case. Importantly, removing instead the marketing variable yields a very similar R-squared of 92%. This indicates that marketing is nearly as important in terms of explaining the size distribution of mutual funds as price or skill.

We are also interested in understanding how do various components of our model contribute to the misallocation of capital to funds. We compute the correlation between model-implied fund size and BG-implied fund size. We can see that in the data is positive but small, at 0.09. If changing one of the components of the model increases this correlation, that means that this change makes capital allocation more efficient. This correlation is at its highest level of 0.59 if we only include fund skill and price. Conversely, removing price or skill but including other (search model) ingredients reduces this correlation, since these are the key elements of the Berk and Green model. At the same time, removing marketing increases the correlation. This means that marketing could potentially account for at least some of the misallocation that we observe in the data. However, this analysis is only suggestive, since we do not compute the optimal response of the funds to the induced change. In what follows, we describe counterfactual experiments that fully take into account the equilibrium behavior of both investors and funds.

[Insert Table 5 Here]

6 Counterfactual Analysis

Section 5.2 documents substantial capital misallocation in the mutual fund industry. In this section, we use our model to quantitatively study the importance of marketing expenses and search costs in shaping the equilibrium fund size and expense ratios. We also investigate how they affect allocational efficiency and investor welfare. First, we explore a counterfactual equilibrium with no marketing.[14] We then investigate the impact of changing search costs on equilibrium marketing expenses. We focus on the most recent year in our sample (2015) for these experiments.

6.1 Welfare measures

We first present how the welfare of different parties in the market are calculated. Fix a year t (the time subscript t will be suppressed in this section). In our model, investor's utility consists

[14] Recently the SEC considered a proposal to improve the regulation of mutual fund distribution fees, in particular, by limiting fund sales charges as a way of protecting retail consumers from unnecessarily high costs. Our counterfactual analysis can be viewed as analyzing welfare consequences of a policy that set the marketing cap at zero.

of two parts, the expected indirect utility provided by the fund that investor chooses and the expected total search costs the investor incurs in order to find this fund. The welfare of investor i with search cost c_i is given by

$$V(c_i) = \frac{\int_{\bar{u}(c_i)}^{+\infty} ud\Psi(u)}{1 - \Psi\left[\bar{u}(c_i)\right]} - c_i \frac{\Psi\left[\bar{u}(c_i)\right]}{1 - \Psi\left[\bar{u}(c_i)\right]}, \tag{17}$$

where $\bar{u}$ is the reservation level of indirect utility (detailed derivation of investor's welfare is provided in the appendix). For a higher level of reservation utility, the investor needs to search more in order to find the desired fund. We see that the expected total search cost $c_i \frac{\Psi[\bar{u}(c_i)]}{1-\Psi[\bar{u}(c_i)]}$ is increasing in $\bar{u}$. In the first term of equation (17), the numerator is the expected indirect utility for the funds with higher than $\bar{u}$ utility level. The denominator adjusts for the fact that the investor will only pick the funds from this part of the distribution. The aggregate measure of investor welfare in this model is derived by integrating across the search cost distribution:

$$U = \int_0^{+\infty} V(c_i) dG(c_i). \tag{18}$$

Fund profits is also part of the total welfare. These include the profits for both active funds and index funds:

$$P = \sum_{j=1}^{N} (p_j - b_j)s_j + s_0 p_0. \tag{19}$$

Here the first part is the total profits for the active funds, the second part is the total profits for the passive funds. In the counterfactual analysis, we assume index fund price is fixed and we resolve the equilibrium for the active funds' prices and marketing expenses. In our counterfactual we assume M stay the same.

If marketing expenses constitute pure payments to labor (e.g., broker commissions) rather than dead weight costs, they should also be considered in the welfare analysis:

$$B = \sum_{j=1}^{N} b_j s_j \tag{20}$$

Our measure of total welfare is the sum of the three components above: $U + P + B$.

6.2 Equilibrium with no marketing

In this simulation, we restrict marketing expenses to zero. We use year 2015's data and the benchmark parameters from column (1) in Table 2. Table 6 provides the comparison between the currently observed equilibrium and the no-marketing equilibrium on some of the key measures. First, the mean expense ratio drops by almost 77 basis points in the counterfactual relative to the current equilibrium. This drop is larger than the decrease in the average marketing expenditure. It indicates fiercer price competition between funds when they cannot attract investors through marketing. To further understand the price changes across funds, we split the funds into four groups based on their marketing expenses in the current equilibrium: (1) funds whose marketing is at the upper bound of 100 bps, (2) funds whose is marketing is at the lower bound of 0, (3) funds whose marketing is between 1 bp to 49 bps and (4) funds whose marketing is between 50 bps and 99 bps. We plot the price differences for all funds between the current equilibrium and the no-marketing equilibrium in Figure 4 panel A. We find that all the funds lower their prices in the no-marketing equilibrium, but the magnitude of change varies substantially across the four groups. Group (1) funds lower their prices by around 100 bps (i.e., roughly their original marketing costs). The most interesting finding is that the group (2) funds in the no-marketing equilibrium also lower their prices by around 30 bps, which necessarily has to come from a reduction in profit margins since these funds do not have any marketing expenses in the current equilibrium. This is mainly due to the effect of competition between funds. A similar but weaker effect is present for most of the funds in groups (2) and (3).

[Insert Figure 4 Here]

Second, we find that the total market share of active funds drops from 74% to 68%. This indicates that marketing is useful for steering investors towards active funds. When funds cannot do marketing, they lose market share since they are less likely to enter investors' information sets. The sampling probability of index funds increases. This is due to the assumption that all the sampling probabilities sum to 1. When active funds cannot do marketing, the index funds are more likely to be "found." In the no-marketing equilibrium, active funds' profits drop by 15 basis points on average. This is resulting from both the shrinking of the total market share and the fall of profit margins.

Investor welfare in the no-marketing equilibrium increases by around 57%. There are three main contributing factors: lower prices, higher alphas, and lower search costs. As assets under management decline, the average alpha of the industry increases from 37 bps to 41 bps, due to the effect of decreasing returns to scale. In Figure 4 panel B, we plot the difference in fund alphas between the no-marketing equilibrium and current equilibrium for different groups of funds. We find that for funds in group (1) alpha increases consistently. This is mainly because in the no-marketing equilibrium their assets under management fall and so, due to decreasing returns to scale, their alphas increase. For other groups of funds, some of the alphas increase, while others decrease.

[Insert Figure 5 Here]

We can compute the total search cost incurred by the investors in the two equilibrium. The total search cost is given by:

$$\int_0^{+\infty} c_i \frac{\Psi\left[\bar{u}(c_i)\right]}{1-\Psi\left[\bar{u}(c_i)\right]} dG(c_i). \tag{21}$$

We find total search costs are lower in the no-marketing equilibrium. In the model, investors search until the expected benefit of finding better funds is smaller than the unit search cost. If investor i has already found fund j with utility u_j, then his incentive to search hinges on both her search cost c_i and the expected possible gain from continuing the search. If there are not too many better funds out there, then investor's incentive to search is weaker. To show that this is indeed the reason why investors search less in the no-marketing equilibrium, we plot the histogram of indirect utilities associated with individual funds in the two equilibria in Figure 5. We find the standard deviation of utility levels in the no-marketing equilibrium is substantially lower than in the current equilibrium. Since the dispersion in available utilities is reduced, so are the expected benefits of searching and, consequently, investors search less. Through the resulting reduction in search costs, investor welfare increases by 17 basis points on average.

[Insert Table 6 Here]

When marketing is eliminated the size of active funds doesn't drop drastically for two reasons. First, there are characteristics in the sampling probability function, besides marketing expenses, which ensure that all active funds sampling probabilities are positive. Second, lowering the size of active funds increases their performance. This effect makes active funds more attractive.

6.2.1 Heterogeneous effect across investors

In our model, we assume different investors have different search costs. In this section, we study the impact of eliminating marketing across investors with different levels of search costs. We focus on the following dimensions: investor welfare, total incurred search cost, gross alpha expected by investors, total expense ratios investors pay, and marketing expenses that investors implicitly pay for as part of their chosen funds' expense ratios (in expectation). Figure 6 panel A shows that for all the search cost levels, in the no-marketing equilibrium, investors achieve a higher level of welfare on average. But the biggest improvements come from the high search cost investors. Their welfare increases roughly by 100 basis points. For the low search cost investors the increase is not very large. This is because the low search cost investors always find the "best" funds available in the market. Figure 6 panel B shows a somewhat non-monotonic relationship between unit search cost and total search costs incurred. For the low search cost investors the total search cost is not very high even though they search a lot since their unit search costs are low. The high search cost investors find it too costly to conduct any search, so they search infrequently, many stopping after the first (free) search. Consequently, high search cost investors' total search cost is also low. The intermediate search cost investors search relatively aggressively and their search costs are non-trivial. So in total they incur the largest total search costs. Comparing the two equilibria, we find that in the no-marketing equilibrium, the intermediate search cost investors incur lower total search costs. This is due to the fact that in the no-marketing equilibrium, average fund quality improves, so that it is easier for investors to find funds that satisfy their reservation levels.

Focusing on Figure 6 panel C and panel D, we find that in general, high search cost investors get lower alpha funds, pay high prices and high marketing expenses. This is simply because high search cost investors don't search very much. An interesting fact is that for the very low search investors, the funds they invest in have positive net alphas. In Berk and Green's model, since investors have zero search cost, in equilibrium, all the funds have zero net alphas. But in our model, since all investors incur a positive search cost, the low search investors are able to find funds that are both skilled and cheap, but not found by enough other investors, so that their performance is not fully eroded by decreasing returns to scale. At the same time, these high-skill funds do not find it optimal to increase their expense ratios (and thus drive net alphas towards zero) because that would make these funds less attractive to the more discerning (low search cost) investors, whose choices are very sensitive to fees.

[Insert Figure 6 Here]

6.3 Allocational efficiency

It is also interesting to consider the consequences of restricting marketing on capital allocation within the mutual fund sector. On the one hand, we see that average (gross) fund alpha increases in the no-marketing equilibrium, suggesting that some highly skilled funds might be "too small," operating below their efficient scale. Indeed, since we show that in the current equilibrium highly skilled funds benefit more from marketing, ceteris paribus, it is reasonable to expect that without the ability to do any marketing these funds might be disproportionately hurt by the imposed constraint. On the other hand, marketing is an important driver of costs, which are in turn a major determinant of net alphas (and indirect utilities) enjoyed by investors.

In keeping with our initial approach, we compare fund size distribution implied by the frictionless benchmark in the style of Berk and Green (2004) and that generated by our search model counterfactual. Figure 7 provides the comparison for the year 2015. Panel A displays the direct analogue of Figure 1 restricted to the data for 2015: the BG-implied values are computed using the posterior distribution of fund alphas as well as their observed expense ratios and the estimated decreasing returns to scale parameter (the fund size in the data is consistent with the search model by construction). Panel B presents the analogues of these values in the counterfactual equilibrium with no marketing. That is, the "counterfactual" plot uses the actual fund size distribution produced by the counterfactual equilibrium simulation, where as the "BG-implied" values are recomputed using the counterfactual expense ratios for the corresponding funds.

We observe that in the no-marketing case the two lines are much closer to each other than in the current equilibrium. This is true only in small part due to the steepening in the relationship between log size and net skill, visible mostly in the middle of the skill distribution. The changing BG-implied distribution plays a noticeably more important effect. This is due to the fact that funds are charging substantially lower fees to their investors in the no-marketing equilibrium. Thus the solid black line in the graph shifts upward, closer to the blue line. The shift appears especially pronounced for the lowest-skill funds, even though they are still "too big" in the counterfactual relative to the frictionless model, where as for the highest-skill funds there is not much difference between the two measures. Thus, the overall effect of eliminating marketing expenditures is to improve the efficiency of capital allocation in the active fund industry, at least from the standpoint of net abnormal returns to investors as emphasized in Berk and Green (2004).

6.4 Reducing search costs

Last, we examine the impact of search costs on equilibrium market outcomes with special attention to marketing expenses. Because of search costs, competing on marketing could be a potential profitable strategy for some funds, since they essentially just need to be sampled by the least-discerning high-cost investors frequently enough. But with the emergence of the Internet, advancement in search technologies (e.g., Google), more transparent comparison (e.g., services like Morningstar and Lipper), and better investor education, we would expect the search frictions to decline over time. In order to analyze the potential impact of new technologies we consider a counterfactual equilibrium where we set the mean search cost to 35 bps or 20 bps.

Given the new search cost, funds reoptimize their prices and marketing expenses. We find that as the average search cost decreases from 39 bps to 35 bps, mean marketing expenses drop from 61 bps to 44 bps. But when the mean search cost further drops to 20 bps, the equilibrium marketing expenses become zero. Notice that the regulatory cap is still held at 100 bps. The intuition is as follows: low search costs render marketing less profitable. In the model with high mean search cost, there exists a large fraction of investors with very high search costs. A subset of funds specifically exploit these "unsophisticated" investors. Those funds invest aggressively in marketing so as to enter more of the high search cost investors' choice sets. Since such investors will not search much, they do end up investing with those funds even if they are not very skilled and quite expensive. But when mean search cost drops to sufficiently low level, this strategy is no longer profitable, since the model presumes there are fewer investors who find it too costly to continue searching for a better fund. Therefore, when search costs are not very high, funds will not invest in marketing and instead compete on price. This result provides a new perspective on the recent evolution of the asset management industry documented by Stambaugh (2014): declining fees charged by active funds coincident with the growth in passive index funds. From the standpoint of our model, both trends can be seen as resulting from falling search costs, due to a combination of information technology and growing investor sophistication.

[Insert Table 7 Here]

7 Concluding Remarks

The question whether actively-managed mutual funds exhibit skill - i.e., persistent outperformance - has a long history in financial economics, since it is central to the debate about informational efficiency of securities markets in the sense of Fama. While there is still substantial debate about the ability of an "average" fund manager to generate abnormal returns (before or after fees are taken into account), perhaps one of the most robust findings in the literature is that investors' flows are much less sensitive to past bad performance than to outperformance (Ippolito 1992, Carhart 1997, Chevalier and Ellison 1997, Sirri and Tufano 1998, etc.). This evidence hints that the market for mutual funds may not be efficient at allocating capital across funds because bad funds aren't punished sufficiently for poor performance, and therefore underperforming managers control more assets than justified by their level of skill. Capital misallocation in the mutual fund industry could potentially lead to inefficiencies in capital allocation across firms, distorting real investment (van Binsbergen and Opp 2016). It is therefore important to understand quantitatively how much capital is misallocated in the mutual fund industry. By estimating the Berk and Green model, we find that in the the U.S. equity mutual funds data, from year 1964 to year 2015, all but the best-performing decile of mutual funds are "too large" relative to the optimal scale predicted by the BG model. These results indicate that there exist substantial frictions in the mutual fund market.

In our paper, we view mutual fund marketing expenses as purely informative (e.g., Butters 1977). It is possible that a portion of these marketing expenses serves a persuasive function in ways highlighted in the theoretical literature: e.g., firms may find it profitable to steer investors toward non-price attributes (Mullainathan et al. 2008, Gabaix and Laibson 2006, Carlin 2009, Ellison and Ellison 2009). But to be able to separate the informative effect from the persuasive effect of marketing would require information about investors' actual choice sets, which is not available. Thus, by making the assumption that all marketing is informative, our welfare analysis results provide an upper bound on the social value of mutual fund marketing.[15] Relaxing this assumption in order to understand the possible welfare loss from "persuasive" marketing is a fruitful venue for future research.

[15] Even if marketing is purely informative, due to the externality of marketing in our model, marketing investment can still be excessive. Fund i's marketing investment could decrease fund j's probability of being known. In a Nash equilibrium, funds will not take the externality into consideration when deciding the marketing investment levels. All of the funds might be better off if they agree on a lower level of marketing investment - But of course, this agreement is fragile since deviation is profitable.

References

Bailey, Warren, Alok Kumar, and David Ng, “Behavioral biases of mutual fund investors,” *Journal of Financial Economics*, 2011, *102* (1), 1 – 27.

Barber, Brad M, Terrance Odean, and Lu Zheng, “Out of sight, out of mind: The effects of expenses on mutual fund flows,” *The Journal of Business*, 2005, *78* (6), 2095–2120.

Baye, Michael R and John Morgan, “Price Dispersion in the Lab and on the Internet: Theory and Evidence,” *RAND Journal of Economics*, 2004, pp. 449–466.

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.

Berry, Steven, James Levinsohn, and Ariel Pakes, “Automobile prices in market equilibrium,” *Econometrica: Journal of the Econometric Society*, 1995, pp. 841–890.

Brown, Stephen J. and William N. Goetzmann, “Performance Persistence,” *Journal of Finance*, 1995, *50* (2), 679–698.

Butters, Gerard R, “Equilibrium distributions of sales and advertising prices,” *The Review of Economic Studies*, 1977, pp. 465–491.

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

Carlin, Bruce I, “Strategic price complexity in retail financial markets,” *Journal of financial Economics*, 2009, *91* (3), 278–287.

Chalmers, John and Jonathan Reuter, “Is Conflicted Investment Advice Better than No Advice?,” Working Paper 18158, National Bureau of Economic Research June 2012.

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.

Christoffersen, Susan EK, Richard Evans, and David K Musto, “What do consumers' fund flows maximize? Evidence from their brokers' incentives,” *The Journal of Finance*, 2013, *68* (1), 201–235.

Egan, Mark, "Brokers vs. Retail Investors: Conflicting Interests and Dominated Products," 2017.

——, **Gregor Matvos, and Amit Seru**, "The market for financial adviser misconduct," Technical Report, National Bureau of Economic Research 2016.

Ellison, Glenn and Sara Fisher Ellison, "Search, obfuscation, and price elasticities on the internet," *Econometrica*, 2009, *77* (2), 427–452.

Elton, Edwin J, Martin J Gruber, and Christopher R Blake, "A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar mutual fund databases," *The Journal of Finance*, 2001, *56* (6), 2415–2430.

Gabaix, Xavier and David Laibson, "Shrouded attributes, consumer myopia, and information suppression in competitive markets," *The Quarterly Journal of Economics*, 2006, *121* (2), 505–540.

Gallaher, Steven, Ron Kaniel, and Laura T Starks, "Madison Avenue meets Wall Street: Mutual fund families, competition and advertising," *Working paper*, 2006.

Garleanu, Nicolae B and Lasse H Pedersen, "Efficiently inefficient markets for assets and asset management," Technical Report, National Bureau of Economic Research 2015.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny, "Money doctors," *The Journal of Finance*, 2015, *70* (1), 91–114.

Greenwood, Robin and Andrei Shleifer, "Expectations of Returns and Expected Returns," *Review of Financial Studies*, 2014, *27* (3), 714–746.

Guercio, Diane Del and Jonathan Reuter, "Mutual Fund Performance and the Incentive to Generate Alpha," *The Journal of Finance*, 2014, *69* (4), 1673–1704.

Gurun, Umit G, Gregor Matvos, and Amit Seru, "Advertising expensive mortgages," *The Journal of Finance*, 2016, *71* (5), 2371–2416.

Hastings, Justine, Ali Hortaçsu, and Chad Syverson, "Advertising and competition in privatized social security: The case of Mexico," *Econometrica*, 2016.

Hendricks, Darryll, Jayendu Patel, and Richard Zeckhauser, "Hot hands in mutual funds: Short-run persistence of relative performance, 1974–1988," *The Journal of finance*, 1993, *48* (1), 93–130.

Honka, Elisabeth, Ali Hortaçsu, and Maria Ana Vitorino, "Advertising, consumer awareness, and choice: Evidence from the US banking industry," *RAND Journal of Economics*, 2016.

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, Kelsey D. Wi, and Hong Yan, “Participation Costs and the Sensitivity of Fund Flows to Past Performance,” *The Journal of Finance*, 2007, *62* (3), 1273–1311.

Ibert, Markus, Ron Kaniel, Stijn Van Nieuwerburgh, and Roine Vestman, “Are Mutual Fund Managers Paid For Investment Skill?,” Working Paper 23373, National Bureau of Economic Research April 2017.

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.

Kaniel, Ron and Robert Parham, “WSJ Category Kings–The impact of media attention on consumer and mutual fund investment decisions,” *Journal of Financial Economics*, 2016.

Kennan, John, “Uniqueness of positive fixed points for increasing concave functions on Rn: An elementary result,” *Review of Economic Dynamics*, 2001, *4* (4), 893–899.

Linnainmaa, Juhani T, Brian T Melzer, and Alessandro Previtero, “The misguided beliefs of financial advisors,” *Unpublished manuscript*, 2016.

Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer, “Coarse thinking and persuasion,” *The Quarterly journal of economics*, 2008, *123* (2), 577–619.

——, Markus Noeth, and Antoinette Schoar, “The market for financial advice: An audit study,” Technical Report, National Bureau of Economic Research 2012.

Pástor, Luboš and Robert F Stambaugh, “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.

Radner, Roy, “Collusive behavior in noncooperative epsilon-equilibria of oligopolies with long but finite lives,” *Journal of economic theory*, 1980, *22* (2), 136–154.

Reuter, Jonathan and Eric Zitzewitz, “Do ads influence editors? Advertising and bias in the financial media,” *The Quarterly Journal of Economics*, 2006, *121* (1), 197–227.

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

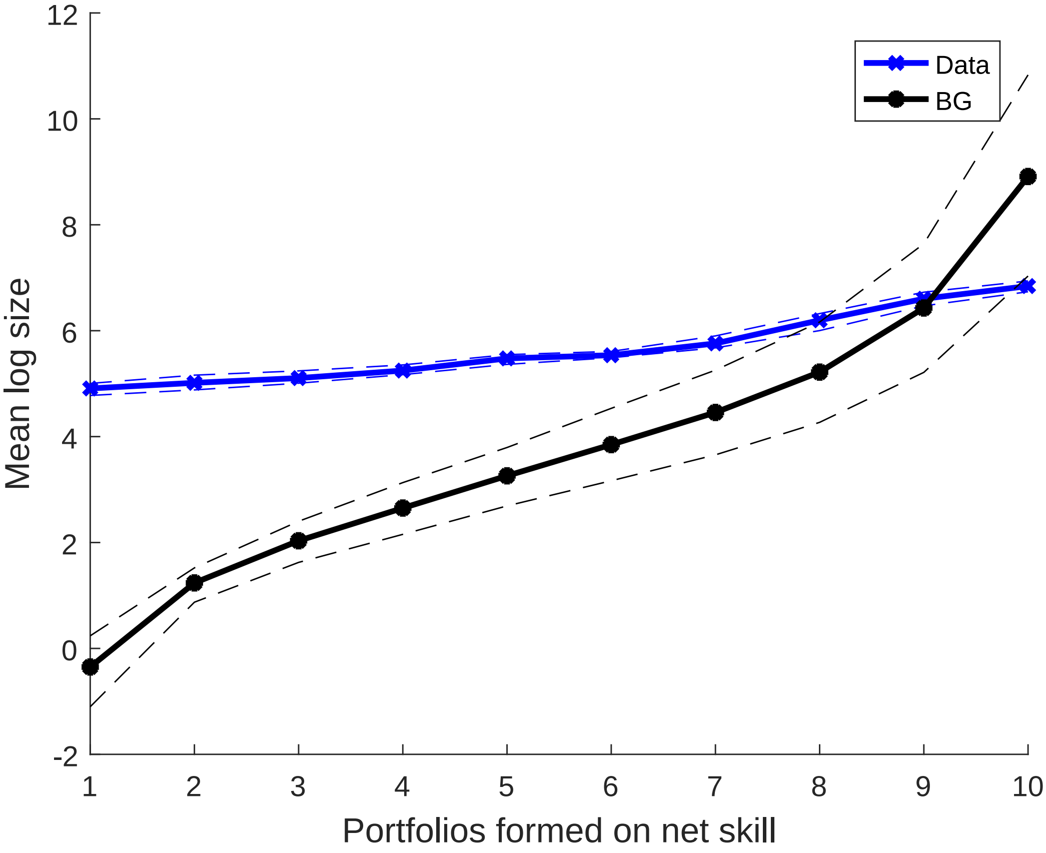
Stambaugh, Robert F., “Investment Noise and Trends,” *The Journal of Finance*, 2014, *69* (4), 1415–1453.

van Binsbergen, Jules and Christian Opp, “Real anomalies: Are financial markets a sideshow,” *Manuscript, University of Pennsylvania*, 2016.

Wooldridge, Jeffrey M, *Econometric analysis of cross section and panel data*, MIT press, 2010.

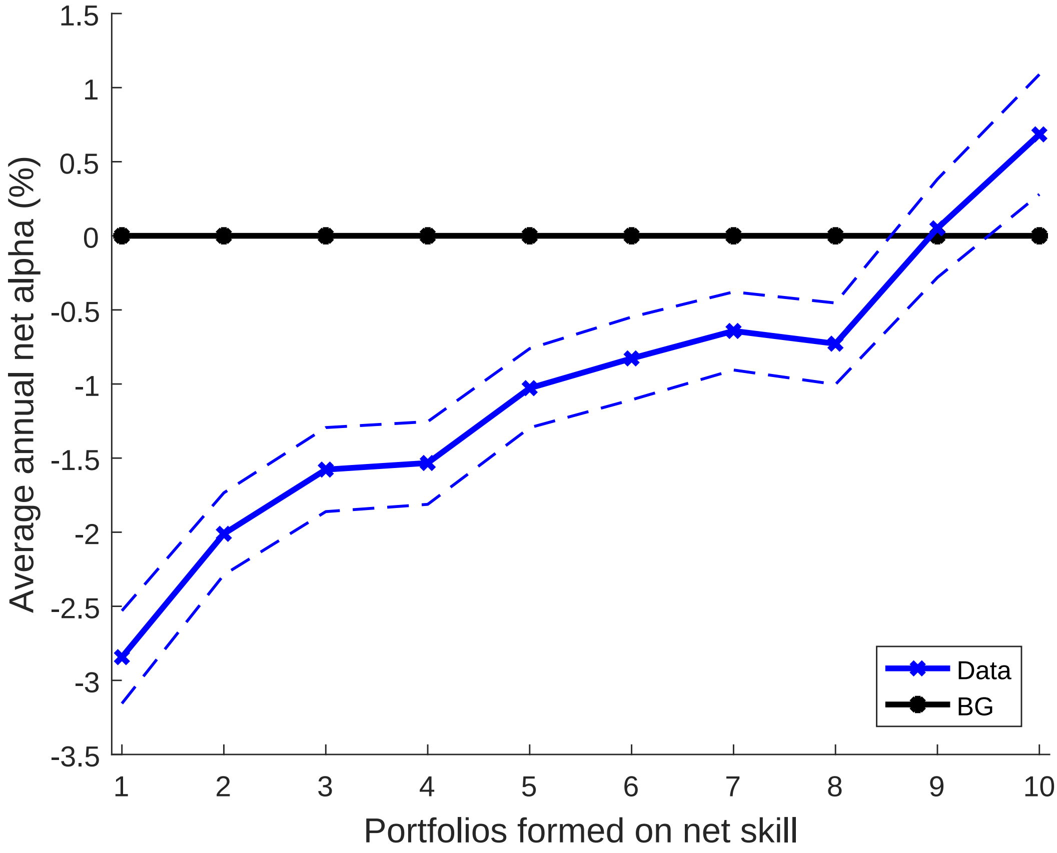
Yan, Xuemin, “Liquidity, investment style, and the relation between fund size and fund performance,” *Journal of Financial and Quantitative Analysis*, 2008, pp. 741–767.

Figure 1: Capital (mis)Allocation in Mutual Funds: Size vs. Net Skill



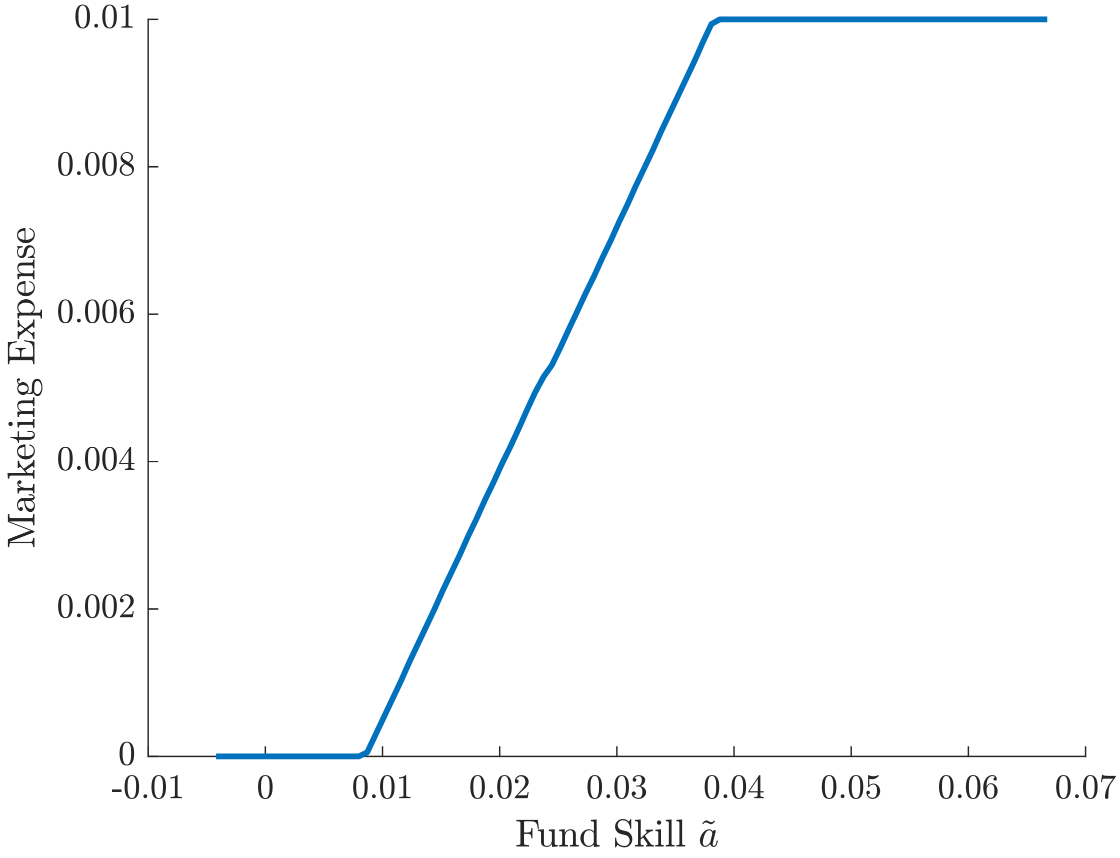
The figure plots the mean of log fund size (fund size is measured in millions of dollars) for portfolios of funds formed on net skill (defined as posterior estimate of fundamental skill level $\tilde{a}$ minus expense ratio p). We compute fund size according to the generalized version of the Berk and Green (2004) model that we estimate using the ratio between net skill and the degree of decreasing returns to scale η: $\log(s_{j,t}^{BG}) = \frac{\widetilde{a}_{j,t} - p_{j,t}}{\eta}$. The black line plots the mean of the Berk and Green model-implied fund sizes for each portfolio (BG). The blue line plots the mean of log fund size in the data for each portfolio. Portfolio 1 has the lowest net skill while portfolio 10 has the highest net skill. 95 percentile confidence bounds are indicated by dashed lines.

Figure 2: Capital (mis)Allocation in Mutual Funds: Net Alpha vs. Net Skill



The figure plots the average annual net alpha for portfolios of funds formed on net skill (defined as posterior estimate of fundamental skill level $\tilde{a}$ minus expense ratio p). The black line plots the Berk and Green (2004) model-implied net alpha for each portfolio (BG). The blue line plots the mean of net alpha in the data for each portfolio. Portfolio 1 has the lowest net skill while portfolio 10 has the highest net skill. 95 percentile confidence bounds are indicated by dashed lines.

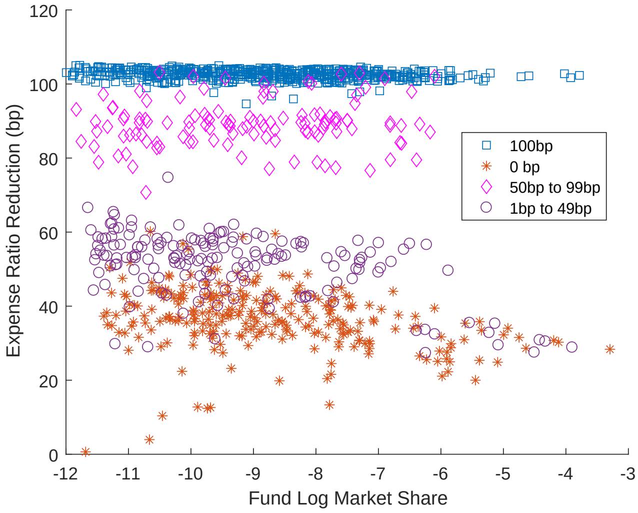
Figure 3: Marketing and Skill



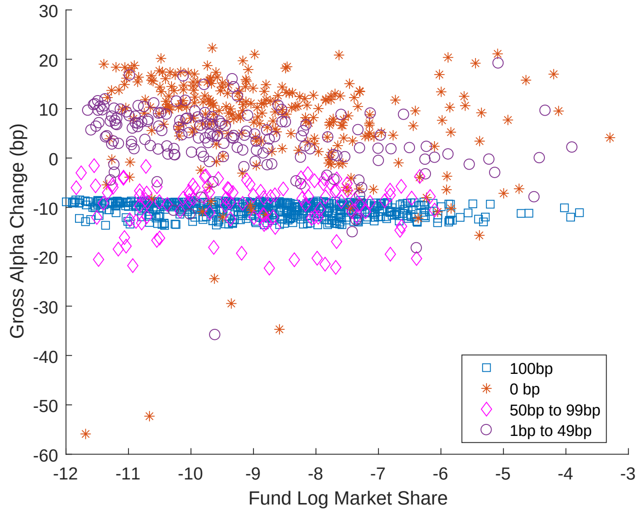
This figure plots the relationship between fund's skill $\tilde{a}$ and model implied fund's marketing expense. We introduce a hypothetical fund in year 2015. The fund has average characteristics, $\xi = 0$, $\zeta = 0$, and $\omega = 0$. We vary the posterior belief about its skill $\tilde{a}$ and calculate the associated optimal marketing expense, given the choices of the other funds observed in the data.

Figure 4: Change From Current Equilibrium to No-Marketing Equilibrium

Panel A: Expense Ratio Reduction from Current Equilibrium to No-Marketing Equilibrium



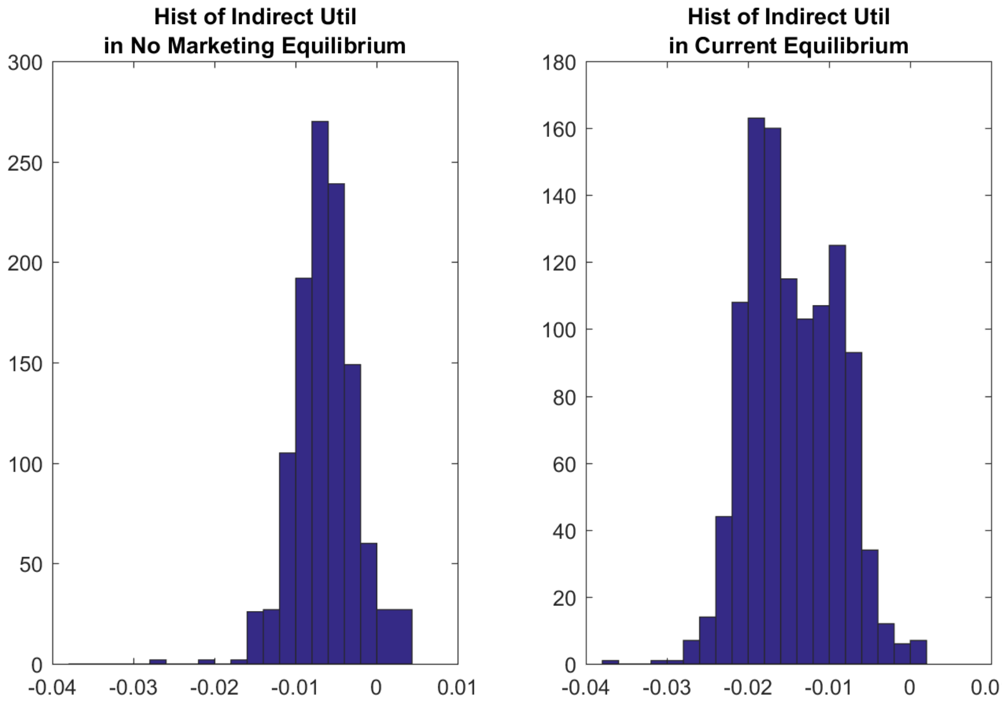
Panel B: Gross Alpha Change From Current Equilibrium to No-Marketing Equilibrium



Panel A plots the expense ratio reduction from the current equilibrium (which allows marketing) to no-marketing equilibrium. The x-axis is fund size (here we use log market share). The y-axis is the expense ratio reduction. More specifically, it is the current equilibrium price minus the no-marketing equilibrium price.

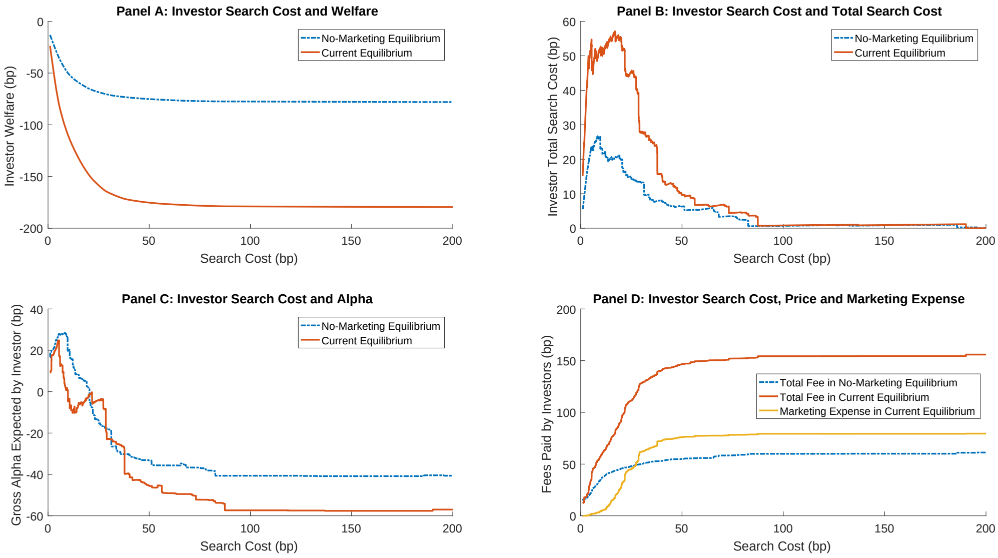
Panel B plots the gross alpha change from current equilibrium to no-marketing equilibrium. The x-axis is fund size (here we use log market share). The y-axis is the gross alpha change. More specifically, it is $\alpha^{old} - \alpha^{new}$. Old refers to the current equilibrium and new refers to no-marketing equilibrium. We split funds into 4 groups based on their old marketing expenses. Group 1, indicated by squares, has $b^{old} = 100$ bps. Group 2, indicated by asterisk, has $b^{old} = 0$ bp. Group 3, indicated by diamond, has $b^{old} \in [1, 49]$ bps. Group 4, indicated by circle, has $b^{old} \in [50, 99]$ bps.

Figure 5: Indirect Utilities: Current Equilibrium vs. No-Marketing Equilibrium



The left panel shows histogram of indirect utility for funds in no-marketing equilibrium. The right panel shows histogram of indirect utility for funds in current equilibrium (which allows marketing). The x-axis is utility level. The y-axis is frequency. Indirect utility is defined in equation (5).

Figure 6: Heterogeneous Effect Across Investors



Panel A plots the investor's welfare against investor's search cost level. The x-axis is investor's search cost. The investor's welfare is in unit of bp. The y-axis is investor's welfare defined as indirect utility provided by chosen fund minus total incurred search cost. For the expression of investor's welfare as a function of search cost, please refer to equation 17.

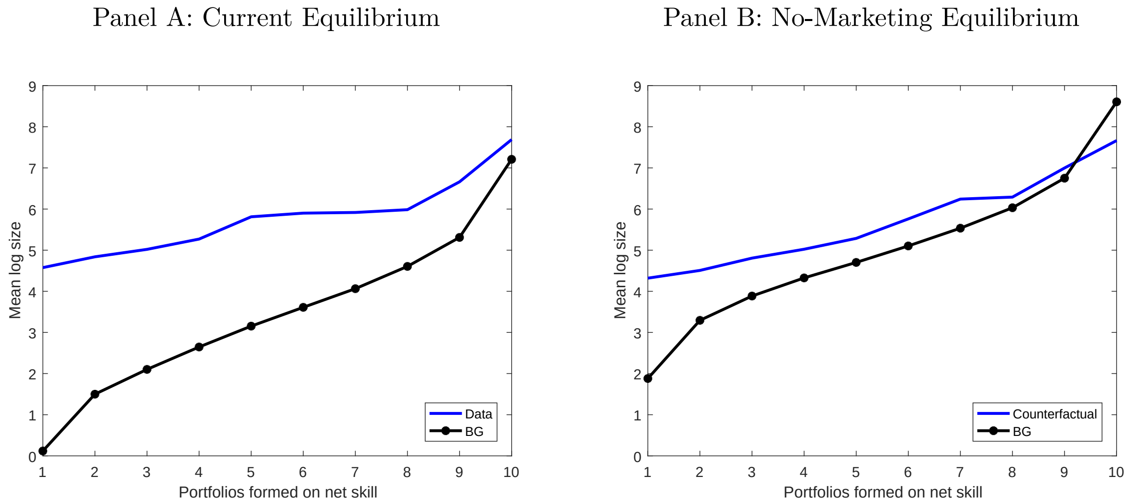
Panel B plots the investor's expected total search cost against investor's search cost level. The x-axis is investor's search cost. The y-axis is investor's total incurred search cost. Total incurred search cost is defined as $c_i \frac{\Psi[\bar{u}(c_i)]}{1-\Psi[\bar{u}(c_i)]}$, where c_i is the search cost level, $\Psi[\bar{u}(c_i)]$ is the probability of sampling a fund that delivers the investor an indirect utility smaller or equal to $\bar{u}(c_i)$. For detailed derivation of the expression, please check the appendix.

Panel C plots the gross alpha expected by investors against investor's search cost level. The x-axis is investor's search cost. The y-axis is the gross alpha that investor gets.

Panel D plots the price and marketing expenses that investors paid against investor's search cost level. The x-axis is investor's search cost. The y-axis is the fees that investor paid.

Solid line stands for the current equilibrium and dash line stands for no-marketing equilibrium. For details on investor search process, please refer to section 2.2.

Figure 7: Capital (mis)Allocation: Counterfactual



Panel A plots the mean of log fund size (fund size is measured in millions of dollars) for portfolio of funds formed on net skill (defined as fund skill level $\tilde{a}$ minus expense ratio p) for the current equilibrium, using data on mutual funds in the year 2015.

Panel B plots the mean of log fund size (fund size is measured in millions of dollars) for portfolio of funds formed on net skill for the no-marketing equilibrium. The expense ratio is outcome of the counterfactual experiment.

We compute the Berk and Green model implied fund size using the ratio between net skill and degree of decreasing returns to scale. The black line plots the mean of log fund size for each portfolio implied by the Berk and Green (2004) model. The blue line plots the mean of log fund size in the data for each portfolio. We construct ten portfolios based on the deciles of net skill. Portfolio 1 has the lowest net skill while portfolio 10 has the highest net skill.

Table 1: Investor Beliefs and Manager Skill (BG)

Parameters	Description	1964-2015
η	Decreasing returns to scale (bp)	48
		(4)
μ	Mean of prior (%)	3.05
		(0.25)
κ	SD of prior (%)	2.41
		(0.12)
δ	SD of realized alpha (%)	7.62
		(0.05)
ρ	Skill persistence	0.94
		(0.02)
LL		1.12
Num of Obs		27,621

This table presents the estimates of the fund performance related parameters. The standard errors are in the parentheses. η is the decreasing returns to scale parameter. μ is the mean of manager's *ex ante* skill distribution. κ is the standard deviation of this distribution. δ is the standard deviation of the idiosyncratic noise in the realized alpha. ρ is the persistence of the manager's skill. LL stands for log likelihood.

Table 2: Search Model Parameters

		(1)	(2)	(3)	(4)
Parameters	Description	Interior	Lower	Upper	All
λ	Mean search cost (bp)	39	39	39	39
		(4)	(4)	(4)	(4)
γ	Alpha coef	0.415	0.415	0.416	0.415
		(0.030)	(0.031)	(0.030)	(0.030)
θ	Marketing coef	113.11	111.22	133.18	122.56
		(7.334)	(7.291)	(8.797)	(7.397)
β_1	Number of family funds coef	0.4048	0.403	0.381	0.393
		(0.026)	(0.026)	(0.026)	(0.026)
β_2	Log fund age coef	1.032	1.032	1.032	1.032
		(0.037)	(0.038)	(0.037)	(0.036)
Year FE		Yes	Yes	Yes	Yes

This table presents the estimates of the structural search model. We use the data from 2001 to 2015. The four columns are corresponding to four sets of moment conditions as we described in the estimation section. In column (1), we use the funds that are not constrained in their marketing expenses to estimate the model. In column (2), we use the funds whose marketing expenses are 0 to estimate the model. In column (3), we use the funds whose marketing expenses are 100 bps to estimate the model. In column (4), we use all the funds to estimate the model.

Table 3: Change in Size when Marketing Expenses Increase by 1 bp

	$\theta = 113.11$	111.22	133.18
Panel A: Sort by Size			
Big Funds	0.8735	0.8904	1.043
Intermediate Size Funds	0.8794	0.8965	1.050
Small Funds	0.9085	0.9261	1.085

Panel B: Sort by Skill			
High Skill Funds	0.9670	0.9858	1.155
Intermediate Skill Funds	0.8987	0.9161	1.073
Low Skill Funds	0.8154	0.8311	0.973

Panel C: Sort by Original Marketing Expense			
Binding at Lower Bound	0.9554	0.9739	1.141
Non Binding	0.8990	0.9165	1.073
Binding at Upper Bound	0.8413	0.8575	1.004

This table provides the percentage changes in funds size for various groups of funds if marketing expense increases by 1 bp. In the table, we use parameters from Table 2 column (1) besides θ. For θ, we use the estimated value of θ from column (2), (1) and (3) respectively in Table 2. In panel A, we sort funds by size. Big funds are funds in the top 10 percentile. Small funds are funds in the bottom 10 percentile. Intermediate size funds are the rest. In panel B, we sort funds by skill level. High Skill funds are funds in the top 10 percentile. Low skill funds are funds in the bottom 10 percentile. Intermediate skill funds are the rest of the funds. In panel C, we sort funds by marketing expenses. Binding at Lower Bound funds are funds who originally choose 0 marketing expenses. Binding at Upper Bound are funds who originally choose 1% marketing expenses. Non binding funds are the rest of the funds.

Table 4: Change in profits when Marketing Expense Increase by 1 bp

	$\theta = 113.11$	111.22	133.18
Panel A: Sort by Size			
Big Funds	-0.4317	-0.4150	-0.2645
Intermediate Size Funds	-0.0311	-0.0143	0.1381
Small Funds	0.0850	0.1024	0.2602

Panel B: Sort by Skill			
High Skill Funds	-0.1267	-0.1081	0.0597
Intermediate Skill Funds	-0.3358	-0.3186	-0.1634
Low Skill Funds	-0.2128	-0.1972	-0.0567

Panel C: Sort by Original Marketing Expense			
Binding at Lower Bound	-0.2100	-0.1917	-0.0263
Non Binding	-0.1722	-0.1550	0.0006
Binding at Upper Bound	-0.4180	-0.4019	-0.2569

This table provides the percentage changes in funds' profits for various groups of funds if marketing expense increases by 1 bp. In the table, we use parameters from Table 2 column (1) besides θ. For θ, we use the estimated value of θ from column (2), (1) and (3) respectively in Table 2. In panel A, we sort funds by size. Big funds are funds in the top 10 percentile. Small funds are funds in the bottom 10 percentile. Intermediate size funds are the rest. In panel B, we sort funds by skill level. High Skill funds are funds in the top 10 percentile. Low skill funds are funds in the bottom 10 percentile. Intermediate skill funds are the rest of the funds. In panel C, we sort funds by marketing expenses. Binding at Lower Bound funds are funds who originally choose 0 marketing expenses. Binding at Upper Bound are funds who originally choose 1% marketing expenses. Non binding funds are the rest of the funds.

Table 5: Quantifying the Importance of Sampling Probability Components

Specification	ξ	age	num of family funds	marketing	skill	price	R^2	Correlation between s^{BG} and s^{Model}
(1)		Y	Y	Y	Y	Y	0.5169	0.2988
(2)			Y	Y	Y	Y	0.2122	0.4988
(3)				Y	Y	Y	0.1614	0.5698
(4)					Y	Y	0.1152	0.5947
(5)	Y	Y	Y		Y	Y	0.9157	0.1456
(6)	Y	Y	Y	Y		Y	0.9008	0.0301
(7)	Y	Y	Y	Y	Y		0.9005	0.0776
Data							1	0.0901

This table presents results regarding quantifying the importance of sampling probability components. Two key measures are reported: R squared of data size on model predicted size and correlation between restricted model implied fund size and Berk and Green model implied fund size. We first compute the model predicted market share with some of the components in sampling probability being removed. The column name indicates the variable we remove. Then we regress log market share in the data onto log market share as predicted by the model and a constant. We report the R squared of each regression in the table. The data period is from 2001 to 2015.

Table 6: Summary of Outcomes for Current Equilibrium and No-Marketing Equilibrium

	Current	No-Marketing
Mean price (bp)	160.27	82.96
Mean marketing (bp)	61.29	0
Mean alpha (bp)	37.24	41.07
Total share of active funds	0.74	0.67
Mean sampling prob (%)	0.085	0.078
Sampling prob for low price funds (%)	0.042	0.14
Sampling prob for index funds (%)	5.91	13.66
Investor welfare (bp)	-140.72	-61.25
Active funds average profits (bp)	57.51	42.19
Passive funds profits (bp)	2.32	2.86
Total Welfare	-37.37	-16.20
Investor's Search Cost (bp)	29.09	12.15

This table provides various measures of the mutual fund industry under current and no-marketing equilibrium. Mean price, mean marketing and mean alpha are the arithmetic average of price, marketing expenses and alpha for all active funds, respectively. Total share of active funds is the market share of all active funds. The rest of the market share belongs to index funds. Sampling prob for low price funds is the mean sampling probability for the funds whose prices are below the mean price. Investor welfare is defined in equation (18). Active funds average profits is the mean of price minus marketing expenses for all active funds. Passive funds profits is defined similarly. Total welfare is the sum of investor welfare, funds' total profits and total marketing expenses. Investor's search cost is the average total incurred search costs.

Table 7: Summary of Outcomes for Different Search Costs

	Low λ	Mid λ	High λ
	20 bps	35 bps	39bps
Mean price (bp)	58.52	136.24	160.27
Mean marketing (bp)	0	44.78	61.29
Mean alpha (bp)	38.94	39.00	37.24
Total share of active funds	0.64	0.71	0.74
Mean sampling prob (%)	0.07	0.08	0.08
Sampling prob for low price funds (%)	0.15	0.04	0.042
Sampling prob for index funds (%)	13.66	7.16	5.91
Investor welfare (bp)	-48.42	-118.41	-140.72
Active fund average profits (bp)	31.97	51.46	57.51
Passive fund profits (bp)	3.15	2.58	2.32
Total welfare (bp)	-13.98	-33.04	-37.37
Investor's search cost (bp)	9.18	25.75	29.09

This table presents various measures of the mutual fund industry under different search costs distributions. In the top row, there are three levels of mean search costs: 20bps, 35bps and 39bps. 39 bps is our estimated value from the data. Mean price, mean marketing and mean alpha are the arithmetic average of price, marketing expenses and alpha for all active funds, respectively. Total share of active funds is the market share of all active funds. The rest of the market share belongs to index funds. Sampling prob for low price funds is the mean sampling probability for the funds whose prices are below the mean price. Investor welfare is defined in equation (18). Active funds average profits is the mean of price minus marketing expenses for all active funds. Passive funds profits is defined similarly. Total welfare is the sum of investor welfare, funds' total profits and total marketing expenses. Investor's search cost is the average total incurred search costs.

Appendix

Investor beliefs

We use the Kalman filter to derive investor belief about manager skill. Let $y_{j,t} \equiv r_{j,t} + D(s_{j,t}; \eta)$. By (1), we have

$$y_{j,t} = a_{j,t} + \varepsilon_{j,t}.$$

We can treat this as the measurement equation in a state space representation. The state equation is a simple AR(1) process for $a_{j,t}$ as specified in (2). Obtaining Equation (3) and (4) is simply a matter of applying the Kalman filter. In particular, $\tilde{a}_{j,t}$ is the one period ahead prediction of the state, and $\tilde{\sigma}_{j,t}$ is the variance of that prediction.

Optimality of cut-off strategy

Here we provide a few details on how to derive the optimal search strategy for the investors. Fix an investor in a period. For notational simplicity, we suppress the subscript i and subscript t. The Bellman equation for the dynamic problem is

$$V(u^*) = \max\left\{u^*, \quad -c + \int_{-\infty}^{+\infty} V\left(\max\{u^*, u\}\right) d\Psi(u)\right\}.$$

Consider a cutoff strategy that stops at any $u > \bar{u}$. With such a strategy, $V(u^*) = u^*$ for all $u^* > \bar{u}$. On the other hand, the value for $u^* \leq \bar{u}$ should be given by

$$\begin{aligned} V(u^*) &= \sum_{t=0}^{+\infty} \Psi(\bar{u})^t \left[1 - \Psi(\bar{u})\right] \left[\frac{\int_{(\bar{u},\infty)} u d\Psi(u)}{1 - \Psi(\bar{u})} - (t+1)c\right] \\ &= \sum_{t=0}^{+\infty} \Psi(\bar{u})^t \int_{(\bar{u},\infty)} u d\Psi(u) - c\left[1 - \Psi(\bar{u})\right] \sum_{t=0}^{+\infty} \Psi(\bar{u})^t (t+1) \\ &= \frac{1}{1 - \Psi(\bar{u})} \int_{(\bar{u},\infty)} u d\Psi(u) - c\left[1 - \Psi(\bar{u})\right] \left[1 + 2\Psi(\bar{u}) + 3\Psi(\bar{u})^2 + 4\Psi(\bar{u})^3 + ...\right] \\ &= \frac{1}{1 - \Psi(\bar{u})} \left[\int_{(\bar{u},\infty)} u d\Psi(u) - c\right]. \end{aligned} \tag{22}$$

On the right side of the first line, $\Psi(\bar{u})^t \left[1 - \Psi(\bar{u})\right]$ is the probability that the investor does not stop for t periods and then stops. Multiplying this probability is the expectation of the sampled u that triggers the stop minus the incurred search costs of $t + 1$ periods.

Most importantly, notice that (22) is a constant that does not depend on u^*. In addition, we must have $V(\bar{u}) = \bar{u}$. Equating (22) with $\bar{u}$ gives us the expression for $\bar{u}$ that we gave in the main text:

$$c = \int_{(\bar{u},\infty)} (u - \bar{u}) d\Psi(u).$$

With $\bar{u}$ thus defined, the value function can be written as

$$V(u^*) = \max\{u^*, \bar{u}\}.$$

We can verify that it satisfies the Bellman equation, as for $u^* \leq \bar{u}$,

$$\begin{aligned}
-c + \int_{-\infty}^{+\infty} V\left(\max\{u^*, u\}\right) d\Psi(u) &= -c + \int_{-\infty}^{+\infty} \max\{u, \bar{u}\} d\Psi(u) \\
&= -c + \bar{u} + \int_{(\bar{u},\infty)} (u - \bar{u}) d\Psi(u) \\
&= \bar{u},
\end{aligned}$$

and for $u^* > \bar{u}$,

$$\begin{aligned}
-c + \int_{-\infty}^{+\infty} V\left(\max\{u^*, u\}\right) d\Psi(u) &= -c + \int_{-\infty}^{+\infty} \max\{u, u^*\} d\Psi(u) \\
&= -c + u^* + \int_{(u^*,\infty)} (u - u^*) d\Psi(u) \\
&< u^*.
\end{aligned}$$

Market shares

To facilitate subsequent derivations, here we define a fund-specific cutoff $f_j,\ j = 0, 1, ..., N$, where

$$f_j = \sum_{k=0}^{N} \psi_k (u_k - u_j) \cdot \mathbf{1}\{u_k > u_j\}.$$

Notice that $u_j = \bar{u}(f_j)$. So, if $c_i > f_j$, then $u_j > \bar{u}(c_i)$. In other words, if an investor's search cost is larger f_j, he will stop searching once he finds fund j. With these funds' specific cutoffs, we can derive closed-form expressions for market shares, first for the fund with the lowest utility, then for the fund with the second lowest utility, etc. Let τ be a permutation on $\{0, 1, ..., N\}$ such that $u_{\tau(0)} \leq u_{\tau(1)} \leq ... \leq u_{\tau(N)}$. As a result, $f_{\tau(0)} \geq f_{\tau(1)} \geq ... \geq f_{\tau(N)}$.

Any investor who has a search cost that is higher than $f_{\tau(0)}$ will not make a second search beyond the free search. Then among all of these investors, with $\psi_{\tau(0)}$ probability, they will find fund $\tau(0)$ (the "worst" fund). Nevertheless, they will invest in fund $\tau(0)$. No one else will invest with fund $\tau(0)$. So the market share for fund $\tau(0)$ is

$$s_{\tau(0)} = \psi_{\tau(0)} \left[1 - G(f_{\tau(0)})\right],$$

where G is the c.d.f. for the distribution of c_i in the population.

Two kinds of investors will buy fund $\tau(1)$. The first kind is the investors with $c_i > f_{\tau(0)}$ that find fund $\tau(1)$ in the free search. They have no choice but to invest. The second kind is investors with $f_{\tau(0)} \geq c_i > f_{\tau(1)}$. For these investors to invest in fund $\tau(1)$, they could have found it in the free search, or have found $\tau(0)$ in the free search and $\tau(1)$ in the second search, or have found $\tau(0)$ in the first two searches and $\tau(1)$ in the third search, and so forth. The total probability of these events is $\psi_{\tau(1)} + \psi_{\tau(0)}\psi_{\tau(1)} + \psi_{\tau(0)}^2\psi_{\tau(1)} + ... = \frac{\psi_{\tau(1)}}{1-\psi_{\tau(0)}}$. So the market share for fund $\tau(1)$ is

$$\begin{aligned} s_{\tau(1)} &= \psi_{\tau(1)} \left[1 - G(f_{\tau(0)})\right] + \frac{\psi_{\tau(1)}}{1-\psi_{\tau(0)}} [G(f_{\tau(0)}) - G(f_{\tau(1)})] \\ &= \psi_{\tau(1)} \left[1 + \frac{\psi_{\tau(0)} G(f_{\tau(0)})}{1-\psi_{\tau(0)}} - \frac{G(f_{\tau(1)})}{1-\psi_{\tau(0)}}\right]. \end{aligned}$$

We can follow this line of deduction to obtain closed-form expressions for the market shares of all funds. For $j \geq 2$,

$$\begin{aligned} s_{\tau(j)} = \psi_{\tau(j)} \Bigg[1 + \sum_{k=0}^{j-1} & \frac{\psi_{\tau(k)} G(f_{\tau(k)})}{\left(1 - \psi_{\tau(0)} - ... - \psi_{\tau(k-1)}\right)\left(1 - \psi_{\tau(0)} - ... - \psi_{\tau(k)}\right)} \\ & - \frac{G(f_{\tau(j)})}{1 - \psi_{\tau(0)} - ... - \psi_{\tau(j-1)}}\Bigg]. \end{aligned}$$

Investor welfare

The previous section provides the proof that the optimal search strategy is a cutoff strategy. In this section we compute investor i's welfare for a given search cost c_i. First we denote $\bar{u}(c_i)$ as the reservation level of utility for the investor i. Investor i will only accept the funds which provide utilities higher or equal to $\bar{u}(c_i)$, , so the expected utility for the potentially accepted funds are

$$\frac{\int_{\bar{u}(c_i)}^{+\infty} u d\Psi(u)}{1 - \Psi\left[\bar{u}(c_i)\right]}.$$

As to the search cost, the probability that the investor will conduct t searches beyond the free search is $(1-\Phi)\Phi^t$, so the expected total search cost is

$$\begin{aligned} c\left[1 - \Psi(\bar{u})\right] \sum_{t=0}^{+\infty} \Psi(\bar{u})^t t &= c\left[1 - \Psi(\bar{u})\right] \{\left[\Psi(\bar{u}) + \Psi(\bar{u})^2 + \Psi(\bar{u})^3 + ...\right] + \\ & \qquad \left[\Psi(\bar{u})^2 + \Psi(\bar{u})^3 + ...\right] + ...\} \\ &= c\left[1 - \Psi(\bar{u})\right] \{\frac{1}{1-\Psi(\bar{u})} + \frac{\Psi(\bar{u})}{1-\Psi(\bar{u})} + ...\} \\ &= c \frac{\Psi(\bar{u})}{1-\Psi(\bar{u})} \end{aligned}$$

where $\bar{u}$ is $\bar{u}(c_i)$. Combining the two parts together, we have the expression for investor's expected welfare.

Frictionless case

Here we derive the limiting case of our model when the search costs go to zero, $\lambda \to 0$. We fix a time period t and suppress the subscript t throughout the derivation.

First, notice that the active funds must provide the same utility, $u_j = u'$ for some u' for all $j \in \{1, ..., N\}$. To see this, suppose that some j has a utility that is strictly smaller than another fund. Because the investors do no incur search cost, no one will buy j. This means $s_j \to 0$, which under the log specification of the decreasing returns to scale, implies that $u_j \to +\infty$, a contradiction. By the same argument, one can show that $u' \geq -p_0$.

Let us first look at the case that $u' > -p_0$ for all $j \in \{1, ..., N\}$. The outside good will have zero market share. So

$$\sum_{j=1}^{N} s_j = 1.$$

In addition, from the utility specification (5), we have

$$s_j = e^{\frac{1}{\eta}\tilde{a}_j - \frac{1}{\eta\gamma}(p_j + u')}.$$

Putting the two above equations together, we can find the solution for u' and plug it back into the last equation to obtain:

$$s_j = \frac{e^{\frac{1}{\eta}\tilde{a}_j - \frac{1}{\eta\gamma}p_j}}{\sum_{k=1}^{N} e^{\frac{1}{\eta}\tilde{a}_k - \frac{1}{\eta\gamma}p_k}}. \tag{23}$$

Next consider the case where $u' = -p_0$. The size of an active fund will be at the point where the decreasing returns to scale drives its utility to be the same as the index fund: this is the key idea of Berk and Green (2004). From the utility specification (5), we have

$$s_j = e^{\frac{1}{\eta}\tilde{a}_j - \frac{1}{\eta\gamma}(p_j - p_0)}. \tag{24}$$

For this case, we must have $\sum_{j=1}^{N} s_j \leq 1$, which translates into

$$-p_0 \geq \eta\gamma \log\left(\sum\nolimits_{k=1}^{N} e^{\frac{1}{\eta}\tilde{a}_k - \frac{1}{\eta\gamma}p_k}\right). \tag{25}$$

In other words, if this condition on the prices holds, then the market shares are given by (24), otherwise the market shares are given by (23).

We can derive the pricing behavior of funds given these market share equations. Each fund chooses p_j to maximize $s_j(p_j - b_j - mc_j)$. Suppose that condition (25) holds so that s_j is given by (24), then the first order condition implies a uniform markup of $\eta\gamma$ across the active funds, or

$$p_j = \eta\gamma + b_j + mc_j.$$

If these prices satisfy condition (25), then we have a Nash-Bertrand equilibrium in which the index fund has a positive market share.

Uniqueness of the fixed point

In this section, we show that the fixed point defined as

$$\boldsymbol{F}_t\left[\boldsymbol{p}_t, \boldsymbol{b}_t, \tilde{\boldsymbol{a}}_t - \eta \log\left(M_t \boldsymbol{s}_t\right), \boldsymbol{x}_t, \boldsymbol{\xi}_t, p_{0,t}; \Theta\right] = \boldsymbol{s}_t$$

is unique. For notational simplicity, we suppress the subscript t. We use a result from Kennan (2001), which provides the uniqueness of a fixed point under R-concavity and the quasi-increasing condition. We first need to show that $\boldsymbol{F}\left[\boldsymbol{p}, \boldsymbol{b}, \tilde{\boldsymbol{a}} - \eta \log\left(M\boldsymbol{s}\right), \boldsymbol{x}, \boldsymbol{\xi}, p_0; \Theta\right] - \boldsymbol{s}$ as a function of $\boldsymbol{s}$

is strictly R-concave, i.e., for any $0 < z < 1$, we have

$$\boldsymbol{F}\left[\boldsymbol{p}, \boldsymbol{b}, \tilde{\boldsymbol{a}} - \eta \log\left(M\boldsymbol{s}\right), \boldsymbol{x}, \boldsymbol{\xi}, p_0; \Theta\right] > z\boldsymbol{s}. \tag{26}$$

Notice that $\tilde{\boldsymbol{a}} - \eta \log\left(zM\boldsymbol{s}\right) = \tilde{\boldsymbol{a}} - \eta \log\left(M\boldsymbol{s}\right) + \eta \log(z^{-1})$, which increases the utility for all the active funds by the same amount $\eta \log(z^{-1})$. This is equivalent to lowering the utility of the outside good (i.e., index fund) by $\eta \log(z^{-1})$. So in the following, we show that lowering the utility of the outside good increases the market share of every active fund. This will imply (26).

Recall that we have for $j = 0, 1, 2, ..., N$, the market share for $\tau(j)$ equals the summation of $j+1$ terms:

$$\begin{aligned} F_{\tau(j)} = \psi_{\tau(j)}[1 - G(f_{\tau(0)})] + \frac{\psi_{\tau(j)}}{1 - \psi_{\tau(0)}}[G(f_{\tau(0)}) - G(f_{\tau(1)})] + \\ ... + \frac{\psi_{\tau(j)}}{1 - \psi_{\tau(0)} - ... - \psi_{\tau(j-1)}}[G(f_{\tau(j-1)}) - G(f_{\tau(j)})]. \end{aligned}$$

where

$$f_j = \sum_{k=0}^{N} \psi_k (u_k - u_j) \cdot \mathbf{1}\{u_k > u_j\}.$$

Suppose that there is a small incremental on u_0. Formally, let $u_0' = u_0 + \Delta$, $u_j' = u_j$ for all $j \neq 0$. Consider the case where u_0 is not equal to any $u_j, j \neq 0$. Then we can take Δ small enough such that the ranking of $\{u_j\}_{j=0}^{N}$ and the ranking of $\{u_j'\}_{j=0}^{N}$ are identical, which means that the same permutation τ can be used. Let k be such that $\tau(k) = 0$, i.e., the index fund is ranked at the kth position. We have

$$f_j' = \begin{cases} f_j + \psi_0 \Delta, & \text{if } u_j < u_0; \\ f_j, & \text{if } u_j > u_0; \\ f_0 - (1 - \psi_{\tau(0)} - ... - \psi_{\tau(k)})\Delta & \text{if } j = 0. \end{cases}$$

Then, for a general $j \neq k$, $F_{\tau(j)}'$ is the summation of $j+1$ terms:

$$\begin{aligned} F_{\tau(j)}{}' = \psi_{\tau(j)}[1 - G(f_{\tau(0)} + \psi_0\Delta)] + \frac{\psi_{\tau(j)}}{1 - \psi_{\tau(0)}}[G(f_{\tau(0)} + \psi_0\Delta) - G(f_{\tau(1)} + \psi_0\Delta)] + ... \\ + \frac{\psi_{\tau(j)}}{1 - \psi_{\tau(0)} - ... - \psi_{\tau(k-1)}}\left[G(f_{\tau(k-1)} + \psi_0\Delta) - G\left(f_{\tau(k)} - (1 - \psi_{\tau(0)} - ... - \psi_{\tau(k)})\Delta\right)\right] \\ + \frac{\psi_{\tau(j)}}{1 - \psi_{\tau(0)} - ... - \psi_{\tau(k)}}\left[G\left(f_{\tau(k)} - (1 - \psi_{\tau(0)} - ... - \psi_{\tau(k)})\Delta\right) - G(f_{\tau(k+1)})\right] + ... \\ + \frac{\psi_{\tau(j)}}{1 - \psi_{\tau(0)} - ... - \psi_{\tau(j-1)}}[G(f_{\tau(j-1)}) - G(f_{\tau(j)})]. \end{aligned}$$

Hence,

$$\begin{aligned}\lim_{\Delta\to 0}\frac{F'_{\tau(j)}-F_{\tau(j)}}{\Delta} &= -\psi_{\tau(j)}G'(f_{\tau(0)})\psi_0+\frac{\psi_{\tau(j)}\psi_0}{1-\psi_{\tau(0)}}[G'(f_{\tau(0)})-G'(f_{\tau(1)})]+...\\ &+\frac{\psi_{\tau(j)}}{1-\psi_{\tau(0)}-...-\psi_{\tau(k-1)}}[\psi_0 G'(f_{\tau(k-1)})+G'(f_{\tau(k)})\cdot(1-\psi_{\tau(0)}-...-\psi_{\tau(k)})]\\ &+\frac{-\psi_{\tau(j)}}{1-\psi_{\tau(0)}-...-\psi_{\tau(k)}}G'(f_{\tau(k)})\cdot(1-\psi_{\tau(0)}-...-\psi_{\tau(k)}).\end{aligned}$$

Combining the last two terms, we have

$$\begin{aligned}\lim_{\Delta\to 0}\frac{F'_{\tau(j)}-F_{\tau(j)}}{\Delta} &= -\psi_{\tau(j)}\psi_0 G'(f_{\tau(0)})+\frac{\psi_{\tau(j)}\psi_0}{1-\psi_{\tau(0)}}[G'(f_{\tau(0)})-G'(f_{\tau(1)})]+...\\ &+\frac{\psi_{\tau(j)}\psi_0}{1-\psi_{\tau(0)}-...-\psi_{\tau(k-1)}}[G'(f_{\tau(k-1)})-G'(f_{\tau(k)})].\end{aligned}$$

Under the exponential specification of G, we know that (i) $G'>0$; (ii) $G'(f_{\tau(k-1)})-G'(f_{\tau(k)})<0$. With these two facts, it is easy to see that $\lim_{\Delta\to 0}\frac{F'_{\tau(j)}-F_{\tau(j)}}{\Delta}<0$. So we have essentially shown that when u_0 does not equal the utility of any other fund,

$$\frac{\partial F_j}{\partial u_0}<0,\ \forall j=1,...,N.$$

Because there are only finite points at which u_0 becomes equal to the utility of some other fund, the above result implies that F_j is strictly decreasing in u_0 for all $j\neq 0$. In words, lowering the utility of the outside good increases the market share of every active fund, which is what we started out to show.

The second condition that we need to show in order to apply the result in Kennan (2001) is that $\boldsymbol{F}[\boldsymbol{p},\boldsymbol{b},\tilde{\boldsymbol{a}}-\eta\log(M\boldsymbol{s}),\boldsymbol{x},\boldsymbol{\xi},p_0;\Theta]$, as a function of $\boldsymbol{s}$, is strictly radially quasiconcave. That is, for any $\boldsymbol{s}$ and $\boldsymbol{s}'$ where $s_j=s'_j$ but $s'_k\geq s_k$ for all $k\neq j$, we have

$$F_j\left[\boldsymbol{p},\boldsymbol{b},\tilde{\boldsymbol{a}}-\eta\log\left(M\boldsymbol{s}'\right),\boldsymbol{x},\boldsymbol{\xi},p_0;\Theta\right]\geq F_j\left[\boldsymbol{p},\boldsymbol{b},\tilde{\boldsymbol{a}}-\eta\log\left(M\boldsymbol{s}\right),\boldsymbol{x},\boldsymbol{\xi},p_0;\Theta\right].$$

In other words, we need to show that when the utilities of all but one active fund decrease, the market share of this one active fund increases. To prove this, we only need to apply a similar arguement as above to show that

$$\frac{\partial F_j}{\partial u_k}<0,\ \forall j,k=1,...,N\text{ and }j\neq k.$$

except for possibly a finite set of points.

Lastly, by Theorem 1 from Kennan (2001) we show that if a positive fixed point exists, it is unique.

Computation and estimation

Let $s_{j,t}$ be the observed share for fund j in period t. Given a set of parameters, we can find the $\boldsymbol{\xi}_t$ by matching our model predicted shares with the observed shares:

$$H_{j,t}(\boldsymbol{p}_t, \boldsymbol{b}_t, \tilde{\boldsymbol{a}}_t, \boldsymbol{x}_t, \boldsymbol{\xi}_t, p_{0,t}; \Theta) = s_{j,t}, \tag{27}$$

where $\tilde{\boldsymbol{a}}_t$ is obtained from (3) using the parameter values estimated in Section 4.1. Solving for $\boldsymbol{\xi}_t$ can be done in a similar fashion as the contraction mapping in Berry et al. (1995). However, because $H_{j,t}$ requires fixed-point iteration to evaluate, this is computationally costly. Instead, solving for $\boldsymbol{\xi}_t$ from

$$F_{j,t}\left[\boldsymbol{p}_t, \boldsymbol{b}_t, \tilde{\boldsymbol{a}}_t - \eta \log(M_t \boldsymbol{s}_t), \boldsymbol{x}_t, \boldsymbol{\xi}_t, p_{0,t}; \Theta\right] = s_{j,t} \tag{28}$$

is generally faster, because $F_{j,t}$ has closed-form expressions as derived in Section 2.3. This amounts to plugging the observed $\boldsymbol{s}_t$ in to the left hand side and searching for the value $\boldsymbol{\xi}_t$ that makes $F_{j,t}$ equal to the observed $s_{j,t}$ for each j. Given the definition of $H_{j,t}$ in (10), solving (27) and solving (28) are equivalent.

Components of Sampling Probability

There is a way to see the impact of various components of our model on capital misallocation. We redraw Figure 1 but remove each of the components (one at a time): ξ, fund age, fund family size, and marketing expenses. The solid black line is the BG model-implied "efficient" fund size. The blue line is the data. The dashed line plots the new fund size as predicted by the the "restricted" model (in the sense of eliminating a particular heterogeneity but keeping the current equilibrium, i.e., without re-solving for all the funds' best response strategies. In the first figure, all the portfolios shift upwards in parallel. This is due to the noise introduced into fund size, which raises the log size on average (by the Jensen's inequality). In the second figure, all the portfolios' average sizes shift downwards because fund age is useful for informing investors. In the third figure we can see the counterfactual is similar to the data, this means that fund family size is not very important in affecting fund size. The last figure plots the counterfactual fund size when there is no marketing. Interestingly, we see that the dashed line becomes steeper than observed in the data and thus closer to the "efficient" allocation. This means that marketing helps less skilled funds attain larger market share than they would otherwise.

[Insert Figure A6 Here]

Standard errors

The standard errors can be computed by parametric bootstrap. The only element that we have to take as exogenous in the simulation is the existence of the funds over time (we do not have a model of entry and exit). The shocks that we need to generate include $\nu_{j,t}$, $\varepsilon_{j,t}$, $\xi_{j,t}$, $\zeta_{j,t}$, and $\omega_{j,t}$. The latter two shocks are highly correlated (as explained in Section 4.2,) and each shows persistence over time. One way to incorporate these is by using a VAR process. We can start at year $t = 1$, first take the $\tilde{a}_{j,1}$ as the prior beliefs, then compute the equilibrium prices, marketing expenses, and market shares, given the prior beliefs and a set of randomly drawn $\xi_{j,1}$'s. After this, we can move on to $t = 2$, first compute the belief $\tilde{a}_{j,2}$ based on the simulated $r_{j,1}$ and $s_{j,1}$, then compute the equilibrium given these beliefs and a set of $\xi_{j,2}$'s, and so on until the last period T. This provides us with a panel of simulated data on which we can apply our estimation algorithm. We can run Monte Carlo experiments to verify that our estimator is able to recover the "true" parameters.

Data Appendix

In this appendix, we describe our dataset construction procedure. Our raw data come from CRSP Survivor-Bias-Free US mutual fund dataset and Morningstar.

CRSP mutual fund data

In this step, we follow Berk and van Binsbergen's (hereafter BB) procedure as close as possible.

crsp_fundno and ticker mapping

We merge CRSP and Morningstar datasets based on both tickers and CUSIPs. In CRSP, the unique identifier for each fund's share classes is crsp_fundno. Our *goal* is that after the cleaning procedures: there is a one to one mapping between crsp_fundno and ticker. And there is a one to one mapping between crsp_fundno and CUSIP.

We download the annual fund summary dataset from CRSP through Wharton Reserach Data Service (WRDS). The data span 1961 Jan to Dec 2015. There are 505,073 observations.

1. Out of 505,073 observations, there are 400 observations with same {crsp_fundno, year} as other observations. These duplications all due to multiple reports in the same year. Out of the 400 observations, there are 200 distinct crsp_fundno. We keep the observation with non missing expense ratio information and delete the other one. Now we have 504,808 observations. After this step, we don't have any observations with identify crsp_fundno and year.

2. Out of 504,808 observations left, we have 86,793 obs that missing ticker. We will follow BB's steps to fill those.

3. First we identify all the unique pairs of {crsp_fundno, ticker}. Here we first delete the observations with missing tickers. Then we got 53,278 unique pairs. We find that there are 5,425 pairs of which have the same crsp_fundno but more than one ticker. We follow BB's procedure: we keep the latest ticker which is the ticker with most recent year. Then we back fill all the crsp_fundno with that ticker. That give us 2,595 unique pairs between {crsp_fundno, ticker}. Then add back the non duplicated cases, we have 50,448 unique {crsp_fundno, ticker} pairs.

4. Up to this point, for each crsp_fundno, there is only one ticker. But for each ticker there could be multiple crsp_fundnos. Now we identify the tickers that have multiple crsp_fundnos. There are actually 4,343. We follow BB to leave them as missing. Now we get 42,436 unique pairs of {crsp_fundno, ticker}.

Feature: now our dataset have the one to one mapping between crsp_fundno and ticker.

crsp_fundno and CUSIP mapping

According to Pastor, Stambaugh and Taylor (hereafter PST), CUSIP can match a lot of Morningstar funds to CRSP funds. So we also clean the CUSIP in CRSP. In general we conduct the exact same procedures as we did with the ticker. So in the following, we only report some key statistics.

1. Out of 505,073 observations, there are 120,837 observations with missing CUSIPs. After we do the back fill, observation with missing CUSIP reduced to 29,436.
2. Next we identify the CUSIPs that has been used by multiple crsp_fundno. There are 494 such CUSIPs. We set them to missing.
3. Lastly we have 53,297 unique pairs of {crsp_fundno, cusip}.

Feature: now our dataset have the one to one mapping between crsp_fundno and CUSIPs.

We append the above two dataset together. Now a fund at least have a ticker or a CUSIP. They could have both. This leave us with 54,911 funds.

We merge this dataset to the annual dataset from CRSP.

"Effective" 12b-1 fees

For fund j in year t, if a C share class exists, we replace all the other share classes expense ratios and 12b-1 fees with the C share class's data. The C share class is the class that charges no front load fees but has higher expense ratios and 12b-1 fees. We replace other share classes expense ratios and 12b-1 fees with the C share class's data on the assumption that to mutual fund investors, since in our model all investors have the same investment horizons, whereas in the data different share classes might be tailored to the demands of investors with different horizons. If no C share class exists in the fund, then for all the other share classes, we take the sum of the share class's 12b-1 fees and the annualized front load for that share class and use it as the effective 12b-1 fees. For this case, we also increase the expense ratio by the amount of the annualized front load. Following Sirri and Tufano (1998), we annualize the front load by dividing it by 7, implicitly assuming that it is amortized over 7 years. Lastly, within a fund, across share classes, we weigh the effective 12b-1 fees by the AUM of each share class to get the fund-level effective 12b-1 fees. We do the same for the expense ratio. In Figure A2, we plot the histogram of the effective 12b-1 fees. We can see that in about 45.7% of the observations marketing expenses are at the upper bound of 1% (the cap imposed by the SEC on 12b-1 fees). About 23.7% of the observations are at the lower bound of zero.

In Figure A3 panel A, we plot the ratio of marketing expenses to total fees for active funds. We can see that this ratio is relatively stable from 1992 to 2015, at around 41%. Figure A3 panel B plots the aggregate time series of the total amount of marketing expenses in billion dollars. The mean is around 8.5 billion dollars per annum. Marketing expenses are substantial both in absolute terms and relative to the industry's total revenues.

[Insert Figure A2 Here]

[Insert Figure A3 Here]

Front Load

Since we define our C share class funds as funds that charge no front load, we need the information about fund's front load. For the front load data set we downloaded from CRSP. The total observations is 101,848.

1. In CRSP mutual fund front load dataset, for each crsp_fundno there is a pricing schedule for the front load. For each pricing schedule we only keep the maximum front load.
2. Then we delete the observations with front load equal to 0. That leave us with 19,626 observations.
3. We delete obs with front load smaller than 0. There are 30 of them.
4. There are 288 cases that a fund have more than one change in front load in one year. We choose to delete them.
5. We expand the front load dataset to a crsp_fundno year style. Because in the raw dataset, each entry have start year and end year. That gives us 108,818 entries.
6. We merge this back to the dataset generated in the above step.

Rear load

For the rear load data set we first download from CRSP. The total observartion is 151,194.

1. In CRSP mutual fund rear load dataset. For each crsp_fundno there is a pricing schedule for the rear load. For each pricing schedule we only keep the maximum front load.
2. Then we delete the obs with rear load equal 0. That leave us with 28,216 observations.
3. We delete obs with rear load smaller than 0. That delete 33 more obs.
4. There are 370 obs have more than one change in rear load in one year. We choose to delete them.
5. Now we expand the rear load dataset to a crsp_fundno year style. Because in the raw dataset, each entry have start year and end year. That gives us 162,099 obs.
6. We merge this back to the dataset generated in the above step.

Morningstar data

We start from the fund_ops file that we got from Morningstar. This dataset contains the fund_name, Morningstar category etc. It has 55,571 obs.

We first identify the non domestic well-diversified equity mutual fund. We follow the method provided in PST data appendix.

1. We first identified the observation with duplicated fund_names. And delete them. We also delete the funds with no MS category which is about additional 661 funds.

2. Then we identify the bond fund, international fund, sector fund, target date fund, real estate fund, other non-equity fund. The definition and methods are provided in PST. We attach the relevant pages in PST at the end of this file. Now we are left with 23592 funds.

3. We delete the funds with neither a ticker nor a CUSIP. We left with 21,580 funds.

For Morningstar, we utilize their information on the fund's category and whether a fund is index fund, fund family and portfolio id.

Merge between CRSP and Morningstar

In Morningstar, the unique identifier is secid. The goad is to get one to one mapping between crsp_fundno and secid. For details on secid, please check PST.

We use the CRSP dataset that have unique pairs between crsp_fundno and ticker or CUSIP to merge with Morningstar. First we merge on ticker. We get 12,412 matches.

A small issue in the Morningstar dataset is that the cusip is 9 digit while in CRSP, the CUSIP is 8 digit. Following the instruction on WRDS, we get rid of the last digit of CUSIP in Morningstar dataset.

Then we merge on CUSIP. We got 17,488 matches.

Finally we take the union of the two types of matches. We have 17,658 matches in total. Or 17,658 unique crsp_fundno and secid pairs.

We merge the above data to annual dataset from CRSP and keep the merged observations.

Correcting expense ratio, 12b-1, turn over and management fees

As mentioned in PST, the timing information for expense ratio, 12b-1 fee etc are not accurate in Morningstar. So we use CRSP dataset for those information. CRSP have 12b-1 fees information starting from 1992. We restrict our dataset to 1992 onwards. There are missing values in expense ratio, 12b-1, turn over and management fees. We want to fill in as many as possible.

So for the fund with missing value X, X can be {expense ratio, 12b-1, turn over and management fee}, we use the time series mean of the fund's X to replace the missing value. For example, if the fund miss expense ratio in 1996, we use the fund's lifetime average expense ratio to fill in the missing value in 1996.

Also for the X, we set -99 to missing and get the time series mean of it and replace the missing ones.

Follow the literature, our final dataset is at the fund level not share class level. For a lot of funds, there are many share classes. Different share classes are corresponded to different fees structure. For example, usually A charge front loads and a lower expense ratio. C share charges not loads but higher 12b-1 fee. To make the aggregation of 12b-1 fees reasonable, we make the following treatment. If a fund has a C share class, in a given year, then we set all the other share classes' 12b-1 fees and expense ratio to C share's 12b-1 fees and expense ratio. This treatment is based on the assumption that across different share classes, mutual funds should spend same amount in marketing. An alternative way would be to annualized all the loads and add them back to the 12b-1 fees before aggregating the share classes. The results are not very different if we use this method instead.

For the funds with no C share classes in a given year, we annualize the front load with 7 years and add it back to 12b-1 fee. In this case, we also increase the expense ratio by the amount of annualized front load.

Finally since there is a cap on 12b-1 fees at 1%, we set the upper bound of 12b-1 fees in our dataset to 1%.

We keep the observation with expense ratio smaller than 10 % and larger than 10 basis points. We also require expense ratio 5 basis point larger than 12b-1 fee.

Correct TNA and Return

As pointed out in PST, before 1993, a lot of the funds in CRSP dataset report their assets under management at quarterly or even annual frequency. But most of the funds report their return at monthly frequency. When we aggregate monthly returns across all the share classes, we need the monthly tna information. So we do the following correction.

1. For the funds who report their tna at quarterly frequency, we replace the missing value of tna with the tna in that quarter. For example, if fund report tna at month 3 for quarter 1, we replace month 1 and month 2 tna as month 3 tna.

2. For the funds who report tna at annual frequency, we replace month 1 to month 11 tna as month 12's tna.

3. If there is no tna information for any month in a year. We delete this year.

After this correction, we have 2,018,242 observations with non missing monthly return and tna.

General Cleaning

We drop the institutional shares which are identified in the following ways: 1 CRSP inst_fund is 'Y'. 2 fund name contains 'Institutional Shares', 'Institutional Class', 'Inst'.

We merge the annual data set developed above with monthly dataset on fund's return, assets under management from CRSP.

Following PST, we exclude fund/month observations with expense ratios below 10 basis points per year, since it is extremely unlikely that any actively managed funds would charge such low fees. We exclude observations with lagged fund size below $15 million.

Identifying Index Funds

In order to identify index funds, we use a simple two steps procedures following PST:

1. If either CRSP or Morningstar indicate an index fund, we label this fund as index fund.

2. If a fund's name contains words 'Index' or 'index', we label it as index fund.

As a result of this procedure, we have 1,480 index funds.

Vanguard Index Fund

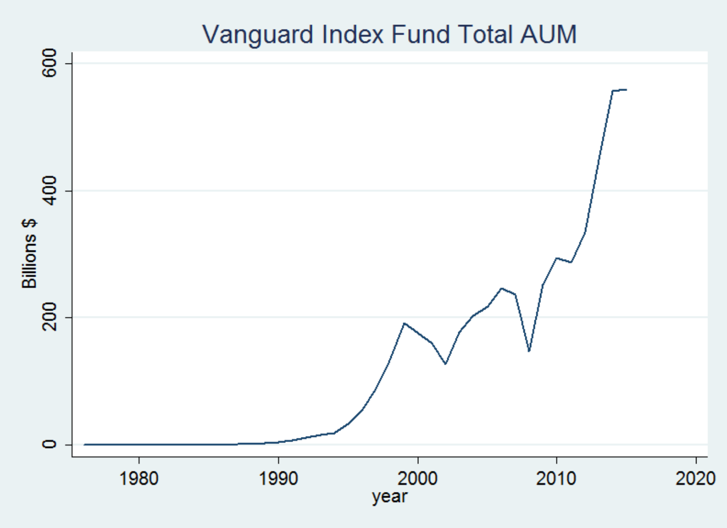
As proposed in BB, index funds from Vanguard are the most accessible index funds to the average investors. We further label whether an index fund is from Vanguard by checking with the fund name or the management firm name contains 'Vanguard'. If it does, we label it as Vanguard index fund.

In the final dataset, we use all of the index funds from Vanguard, combined, as the outside good. In each month, we get the total asset under management, asset weighted mean of management fee, returns, expense ratios, turnover ratios, and 12b-1 fees. Then for each month we only keep one observation for the Vanguard index fund.

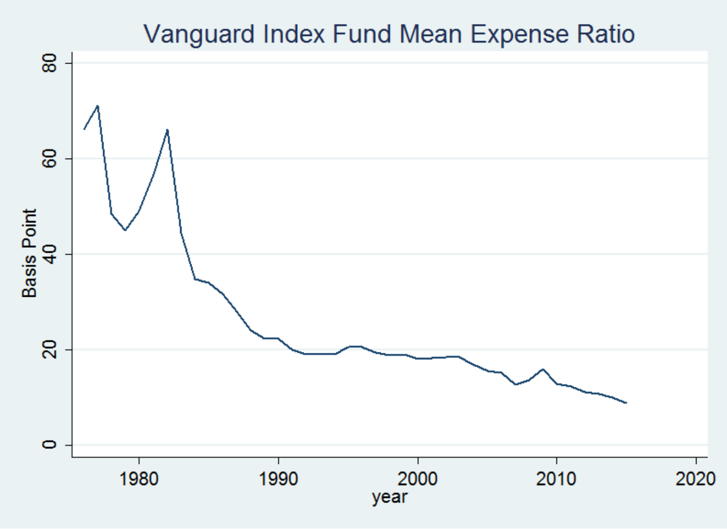
A Additional figures and tables

Figure A1: AUM and Price of Vanguard Index Funds

Panel A: Total Asset Under Management of all the Vanguard Equity Index Funds



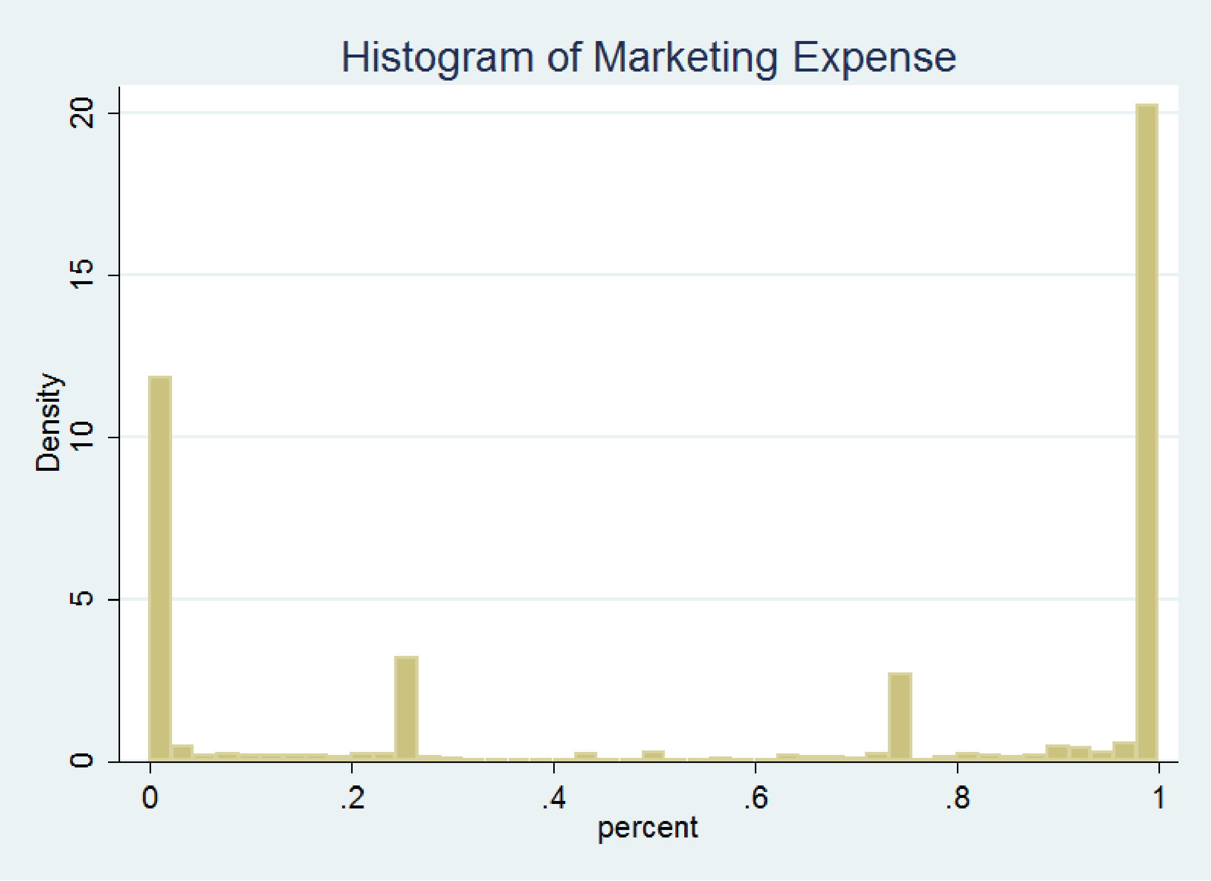
Panel B: Value Weighted Mean Expense Ratio of Vanguard Equity Index Fund



Panel A plots the total asset under management (AUM) of all the Vanguard equity index funds. The data series starts from 1975 when Vanguard launching its S&P 500 index fund. In year 2015, the total AUM reaches 600 billions of dollars. We find out all Vanguard funds by fund names (contains 'Vanguard' or 'vanguard'). We manually screen out the non equity index funds by checking fund's prospectus. The funds' asset under management are inflated to 2015 dollars by using Consumer Price Index downloaded from FRED.

Panel B plots the asset weighted mean expense ratio for all the Vanguard equity index funds. The unit is in basis point. This series covers from 1975 to 2015. We can see a significant drop from over 60 bp in 1975 to under 10 bp in 2015.

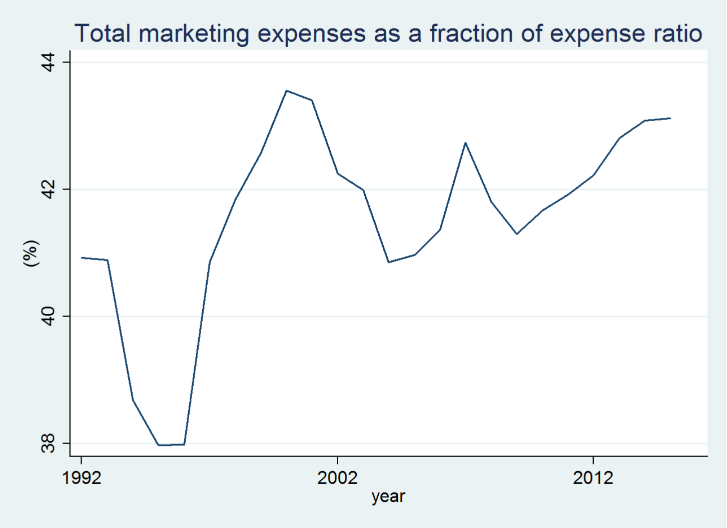
Figure A2: Histogram of Effective Marketing Expenses



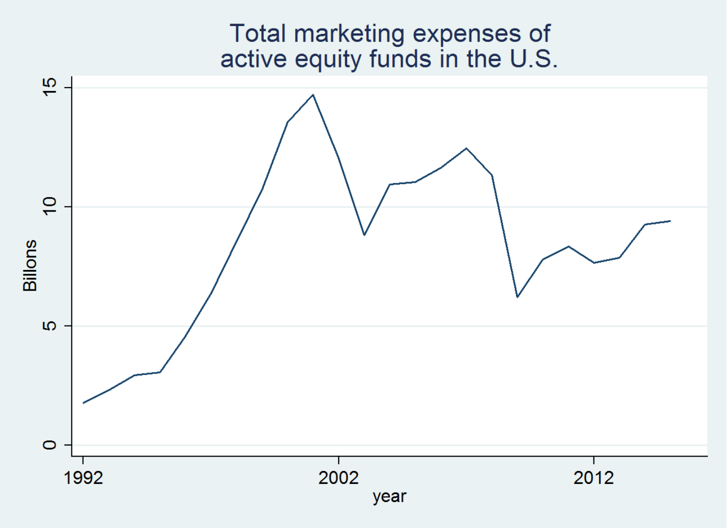
The figure plots the histogram of effective marketing expenses for the main sample covering 1964 to 2015. We define the marketing expense in the following way: For fund j in year t , if a C share class exists, we replace all the other share classes expense ratios and 12b-1 fees with the C share class's data. If no C share class exists in the fund, then for all the other share classes, we take the sum of the share class's 12b-1 fee and the annualized front load for that share class and use it as the effective 12b-1 fee. For this case, we also increase the expense ratio by the amount of the annualized front load. Lastly, within a fund, across share classes, we aggregate the effective 12b-1 fee by the AUM of each share class to get the fund level effective 12b-1 fee. About 45.7% of the observations are binding at the upper bound, 1% level. And about 23.7% of the observations are binding at 0%.

Figure A3: Total Marketing

Panel A: Total marketing expenses as a fraction of expense ratio for active funds



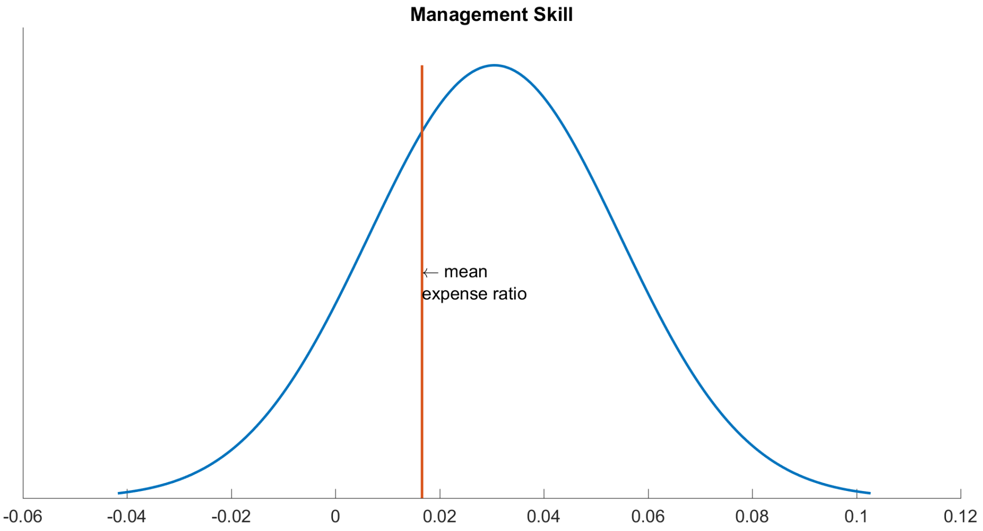
Panel B: Total marketing expenses for the active funds



Panel A plots the ratio between total marketing expenses and total expense ratios for active funds.

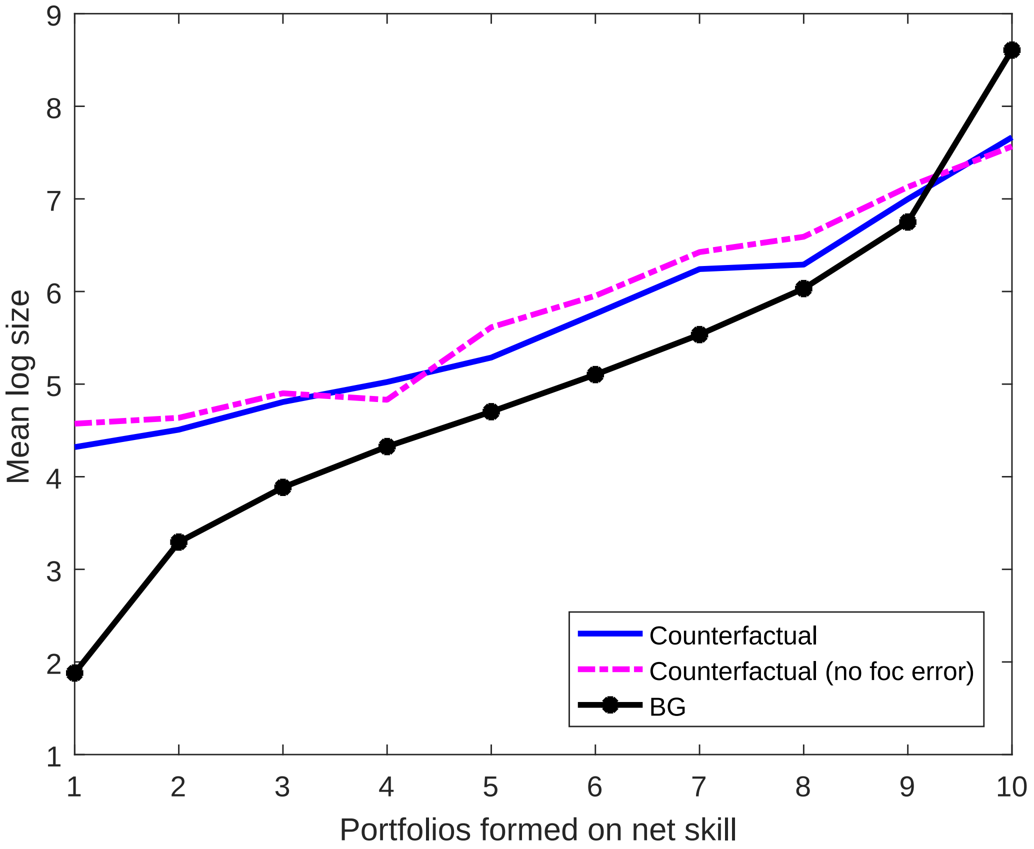
Panel B plots the total marketing expenses (in billions dollars) for all the active equity mutual funds in the U.S. from 1992 to 2015. The funds' asset under management are inflated to 2015 dollars by using Consumer Price Index downloaded from FRED.

Figure A4: Prior Distribution of Management Skill



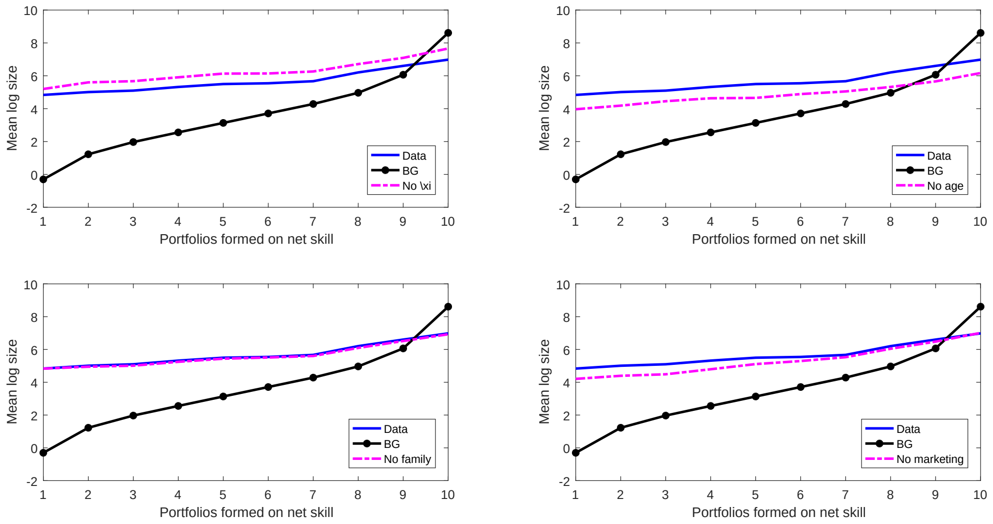
The figure presents the prior distribution of management skill (a). The vertical line marks the mean of expense ratio in our data, 1.66%. Approximately 71% of funds have management skills higher than the mean expense ratio. The parameter values are $\mu = 0.0305$ (mean of prior) and $\kappa = 0.0241$ (std of prior).

Figure A5: Capital (mis)Allocation in the No-Marketing Equilibrium: Size vs. Net Skill



The figure plots the mean of log fund size (fund size is measured in million dollars) for portfolio of funds formed on net skill for the no-marketing equilibrium. The expense ratio is outcome of the counterfactual experiment. We compute fund size implied by the generalized Berk and Green (2004) model using the ratio between net skill and the degree of decreasing returns to scale (BG). The black line plots the mean of log fund size implied by BG. The blue line plots the mean of log fund size generated by our search model in the counterfactual equilibrium for each portfolio. The pink line plots the mean of log fund size generated by our search model in the counterfactual equilibrium with *no foc errors* for each portfolio.

Figure A6: Counterfactual (restricted model)



The four figure plots the mean of log fund size (fund size is measured in million dollars) for portfolio of funds formed on net skill (defined as fund skill level $\tilde{a}$ minus expense ratio p). We compute the Berk and Green model implied fund size using the ratio between net skill and degree of decreasing returns to scale. The fund skills are estimated using our generalized Berk and Green model. For more details, please check section 2.1. The black line plots the mean of log Berk and Green model implied fund size for each portfolio. The blue line plots the mean of log fund size in the data for each portfolio. The purple dash line plots the mean of log restricted model implied fund size for each portfolio. In top left figure, purple line plots fund size when there is no ξ. In top right figure, purple line plots fund size when there is no age. In bottom left figure, purple line plots fund size when there is no fund family size. In bottom right figure, purple line plots fund size when there is no marketing. The ten portfolios are formed on net skill. Portfolio 1 has the lowest net skill while portfolio 10 has the highest net skill.

Table A1: Data Definition

Variable	Definition
Fund AUM	Fund's total net asset under management at the beginning of each year, in unit of millions of dollars
Fund expense ratio	The ratio between operating expenses that shareholders pay to the fund and the fund's AUM
Actual 12b1	Reported as the ratio of the AUM attributed to marketing and distribution costs
Management fee	The ratio of the AUM attributed to fund management costs
Fund turnover	Minimum (of aggregated sales or aggregated purchases of securities), divided by the average 12-month AUM of the fund
Total market	Sum of all funds' AUM including both active funds and index fund
Market share	Ratio between fund's AUM and total market in the same year
Age	Number of years fund is in the sample prior to given year
Family size	Number of funds in the same fund family
CAPM α	Outperformance estimated by CAPM
FF3 α	Outperformance estimated by Fama French 3 factor model
FFC α	Outperformance estimated by Fama French and Carhart model
FF5 α	Outperformance estimated by Fama French 5 factor model
New	Dummy which equals 1 if fund is new in the current period
Index fund price	Fund expense ratio of the index fund

This table presents the data definition of all the variables used in the paper. For detailed data construction process, please check the data appendix.

Table A2: Summary Statistics

				Percentiles		
	Num of Obs	Mean	Stdev	25%	50%	75%
FF5 α (%)	27,621	0.54	7.98	-3.47	0.07	3.79
Fund AUM (million \$)	27,621	1339	4791	82	254	886
Fund exp ratio (%)	27,621	1.66	0.53	1.23	1.75	2.05
Marketing expense (%)	27,621	0.61	0.44	0.01	0.89	1.00
Market share	27,621	0.0018	0.0066	0.0001	0.0002	0.0009
Age	27,621	11.46	10.3	4	8	16
New dummy	27,621	0.0827	0.2755	0	0	0
Family size	27,621	12.08	13.15	3	7	17
Index fund price (%)	27,621	0.17	0.09	0.13	0.17	0.19
Total market AUM (trillion \$)	27,621	1.54	0.75	1.26	1.77	2.13
Family AUM (million \$)	27,621	27,826	77,700	729	4,920	15,787
FFC α (%)	27,621	0.55	7.86	-3.41	0.25	3.97
FF3 α (%)	27,621	0.65	8.13	-3.39	0.25	4.09
CAPM α (%)	27,621	0.97	9.68	-3.76	0.45	4.92

This table presents summary statistics for our sample of U.S. equity mutual funds. For detailed variable definitions see table A1. The sample period is from 1964 to 2015. Our unit of observation is fund/year.

Table A3: Summary of Outcomes for Current Equilibrium and No-Marketing Equilibrium

	Current	No-Marketing	No-Marketing (no foc errors)
Mean price (bp)	160.27	82.96	79.85
Mean marketing (bp)	61.29	0	0
Mean alpha (bp)	37.24	41.07	34.55
Total share of active funds	0.74	0.67	0.66
Mean sampling prob (%)	0.085	0.078	0.078
Sampling prob for low price funds (%)	0.042	0.14	0.20
Sampling prob for index funds (%)	5.91	13.66	13.66
Investor welfare (bp)	-140.72	-61.25	-66.26
Active funds average profit (bp)	57.51	42.19	48.45
Passive funds average profit (bp)	2.32	2.86	3.01
Total Welfare	-37.37	-16.20	-14.79
Investor's Search Cost (bp)	29.09	12.15	10.48

This table provides various measures of the mutual fund industry under current and no marketing equilibrium. Additionally, we provide those measures for the no marketing equilibrium with no foc errors. Mean price, mean marketing and mean alpha are the arithmetic average of price, marketing expenses and alpha for all active funds, respectively. Total share of active funds is the market share of all active funds. The rest of the market share belongs to index funds. Sampling prob for low price funds is the mean sampling probability for the funds whose prices are below the mean price. Investor welfare is defined in equation 18. Active funds average profit is the mean of price minus marketing expenses for all active funds. Passive funds average profit is defined similarly. Total welfare is the sum of investor welfare, funds' total profits and total marketing expenses. Investor's search cost is the average total incurred search costs.