Marketing Mutual Funds*

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

September 12, 2018

Abstract

Marketing and distribution expenses are responsible for about a third of the cost of active management in the mutual fund industry. We develop and estimate a structural model of mutual fund marketing with learning about unobserved skill and costly investor search. Our estimates suggest that marketing is nearly as important as performance and fees for determining fund size. Eliminating marketing substantially improves welfare, as capital shifts towards cheaper funds and competition decreases fees. Average alpha increases as active funds shrink, and capital allocation becomes more closely aligned with manager skill net of fees. Declining investor search costs over time imply a reduction in marketing expenses and management fees as well as a shift towards passive investing, as observed empirically.

Keywords: marketing, advertising, 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, Tony Cookson, Mark Egan, Rich Evans, Joao Gomes, Marcin Kacperczyk, Ron Kaniel, Juhani Linnainmaa, Dmitry Livdan, Gregor Matvos, Igor Makarov, David Musto, Stijn van Nieuwerburgh, Amit Seru, Jesse Shapiro, Clemens Sialm, Kent Smetters, Rob Stambaugh, Luke Taylor, Jessica Wachter, and Eric Zitzewitz, as well as audiences at ANU, BU, John Cochrane's Birthday Conference at Hoover, ESSFM Gerzensee, GSU CEAR Conference, LBS, LSE, Marketing Science Conference 2017, MIT Sloan, RCFS Nassau Conference, SFS Cavalcade North America 2018, UBC Winter Conference, UT Dallas, University of Sydney, and Wharton. Authors are listed in alphabetical order.

†The Wharton School, University of Pennsylvania and NBER

‡Guanghua School of Management, Peking University

§Marshall School of Business, University of Southern California

1 Introduction

In 2015, U.S. mutual fund industry managed a total of $16 trillion of investor assets, generating fee revenue on the order of $100 billion.[1] Historically, about one third of this revenue has been designated as expenditures associated with marketing, largely consisting of sales loads and distribution costs (known as 12b-1 fees). Positive relationship between funds' marketing efforts and investor flows is well-documented (e.g., Sirri and Tufano 1998, Barber, Odean, and Zhang 2005, Gallaher, Kaniel, and Starks 2006, Reuter and Zitzewitz 2006, Bergstresser, Chalmers, and Tufano 2009, Chalmers and Reuter 2012, 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?[2] 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, based on the Bayesian forecast of skill net of fees, are actually smaller than is "efficient," and thus are able to outperform.[3] 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 that capture funds' "brand value," 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 our structural 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. Even though marketing in our model is informative, and indeed complementary with fund skill, counterfactual analysis indicates that it reduces welfare due to the positional arms race that it entails.

Following Hortaçsu and Syverson (2004), we model impediments to optimal allocation of

[1] Berk and Green (2004), Chen, Hong, Huang, and Kubik (2004), Pollet and Wilson (2008), Pástor and Stambaugh (2012), Berk and van Binsbergen (2015), Pástor, Stambaugh and Taylor (2015), among others, focus on the relationship between mutual fund size, performance, and fees. Understanding investor demand for (active) asset management is central to a broader debate about the size and value added of the finance industry as a whole, e.g. French (2008), Cochrane (2013), Greenwood and Scharfstein (2013), Malkiel (2013), Philippon and Reshef (2013), Gennaioli, Shleifer and Vishny (2014), and Philippon (2015).

[2] For example, Stigler (1961) argued that advertising serves to reduce information costs for consumers, thus helping reduce price dispersion. While commissions/kickbacks to financial intermediaries constitute a much larger portion of mutual fund marketing expenses than does advertising, it might seem reasonable to extend his classic logic to this more general notion of marketing, whereby it lowers effective search costs for imperfectly informed investors. In contrast, Mullainathan, Schwartzstein and Shleifer (2008) interpret empirical patterns of mutual fund advertising using a model of marketing that is purely persuasive, rather than informative.

[3] Following Jensen (1968), the bulk of 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).

capital to mutual funds as a search friction, whereby investors randomly sample and evaluate funds until deciding to invest in one of the funds drawn. Heterogeneity in search costs faced by investors captures the wide variation in financial sophistication (and perhaps even cognitive ability) required to consider and analyze the different investment alternatives. 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.[4] Mutual fund performance is determined by managerial skill as well as decreasing returns to scale. Investors care about funds' expected performance and expense ratios, but must learn about latent fund skill from past performance. Hence, our model nests Berk and Green (2004) as a limiting special case where search costs go to zero. Our key innovation is allowing mutual funds to influence the likelihood of being observed by investors through costly 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. 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 search further (in practice, this could mean investors who lack financial sophistication following a broker's recommendation without questioning it, consulting other sources, etc.). In comparison, Hortaçsu and Syverson (2004) estimate the mean search cost for an average investor of S&P 500 index fund between 11 and 20 basis points. This difference 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 relatively low-cost passive funds). 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 narrow confines of the S&P 500 index fund sector.

In order to verify that our estimates are not overly influenced by the assumption of vertical differentiation that we adopt from Berk and Green (2004) – that is, the assumption that investors only care about the fund's (expected) abnormal return and expense ratio and there is no heterogeneity across investors – we estimate an extended model that allows for horizontal differentiation. Specifically, we consider the possibility that different investors pursue funds that follow different investment "styles" or factors. Accordingly, in the model, we allow some investors to only sample funds from a particular Morningstar Style Box. We find that allowing

[4]Indeed, Bronnenberg, Dubé, Gentzkow and Shapiro (2015) find that informed retail consumers are less likely to pay the brand premium that is associated with heavily-marketed products. In the mutual fund context, Choi, Laibson and Madrian (2010) show experimentally that financial literacy helps investors avoid high-cost index funds.

for this important dimension of heterogeneity only slightly reduces the estimated magnitudes of investors' search costs, as well as their sensitivity to fund fees.

Our baseline 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 roughly 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 fund's size, 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. The degree of this complementarity between skill and marketing depends on the effectiveness of marketing expenses in determining the likelihood that the fund will be sampled by investors, which is a key structural parameter that we estimate. Using the aggregate demand curves implied by our estimates we find that marketing expenses alone can explain close to 10% of the variation in mutual fund size; this explanatory power is comparable to either fund manager skill or fund price.

While our model suggests that it is optimal for more highly skilled funds to invest more in marketing, this does not imply that the resulting allocation of capital to funds is efficient. We 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. Specifically, we simulate the impact of preventing funds from doing any marketing by solving for the new equilibrium size distribution and funds' fees choices using the estimated model parameters. We find that if the marketing is set to zero (instead of being capped at one percent of assets, as per current regulation), 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 result indicates that preventing funds from competing on non-price attributes (such as marketing) significantly intensifies 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 the decrease in fund size. The increase in investor welfare is substantially greater than the decline in aggregate mutual fund profits (or even total fee revenue). Thus, in the aggregate marketing reduces welfare. The reason for this inefficiency is that funds engage in a wasteful "arms race" as they compete for the same pool of investors via marketing. For the parameter values that we estimate this effect dominates the positive effect of complementarity between marketing and skill, even though in general this need not be true in our model.

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.[5]

In addition, we examine the impact of search costs on equilibrium market outcomes. With the advancement in information technology and the emergence of services enabling more transparent comparison between funds, we would expect the search frictions to decline 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 the mean search cost decreases from 39 bps to 35 bps, marketing expenses drop from 61 bps to 44 bps, on average. However, when the 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. In the model 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 are willing to invest with high fee funds, justifying these funds' marketing expenses. But when mean search cost drops to a sufficiently low level, this strategy is no longer effective, as more and more investors are willing to search for a "better" (in particular, cheaper) fund, intensifying price competition. Sufficiently low search costs render marketing unprofitable, since it becomes easy for most investors to find low-cost funds that don't expend resources on marketing themselves. Thus 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).

While we abstract from the specifics of the interaction between a mutual fund, its sales force (such as brokers and financial advisors), and individual investors, our model of marketing effort is motivated by a growing literature examining the role of retail 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, at least in terms of portfolio returns. 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-fee funds yielding 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 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

[5]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. However, empirical estimates by Linnainmaa, Melzer, and Previtero (2018) suggest that net gains to financial advice that trade off certainty equivalent benefit of increased risky asset holdings against the cost of advise are likely to be small.

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 (2018b), who show that financial advisors tend to commit common investment mistakes in their own portfolios. Choi and Robertson (2018) present survey-based evidence that the belief in the ability of active managers to deliver above-market returns and the recommendation of a financial advisor are by far the most important drivers of mutual fund investment decisions by individual investors.

More closely related to our work, Hastings, Hortaçsu and Syverson (2016) study the role of marketing efforts on observed market outcomes in Mexico's privatized retirement savings system. 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 observed 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. Jiang and Xiaolan (2018) show that a larger share of marketing/sales employees in a fund family labor force is associate with a faster growth of assets under management and a greater convexity of the flow-performance relationship (i.e., a higher sensitivity of fund size to outperformance than to poor performance).

Our paper is related to the literature that aims to understand the observed underperformance of active funds given investors' uncertainty about unobserved skill of asset managers. Baks, Metrick and Wachter (2001) show that investors will find some active funds attractive even if they have skeptical prior beliefs about managerial ability. Pástor and Stambaugh (2012) develop a tractable model of the active management industry as a whole. They explain the popularity of 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.[6]

This paper is also 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.[7]

[6]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.

[7]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).

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 (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 is drawn from the stationary distribution $\mathcal{N}(\mu, \kappa^2)$. Parameter ρ captures the persistence of the skill level. 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 to the fund manager herself: it is a hidden state. Let $\widetilde{a}_{j,t}$ be investors' (and manager'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. In the special case of $\rho = 1$ these coincide with the expressions derived by Berk and Green (2004) in their Proposition 1. The difference between our updating rule and theirs is that in the Berk and Green (2004) 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 greater weight. This allows us to capture the fact that fund managers and/or their strategies change over time, and that investors might therefore *rationally* place a larger weight on recent history.[8]

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

[8]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 over-extrapolation - e.g. Bailey, Kumar and Ng (2011), Greenwood and Shleifer (2014), Bordalo, Gennaioli, La Porta and Shleifer (2017). There is also some evidence in the literature that active managers' ability to deliver alpha varies over time, in particular with the state of the economy - e.g., Kacperczyk, Van Nieuwerburgh and Veldkamp (2014).

to be zero.

Fix a time period t. Investor i pays search cost c_i to sample one fund from the distribution of funds (this distribution is known to all investors, while specific fund identities are not). The search costs are different across the population of investors and follow a continuous distribution G (which we parameterize as exponential, with its mean given by 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 of 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

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

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

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.

Importantly, marketing expenditures are only incurred by the fund when it attracts investment. This is intuitive if we think of marketing as a commission paid to a broker or advisor - while a promise of a kickback might increase the probability the fund is recommended, if the client chooses not to invest in the fund, the commission is not paid. While we do not model this intermediated relationship between investors and funds in detail, our specification of the sampling probability function is meant to capture it in a somewhat reduced form.[9] In addition, it captures other aspects of marketing that are not directly tied to the marketing expenses that reported a part of the fund expense ratio, such as "brand," which might be captured by both observed and unobserved characteristics of the fund. For example, older funds may be more prominent and have better "name recognition" among investors, where as larger fund families may benefit from economies of scale in some forms of marketing, e.g. due to fixed costs of advertising or availability of an in-house sales force.[10]

We use $\boldsymbol{p}_t$ to denote the vector that collects $p_{j,t}$ for $j = 1, ..., N$; 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.

Active funds maximize profits each period t, which 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

[9] The role of commissions/kickbacks in investment management is analyzed theoretically by Inderst and Ottaviani (2012) and Stoughton, Wu and Zechner (2011).

[10] This is consistent with evidence in Jiang and Xiaolan (2018) that larger fund families spend a higher fraction of their wage bill on employees in sales and marketing, despite charging lower expense ratios on their funds, on average.

other funds' choices $p_{-j,t}$ and $b_{-j,t}$ (as well as its own and other funds' estimated skill levels and exogenous characteristics).[11] Passive fund expense ratio $p_{0,t}$ is set exogenously. 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.

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 Pástor, Stambaugh and Taylor (2015).

To compute the annual realized alpha $r_{j,t}$, we start with monthly return data. We first augment each fund's monthly net return with 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 Pástor, 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.[12] We aggregate all index funds from Vanguard to build a single index fund. We choose Vanguard because, as argued by Berk and van Binsbergen (2015), Vanguard index funds have been the most accessible index funds for retail investors historically. 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 all included 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 (the units are millions of 2015 dollars). 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 Pástor, Stambaugh and Taylor (2015). In

[11] 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).

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

our dataset, there is a huge skewness in fund's AUM. From the summary statistics, we can see the mean of funds' 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, 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 funds. The second summation is across all periods in which fund j exists. We maximize this likelihood with respect to μ, κ, δ, ρ, and η to obtain the estimates. Our MLE estimation does not rely on the assumptions and structure of our search model - only on the specification of the exogenous skill process. Therefore our estimates of skill and decreasing returns to scale parameters are valid even if fund market shares are determined by some other model, for example Berk and Green (2004), which we use as a benchmark.

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 $\boldsymbol{F}_t\left[\boldsymbol{p}_t,\boldsymbol{b}_t,\widetilde{\boldsymbol{a}}_t-\eta\log\left(M_t\boldsymbol{s}_t\right),\boldsymbol{x}_t,\boldsymbol{\xi}_t,p_{0,t};\Theta\right]=\boldsymbol{s}_t$ (which is described in equation (9)) for $\boldsymbol{\xi}_t$ (given a set of parameter values and observed fund choices).[13].

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

$$\sum\nolimits_{t=1}^{T}\sum\nolimits_{j\in t}\xi_{j,t}\begin{pmatrix}\boldsymbol{x}_{j,t}\\ \widetilde{a}_{j,t}\end{pmatrix}=\boldsymbol{0}. \tag{12}$$

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

[13]The solution to this equation can be found by iteration similar to the contraction mapping approach in Berry, Levinsohn and Pakes (1995).

firms' pricing or marketing choices, for instance, Berry, Levinsohn and Pakes (1995) (hereafter BLP). The BLP price instruments that are based on competitiveness measures of a particular market (e.g., the number of products competing in the market) rely on the assumption that the product characteristics of the other firms are exogenous to the unobserved characteristic $\xi_{j,t}$. But in our mutual fund industry setting this assumption is unlikely to hold, since introduction and closure of funds by fund families can respond swiftly to shifts in investor demand for mutual funds, thus using, for example, the number of funds in the market as an instrument would violate the exclusion restriction. 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.[14] 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:

$$s_{j,t} + \left(\frac{\partial H_{j,t}}{\partial p_{j,t}} \cdot e^{\zeta_{j,t}} \right) (p_{j,t} - b_{j,t}) = 0, \tag{13}$$

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 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 $0 \leq b_{j,t} \leq \overline{b}$. We let

$$-s_{j,t} + \left(\frac{\partial H_{j,t}}{\partial b_{j,t}} \cdot e^{\omega_{j,t}} \right) (p_{j,t} - b_{j,t}) \begin{cases} \leq 0, & \text{if } b_{j,t} = 0; \\ \geq 0, & \text{if } b_{j,t} = \overline{b}; \\ = 0, & \text{otherwise.} \end{cases} \tag{14}$$

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). Indeed, unlike expense ratios, marketing expenses observed in the data tend to cluster at "round" numbers, such as 75 basis points. Some of these values are salient due to regulation

[14] 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.

(e.g., only funds with zero sales loads and 12b-1 fees that do not exceed 25 basis points can be designated as "no load"). Clearly, there is no reason that such clustering would be optimal given the objective in (11). 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).[15]

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 (13) and (14), 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,

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

$$\sum\nolimits_{t=1}^{T} \sum\nolimits_{j \in t} \omega_{j,t} = 0. \tag{16}$$

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

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

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

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

where

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

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 (16):

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

[15] 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 (13) and (14). 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.

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, $\overline{b}$, 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 robustness 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 (16) are not too severely impacted by the truncation.[16]

Our GMM estimation is just identified, since there are five unknown parameters (not counting the year fixed effects) 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 (15) and (16), respectively. We conduct this second-stage estimation in one step using the identity weighting matrix. Standard errors are estimated via parametric bootstrap (described in the Appendix), which allows for arbitrary correlation between error terms, in particular $\zeta_{j,t}$ and $\omega_{j,t}$.

5 Results

5.1 Fund performance

Table 1 reports estimates of the fund performance-related parameters using our full sample.[17] The magnitude of decreasing returns to scale parameter η is 0.0048, and it is statistically significant. Since the standard deviation of log fund size is 1.628, a one standard deviation *increase* in log fund size is associated with approximately a 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

[16] We analyze how sensitive our parameter estimates are to the violations of these two key moment conditions using the method proposed by Andrews, Gentzkow and Shapiro (2017) in the Appendix. We show that reasonable deviations from the assumption that these moment conditions hold with equality, even at the boundaries, implies negligible changes in the estimated parameter values.

[17] 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.

estimate the model using linear (rather than logarithmic) specification that is similar to Pástor, Stambaugh and Taylor (2015) and obtain estimates broadly consistent with theirs.[18]

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), who argue that the lack of performance persistence does not imply absence of fund manager skill, as long as capital flows to outperforming funds and if the resulting increase fund size erodes performance. Here we take a different approach by estimating a version of the Berk and Green (2004) 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. 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. To the extent that the skill of a mutual fund is partially due to the fund's manager, a change in the management team might affect the skill level of the fund. Fidelity Magellan's Peter Lynch is a case in point: during his tenure at the helm of the fund from 1977 to 1990 Magellan achieved 14 consecutive years of positive alpha according to our measure of performance. After Peter Lynch's departure, the fund's performance becomes less impressive, consistent with reversion towards the mean.

[Insert Table 1 Here]

5.2 Asset misallocation

Equipped with estimated parameters of the fund skill distribution, we compute the implied investor beliefs about each fund's skill level at each point in time. We can then derive corresponding fund size implied by 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

[18]There is some disagreement in the literature regarding the role of fund size in eroding performance due to decreasing returns to scale, since estimates of the latter can potentially suffer from omitted variable bias, since both fund size and observed performance are correlated with underlying fund skill, which is unobservable - e.g., see discussion in Pástor, Stambaugh and Taylor (2015) and Reuter and Zitzewitz (2015). Our estimation approach is not subject to this concern, however, since it does not rely on the correlation between fund size and performance across funds to identify η. Indeed, we verify that our estimator is able to recover the true value of this parameter, which controls the strength of decreasing returns, via Monte Carlo simulation.

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 $s_{j,t}^{BG}$, should be such that the decreasing returns to scale exactly offsets the investor belief less fund expense ratio, denoted as "net skill": $D(s_{j,t}^{BG};\eta) = \widetilde{a}_{j,t} - p_{j,t}$. So, with a log specification for $D(\cdot)$, we have

$$\log(s_{j,t}^{BG}) = \frac{\widetilde{a}_{j,t} - p_{j,t}}{\eta}. \tag{17}$$

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 the data, we construct ten portfolios of mutual funds sorted on net skill. We then compute the mean of log size in the data and in the BG model for each portfolio.[19] 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 can raise their fees and thus extract the rents associated with their ability. As a result, net alpha of these funds should be zero in the future, on average. We can test this prediction of the 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 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

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

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 (16). 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.[20] 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.[21] Our estimated average 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). This is because 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 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

[20] We also estimate our model using three sub-samples: 2001-2005, 2006-2010, 2011-2015. The results are in table A4. There is a decreasing pattern of mean search cost which is consistent with the observed growth in passive index funds over the last two decades highlighted by Stambaugh (2014).

[21] Hortaçsu and Syverson (2004) estimated two variants of their search model: one in which sampling probabilities are different across funds, and another one where they are identical. We view our model as being closer to the former. The estimation results for that version of the model are reported in Table III of their paper. The log mean search cost is around -6.17 to -6.78, which implies mean search costs ranging from 11 bps to 20 bps.

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 (16) 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 Accounting for variation in fund size

In this section, we quantify the role of individual ingredients of our search model in explaining the empirical size distribution of mutual funds. Our method is as follows: we first set a particular component in the investor utility function (5) or in the sampling probability function (6) to a constant. For example, by setting all funds' marketing expenses b to zero, we are effectively removing all the explanatory power of marketing expenses from the model. We then compute synthetic market shares using investor demand curves implied by our estimated model, as summarized in equation (9), and compare them with actual market shares. Specifically, let

$$\boldsymbol{s}_t^* = \boldsymbol{F}_t \left[\boldsymbol{p}_t^*, \boldsymbol{b}_t^*, \widetilde{\boldsymbol{a}}_t^* - \eta \log \left(M_t \boldsymbol{s}_t \right), \boldsymbol{x}_t^*, \boldsymbol{\xi}_t^*, p_{0,t}; \Theta \right], \tag{18}$$

where the arguments with the asterisks are equal to their empirical values if the variable is "included" in the specification, and otherwise set to zero (for fund skill $\tilde{a}_j$ and expense ratio p_j we use sample means when the variables are "not included" in the specification.[22]). Importantly, in this exercise we are not recomputing the whole equilibrium, since we keep other variables fixed (rather than solving for every fund's best response).

[22] An explicit recursive expression for the market share function $\boldsymbol{F}$ is provided in the Appendix.

We regress the log of market share in the data $s_{j,t}$ on these synthetic log market shares $s^*_{j,t}$ and report the R-squared of these regressions in Table 5. The lower the R-squared, the more important that argument is for explaining the size distribution in the data. Among all of them, the unobserved characteristics of the fund ξ constitute the most important component (responsible for almost half of the R-squared). This is intuitive since 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 fund 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 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 observed misallocation of capital to funds. We compute the correlation between the synthetic fund size $M_t s^*_{j,t}$ and the BG-implied fund size $s^{BG}_{j,t}$, as defined in equation (17). We can see that in the data (or, equivalently, our unrestricted model, which matches the data by construction) the correlation 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 (in the sense of BG). 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 result suggests that marketing could potentially account for at least some of the misallocation that we observe in the data. However, it is possible that if funds were not able to market they would respond in other ways that might either improve or further harm allocational efficiency. In Section 7 below we describe counterfactual experiments that fully take into account the equilibrium behavior of both investors and funds.

[Insert Table 5 Here]

6 Horizontal differentiation

We have so far assumed that mutual funds in our sample are vertically differentiated. That is, in a given year, all funds can be ranked by the utility they deliver to investors, which is based on the combination of their expected performance and fees, and any two funds with identical skills and fees are viewed by all investors as perfect substitutes. This is consistent with the assumptions of the BG model, where investors only care about expected (abnormal) returns net of fees. However, in reality funds differ on a variety of attributes that investors may not be indifferent about, other than performance and fees. Importantly, investors may have a preference for funds that provide exposure to particular risk factors (e.g., as suggested, by Cochrane 1999) or funds that pursue certain investment styles (e.g., as in Barberis and Shleifer 2003).[23] Indeed,

[23]There is also a growing literature that analyses the role of active funds in allowing investors to hedge against particular states of the world, e.g. Glode (2011), Savov (2014), Kacperczyk et al. (2014), and Polkovnichenko,

even within the diversified equity mutual fund sector most funds specialize in a certain subset of stocks defined by key characteristics (e.g., large value, small growth, mid-cap blend, etc.) This type of heterogeneity matters because it can influence our estimates of the search cost distribution; some funds that are "too large" (or "too small") relative to their observed skill might be so because they pursue a popular (or shunned) investment style, rather than due to high investor search costs.

In order to address this concern, we extend our model by treating funds in different Morningstar "style box" categories (e.g., Large Cap Blend, Large Cap Growth, Large Cap Value, Mid Cap Blend, Mid Cap Growth, Mid Cap Value, Small Cap Blend, Small Cap Growth, and Small Cap Value) as different goods. Estimating an arbitrary elasticity between these goods is not feasible in our setting, as it would not be separately identified from the search cost parameter. Instead, we consider the polar opposite to the case of perfect vertical differentiation, whereby certain subsets of investors only consider funds within a particular style box. Specifically, we assume there are K categories of funds and a certain proportion of investors Υ_k, where $k \in \{1...K\}$, only consider funds in category k. We also assume that there potentially exist some investors who consider all the funds, and we denote their proportion as Υ_0 and we have

$$\sum_{k=0}^{K} \Upsilon_k = 1. \tag{19}$$

Type-0 investors behave exactly like the investors in section 2.2. For type-k investors, all their decision rules are the same as type 0 investors. The only difference is that they consider funds from category k exclusively, in addition to the index fund. Funds in sector k offer indirect utility

$$u_{j,t} = \gamma \widetilde{r}_{j,t} - p_{j,t},$$

where

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

Type-k investors conduct sequential costly search to decide in which fund they want to invest, in the same fashion as described in section 2.2, except that their search is "directed" towards funds in sector k. Thus, we can solve for the market share of each fund within category k, $s^k_{j,t}$ as a function of the utilities $\{u_{j,t}\}_{j\in k,t}$, sampling probabilities $\{\psi_{j,t}\}_{j\in k,t}$, and the distribution of search costs $G(\cdot)$, where $j \in k, t$ indicates the fund belonging to category k during time period t. Within sector k, we have $\sum_{j\in k,t} s^k_{j,t} = 1$ where $j = 0$ denotes the index fund; we also have $\sum_{j\in k,t} \psi_{j,t} = 1$. The index fund's market share in sector k is computed as the residual of all the other funds in that sector,

$$s^k_{0,t} = 1 - \sum_{j\in k,t} s^k_{j,t} \tag{20}$$

To convert the above category-specific market share into the total market share we defined in section 2.3, we can use the following relationships. For active funds,

$$s_{j,t} = s^k_{j,t} \cdot \Upsilon_k + s^0_{j,t} \cdot \Upsilon_0; \tag{21}$$

Wei and Zhao (2012).

where $s^0_{j,t}$ denotes the market share of fund j among funds chosen by type-0 investors. For the index fund,

$$s_{0,t} = \sum_{k=0}^{K} s^k_{0,t} \cdot \Upsilon_k. \tag{22}$$

6.1 Estimation

In addition to the parameters in the search model described in section 4.2, we now have K additional parameters $\{\Upsilon_1 ... \Upsilon_K\}$ to estimate.[24] The additional moment conditions are that ξ is mean-zero within each sector. Let $j \in k, t$ denote any active fund j in category k that is alive in period t, as indicated above. The sample version of the moment conditions (for all k) is

$$\sum\nolimits_{t=1}^{T} \sum\nolimits_{j \in k,t} \xi_{j,t} = 0. \tag{23}$$

The intuition for these additional moment conditions is as follows: in the benchmark model, taking funds' utilities and sampling probabilities as given, ξ works as the residual to make model-predicted market shares equal their data counterparts. When we group funds into different sectors, as Υ_k increases, the $\{\xi_j\}_{j \in k}$ would need to decrease to match model-predicted shares with observed shares. So we can pin down Υ by imposing that ξ is mean zero within each sector. In terms of identification, Υ works like sector fixed effects or sector-specific dummies (since (23) is equivalent to including indicator functions for $j \in k, t$ in $\boldsymbol{x}_{j,t}$). Of course, in terms of the modeling, this is different than simply adding dummies for sector k in the baseline model as estimated in Section 4.2, since the specifications of market shares differ, implying different price elasticities. With heterogeneous sectors, each fund competes for investors more intensely with funds that belong to the same style box than with other funds. In the extreme case $\Upsilon_0 = 0$, a fund effectively does not compete with funds outside its style category.

Since ξ is mean-zero across all funds and now we also require ξ to be mean-zero in each sector, one of the new moment conditions is redundant. If (23) is satisfied in $K - 1$ sectors, then it will automatically satisfy for sector K. We address this under-identification problem by setting Υ_0, the proportion of investors who consider all funds exogenously. In the estimation results reported below we set $\Upsilon_0 = 0.4$ and, as a robustness check, we also consider $\Upsilon_0 = 0.2$.

6.1.1 Results

In order to accommodate the potentially changing importance of investment styles over time, we estimate the extended model over three subperiods: 2001-2005, 2006-2010, and 2010-2015. Parameter estimates are reported in table 6. Estimates of our baseline model parameters for the three corresponding subperiods are provided in table A4. After allowing for heterogeneity, the search costs are somewhat lower in each than under the baseline model estimates over the same sub-samples. This result is intuitive: search cost and preference for particular categories of funds can both help explain the fact that some 'bad' (low alpha) funds have sizable market shares, either because some investors have very high search costs so that they settle with 'bad' funds or because some investors just want to invest in certain types of funds even though these

[24] Since $\{\Upsilon_0 ... \Upsilon_K\}$ sum to 1, we only need to estimate K out of $K + 1$ parameters.

funds might underperform factor-based benchmarks. However, the differences between the two sets of search cost estimates are small: e.g., for the 2001-2005 period, it falls from 49 basis points to 43 basis points when we introduce heterogeneity in (preference towards) fund styles. This suggests that this (most natural) source of horizontal differentiation introduces at most a mild upward bias to the estimate of the average search cost under our baseline model.

Another interesting finding is that γ, the coefficient in front of performance in the investor's utility function is generally larger in the model allowing for heterogeneity. This result is also reasonable: for instance, some investors might have a preference for, say, large-cap funds, despite the fact that they might (hypothetically) exhibit lower alpha than small-cap funds, which, in turn, might have higher expense ratios, on average. Our baseline model does not account for this effect and will instead attribute the larger size of such "disproportionately expensive" funds to investors' insensitivity to performance (relative to fees). Here, too, however, the effect is relative small: when allowing for heterogeneity γ increases from 0.41 to 0.47 in the 2006-2010 subperiod (and by even less in the other subperiods).

All of the other parameter estimates remain essentially the same as in the baseline model, which is intuitive. All our results are quantitatively similar for the two exogenously set values of the residual share of investors considering all funds, $\Upsilon_0 = 0.4$ and $\Upsilon_0 = 0.2$. This fact gives us confidence that assuming vertical differentiation in our baseline model does not substantially bias our parameter estimates in the baseline model, thus allowing us to conduct meaningful counterfactual experiments. That is not to say that horizontal differentiation in itself is not interesting or important. In fact, there is quite a bit of heterogeneity in fund size along the style dimension, as evidenced by our estimates of Υ_k parameters that capture the fraction of investors considering only funds pursuing a particular style. Figure 6 displays these estimates for the nine Morningstar style box categories, as well as for the fraction considering all funds, for the two possible values that we consider. There is a consistent monotonically declining pattern of Υ_k as stock size that the funds invest in declines, indicating that investors generally prefer large-cap funds, with small-cap funds being least popular. There is less of a consistent preference for value versus growth funds, however, although the latter are the least "popular" categories within mid-cap and small-cap stocks.

[Insert Table 6 Here]

[Insert Figure 6 Here]

7 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.[25] We then investigate the impact of changing search costs on equilibrium

[25]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.

marketing expenses. We focus on the most recent year in our sample (2015) for these experiments and abstract from horizontal differentiation between funds, relying on the parameter estimates for the baseline model.

7.1 Welfare measures

In order to analyze the welfare impact of marketing we need to define welfare for all the agents in the model. Fix a year t (the time subscript t will be suppressed in this section). In our model, investor's utility consists 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

$$V(c_i) = \frac{\int_{\bar{u}(c_i)}^{+\infty} u d\Psi(u)}{1 - \Psi\left[\bar{u}(c_i)\right]} - c_i \frac{\Psi\left[\bar{u}(c_i)\right]}{1 - \Psi\left[\bar{u}(c_i)\right]}, \tag{24}$$

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 (24), 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:

$$U = \int_0^{+\infty} V(c_i) dG(c_i). \tag{25}$$

Fund profits are also part of the total welfare. These include the profits for both active funds and index funds (even though the latter are not maximizing agents in our model):

$$P = \sum_{j=1}^{N} (p_j - b_j) s_j + s_0 p_0. \tag{26}$$

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 the total size of the mutual fund market M stays the same.

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

$$B = \sum_{j=1}^{N} b_j s_j. \tag{27}$$

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

7.2 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 7 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 aggregate search cost incurred by the investors in the two equilibrium.

The aggregate search cost is given by:

$$\int_0^{+\infty} c_i \frac{\Psi\left[\bar{u}(c_i)\right]}{1-\Psi\left[\bar{u}(c_i)\right]} dG(c_i). \tag{28}$$

We find aggregate 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 her 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 7 Here]

When marketing is eliminated the size of active funds doesn't drop drastically for two reasons. First, exogenous fund characteristics in the sampling probability function ensure that all active funds sampling probabilities are positive. Second, reducing the size of active funds improves their expected performance. This effect makes active funds more attractive.

The increase in investor welfare in the counterfactual equilibrium exceeds both the decline in fund profits P and the loss of marketing "revenue" B, resulting in a total welfare increase. Thus, even though marketing in our model is purely informative, it can still be excessive from the social standpoint. To some extent, this is due to the fact that the low cost index fund is not on an equal footing with the active funds since it does not engage in marketing. But, even in the absence of the low cost index fund alternative, competition on marketing (rather than just price) generates overinvestment in marketing and, as a result, excessive fees borne by investors. This is due to the positional externality induced by the inelastic aggregate demand for funds. As an example, fund i's marketing investment could decrease fund j's probability of being known by increasing its own sampling probability. In a Nash equilibrium, funds do not take the externality into consideration when choosing their individual marketing investment levels. All of the funds might be better off if they could agree on a lower level of marketing investment. Of course, such an agreement would be fragile, since a deviation from it by any individual fund is potentially highly profitable. Hence, all fund engage in too much marketing as they compete for market share. This arms race externality is at the root of the welfare results documented above.

7.2.1 Heterogeneous effect across investors

In our model investors are heterogeneous in their ability to sample and screen mutual funds, which we model as exogenous variation in their 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 for each investor: individual 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 7 panel A shows that for all the search cost levels investors achieve a higher level of welfare in the no-marketing equilibrium. 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 7 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 7 panel C and panel D, we find that, in general, high search cost investors invest in lower alpha funds, pay high prices and, implicitly, incur high marketing expenses. This is simply because high search cost investors don't search very much. In contrast, investors who have very low search investors invest in funds that have positive net alphas. In Berk and Green's model, since all investors have zero search costs, in equilibrium, all the funds have zero net alphas. But in our model, since all investors incur a positive search cost, only 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 7 Here]

7.2.2 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 8 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 beliefs about 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 fund size computed under the counterfactual equilibrium, where as the "BG-implied" values are recomputed using the expense ratios in the counterfactual equilibrium.

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).

7.3 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.

[Insert Table 8 Here]

Thus our model 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. Indeed, Table A4 in the Appendix demonstrates that if we estimate our model over subsamples representing different time periods, average search costs decline from 49 bps in 2001-2005, to 39 bps in 2006-2010, and to 20 bps in 2011-2015.

8 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 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 market for mutual funds.

In our paper, we view mutual fund marketing expenses as purely informative (e.g., Butters 1977). It is possible that a portion of these 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). Separating the informative effect from the persuasive effect of marketing one would require information on investors' actual choice sets. 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. Relaxing this assumption in order to understand the possible welfare loss from "persuasive" marketing is a fruitful venue for future research.

References

Andrews, Isaiah, Matthew Gentzkow, and Jesse M Shapiro, "Measuring the sensitivity of parameter estimates to estimation moments," *The Quarterly Journal of Economics*, 2017, *132* (4), 1553–1592.

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

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

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.

Barberis, Nicholas and Andrei Shleifer, "Style investing," *Journal of Financial Economics*, 2003, *68* (2), 161–199.

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.

Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer, "Diagnostic Expectations and Stock Returns," Working Paper 23863, National Bureau of Economic Research September 2017.

Bronnenberg, Bart J., Jean-Pierre Dubé, Matthew Gentzkow, and Jesse M. Shapiro, "Do Pharmacists Buy Bayer? Informed Shoppers and the Brand Premium *," *The Quarterly Journal of Economics*, 2015, *130* (4), 1669–1726.

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.

Choi, James J. and Adriana Z. Robertson, "What Matters to Individual Investors? Evidence from the Horse's Mouth," 2018. Yale School of Management working paper.

____ , David Laibson, and Brigitte C. Madrian, "Why Does the Law of One Price Fail? An Experiment on Index Mutual Funds," *The Review of Financial Studies*, 2010, *23* (4), 1405–1432.

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.

Cochrane, John H, "Portfolio Advice for a Multifactor World," Working Paper 7170, National Bureau of Economic Research June 1999.

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

Egan, Mark, "Brokers vs. Retail Investors: Conflicting Interests and Dominated Products," *The Journal of Finance, Forthcoming*, 2018.

____ , Gregor Matvos, and Amit Seru, "The market for financial adviser misconduct," *Journal of Political Economy, Forthcoming*, 2018.

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.

French, Kenneth R., "Presidential Address: The Cost of Active Investing," *The Journal of Finance*, 2008, *63* (4), 1537–1573.

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," *Journal of Finance, forthcoming*, 2018.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny, "Finance and the Preservation of Wealth," *Quarterly Journal of Economics*, 2014, *129* (3), 1221–1254.

___ , ___ , and ___ , "Money doctors," *The Journal of Finance*, 2015, *70* (1), 91–114.

Glode, Vincent, "Why mutual funds "underperform"," *Journal of Financial Economics*, 2011, *99* (3), 546–559.

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

___ and David Scharfstein, "The Growth of Finance," *Journal of Economic Perspectives*, May 2013, *27* (2), 3–28.

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?," *The Review of Financial Studies*, 2017, *31* (2), 715–772.

Inderst, Roman and Marco Ottaviani, "Competition through commissions and kickbacks," *The American Economic Review*, 2012, *102* (2), 780–809.

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.

Jensen, Michael C, "The Performance of Mutual Funds in the Period 1945-1964," *Journal of Finance*, June 1968, *23* (2), 389–416.

Jiang, Wenxi and Mindy Zhang Xiaolan, "Growing Beyond Performance," *working paper*, 2018.

Kacperczyk, Marcin, Stijn Van Nieuwerburgh, and Laura Veldkamp, "Time-varying fund manager skill," *The Journal of Finance*, 2014, *69* (4), 1455–1484.

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, "Financial Advisors and Risk-Taking," *Unpublished manuscript*, 2018.

____ , ____ , and ____ , "The misguided beliefs of financial advisors," *Journal of Finance, forthcoming*, 2018.

Malkiel, Burton G., "Asset Management Fees and the Growth of Finance," *Journal of Economic Perspectives*, May 2013, *27* (2), 97–108.

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.

Philippon, Thomas, "Has the US Finance Industry Become Less Efficient? On the Theory and Measurement of Financial Intermediation," *American Economic Review*, April 2015, *105* (4), 1408–38.

___ **and Ariell Reshef**, "An International Look at the Growth of Modern Finance," *Journal of Economic Perspectives*, May 2013, *27* (2), 73–96.

Polkovnichenko, Valery, Kelsey D Wei, and Feng Zhao, "Cautious risk-takers: Investor preferences and demand for active management," 2012.

Pollet, Joshu M. and Mungo Wilson, "How Does Size Affect Mutual Fund Behavior?," *The Journal of Finance*, 2008, *63* (6), 2941–2969.

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.

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

Savov, Alexi, "The price of skill: Performance evaluation by households," *Journal of Financial Economics*, 2014, *112* (2), 213–231.

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.

Stigler, George J., "The Economics of Information," *Journal of Political Economy*, 1961, *69* (3), 213–225.

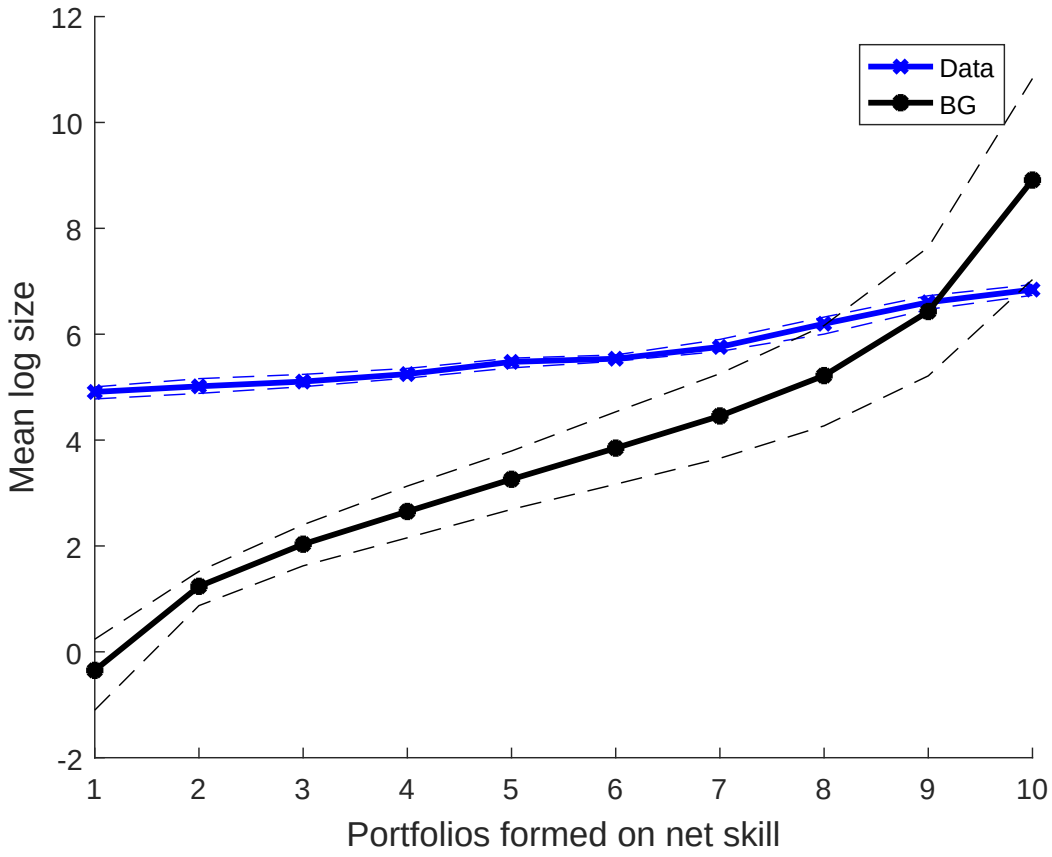
Stoughton, Neal M., Youchang Wu, and Josef Zechner, "Intermediated Investment Management," *The Journal of Finance*, 2011, *66* (3), 947–980.

van Binsbergen, Jules and Christian Opp, "Real anomalies: Are financial markets a sideshow," *Journal of Finance, forthcoming*, 2018.

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

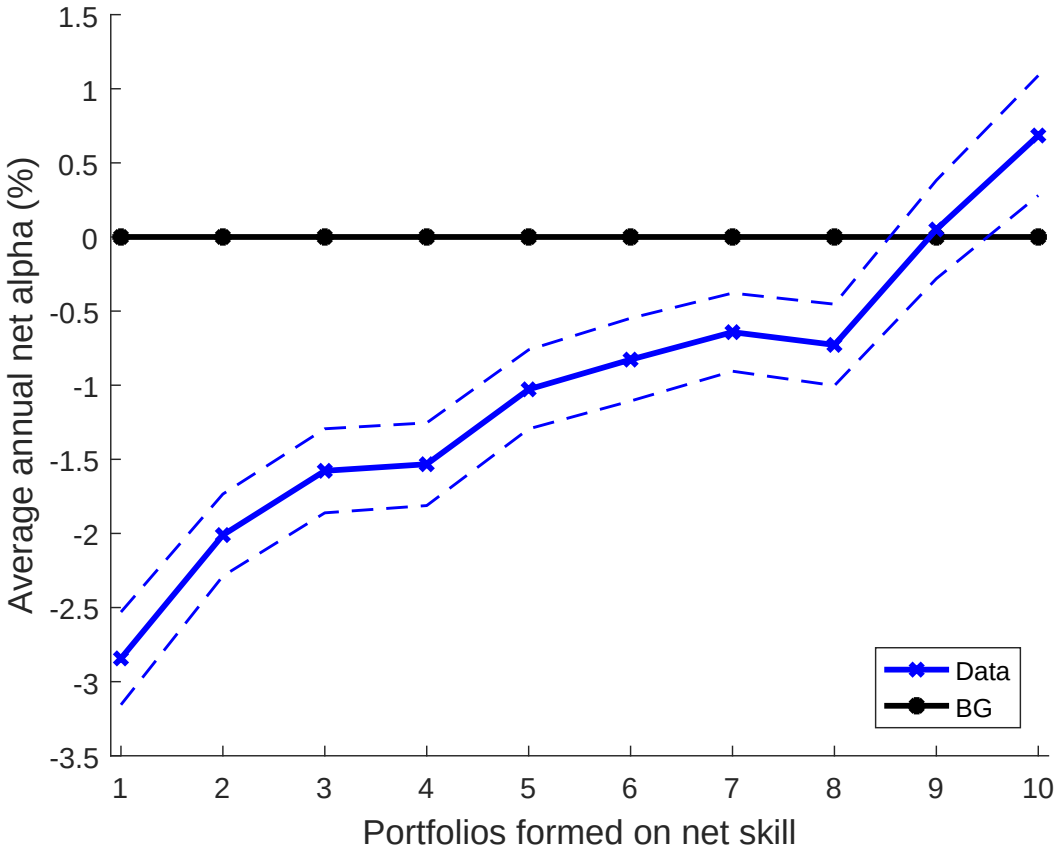
Yan, Xuemin, "Liquidity, investment style, and the relation between fund size and fund performance," *Journal of Financial and Quantitative Analysis*, 2008, pp. 741–767.

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



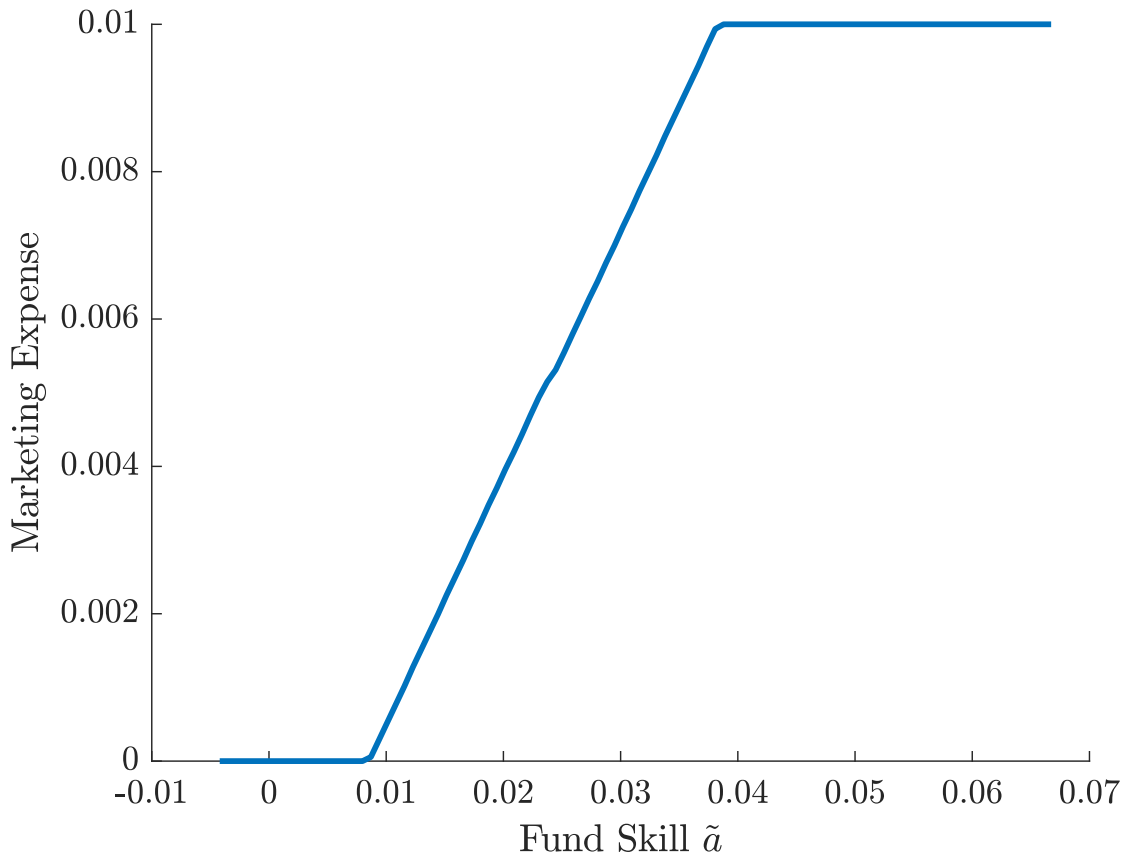
This figure plots the mean of log fund size (measured in millions of dollars) for portfolios of funds formed on net skill (defined as posterior belief about the fundamental skill level $\tilde{a}$ minus expense ratio p). We compute fund size according to the generalized version of the Berk and Green (2004) model that we estimate: $\log(s_{j,t}^{BG}) = \frac{\widetilde{a}_{j,t}-p_{j,t}}{\eta}$, where η captures decreasing returns to scale. 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



This figure plots the average annual net alpha for portfolios of funds formed on net skill (defined as posterior belief about the 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.

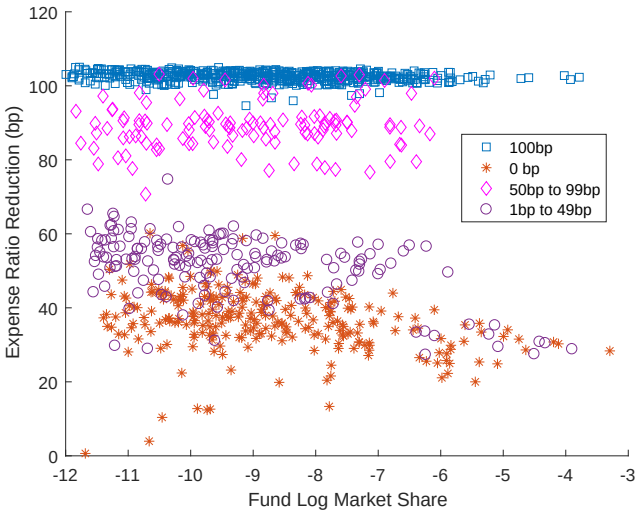
Figure 3: Marketing and Skill



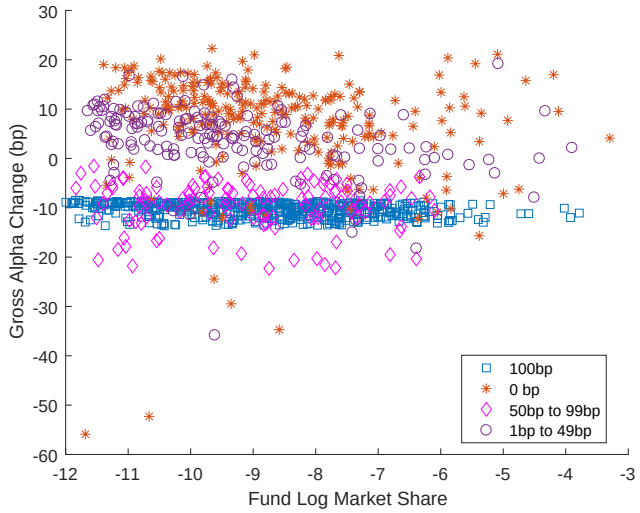
This figure plots the relationship between fund's skill $\tilde{a}$ and model-implied fund's marketing expense, for a fund in year 2015 that has average characteristics, $\xi = 0$, $\zeta = 0$, and $\omega = 0$. We vary the posterior belief about its skill $\tilde{a}$ and calculate the associated optimal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

Figure 4: Price and Performance: Current Equilibrium vs. No-Marketing Equilibrium

Panel A: Expense Ratio Reduction

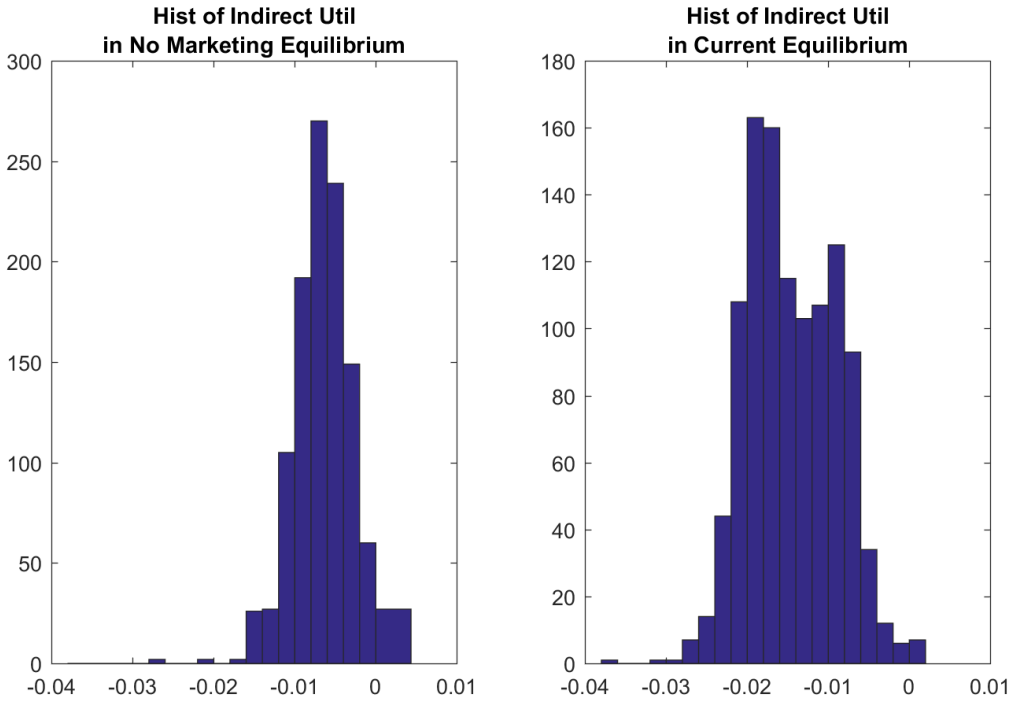


Panel B: Change in Gross Alpha



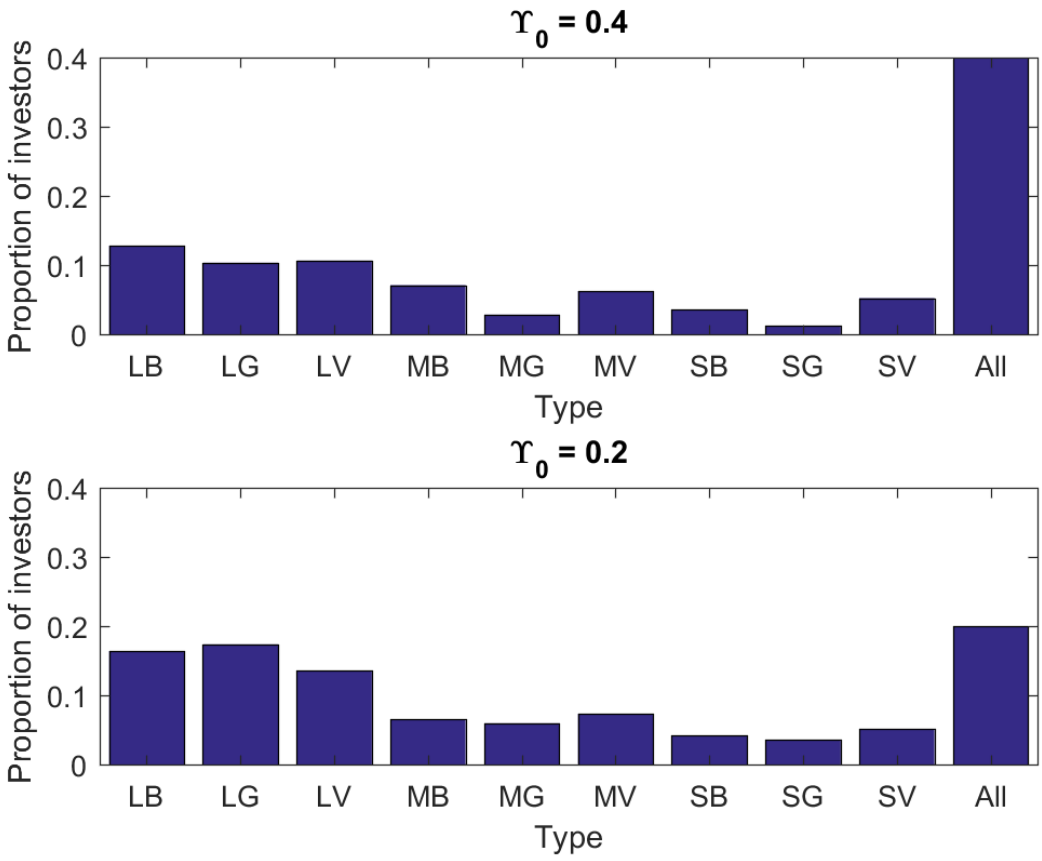
Panel A plots the expense ratio reduction as a result of moving from the current equilibrium (which allows marketing) to the counterfactual no-marketing equilibrium. The x-axis is fund size (here we use log market share). The y-axis is the current equilibrium price minus the no-marketing equilibrium price. Panel B plots the differences in gross alpha between the current equilibrium and the no-marketing equilibrium. The x-axis is fund size (here we use log market share). The y-axis is the gross alpha changes. We split funds into 4 groups based on their current marketing expenses. Group 1, indicated by squares, marketing expenses of 100 bps. Group 2, indicated by asterisk, marketing expenses of 0 bp. Group 3, indicated by diamond, marketing expenses between 1 and 49 bps. Group 4, indicated by circle, marketing expenses between 50 and 99 bps.

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



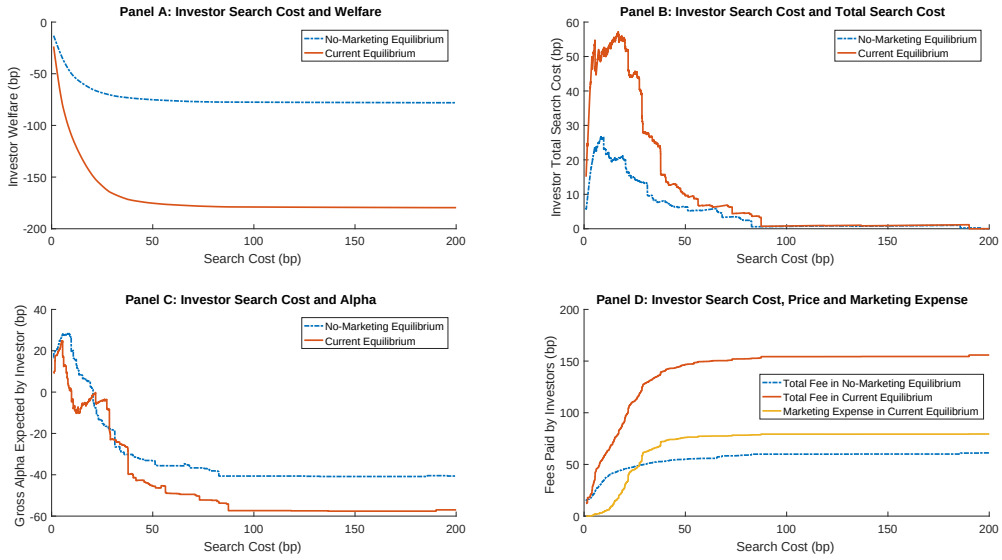
The left panel shows the histogram of indirect utilities associated with funds in the no-marketing equilibrium. The right panel shows the histogram of indirect utilities associated with funds in current equilibrium (which allows marketing). The x-axis is utility level. The y-axis is frequency. Indirect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

Figure 6: Investor Preferences for Fund Styles



The figure plots the histogram of different types of investors for the period of $2011 - 2015$ for $\Upsilon_0 = 0.4$ and $\Upsilon_0 = 0.2$, respectively. LB stands for Large Blend; LG stands for Large Growth; LV stands for Large Value; MB stands for Mid-Cap Blend; MG stands for Mid-Cap Growth; MV stands for Mid-Cap Value; SB stands for Small Blend; SG stands for Small Growth; SV stands for Small Value. All indicates the type-0 investors who consider all types of funds.

Figure 7: Investor Heterogeneity



Panel A plots investor welfare as a function of unit search cost levels. The x-axis is investor's unit search cost, c_i in basis points. The investor welfare is in unit of bp. The y-axis is investor welfare defined as indirect utility provided by chosen fund minus total incurred search cost. For the expression of investor welfare as a function of search cost, please refer to equation (24).

Panel B plots investor's expected total search cost as a function of unit search cost levels. Expected total 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, $\Psi\left[\bar{u}(c_i)\right]$ is the probability of sampling a fund that delivers the investor an indirect utility smaller or equal to $\bar{u}(c_i)$ (detailed derivation is in appendix.)

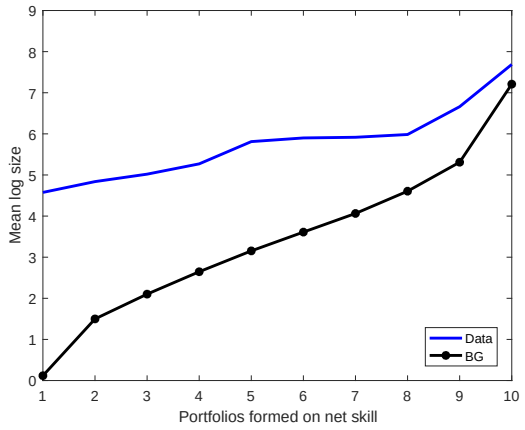
Panel C plots the gross alpha expected by investors as a function of unit search cost levels.

Panel D plots the expense ratios and marketing expenses investors incur as a function of the unit search cost levels.

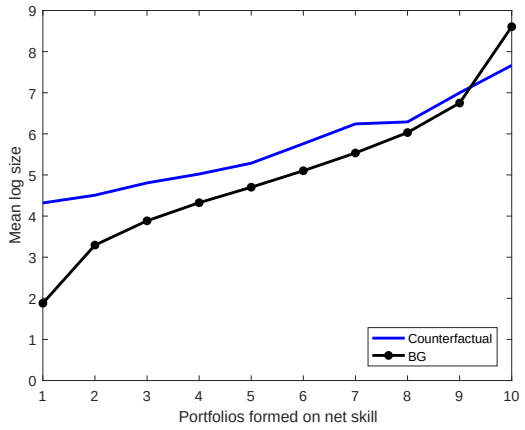
Solid line stands for the current equilibrium and dashed line stands for no-marketing equilibrium.

Figure 8: Capital (mis)Allocation: Counterfactual

Panel A: Current Equilibrium



Panel B: No-Marketing Equilibrium



Panel A plots the mean of log fund size (measured in millions of dollars) for portfolios of funds formed on net skill (defined as posterior belief about the fundamental 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 (measured in millions of dollars) for portfolios of funds formed on net skill for the no-marketing equilibrium.
We compute fund size according to the generalized version of the Berk and Green (2004) model that we estimate: $\log(s_{j,t}^{BG}) = \frac{\widetilde{a}_{j,t} - p_{j,t}}{\eta}$, where η captures 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 or in the counterfactual 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 Parameters

Parameters	Description	1964-2015
η	Decreasing returns to scale (%)	0.48
		(0.04)
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)
Num of Obs		27,621

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

Table 2: Search Model Parameters

		(1)	(2)	(3)	(4)
Parameters	Description	Interior	Lower	Upper	All
λ	Mean search cost (%)	0.39	0.39	0.39	0.39
		(0.04)	(0.04)	(0.04)	(0.04)
γ	Alpha weight	0.41	0.41	0.41	0.41
		(0.03)	(0.03)	(0.03)	(0.03)
θ	Marketing effectiveness	113.11	111.22	133.18	122.56
		(7.33)	(7.29)	(8.79)	(7.39)
β_1	Family size effect	0.40	0.40	0.38	0.39
		(0.03)	(0.03)	(0.03)	(0.03)
β_2	Fund age effect	1.03	1.03	1.03	1.03
		(0.04)	(0.04)	(0.04)	(0.04)
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 correspond to four sets of moment conditions as described in Section 4. In column (1), we use the funds that are not constrained in their marketing expenses to estimate the model. In column (2), we use only 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 of 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 fund size resulting from a 1 basis point increase in marketing expenses for various groups of funds using parameters from Table 2 column (1) except θ. For θ, we use the estimated value of θ from columns (2), (1) and (3) respectively in Table 2. In panel A, we sort funds by size. "Big Funds" are funds in the top decile. "Small Funds" are funds in the bottom decile. "Intermediate Size Funds" are the rest. In panel B, we sort funds by skill level. "High Skill Funds" are funds in the top decile. "Low Skill Funds" are funds in the bottom decile. "Intermediate Skill Funds" are the rest. In panel C, we sort funds by marketing expenses. "Binding at Lower Bound" are funds with no marketing expenses. "Binding at Upper Bound" are funds whose marketing are at the upper bound of 1%. "Non-Binding" are the rest.

Table 4: Change in Profits when Marketing Expenses 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 fund profits resulting from a 1 basis point increase in marketing expenses for various groups of funds using parameters from Table 2 column (1) except θ. For θ, we use the estimated value of θ from columns (2), (1) and (3) respectively in Table 2. In panel A, we sort funds by size. "Big Funds" are funds in the top decile. "Small Funds" are funds in the bottom decile. "Intermediate Size Funds" are the rest. In panel B, we sort funds by skill level. "High Skill Funds" are funds in the top decile. "Low Skill Funds" are funds in the bottom decile. "Intermediate Skill Funds" are the rest. In panel C, we sort funds by marketing expenses. "Binding at Lower Bound" are funds with no marketing expenses. "Binding at Upper Bound" are funds whose marketing are at the upper bound of 1%. "Non-Binding" are the rest.

Table 5: Quantifying the Importance of Search Model 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 various components of the search model in explaining the size distribution of funds. Our method is as follows: we first set a particular component in the utility function (5) or the sampling probability equation (6) equal to zero. Then we compute synthetic market shares of all funds using equation (9). 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 synthetic log market shares and report the R-squared. We also report the correlation between the synthetic fund size and the generalized Berk and Green model-implied fund size. "Y" indicates that we include this component in the sampling probability. Blank means that we remove this component. The data period is from 2001 to 2015.

Table 6: Search Model Parameters for Different Sub-Samples with Investor Heterogeneity

		$\Upsilon_0 = 0.4$			$\Upsilon_0 = 0.2$		
Parameters	Description	2001-2005	2006-2010	2011-2015	2001-2005	2006-2010	2011-2015
λ	Mean search cost (%)	0.45	0.35	0.17	0.43	0.37	0.18
		(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
γ	Alpha weight	0.42	0.47	0.56	0.44	0.47	0.56
		(0.05)	(0.05)	(0.54)	(0.05)	(0.05)	(0.54)
θ	Marketing effectiveness	112.24	125.16	143.14	113.25	125.10	143.01
		(12.66)	(12.48)	(13.21)	(12.55)	(12.88)	(13.11)
β_1	Family size effect	0.44	0.36	0.39	0.44	0.36	0.39
		(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
β_2	Fund age effect	1.08	1.07	1.10	1.08	1.07	1.10
		(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Year FE		Yes	Yes	Yes	Yes	Yes	Yes

This table presents the estimates of the structural search model for different sub-samples with investor heterogeneity. In this table, we choose Υ_0, the proportion of investors who consider all the funds to be 0.4 and 0.2. The moment conditions are described in section 6.1.

Table 7: 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 (%)	74	68
Mean sampling prob (%)	0.085	0.078
Mean sampling prob for low price funds (%)	0.042	0.14
Mean sampling prob for index fund (%)	5.91	13.66
Investor welfare (bp)	-140.72	-61.25
Active fund profits (bp)	57.51	42.19
Index fund profits (bp)	2.32	2.86
Total welfare	-37.37	-16.20
Aggregate investor 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. Mean sampling prob for low price funds is the mean sampling probability for the funds whose prices are below average. Investor welfare is defined in equation (25). Active fund profits are defined as $\sum_{j=1}^{N}(p_j - b_j)s_j$, which are the total profits for all active funds. Index fund profits are defined similarly. Total welfare is the sum of investor welfare, funds' total profits and total marketing expenses. Aggregate investor search cost is described in equation (28).

Table 8: 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 (%)	64	71	74
Mean sampling prob (%)	0.07	0.08	0.08
Mean sampling prob for low price funds (%)	0.15	0.04	0.042
Mean sampling prob for index fund (%)	13.66	7.16	5.91
Investor welfare (bp)	-48.42	-118.41	-140.72
Active fund profits (bp)	31.97	51.46	57.51
Index fund profits (bp)	3.15	2.58	2.32
Total welfare (bp)	-13.98	-33.04	-37.37
Aggregate investor 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. Mean sampling prob for low price funds is the mean sampling probability for the funds whose prices are below average. Investor welfare is defined in equation (25). Active fund profits are defined as $\sum_{j=1}^{N}(p_j - b_j)s_j$, which are the total profits for all active funds. Index fund profits are defined similarly. Total welfare is the sum of investor welfare, funds' total profits and total marketing expenses. Aggregate investor search cost is described in equation (28).

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

$$\begin{aligned}
V(u^*) &= \sum_{t=0}^{+\infty} \Psi(\bar{u})^t \left[1 - \Psi(\bar{u})\right] \left[\frac{\int_{(\bar{u},\infty)} u d\Psi(u)}{1 - \Psi(\bar{u})} - (t+1)c\right] \\
&= \sum_{t=0}^{+\infty} \Psi(\bar{u})^t \int_{(\bar{u},\infty)} u d\Psi(u) - c\left[1 - \Psi(\bar{u})\right] \sum_{t=0}^{+\infty} \Psi(\bar{u})^t (t+1) \\
&= \frac{1}{1 - \Psi(\bar{u})} \int_{(\bar{u},\infty)} u d\Psi(u) - c\left[1 - \Psi(\bar{u})\right] \left[1 + 2\Psi(\bar{u}) + 3\Psi(\bar{u})^2 + 4\Psi(\bar{u})^3 + ...\right] \\
&= \frac{1}{1 - \Psi(\bar{u})} \left[\int_{(\bar{u},\infty)} u d\Psi(u) - c\right]. \qquad (29)
\end{aligned}$$

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 (29) is a constant that does not depend on u^*. In addition, we must have $V(\bar{u}) = \bar{u}$. Equating (29) 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 $j \geq 2$,

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

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 is

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

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

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

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. Also, since M_t, the total size of the market, is exogenously given in our model, we normalize it to 1 here to simplify the notation.

First, consistently with Berk and Green (2004) intuition, all active funds must provide the same utility, $u_j = u'$. To see this, suppose that some fund j has a utility that is strictly smaller

than another fund. Because the investors do not incur search costs, no one will invest in fund 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 $u' \geq -p_0$, where $-p_0$ is the utility provided by the index fund.

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':

$$u' = \eta\gamma \log\left(\sum\nolimits_{k=1}^{N} e^{\frac{1}{\eta}\tilde{a}_k - \frac{1}{\eta\gamma}p_k}\right),$$

and plug it back into the last equation to obtain:

$$s_j = \frac{e^{\frac{1}{\eta}\tilde{a}_j - \frac{1}{\eta\gamma}p_j}}{\sum_{k=1}^{N} e^{\frac{1}{\eta}\tilde{a}_k - \frac{1}{\eta\gamma}p_k}}. \tag{30}$$

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

$$s_j = e^{\frac{1}{\eta}\tilde{a}_j - \frac{1}{\eta\gamma}(p_j - p_0)}. \tag{31}$$

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

$$-p_0 \geq \eta\gamma \log\left(\sum\nolimits_{k=1}^{N} e^{\frac{1}{\eta}\tilde{a}_k - \frac{1}{\eta\gamma}p_k}\right). \tag{32}$$

In other words, if this condition on the prices holds, then the market shares are given by (31), otherwise the market shares are given by (30).

We can derive the pricing behavior of funds given these market share equations. Each fund chooses p_j to maximize $s_j(p_j - b_j)$. Suppose that condition (32) holds so that s_j is given by (31), then the first order condition implies a uniform markup of $\eta\gamma$ across the active funds, or

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

If these prices satisfy condition (32), then we have a Nash 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

$$\boldsymbol{F}\left[\boldsymbol{p}, \boldsymbol{b}, \tilde{\boldsymbol{a}} - \eta \log\left(M\boldsymbol{s}\right), \boldsymbol{x}, \boldsymbol{\xi}, p_0; \Theta\right] > z\boldsymbol{s}. \tag{33}$$

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 (33).

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

argument 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:

$$H_{j,t}(\boldsymbol{p}_t, \boldsymbol{b}_t, \tilde{\boldsymbol{a}}_t, \boldsymbol{x}_t, \boldsymbol{\xi}_t, p_{0,t}; \Theta) = s_{j,t}, \tag{34}$$

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

$$F_{j,t}\left[\boldsymbol{p}_t, \boldsymbol{b}_t, \tilde{\boldsymbol{a}}_t - \eta \log(M_t \boldsymbol{s}_t), \boldsymbol{x}_t, \boldsymbol{\xi}_t, p_{0,t}; \Theta\right] = s_{j,t} \tag{35}$$

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 (34) and solving (35) are equivalent.

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 run Monte Carlo experiments to verify that our estimator is able to recover the "true" parameters.

Search model parameter sensitivity

In order to verify that our model estimation is well-specified, we report sensitivities of our parameter estimates to two key moments using a local measure developed in Andrews, Gentzkow

and Shapiro (2017). This measure helps us assess how much the parameters change if moment conditions are violated. Many of the results in this paper rely on correctly estimating the demand effects of price and marketing. In our estimation, the demand elasticities of price and marketing are identified from two behavioral assumptions (equation (15) and (16)). These two moment conditions require that *on average* mutual funds are setting their expense ratios and marketing expenditures given these price and marketing elasticities. The latter of the optimality conditions, in particular, might be violated on average if either the upper or the lower limit on the marketing expenditure is binding. Here we show how the parameter estimates will change if there are small systematic deviations from these optimal behaviors.

The results are provided in table A6. The numbers can be interpreted as the percentage bias of the parameter estimates if a moment is violated by 1% of the standard deviation of ω or ζ. For example, if the average ζ is not equal to zero but rather to 1% of the estimated standard deviation of ζ, then the estimate for λ would be downward biased by 0.65% from its true value, approximately. Overall, we find that the mean search cost λ and the weight of gross outperformance in the utility function γ are somewhat sensitive to the violations of the pricing moment condition (with sensitivities of -0.64 and -0.54, respectively, when using all funds in the sample). The other parameters are insensitive to the violations of this moment condition. The effect of marketing expenditure on sampling probability θ is also somewhat sensitive to the violation of marketing moment condition (with sensitivity of 0.32), while other parameters are not.

The latter sensitivity can be used to assess the degree to which bounds on marketing expenditures effect our parameter estimates. When funds are constrained in their ability to market by the SEC-imposed cap on 12b-1 fees (so that $\sum \omega_{j,t} > 0$), but the econometrician (mistakenly) assumes the moment condition holds with equality, the demand effects of marketing will be exaggerated, which manifests in the model as a smaller coefficient in front of the marketing expenses. However, this effect is not very large, as a 1% of the standard deviation increase in the moment condition translates into roughly one third of a percentage point increase in the marketing coefficient. More importantly, the effect is essentially the same for funds that are not at the upper bound (and similar for those at the lower bound of zero), suggesting that the impact of the binding constraints on the estimation of search model parameters is small.

[Insert Table A6 Here]

Data Appendix

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

Matching between CRSP and Morningstar

Our goal is to merge the CRSP mutual fund dataset with the Morningstar dataset. The identifiers that are common across these two datasets are ticker and CUSIP. However, in CRSP, the unique identifier is crsp_fundno and in Morningstar, it is secid. Both identify a unique share class, not a fund (for example, C share class of X fund). In this section, we create the one-to-one mapping between crsp_fundno and ticker;, crsp_fundno and CUSIP; secid and ticker; secid and CUSIP. We follow Berk and van Binsbergen's (hereafter BB) procedure as closely as possible.

crsp_fundno and ticker mapping

CRSP data spans from Jan 1961 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 happen due to multiple reports in the same year. Out of the 400 observations, there are 200 distinct crsp_fundnos. We keep the observation with non missing expense ratio information and delete the others. Now we have 504,808 observations.

2. Out of 504,808 observations left, we have 86,793 obs for which ticker is missing. We follow BB's steps to fill those. First, we identify all the unique pairs of {crsp_fundno, ticker}. Here we first delete the observations with missing tickers. Then, we delete the observations with duplicated pairs of {crsp_fundno, ticker}. We have 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 the most recent year. Then, we back fill all the previous ticker with that tickers. This step gives us 2,595 additional unique pairs of {crsp_fundno, ticker}. Then, we add back the pairs from the non duplicated case, we have 50,448 unique {crsp_fundno, ticker} pairs.

3. 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 4,343 of them. We follow BB and treat those pairs as missing. Now we have 42,436 unique pairs of {crsp_fundno, ticker}.

crsp_fundno and CUSIP mapping

According to Pástor, Stambaugh and Taylor 2015 (hereafter PST), CUSIPs can be used to match a number of Morningstar funds to CRSP funds that cannot be matched using tickers. We repeat the procedure above using CUSIPs instead of tickers. Here 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, the number of observations with missing CUSIP is reduced to 29,436.

2. Next, we identify the CUSIPs that have 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}.

We merge the above two datasets together and keep all observations. This leaves us with 54,911 unique crsp_fundnos.

Morningstar data

We start from the fund_ops file from Morningstar. This dataset contains the Morningstar Category, Fund Family name and other information, for the total of 55,571 observations.

We only keep the domestic well-diversified equity mutual funds. We follow the method provided in PST data appendix in identifying this type of funds.

1. We first identify the observations with duplicated fund_name and delete them. We also delete the funds with no Morningstar category which corresponds to additional 661 funds.

2. Then, we identify bond funds, international funds, sector funds, target date funds, real estate funds, other non-equity funds. The definition and method are provided in PST. Now we are left with 23,592 funds.

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

Merge between CRSP and Morningstar

Our goal is to get a one-to-one mapping between crsp_fundno and secid through ticker or CUSIP. For details on secid see PST.

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

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

Finally, we take the union of the two types of matches and we delete the duplicated pairs of {crsp_fundno, secid}. We have 17,658 unique crsp_fundno and secid pairs.

CRSP dataset clean

We merge the above identified 17,658 unique observations with annual CRSP Fund Summary dataset and keep the merged observations. We denote this dataset as the *baseline* dataset. It has 169,488 observations at fund share class/year level.

Correct TNA

As pointed out in PST, before 1993, a lot of the funds in the CRSP dataset report their assets under management (AUM or TNA) at a quarterly or annual frequency. Meanwhile, most of the funds report their returns at monthly frequency. When we aggregate variables such as returns and expense ratios across share classes, we need the monthly TNA information. Starting from the raw dataset of mutual funds monthly returns downloaded from WRDS, we merge it with

the 17,658 unique crsp_fundno and keep the merged observations. This gives us 2,149,498 observations (covers year from 1962 to 2016). Then we do the following correction:

1. If there is no TNA information for a given fund for any month in a year, we delete this year.
2. For the funds who report TNA at an annual frequency (with only one non-missing TNA per year), we replace the other 11 month's TNAs with the non-missing TNA.
3. For the funds who report their TNA at quarterly frequency (with only one non-missing TNA per quarter), we replace the missing values of TNA with the TNA in that quarter. For example, if a fund reports TNA at month 3 for quarter 1, we replace month 1 and month 2 TNA as month 3 TNA.
4. We delete the observations where TNA is zero, negative.
5. We delete the observations with missing TNA or monthly return.
6. We also delete duplicated observations for the same crsp_fundno in the same month.

After this correction, we have 2,018,242 observations with non-missing monthly return and TNA. (In this data appendix, we use TNA and AUM interchangeably.)

Inflation adjustment

To make the TNAs comparable across time, we adjust for inflation using the Consumer Price Index from FRED, Federal Reserve Economic Data provided by St. Louis Fed. The series we used is Consumer Price Index for All Urban Consumers: All Items[26]. This series is at the monthly frequency and it is seasonally adjusted. We convert it to annual frequency by keeping each year's December's value as this year's value for CPI. Then, we use year 2015 as the baseline year to adjust all other year's TNAs. For example, the CPI value in 1970 is 39.6 and the CPI value in 2015 is 238.3. Then, all the monthly TNAs in 1970 are multiplied by 6 (= 238.3/39.6) to make them comparable to TNAs in 2015.

Vanguard Index Fund

As proposed in BB, index funds from Vanguard are the most accessible index funds to the average investors. In our paper, we use all the equity index funds from Vanguard, combined, as the outside good. Within *baseline* dataset, we first drop all the institutional share classes: drop the funds if inst_fund== Y, sharetype == "Inst" or fund name contains "Institutional Shares" or "Institutional Class". Then, we identify passive funds from Vanguard following two steps: 1 find all the index funds. 2 find out the Vanguard index funds.

In order to identify index funds, we use a simple two-step procedure following PST:

1. If either CRSP or Morningstar indicates the fund is an index fund, we label this fund as index fund.

[26]The url is https://fred.stlouisfed.org/series/CPIAUCSL

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

Then, we check whether the fund name or the fund sponsor's name contains "Vanguard". There are 552 such observations, i.e., 552 Vanguard index fund observations.

We fill the missing value of expense ratio using the fund's life time average. Then, we delete the observations with missing expense ratio. This gives us 494 share class/year observations. We merge it with fund monthly dataset which gives us 6,279 share class/month observations.

In each month, we get the total assets under management (inflation adjusted, for details see section "Inflation adjustment"), asset weighted mean of management fees, returns, expense ratios, turnover ratios, and 12b-1 fees. Then, for each month we only keep one observation for the Vanguard index fund. Then, for each year we keep first month's value as the Vanguard fund's annual variable. We have 40 observations from year 1976 to 2015.

Active Funds Cleaning

From the *baseline* dataset, we drop all the institutional share classes.

Then, we drop all the index funds. For methods, please check section "Vanguard Index Fund". Further we only keep the funds with the following Morningstar categories: Large Blend, Large Growth, Large Value, Mid-Cap Blend, Mid-Cap Growth, Mid-Cap Value, Small Blend, Small Growth, Small Value and Aggressive Allocation. We call this dataset *Active Fund* dataset. It has 87,842 observations at share class/year level.

Front Load

To construct "effective" 12b-1 fees, we need information about fund's front load. We downloaded the front load dataset from CRSP. The total number of 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 (i.e., for certain amount of initial investment, the fund will charge certain percentage as the front load.). For each pricing schedule we only keep the maximum level of front load.

2. Then, we delete the observations where front load equals 0. That leaves us with 19,626 observations.

3. We delete observations with front load smaller than 0. There are 30 of them.

4. There are 288 cases where a fund has more than one change in front load in one year. We delete them.

5. We expand the front load dataset to the crsp_fundno year style.This step gives us 108,818 observations.

Next, we merge the above front load data set with the *Active Fund* dataset. We set missing front loads to zero.

"Effective" 12b-1 fees and expense ratio adjustment

We combine fund's 12b-1 fee and front load to create an item we call "effective" 12b-1 fee. For the fund share class with missing value of expense ratio or 12b-1, we use the time series mean of the no-missing expense ratios or 12b-1 fees to replace the missing value (expense ratio or 12b-1 fee smaller than 0 or equal to -99 are set to missing also). We set the everywhere-missing 12b-1 fee to 0 and replace the observation with 12b-1 fee larger than 1% to 1%.

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 mutual fund investors are indifferent towards different share classes. This assumption is valid if all investors have the same investment 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.

The way to identify whether some share classes belong to the same fund is by using MS_fundid from Morningstar. Also we make sure that the "effective" 12b-1 fee is not larger than 1%. We also did the same adjustment to expense ratio. Finally, we drop the observation with expense ratio greater than 5% and we drop the observations with expense ratio smaller than the sum of 5 bps and "effective" 12b-1 fee. We only keep the observations later than 1964 (include 1964).

[Insert Figure A1 Here]

Constructing the Monthly Return Dataset

Starting from *Active Fund* dataset, we merge the Monthly Return Dataset into it. Here the monthly TNAs are already corrected in the "Correct TNA" section. We also convert TNA into real terms using the procedures described in section "Inflation adjustment".

The gross return for each share class is the sum of net return (mret) and one twelfth of the expense ratio. (Here the expense ratio we use is from the raw data instead of the one we generated in the previous section. We replace the missing value of expense ratio by the time series mean. Expense ratio smaller or equal to zero are treated as missing. The reason why we use this unadjusted expense ratio is because we want our gross return to be comparable to other papers which study mutual fund performance in this fashion.)

We aggregate the gross return, net return, 12b-1 fee and expense ratio to fund level by using each share class's TNA as weight. We use MS_fundid as the identifier for a fund (not a share class). For each fund, each month, we only keep one observation.

We also clean the turnover ratio and management fee using the same procedure (in this section) as we clean the expense ratio. Here we further impose that expense ratio is larger than 20 bps. PST uses 15 bps as the lower bound for the expense ratio for active funds. After this step we have 516,849 observations.

General cleaning

Starting from the dataset in the previous step, we keep the observations with TNA>15 million dollars (this is the threshold used in PST). Then, we keep only the cases where the fund in a given year has 12 observations. Next, we keep month 1's observation (to convert the dataset from monthly to annual.). We check whether there is a gap in this annual dataset. For example for fund j, suppose it has the annual data from 1996 to 2000 and 2002 to 2010. Then, we will delete this fund, i.e., all the 14 (=5+9) observations. We also drop the active funds from Vanguard. Now for all the active funds, we have 27,621 observations. It is at fund/year level. Now we append the Vanguard Index Fund data into the above dataset. There is additional 40 fund/year observations. This is the *almost there* dataset.

Performance adjustment

Starting with the *almost there* dataset, we merge it with the Monthly Return Dataset. We keep the merged observations which is 331,452.

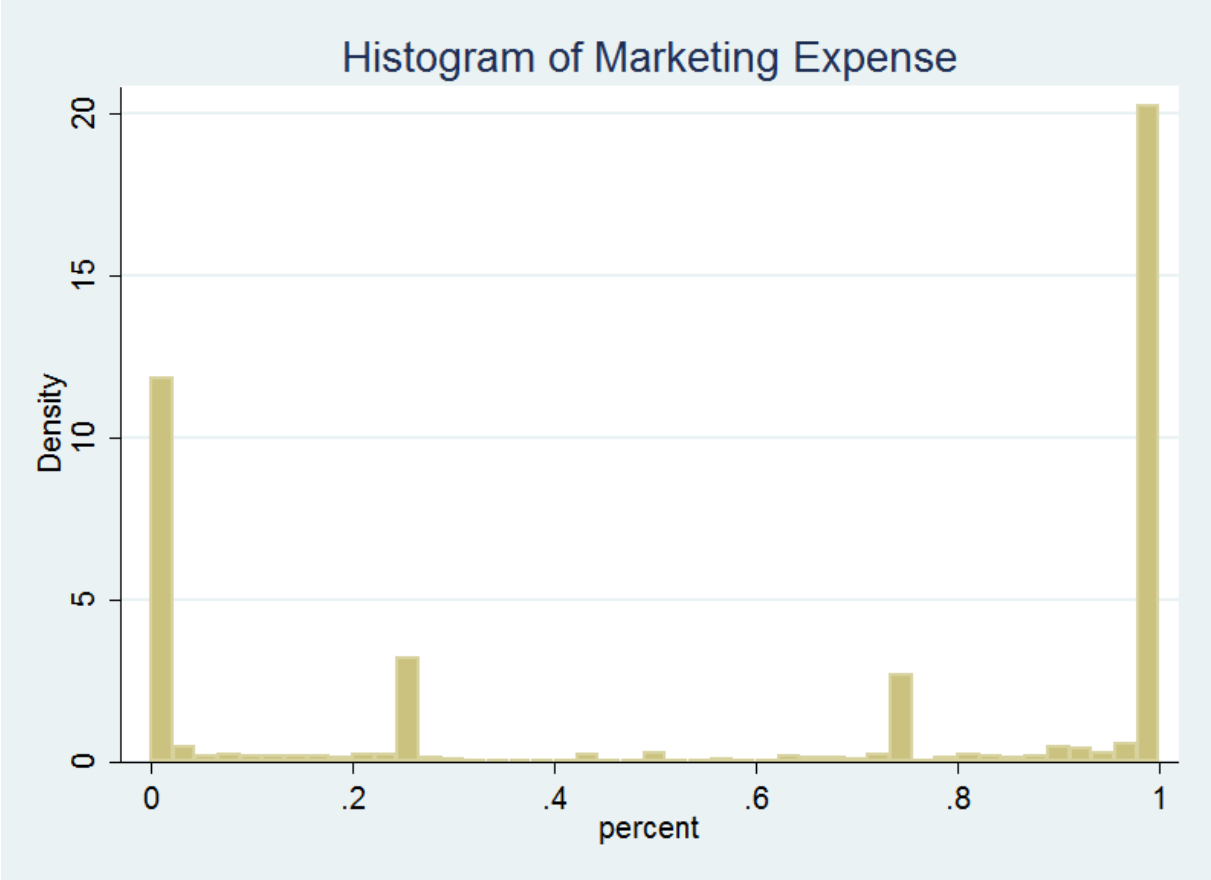
To adjust for risk exposure, we use various versions of asset pricing models. The monthly factor returns are downloaded from Ken French's website. Then, we use the following models to adjust the returns: CAPM, Fama-French 3 factor model, Fama-French-Carhart Model, Fama-French 5 factor model. The dependent variable is fund's excess return which is the difference between gross monthly return of the fund and monthly risk free rate. To increase the accuracy of beta estimation, for each fund we use all the fund's returns to estimate the betas. Then, we subtract the predicted return (betas multiplied by the factors' returns) from the fund's excess return. This gives us the monthly alphas.

Then, we aggregate alphas together by sum across 12 months to get annual alphas. And we merge the annual alphas to the *almost there* dataset to get our final dataset.

Variable definitions

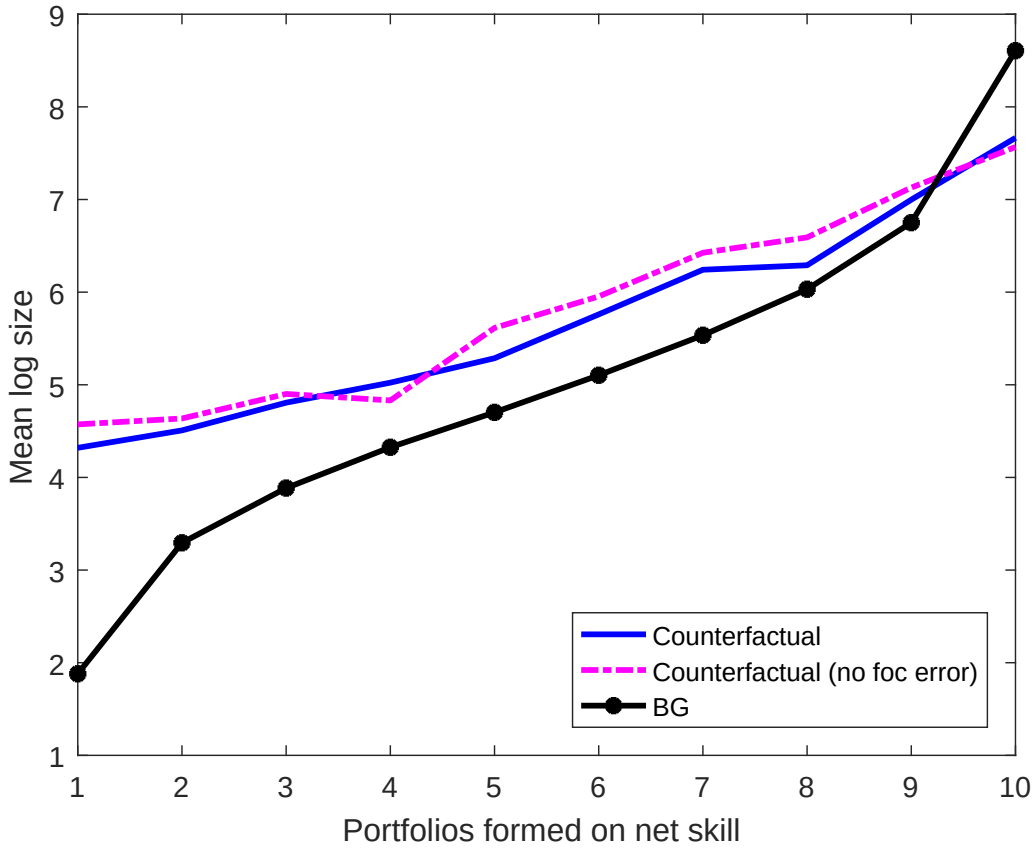
Total market is the sum of AUM for all the funds (active and passive) in a given year.
Market share is the ratio between fund's AUM and Total market.
Index fund price is the expense ratio of the Vanguard index fund.
Family size is the number of funds in the same fund family. Fund family is identified using Family variable from Morningstar. Fund age is the number of years since the fund first appears in the dataset.

Figure A1: Distribution of Marketing Expenses



The figure plots the histogram of effective marketing expenses for the main sample covering 1964 to 2015. We define the marketing expenses 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 A2: 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 purple line plots the mean of log fund size generated by our search model in the counterfactual equilibrium with *no foc errors* for each portfolio.

Table A1: Data Definition

Variable	Definition
Fund AUM	Fund's total net assets 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 expenses (%)	27,621	0.61	0.44	0.01	0.89	1.00
Market share (%)	27,621	0.18	0.66	0.01	0.02	0.09
Age (years)	27,621	11.46	10.3	4	8	16
New funds (%)	27,621	8.27	27.55	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: Investor Beliefs and Manager Skill Parameters (with skill persistence fixed at 1)

Parameters	Description	baseline	$\rho = 1$
η	Decreasing returns to scale (%)	0.48	0.50
		(0.04)	(0.04)
μ	Mean of prior (%)	3.05	3.06
		(0.25)	(0.20)
κ	SD of prior (%)	2.41	1.98
		(0.12)	(0.11)
δ	SD of realized alpha (%)	7.62	7.75
		(0.05)	(0.06)
ρ	Skill persistence	0.94	1
		(0.02)	(NA)
Num of Obs		27,621	27,621

This table presents the estimates of the fund performance-related parameters. The decreasing returns to scale parameter, η, the mean of manager's *ex ante* skill distribution, μ, the standard deviation of this distribution, κ, the standard deviation of the idiosyncratic noise in the realized alpha, δ, and the persistence of the manager's skill, ρ. The standard errors are in the parentheses. Column (1) contains the baseline estimates. Column (2) contains the estimates when ρ is fixed at 1.

Table A4: Search Model Parameters for Different Sub-Samples

Parameters	Description	2001-2005	2006-2010	2011-2015	All
λ	Mean search cost (%)	0.49	0.39	0.20	0.39
		(0.07)	(0.07)	(0.07)	(0.04)
γ	Alpha weight	0.38	0.41	0.51	0.41
		(0.051)	(0.05)	(0.54)	(0.03)
θ	Marketing effectiveness	110.11	124.24	143.16	122.56
		(12.63)	(12.78)	(13.31)	(7.39)
β_1	Family size effect	0.44	0.37	0.39	0.39
		(0.04)	(0.04)	(0.04)	(0.03)
β_2	Fund age effect	1.06	1.01	1.04	1.03
		(0.06)	(0.06)	(0.06)	(0.04)
Year FE		Yes	Yes	Yes	Yes

This table presents the estimates of the structural search model for different sub-samples. We use the data from 2001 to 2015 as the full sample. All columns use the moment condition (iv) described in Section 4.2.

Table A5: Summary of Outcomes for Current Equilibrium and No-Marketing Equilibrium (with/without foc error)

	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 (%)	74	68	66
Mean sampling prob (%)	0.085	0.078	0.078
Mean sampling prob for low price funds (%)	0.042	0.14	0.20
Mean sampling prob for index fund (%)	5.91	13.66	13.66
Investor welfare (bp)	-140.72	-61.25	-66.26
Active funds profits (bp)	57.51	42.19	48.45
Index fund profits (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 fund. Mean 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 25. Active fund profits are defined as $\sum_{j=1}^{N}(p_j - b_j)s_j$, which are the total profits for all active funds. Index fund profits are 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 A6: Sensitivity of Parameter Estimates to Moments

		Interior		Lower		Upper		All	
Parameter	Description	ζ	ω	ζ	ω	ζ	ω	ζ	ω
λ	Mean search cost (bp)	-0.65	-0.01	-0.53	-0.01	-0.59	-0.01	-0.64	-0.01
γ	Alpha weight	-0.55	0.01	-0.45	0.01	-0.50	0.01	-0.54	0.01
θ	Marketing effectiveness	-0.01	0.31	-0.03	0.28	0.02	0.33	0.01	0.32
β_1	Family size effect	0.08	-0.10	0.07	-0.08	0.07	-0.13	0.08	-0.11
β_2	Fund age effect	-0.04	-0.01	-0.03	-0.01	-0.03	-0.01	-0.03	-0.01

In this table, we provide the sensitivity of parameter estimates to two moment conditions: for fund pricing ($\sum_{t=1}^{T}\sum_{j\in t}\zeta_{j,t} = 0$) and marketing expenditures ($\sum_{t=1}^{T}\sum_{j\in t}\omega_{j,t} = 0$), for four subsamples that we use to estimate the model. "Interior" subsample includes only funds that are between the upper and the lower bounds on the marketing expenditures. in their marketing expenses. "Lower" includes only funds whose marketing expenses are zero. "Upper" refers to the funds whose marketing expenses are at 100 bps. "All" refers to all of the funds in our sample. The magnitudes can be interpreted as the percentage bias of a parameter estimate if a moment is violated by 1% of the standard deviation of the corresponding moment condition error, ω or ζ. For example, if the average ζ is not zero but instead equals to 1% of the estimated standard deviation of ζ, then the estimate for λ would be downward biased by 0.64% from its true value (if all funds are used).