Learning from Prospectuses

SIMONA ABIS
Columbia University

ANDREA M. BUFFA
University of Colorado Boulder

APOORVA JAVADEKAR
Indian School of Business

ANTON LINES
Columbia University

June 2021*

Abstract

We study qualitative information disclosure by mutual funds when investors learn from these disclosures in addition to past performance. We show theoretically that fund managers with specialized strategies optimally choose to disclose detailed strategy descriptions, while managers with standardized strategies provide generic descriptions. Generic descriptions lead to errors in benchmarking by investors and thus higher volatility in capital flows. While all fund managers dislike such volatility, those with above-average factor exposures also benefit from benchmarking errors as investors incorrectly ascribe factor returns to managerial skill. The model generates a number of predictions that we are able to test empirically using a comprehensive dataset of fund prospectuses. Consistent with the model's predictions, funds with standardized strategies include more boilerplate in their descriptions, grow larger and have lower flow-performance sensitivity, despite having greater flow volatility.

*Contacts: simona.abis@gsb.columbia.edu, buffa@colorado.edu, apoorva_javadekar@isb.edu, and anton.lines@gsb.columbia.edu. For helpful comments and discussions, we thank seminar participants at Columbia University.

Mutual fund prospectuses provide you with important information so you understand how the fund works and can easily compare it with other funds. If you wish to make an informed investment decision, you should read the prospectus before buying or selling shares in a mutual fund. SEC (2016)

1 Introduction

Prospectuses are the primary sources of information about mutual funds in the United States. The Securities and Exchange Commission requires funds to distribute these documents to investors at regular intervals, and specifies what sort of information they should contain. This includes not only standard historical performance and holdings data, but also narrative descriptions of the funds' investment activities. While the role of fund performance in investor decision-making has been studied extensively (e.g., Chevalier and Ellison (1997); Sirri and Tufano (1998); Baks et al. (2001); Pástor and Stambaugh (2002)), these narrative descriptions—particularly regarding the fund's strategy—provide a potentially important alternative signal. The SEC clearly views them as important, encouraging investors to read prospectuses carefully and not to rely solely on past performance. However, funds have substantial discretion in how much detail they ultimately choose to provide; indeed, Abis and Lines (2020) document large variation in length, complexity and content of funds' strategy descriptions, even though they are all responding to the same regulatory prompt. This variation begs the question: how do funds choose what to disclose, how do investors make use of this information, and how do these decisions interact?

We hypothesize that investors rely on strategy descriptions when deciding how to benchmark fund performance. Benchmarking is one of the most crucial elements in assessing and interpreting fund behavior, spawning a rich academic literature (Carhart (1997), Daniel et al. (1997), Berk and van Binsbergen (2015), Fama and French (2010), among others) and continued practitioner interest in the topic. However, this literature typically treats benchmarks as an exogenous function of the risk profile of the fund, while the interplay between investors' learning and managers' signaling is left largely unexplored.

In this paper we develop and test a model to analyze this interaction. In our framework, investors allocate capital according to their perception of manager skill, as in Berk and Green (2004), which they infer from active fund returns—i.e., fund returns minus benchmark returns. However, to determine the appropriate benchmark they must also learn about the fund's investment strategy by reading its prospectus. Knowing the strategy of the fund allows investors to exclude the component of performance that funds would have delivered in the absence of managerial skill. While most existing models assume that investors observe the fund's true strategy directly, we depart from this assumption and allow a more generalized learning framework where investors simultaneously infer fund skill and fund strategy.[1] This introduces an economic role for fund prospectuses.

For simplicity, we assume prospectuses can be either of two types: *detailed* or *generic*. Detailed prospectuses are fully revealing of a fund's true strategy, while generic prospectuses provide a noisy signal of only the average strategy across all funds in a particular mandate. Our interpretation is that if a particular fund's prospectus provides no specific cues to distinguish it from its peers, investors will (reasonably) default to the assumption that it is following the most common strategy within the mandate. In turn, this benchmark misperception leads to over- or under-allocation of capital as factor returns are mistaken for skill or vice versa, generating additional volatility in capital flows. Furthermore, generic prospectuses also result in lower *flow-performance sensitivity*, since investors' learning rates are slower when the prospectus signal is less precise.

Risk-averse fund managers earn fees proportional to the size of their funds, exhibiting a preference for high capital flows but low flow volatility. When choosing their prospectus type, they take into account investors' inferences about their benchmarks, and therefore face a trade-off between the two aforementioned (opposing) effects on flow volatility, as well as a third, more subtle, effect on the *level* of flows. This level effect occurs as a consequence of our assumption that managers are heterogeneous in their investment strategies, which are modeled as different loadings on an aggregate factor portfolio. Recall that generic prospectuses only provide a signal about the *average* benchmark; therefore, for managers with *above-average* factor loadings, investors perceive a loading that is lower than its true value. Assuming that factor returns are on average positive, the resulting errors will tend to favor these managers since part of the positive factor return is misperceived as alpha. For example, suppose that the fund's true strategy can be characterized by a loading of 1.2 on a particular factor. In the case of a detailed prospectus, investors understand that the factor exposure is exactly 1.2 and ignore the factor-related component of the return entirely. In the case of a generic prospectus, however, they must estimate the factor loading using fund returns, factor returns and the signal from the fund's prospectus. Now, suppose that this estimate turns out to be 0.9. Then the excess exposure of 0.3 times the factor return would be treated as active return by investors and used to form an upward-biased estimate of the manager's skill.

[1] A notable exception is the recent paper by Franzoni and Schmalz (2017), who document that fund flows become less sensitive to fund performance when aggregate risk-factor realizations are extreme, because in those states the investors' ability to learn about managerial skill is reduced.

In equilibrium, managers of funds with more "specialized" strategies (i.e., those with more extreme positive or negative loadings on the aggregate factor portfolio) optimally choose to disclose more information about their strategies, in the form of detailed prospectuses. In contrast, managers following more "standardized" strategies (i.e., those with factor loadings closer to the mandate average) optimally choose to write generic prospectuses. Overall, withholding information about a fund's strategy may lead to higher average capital flows and thus a larger average fund size, but the resulting uncertainty in investors' minds leads to greater volatility in the allocation of their capital. To some extent, this increase in volatility is mitigated by a lower flow-performance sensitivity. Our model illustrates that when a fund's strategy is sufficiently specialized, the downsides of obfuscation dominate and the choice of a detailed prospectus becomes optimal.

The theoretical interplay between fund strategies, fund prospectus choice, and investor learning generates several clear empirical predictions:

(i) *Funds with more specialized strategies* (i.e., funds with factor exposures far from the mandate average) *should, on average, adopt a more detailed prospectus.* This prediction is at the core of our equilibrium separation and, as explained above, reflects the trade-off between higher and more volatile capital flows. It is beneficial for managers to disclose more detail about less common investment strategies because, for these strategies, the lack of transparency about the correct benchmark would excessively raise the volatility of fund flows.

(ii) *Funds with more standardized strategies should, on average, be larger.* This is because the lack of transparency associated with the choice of a generic prospectus causes investors to mistakenly attribute some of the benchmark return to the fund's active return and thereby overestimate manager skill, triggering greater flows of capital.

(iii) *Funds with more specialized strategies should, on average, exhibit lower volatility in fund flows.* The choice of a detailed prospectus should limit investors' benchmarking errors when evaluating fund performance, reducing the inflows and outflows associated with such mistakes.

(iv) *Funds with more specialized strategies should, on average, exhibit higher flow-performance sensitivity.* Prospectuses for these funds should provide more precise signals about the fund's strategy, thus providing investors with a more reliable measure of active returns. Since active returns are highly informative about the skill of the manager in this case, investors should rely more heavily on them when allocating their capital.

To test these predictions, we rely on a comprehensive dataset of mutual fund strategy descriptions, collected manually from fund prospectuses via the SEC's EDGAR online reporting system. We then merge these data to fund returns, holdings and other characteristics from CRSP. Following the guidance of the model, we construct the following empirical proxies:

(i) *Investment mandates.* We use the text-based *Strategy Peer Groups* (SPGs) constructed by Abis and Lines (2020) to separate funds into groups with broadly similar objectives. These peer groups are the output of an unsupervised machine learning algorithm, which clusters funds according to similarities in the distribution of words in the prospectus text.

(ii) *Fund specialization.* In each month, we measure the degree of specialization of a fund's strategy as the sum of squared differences between (i) the weight allocated to each stock in its portfolio and (ii) the corresponding weight for the average fund in the same mandate. Intuitively, this metric can be interpreted as the squared Euclidean distance of the fund from the mandate average, within the vector space of portfolio weights.

(iii) *Prospectus detail.* We measure the level of detail in a fund's prospectus as the percentage of "boilerplate" language that it contains, where a boilerplate phrase is defined as a member of the set of the most frequent four-word phrases (4-grams) in the entire corpus of prospectuses. This definition relies on the assumption that commonly-used language contains less specific information about funds' strategies than rarely-used language. In support of this assumption, we find an inverse relationship between boilerplate and textual measures of length, complexity, and finance-specific content words, providing support for this assumption.

Utilizing these measurement proxies, we confirm the empirical analogue of our model's main prediction: funds with more specialized strategies display a significantly lower percentage of boilerplate in their strategy descriptions. Moreover, by analyzing fund size and capital flow volatility, we demonstrate the empirical existence of a trade-off between the two. Indeed, funds with more generic prospectuses tend to be larger but have more volatile flows than funds with more detailed prospectuses.

Finally, we turn our attention to the model's implications for flow-performance sensitivity. For each fund-month, we estimate the relevant benchmark using rolling one-year regressions of daily fund returns on the daily returns of six Vanguard index funds (broad, growth, large, mid, small and value), in line with the "investible benchmarks" argument of Berk and van Binsbergen (2015) and Buffa and Javadekar (2020b). We then construct a measure of active return as the difference between the fund's excess return and its benchmark excess return. As predicted by the model, we document that future flows are positively related to active performance, but this effect is significantly weaker when the fraction of boilerplate language is high. In other words, the more generic a fund's strategy description, the lower its flow-performance sensitivity.

The remainder of the paper is organized as follows: section 2 reviews the related literature, section 3 introduces the model and presents the equilibrium predictions, section 4 tests these predictions empirically, and section 5 concludes. Appendix A contains the proofs of the model, while appendix B provides examples of high and low boilerplate prospectuses.

2 Related Literature

Our paper contributes to two broad, interrelated areas of research: the study of learning by mutual fund investors and/or managers, and textual analysis of fund disclosures.

The study of learning by *investors* dates back to at least Chevalier and Ellison (1997) and Sirri and Tufano (1998), who show that capital flows respond positively to past fund performance. Theoretical analysis of this behavior originates with Baks et al. (2001), who examine partial-equilibrium optimal allocation across funds when investors have Bayesian prior beliefs about fund alpha. Pástor and Stambaugh (2002) build on this work by allowing for priors over both alpha and the pricing model used to calculate alpha, where investors update using past performance data in accordance with their uncertainty about the pricing model. Jones and Shanken (2005) allow investors to first update their alpha priors based on the observed cross-sectional distribution across all funds, before incorporating performance data on the fund of interest. Huang et al. (2012) examine a market containing both sophisticated Bayesian investors and naive performance-chasing investors; they also report implications for flow-performance sensitivity, which decreases with the volatility of fund returns due to its effects on signal precision.

In a richer equilibrium setting, Berk and Green (2004) show that investor learning about skill drives net alphas to zero when the market for asset management is perfectly competitive, fund size is endogenous, and funds experience decreasing returns to scale. This seminal framework has been extended in a number of ways that are relevant for our work. Some extensions allow learning about common components of skill within fund families (Brown and Wu (2016)) or different funds managed by the same portfolio manager (Choi et al. (2016)). Other extensions allow the precision of learning to vary with macroeconomic uncertainty (Starks and Sun (2016)), average idiosyncratic risk (Harvey and Liu (2019)), or investor uncertainty about funds' exposure to an aggregate risk factor (Franzoni and Schmalz (2017)). Buffa and Javadekar (2020a) endogenize fund manager strategy choice, showing that when investors learn more precisely about skill from stock picking strategies than market timing strategies, skilled managers are incentivized to self-select into the former.

Our paper also extends the baseline Berk and Green (2004) model. Like the aforementioned prior work, we allow investors to learn about both idiosyncratic and common components of skill within fund groups; however, we assume that the relevant groups are fund strategy mandates (measured using the Strategy Peer Groups of Abis and Lines (2020)) rather than fund families or funds with the same manager. As in prior work, investors in our model learn about funds' alphas and their loadings on an aggregate risk factor (which affects their inferences about alpha), but learning about the factor loadings occurs through qualitative fund disclosures rather than past fund returns. Additionally, where prior work typically derives flow-performance sensitivity as a function of fund return volatility, in our model it is also determined by the level of detail provided in the qualitative disclosure. We are the first to analyze funds' disclosure choice and investors' learning from this choice in a rational competitive equilibrium model. Together, these theoretical innovations generate novel predictions for the relationships between strategy specialization, disclosure informativeness, fund size, flow volatility, and flow-performance sensitivity.

Another branch of the literature studies learning by fund *managers*. In a noisy rational expectations equlibrium (NREE) model, Kacperczyk et al. (2016) show that managers with limited information processing capacity choose to learn about the systematic component of asset returns in recessions (high volatility; high price of risk) and the idiosyncratic component in expansions (low volatility; low price of risk). Gârleanu and Pedersen (2018) combine a similar NREE setup with an imperfectly competitive market for investor capital, subject to search frictions. Unlike in the competitive market models based on Berk and Green (2004), informed managers in this model outperform in equilibrium, and the overall efficiency of asset prices is a decreasing function of search costs. Our work is complementary to this branch of the literature, as we assume instead that skill is homogeneous, and that managers internalize investors' optimal behavior directly rather than through a learning process.

The second broad area to which our paper contributes is the empirical examination of mutual fund qualitative disclosures, using textual analysis. While the use of textual analysis in asset pricing goes back to Tetlock (2007), the earliest application in the mutual fund literature is by Hillert et al. (2016), who examine mutual fund shareholder letters. They find that a negative tone in these letters predicts lower fund flows, while a "plain English" writing style predicts higher flows. Similarly, Hwang and Kim (2017) examine shareholder letters of closed-end funds, finding that low "readability" is associated with a greater discount to net asset value. Also using shareholder letters, Zhang (2020) trains a supervised machine learning model to identify informed fund managers, finding that these managers experience higher flows. Relative to these papers, we contribute by examining fund prospectuses, which are investors' primary source of information about fund strategies, instead of shareholder letters, which mainly discuss performance and market conditions. We also focus on the level of detail of the content rather than tone or writing style.

More recently, a few other papers examine mutual fund prospectus text, as we do. Abis (2020) uses supervised machine learning on fund strategy descriptions to identify and test hypotheses about quantitative funds. Kostovetsky and Warner (2020) measure the textual uniqueness of strategy descriptions relative to other funds in the same Morningstar category, and find that this uniqueness positively predicts fund flows in the first three years after inception. Abis and Lines (2020) use strategy descriptions to construct interpretable peer groups for active mutual funds, using the peer groups to test hypotheses of industry organization and investor clientele effects. Akey et al. (2021) use strategy descriptions to categorize ETFs and index funds in order to study the activeness of ostensibly passive investments. The focus of our paper differs from these in that we examine managers' strategic choice of how much to disclose, rather than taking the disclosure as given.

Additionally, several contemporaneous studies analyze prospectus summary sections (or separate "summary prospectus" documents). Krakow and Schäfer (2020) use textual uniqueness relative to other funds in the same family to measure disclosure informativeness, arguing that more similar text indicates greater use of family-specific boilerplate language. They find that more informative funds exhibit higher idiosyncratic risk and better performance, and have weakly higher fund flows. By contrast, using a more direct measure of disclosure detail, we find that funds that disclose more information are actually smaller on average. Sheng et al. (2021) find that summary risk descriptions largely line up with academic risk factors, and that more informative risk disclosures predict worse future performance but, unlike the other studies, have no effect on flows. Tucker and Xia (2020) find that summary prospectuses generally have low readability but, in contrast to Hwang and Kim (2017), lower readability predicts better future performance. Tucker et al. (2020) examine the tone of these summaries. They report that risk tone covaries positively with performance (conditional only on high readability), while strategy tone covaries negatively with performance (conditional only on low readability). The relationship between tone and measured fund risk is ambiguous.

Overall, the difficult-to-interpret and often contradictory nature of many existing empirical results highlights the need for stronger theoretical guidance, which is a major contribution of our paper. Like Abis (2020) and Abis and Lines (2020), we also analyze the *full* content of funds'

Principal Investment Strategy descriptions. Prospectus summaries are much shorter than the full sections and contain more boilerplate language. Even the longer descriptions provided by Morningstar, used by Kostovetsky and Warner (2020), are significantly shorter than the full text (70 words versus over 300 words, on average). Many of the differences in our results compared to the prior literature can be explained by this improvement in the data.

3 Model

3.1 Economic Setting

We consider an economy in which a continuum of investors delegate their money to a continuum of fund managers. Both have a measure of one.

Fund skill. Managerial skill, denoted by α, is constant across managers and is unknown by the investors. All investors have the same prior on α:

$$\alpha \sim \mathcal{N}(\bar{\alpha}, \sigma_\alpha^2) \tag{1}$$

Fund strategy. Fund strategies, denoted by β_i, differ across managers and capture the fund exposure to a risk factor. Moreover, a fund strategy is characterized by two components, a common one, b, and fund specific one, γ_i:

$$\beta_i = b + \gamma_i. \tag{2}$$

The common component b can be interpreted as the most common strategy within a mandate, while γ_i capture the deviation from it. A larger value of γ, in absolute value, indicates a more specialized strategy. We assume that, besides not knowing α, investors do not observe β_i, that is they face uncertainty on which strategy the manager has implemented, and hence uncertainty on how to evaluate her performance. Investors have a prior on the two components of β_i:

$$b \sim \mathcal{N}(\bar{b}, \sigma_b^2), \quad \text{and} \quad \gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2), \tag{3}$$

where $b \perp \gamma_i$ for any i, and $\gamma_i \perp \gamma_j$ for any $j \neq i$.[2]

[2]While in this study the investment strategy of a fund manager is taken as given, Buffa and Javadekar (2020a) endogenize it in a model of investors' learning and heterogenous fund skills.

Fund return. The gross return generated by manager i at time t is given by

$$R_{it} = \alpha + \beta_i F_t + \varepsilon_{it} \tag{4}$$

where $(\alpha + \varepsilon_{it})$ captures the active part of the fund return, while $\beta_i F_t$ captures the passive one. We further assume that ε_{it} and F_t are *iid* with

$$\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2), \quad \text{and} \quad F_t \sim \mathcal{N}(\mu_F, \sigma_F^2). \tag{5}$$

The fund net return at time t is given by

$$r_{it} = R_{it} - \mathcal{C}(q_{it-1}) - f. \tag{6}$$

where f denotes the fees per dollar invested and $\mathcal{C}(q)$ captures the returns to scale of the fund strategy. As in Berk and Green (2004), we maintain that $\mathcal{C}(q) = c \cdot q$, where the fund strategy is characterized by decreasing returns to scale (DRS) if $c > 0$ and by increasing returns to scale if $c < 0$.[3] We interpret $\beta_i F_t$ as the return of manager i's *true* benchmark:

$$r_{it}^B = \beta_i F_t. \tag{7}$$

Given that fund investors might not know the fund strategy β_i, we also define manager i's *perceived/investible* benchmark:

$$\hat{r}_{it}^B = \hat{\beta}_{it-1} F_t, \tag{8}$$

where $\hat{\beta}_{it-1}$ is the perceived fund strategy at the start of time t.[4]

Fund prospectus. The fundamental role of fund prospectus is to provide information about fund strategies. For simplicity, we take the polarized view that fund managers can write only two types of prospectus, a *generic* prospectus and a *detailed* prospectus. Denoting the set of prospectus types by $\mathcal{P}$, we have that $\mathcal{P} = \{g, d\}$. Both types of prospectus convey the fund mandate to the investors. However, while a generic prospectus provides to the fund investors only a noisy signal about the common component of the fund strategy, a detailed prospectus perfectly

[3]For simplicity, we assume that the DRS parameter c is common across all fund strategies and is known by the investors. We note that an alternative interpretation of DRS is that it captures another dimension of managerial skill, which is the ability to scale up the fund strategy.

[4]In an empirical study of the source of fund activeness, Buffa and Javadekar (2020b) construct investible benchmarks to identify the active returns generated in the mutual fund industry.

reveals the strategy (i.e., the sum of the two components). Formally, we assume that reading a generic prospectus is equivalent to receiving a signal s_i^g where

$$s_i^g = b + \eta_i \quad \text{with} \quad \eta_i \sim \mathcal{N}(0, \sigma_\eta^2), \tag{9}$$

while reading a detailed prospectus is equivalent to receiving a signal s_i^d, where

$$s_i^d = b + \gamma_i. \tag{10}$$

Timeline. We consider a timeline with four dates:

- $t = 0$: managers choose which prospectus to adopt;
- $t = 1$: investors learn from prospectus and public signal and allocate their capital;
- $t = 2$: fund returns materialize, investors update their beliefs and re-allocate their capital;
- $t = 3$: fund returns materialize, funds are liquidated and the economy ends.

For ease of exposition, in what follows we refer to the investors who reads generic and detailed prospectus as $\mathcal{I}^g$ and $\mathcal{I}^d$, respectively.

3.2 Allocation of Capital and Prospectus Choice

We consider a competitive capital market, as in Berk and Green (2004), which implies that fund investors provide or remove capital till the point in which their expected utility of investing in an active fund is equal to that of investing (on their own) in the perceived fund's benchmark. This equilibrium condition determines the optimal fund size q_{it} for $t = 1, 2$:

$$q_{it}(p): \quad \mathbb{E}_t^{\mathcal{I}^p}[u(r_{it+1})] = \mathbb{E}_t^{\mathcal{I}^p}[u(\hat{r}_{it+1}^B)]. \tag{11}$$

When investors are risk-neutral, as we maintain in our baseline specification, (11) implies that the expected perceived net active returns are driven to zero by the endogenous allocation of capital, $\mathbb{E}_t^{\mathcal{I}^p}[\hat{r}_{it+1}^A] = 0$, where

$$\hat{r}_{it+1}^A \equiv r_{it+1} - \hat{r}_{it+1}^B = (\alpha + \varepsilon_{it+1}) + (\beta_i - \hat{\beta}_{it})F_{t+1} - c \cdot q_{it} - f. \tag{12}$$

It follows that the equilibrium fund size of a fund adopting prospectus p is equal to

$$q_{it}(p) = \frac{\hat{\alpha}_t(p) - f}{c}, \tag{13}$$

where $\hat{\alpha}_t$ is the perceived skill of the fund manager from the investors' perspective (i.e., the posterior mean of α) after having observed the performance of the fund at time t. Fund flows at time t, denoted by $flow_{it}$, are defined as the relative change in fund size, $flow_{it} \equiv (q_{it} - q_{it-1})/q_{it-1}$.

Taking into account how investors' learning affects fund flows, fund managers choose the prospectus that maximize their ex-ante ($t = 0$) expected utilities defined over managerial fees:

$$p^* = \arg\max_{p \in \mathcal{P}} \ \mathbb{E}_0^{\mathcal{M}_i}[v_i(q_{i1}(p) \cdot f) + \delta \cdot v_i(q_{i2}(p) \cdot f)]. \tag{14}$$

In what follows, we assume that fund managers have mean-variance preference (with a coefficient of risk aversion normalized to 1).

We next discuss the investors' leaning and characterize the posterior distribution of managerial skill as a function of fund prospectus.

3.3 Investors' Learning

Learning from prospectus. At $t = 1$, the investors refine their priors after reading their fund prospectus. Since a generic fund prospectus does not fully reveal the fund strategy, investors $\mathcal{I}^g$ form a posterior (normal) distribution on the fund strategy β_i, with posterior mean and variance equal to

$$\hat{\beta}_{i1}(g) = (1 - \lambda^b)\bar{b} + \lambda^b s_i^g, \tag{15}$$

$$\hat{\sigma}^2_{\beta_i 1}(g) = \sigma_b^2(1 - \lambda^b) + \sigma_\gamma^2, \tag{16}$$

respectively, where $\lambda^b = \sigma_b^2/(\sigma_b^2 + \sigma_\eta^2)$ is the relative weight of the prospectus signal in determining the posterior mean of the common component of each fund strategy b.

Investors $\mathcal{I}^d$, instead, perfectly learn the fund strategy β_i. Therefore, the posterior mean and variance are qual to $\hat{\beta}_{i1}(d) = \beta_i$ and $\hat{\sigma}^2_{\beta_i 1}(d) = 0$, respectively.

Since fund returns have not materialize yet, investors $\mathcal{I}^g$ and $\mathcal{I}^d$ do not update their beliefs about managerial skill. Therefore, the mean and variance of the perceived managerial skill are equal to those of the prior, regardless of the prospectus:

$$\hat{\alpha}_1(p) = \bar{\alpha}, \tag{17}$$
$$\hat{\sigma}^2_{\alpha 1}(p) = \sigma^2_\alpha. \tag{18}$$

For tractability, we assume that investors do not learn form the equilibrium choice of prospectus. To be more precise, they do not update the distribution of the fund strategy based on the type of prospectus adopted by the fund.

Learning from fund returns. At $t = 2$, the investors observe the fund and the factor returns and update their beliefs about managerial skill (and simultaneously about the fund strategy if the fund has adopted a generic prospectus).[5] Given a prospectus of type $p = \{g, d\}$, the forecasted managerial skill, and its posterior variance, are respectively qual to

$$\hat{\alpha}_2(p) \equiv \mathbb{E}_2^{\mathcal{I}^p}[\alpha | R_{i2}, F_2] = (1 - \lambda_p^\alpha)\bar{\alpha} + \lambda_p^\alpha \underbrace{\left(R_{i2} - \mathbb{E}_1^{\mathcal{I}^p}[\beta_i]F_2\right)}_{\text{perceived active return } \hat{r}_{i2}^A}, \tag{19}$$
$$\hat{\sigma}^2_{\alpha 2}(p) \equiv \mathbb{V}\mathrm{ar}_2^{\mathcal{I}^p}[\alpha | R_{i2}, F_2] = \sigma^2_\alpha(1 - \lambda_p^\alpha) \tag{20}$$

where

$$\lambda_p^\alpha = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\varepsilon + \mathbb{V}\mathrm{ar}_1^{\mathcal{I}^p}[\beta_i]F_2^2}. \tag{21}$$

Intuitively, the best forecast of a fund manger's skill is a weighted average of the prior belief and the perceived active return. Similarly, the fund strategy has a posterior mean and variance equal to

$$\hat{\beta}_{i2}(p) \equiv \mathbb{E}_2^{\mathcal{I}^p}[\beta_i | R_{i2}, F_2] = (1 - \lambda_p^\beta)\mathbb{E}_1^{\mathcal{I}^p}[\beta_i] + \lambda_p^\beta(R_i - \mathbb{E}_1^{\mathcal{I}^p}[\beta_i]F_2), \tag{22}$$
$$\hat{\sigma}^2_{\beta_i 2}(p) \equiv \mathbb{V}\mathrm{ar}_2^{\mathcal{I}^p}[\beta_i | R_{i2}, F_2] = \mathbb{V}\mathrm{ar}_1^{\mathcal{I}^p}[\beta_i](1 - \lambda_p^\beta F_2) \tag{23}$$

[5] The investors observe net returns r_{it}. However, since they also observe fund size q_{it} and since the DRS/scale parameter c is common knowledge, they can back out gross returns R_{it}. If c were not known by the investors, then the estimated skill $\hat{\alpha}$ would depend on the estimated scale $\hat{c}$.

where

$$\lambda_p^\beta = \frac{\mathbb{V}\mathrm{ar}_1^{\mathcal{I}^p}[\beta_i]F_2}{\sigma_\alpha^2 + \sigma_\varepsilon^2 + \mathbb{V}\mathrm{ar}_1^{\mathcal{I}^p}[\beta_i]F_2^2}. \tag{24}$$

3.4 Equilibrium

Since the estimated managerial skill at time 1 is independent of fund prospectus (as it is equal to the prior mean $\bar{\alpha}$), the optimal choice of prospectus only depends on the managerial fees at time 2, and is the solution to the following problem:

$$\begin{aligned} p^* &= \arg\max_{p\in\mathcal{P}} \; \mathbb{E}_0^{\mathcal{M}_i}[\hat{\alpha}_2(p)] - \frac{f}{2c}\mathbb{V}\mathrm{ar}_0^{\mathcal{M}_i}[\hat{\alpha}_2(p)] & (25)\\ &= \arg\max_{p\in\mathcal{P}} \; \bar{\alpha} + \mathbb{E}_0^{\mathcal{M}_i}[\lambda_p^\alpha](\alpha-\bar{\alpha}) + \mathbb{1}_{\{p=g\}}\mathbb{E}_0^{\mathcal{M}_i}[\lambda_g^\alpha F_2](\beta_i - \bar{b}) \\ &\qquad - \frac{f}{2c}\mathbb{V}\mathrm{ar}_0^{\mathcal{M}_i}\left[\lambda_p^\alpha\left((\alpha-\bar{\alpha}) + \varepsilon_i + \mathbb{1}_{\{p=g\}}F_2(\beta_i - \bar{b} - \lambda^b(s_i^g - \bar{b}))\right)\right]. & (26)\end{aligned}$$

We first focus only on the effect of fund strategy heterogeneity by setting $\alpha = \bar{\alpha}$ and $b = \bar{b}$. In other words, we focus on the most likely managers in terms of skill and common component of their strategies. Evaluating (26) when $p = d$ yields

$$\bar{\alpha} - \frac{f\sigma_\alpha^4\sigma_\varepsilon^2}{2c(\sigma_\alpha^2+\sigma_\varepsilon^2)^2}, \tag{27}$$

while evaluating (26) when $p = g$ yields

$$\bar{\alpha} + \mathbb{E}_0^{\mathcal{M}_i}\left[\lambda_g^\alpha F_2\right]\gamma_i - \frac{f}{2c}\mathbb{V}\mathrm{ar}_0^{\mathcal{M}_i}\left[\lambda_g^\alpha\left(\varepsilon_i + F_2(\gamma_i - \lambda^b\eta_i)\right)\right]. \tag{28}$$

where

$$\lambda_g^\alpha = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2 + [\sigma_b^2(1-\lambda^b) + \sigma_\gamma^2]F_2^2}. \tag{29}$$

Since λ_g^α depends on the factor return at time 2, it is random from the perspective of the fund manager at time 0. A manager chooses a generic (detailed) prospectus if (28) is larger (smaller) than (27), which, as we discuss below, depends on the fund strategy γ_i.

When fund managers care about the volatility of their fees, they face the following trade-off regarding the choice of prospectus. On one hand, they have the incentive to adopt a generic prospectus to reduce the transparency of the fund strategy. This is because the lack of transparency may induce the investors to make a mistake and estimate a higher managerial skill than what it actually is. In particular, the investors' mistake (in judging the manager's skill), benefits the manager, on average, when the factor exposure is larger than its average ($\gamma_i > 0$) and the factor return is expected to be positive ($\mu_F > 0$), or when the factor exposure is smaller than its average ($\gamma_i < 0$) and the factor return is expected to be negative ($\mu_F < 0$). In both cases, investors allocates excessive capital to the fund manager, thus increasing her management fees. On the other hand, less transparency means that the mistakes of the investors makes the allocation of capital, and consequently the management fees, more volatile, thus reducing the manager's utility. The increase in volatility due to the lack of transparency is particularly severe when the factor exposure γ_i is sufficiently is sufficiently far from its average. Intuitively, therefore, when a fund strategy is not too specialized, the lack of transparency associated with a generic prospectus is beneficial to the manager.

Although the moments in (28) are not available in closed-form, we obtain the following results.

Proposition 1 (Prospectus Choice). *A generic prospectus is optimally adopted if and only if the fund strategy is not too specialized, $\underline{\gamma} < \gamma_i < \overline{\gamma}$, where the two thresholds $(\underline{\gamma}, \overline{\gamma})$ exist if and only if*

$$\frac{c^2\mathbb{E}[\lambda_g^\alpha F_2]^2}{f^2\mathbb{V}\mathrm{ar}[\lambda_g^\alpha F_2]} + \sigma_\varepsilon^2\left((\lambda_d^\alpha)^2 - \mathbb{E}[(\lambda_g^\alpha)^2]\right) - (\lambda^b\sigma_\eta)^2\mathbb{E}[(\lambda_g^\alpha F_2)^2] > 0, \tag{30}$$

and are such that $|\overline{\gamma}| > |\underline{\gamma}|$ if $\mu_F > 0$ and $|\overline{\gamma}| < |\underline{\gamma}|$ if $\mu_F < 0$. The fraction of fund managers adopting a generic prospectus is given by

$$m = \Phi(\overline{\gamma}/\sigma_\gamma) - \Phi(\underline{\gamma}/\sigma_\gamma), \tag{31}$$

where $\Phi(\cdot)$ is the cumulative density function of a standard normal distribution. When a generic prospectus either perfectly reveals the common strategy b ($\sigma_\eta = 0$), or does not provide any information ($\sigma_\eta \to \infty$), the two thresholds $(\underline{\gamma}, \overline{\gamma})$ always exist and are such that $\underline{\gamma} < 0 < \overline{\gamma}$.

Proposition 1 formally shows that the distribution of fund strategies across prospectus is characterized by two thresholds of the specialized exposure γ_i. A generic prospectus is the optimal choice for any γ_i in between the two thresholds. A detailed prospectus is optimal otherwise. When the thresholds do not exist (i.e., when the condition in (30) is not satisfied) a detailed prospectus is always optimal. For the limit cases $\sigma_\eta = 0$ and $\sigma_\eta \to \infty$, we are able to show that the condition in (30) is always satisfied and that the two thresholds have opposite signs. The top-left panel in Figure 1 provides a graphical illustration of the equilibrium cross-sectional distribution of fund prospectuses as a function of the fund strategy.

Proposition 2 (**Fund Size**). *The cross-sectional average of fund size at time 2 is expected to be equal to $(\bar{\alpha} - f)/c$ among funds adopting a detailed prospectus, and is expected to be equal to*

$$\frac{\bar{\alpha} - f}{c} + \frac{\sigma_\gamma}{c}\left(\frac{\phi(\underline{\gamma}/\sigma_\gamma) - \phi(\overline{\gamma}/\sigma_\gamma)}{m}\right)\mathbb{E}[\lambda_g^\alpha F_2] \tag{32}$$

among funds adopting a generic prospectus, where $\phi(\cdot)$ is the probability density function of a standard normal distribution, and m is as in (31). It follows that the expected fund size associated with generic prospectuses is higher than the expected fund size associated with detailed prospectuses. Moreover, it exhibits larger cross-sectional dispersion.

Proposition 2 characterizes the equilibrium fund size as a function of the prospectus. Two key findings emerge. The expected fund size associated with generic prospectuses is higher on average and more heterogeneous across funds. It is higher on average because the average mistake of the investors, $\mathbb{E}[\gamma_i|\underline{\gamma} < \gamma_i < \overline{\gamma}]$, is positive when the factor return is expected to be positive, and it is negative when the factor return is expected to be negative. The intuition is that funds that optimally adopt a generic prospectus tend to benefit from the investor's mistake, and this benefit comes in the form of higher expected fund flows. Everything else equal, this leads to higher AUM.

The larger cross-sectional dispersions in fund size, instead, is a reflection of the heterogeneity in the expected mistakes that the investors make, on average, when evaluating the managers' performance. The more specialized the strategy, the higher the expected mistake. The top-right panel in Figure 1 confirms these findings.

We next turn to fund flows, defined as the relative change in AUM: $flow_{it} \equiv (q_{it} - q_{it-1})/q_{it-1}$.

Proposition 3 (**Fund Flow Volatility**). *The cross-sectional average of fund flow volatility at time 2 is equal to $\sigma_\alpha^2\sigma_\varepsilon/(\bar{\alpha} - f)(\sigma_\alpha^2 + \sigma_\varepsilon^2)$ among funds adopting a detailed prospectus, and equal to*

$$\frac{1}{m(\bar{\alpha} - f)}\int_{\underline{\gamma}/\sigma_\gamma}^{\overline{\gamma}/\sigma_\gamma}\sqrt{\sigma_\varepsilon^2\mathbb{E}[(\lambda_g^\alpha)^2] + (\lambda^b\sigma_\eta)^2\mathbb{E}[(\lambda_g^\alpha F_2)^2] + \mathbb{V}\mathrm{ar}[\lambda_g^\alpha F_2]\sigma_\gamma^2 z^2}\,\phi(z)dz \tag{33}$$

among funds adopting a generic prospectus, where $\phi(\cdot)$ is the probability density function of a standard normal distribution, and m is as in (31).

For funds adopting a detailed prospectus, fund flow volatility is driven only by idiosyncratic component of returns, ε_i. While flow volatility of a fund that writes a generic prospectus is also driven by the idiosyncratic shocks of fund returns, it is also affected by the mistakes that investors may make when evaluating their fund managers. These mistakes generate an additional source of uncertainty, $F_2(\gamma_i - \lambda^b \eta_i)$, which tends to increase the overall flow volatility. However, this is not always the case, because the higher uncertainty is dampened by a lower flow-performance sensitivity λ_g^a, reflecting the slower learning process of the investors that read generic prospectuses. The bottom-left panel in Figure 1 provides an example in which the dampening effect of a lower flow-performance sensitivity might dominate if γ_i is close to 0 (i.e., when the mistakes in inferring the fund strategy are small). In this case, the fund flow volatility associated with a generic prospectus become lower than the flow volatility associated with a detailed prospectus. It follows that the cross-sectional average of fund flow volatility might be higher or lower for generic prospectuses, depending on the distribution of fund strategies.

Proposition 4 (**Flow-Performance Sensitivity**). *The flow-performance sensitivity at time 2,*

$$\frac{\partial flow_{i2}}{\partial \hat{r}_{i2}^A} = \frac{\lambda_p^\alpha}{\bar{\alpha} - f}, \tag{34}$$

is always larger for a fund adopting a detailed prospectus, irrespective of the factor return F_2. Moreover, the difference in flow-performance sensitivity between the two prospectuses is increasing in F_2, σ_γ and σ_b.

Given the central role of learning in our model, Proposition 4 presents our results on the flow-performance sensitivity. The higher information content contained in detailed prospectus makes the inference problem of the investors easier. Indeed, the active returns of a fund with a detailed prospectus are a more precise signals about the manager's skill since the uncertainty on how to benchmark the performance of the manager is completely eliminated. This implies that a given level of active returns triggers a higher flow of capital (inflows or outflows) for funds with a detailed prospectus, compared to funds with a generic one. Moreover, the differential in flow-performance sensitivity becomes larger, the more extreme the factor return is. This is because a large positive or negative factor return makes the learning process of the investors more difficult when the prospectus is generic. The bottom-right panel in Figure 1 illustrate these results.

4 Empirical Analysis

4.1 Data

In this paper we focus on US active equity mutual funds, combining data from four different sources: (i) information about mutual fund characteristics and returns is obtained from the CRSP Survivorship-Bias-Free Mutual Fund dataset; (ii) holdings data are from CRSP and Thomson CDA/Spectrum Mutual Fund Holdings; (iii) stock-level information is from the CRSP security file; and (iv) textual data from mutual fund prospectuses are obtained from the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) of the SEC. After combining these datasets, our final sample consists of $2,995$ unique funds and $320,750$ fund–month observations, and runs from January 2000 to December 2017.

We restrict the sample by applying various filters. First, we limit the focus of our analysis to equity funds, excluding international funds, sector funds, index funds, and underlying variable annuities. Second, we exclude observations prior to each fund's first offer date to avoid incubation bias (Evans (2010)), as well as funds with less than $5 million in Total Net Assets (TNA) (Kacperczyk et al. (2008)) and funds with fewer than 12 months of observations. Finally, we remove funds that hold fewer than 10 stocks or less than 80% of their assets (excluding cash) in US common stocks, on average.

While our analysis is conducted at the fund level, all CRSP data is reported at the share class level. Hence, at each point in time, we aggregate information across all share classes belonging to the same fund, taking the sum of TNA and the average of all other variables (e.g. fees, returns, turnover, ...), weighted by lagged TNA. Following Abis (2020), we identify share classes of the same fund by constructing a comprehensive fund identifier, using the CRSP Class Group identifier, the WFICN identifier in the MFLinks linking table, and fund names. This choice is particularly relevant for matching returns and characteristics to funds' holdings. In fact, the MFLinks linking table excludes many new funds in recent years (Zhu (2020); Shive and Yun (2013)); for that reason we use the Thomson CDA/Spectrum Mutual Fund Holdings dataset from January 2000 to August 2008 and the CRSP Mutual Fund Holdings dataset from September 2008 to December 2017. We then forward-fill holdings to the monthly frequency.[6]

We follow Abis and Lines (2020) for the collection and processing of prospectuses data. Although the EDGAR system has been active since 1994, reliable data can be obtained only as of

[6]Monthly holdings are already available for 42% of the final sample. 90% of the data is forward-filled for at most 1 quarter and 99% is forward-filled for at most 2 quarters. Maximum forward-filling is restricted to 1 year.

2000, which is when our sample starts. We are able to match 31,695 prospectuses to funds of interest. We focus on a specific section of prospectuses, the *Principal Investment Strategies* (*PIS*) section, in which funds are obliged (by law) to disclose the types of assets they hold and the key criteria used in selecting those assets. Since any material change to the management or strategy of a fund must be reported to both the SEC and fund investors, for any month in which a prospectus is not available, we forward-fill using the latest available prospectus.

4.2 Empirical Measures

To test the predictions of our model, we require empirical measures for five conceptual quantities: (1) the investment mandate of each fund; (2) the degree of detail in funds' prospectuses disclosures; (3) the degree of specialization of their strategies relative to their peers; (4) their active returns; and (5) fund flows. The following paragraphs describe the construction of these measures.

4.2.1 Mandates

To maintain consistency with our hypothesis that investors use the *Principal Investment Strategy* (*PIS*) sections of funds's prospectuses to infer their strategies, we also identify the investment mandates using this text. Specifically, we rely on the *Strategy Peer Groups* (*SPGs*) of Abis and Lines (2020), which are obtained using a clustering algorithm that groups together funds with similar descriptions. Applying this methodology to the universe of US active equity funds, the text can be optimally divided into 17 distinct SPGs: *Large Cap*, *Mid Cap*, *Small Cap*, *Fundamental*, *Quantitative*, *Long Term*, *Intrinsic Value*, *Defensive*, *Dividends*, *New Products & Services*, *PE-Ratio*, *Competitive Advantage*, *Foreign (ADR)*, *Foreign (Emerging Markets)*, *Fixed Income*, *Derivatives*, and *Tax-Managed*.[7] Funds belonging to the same SPG display significantly greater similarity in risk factor exposures, stock characteristics, returns, and portfolio holdings.

For a full description of the clustering algorithm (*k-means*), further details on the strategy labels and the strategies they represent, as well as additional validation tests, see Abis and Lines (2020).

[7]Note that while all funds hold at least 80% of their assets in common stock on average, some funds' most distinctive features are secondary asset classes that make up the remaining 20%.

4.2.2 Prospectus Detail

The level of detail of each prospectus is represented by a continuous measure of *boilerplate* $\in [0, 1]$.

First, we identify *boilerplate language* as follows. We collect all four-word phrases (4-grams) from the corpus of prospectuses in our sample, and rank them according to their overall frequency in the corpus. The top 0.1% of these 4-grams (607 distinct phrases) are considered to be boilerplate language. Figure 2 displays the top 100 4-grams, ranked by frequency.

Next, we construct our boilerplate measure at the fund-month level by computing the percentage of all words in the most recently available prospectus that belong to the set of boilerplate language. Specifically, each PIS section is represented by a vector of stemmed words: the *strategy vector*.[8] From this vector, we construct another vector containing the subset of words that belong to one of the identified boilerplate phrases: the *boilerplate vector*.[9]

Our boilerplate measure for fund j at time t is then computed as:

$$Boiler_{j,t} = \frac{L^B_{j,t}}{L^S_{j,t}}, \tag{35}$$

where $L^B_{j,t}$ is the length of the *boilerplate vector* for fund j at time t, and $L^S_{j,t}$ is the length of its *strategy vector*. Appendix B provides an example of high and low boilerplate prospectuses and additional details on the measure's construction.

In the context of the model, our interpretation is that *high* levels of *Boiler* correspond to *generic* prospectuses, and *low* levels correspond to *detailed* prospectuses. To validate this interpretation, we examine how the measure is related to the length, textual complexity, and informative financial content of the PIS sections. Length is measured using a simple word count. Textual complexity is measured using the Flesch-Kincaid grade-level complexity score (Kincaid et al. (1975)), which is calibrated to show the estimated number of years of schooling required to understand the text.

To measure informative financial content, we start from *Campbell Harvey's Hypertextual Finance Glossary*, which provides a comprehensive list of finance-related words and short phrases.[10] For each PIS section, we construct the $financial\ vector$ as the subset of words in the *strategy vector* (defined above) that can also be found in the Harvey dictionary. However, since the prospectus is comprised mostly of financial text, the simple ratio between the length of the $financial\ vector$

[8] Stop words, symbols, numbers, websites, and punctuation are excluded. Words are stemmed using the Porter stemmer.

[9] Note: words must appear in the correct order to be considered part of a boilerplate phrase.

[10] https://people.duke.edu/∼charvey/Classes/wpg/glossary.htm

and that of the *strategy vector* would not necessarily identify *relevant* financial content. Thus, we first exclude from the *financial vector* all words that appear in more than 75% of all PIS descriptions,[11] then weight the remaining terms by the inverse of their frequency in the entire corpus. Thus, our measure of relevant financial content is defined as follows:

$$Fin^a_{j,t} = \frac{\sum_{i=1}^{L^F_{j,t}} \omega_i}{L^S_{j,t}}, \qquad \text{for} \qquad \omega_i = \frac{1}{(1 + Coverage_i)^a}. \tag{36}$$

$Coverage_i$ indicates the percentage of all PIS sections containing financial term i. $L^F_{j,t}$ and $L^S_{j,t}$ represent the length of the *fianancial vector* and *strategy vector*, respectively, for fund j in month t. The variable a controls the strength of the "penalty" applied to commonly used terms; the greater a, the greater the penalty. Note that setting $a = 0$ is equivalent to not applying any weighting. We compute the measure for $a = [0, 1, 2, 4]$. Finally, we normalize $Fin^a_{j,t}$ by its average value across funds in the same SPG.

We assess the validity of *Boiler* as a measure of prospectus detail by running the following regression:

$$Y_{j,t} = \alpha + \beta Boiler_{j,t} + \gamma' X_{j,t} + \varepsilon_{j,t}. \tag{37}$$

where $Y_{j,t}$ is either the length, complexity, or financial content measure ($[Length_{j,t}, Complexity_{j,t}, Fin^a_{j,t}]$) for fund j in month t. $X_{j,t}$ is a vector of control variables: the log of fund size and age, the turnover and expense ratios, net fund flows and flow volatility, and style control variables constructed as the average loading of the stocks held by fund j at time t on the Fama and French (2015) size, value, momentum, investment and profitability factors.

The results are reported in table I. Columns (3)-(6) show that *Boiler* is inversely related to PIS length, textual complexity, and all variants of the financial content measure (Fin^a_{jt}) for $a > 0$. This indicates that this simple measure is indeed effective at capturing the presence of relevant financial content (non-standardized financial terminology) in strategy descriptions. We therefore use boilerplate as a proxy for the level of detail (informativeness) of prospectuses.

4.2.3 Fund Strategy Specialization

Having established our measure of fund strategy mandates, we now turn to funds' specialization within each mandate. Our measure of strategy specialization is constructed as the log of the sum

[11]This filter only excludes the following 9 stemmed terms: invest, fund, compani, secur, market, stock, asset, cap, manag

of squared differences between the elements of each fund's portfolio weight vector (expressed in percentage of TNA) and that of the average portfolio weight vector for funds in the same SPG:

$$Specialization_{j,t} = ln \left[\sum_{i=1}^{N_{j,t}} \left(w_{i,j,t} - \bar{w}_{i,j,t}^{SPG_{j,t}} \right)^2 \right], \tag{38}$$

where $w_{i,j,t}$ is the percentage of TNA allocated by fund j to stock i in month t; while $\bar{w}_{i,j,t}^{SPG_{j,t}}$ is the average weight allocated to stock i by funds in the same SPG as fund j at time t.

4.2.4 Active Returns

We compute each fund's active performance in month t as the difference between the fund's excess return ($R_{j,t}$) and the excess return on a passive benchmark. First, we estimate time-varying loadings for each fund of interest on the returns of seven passive Vanguard funds (*Broad*, *Large*, *Mid*, *Small*, *Value*, *Growth*, and *Momentum*), using the following rolling regressions:

$$R_{j,t} = \alpha + {\beta_{j,t}}' V_t + \epsilon_{j,t}, \tag{39}$$

where $V_t = [R_t^{Broad}, R_t^{Large}, R_t^{Mid}, R_t^{Small}, R_t^{Value}, R_t^{Growth}, R_t^{Momentum}]$.

We then construct the active return, $\hat{\alpha}$, using estimated loadings as of $t-1$ so that the benchmarks are investible in real time:

$$\hat{\alpha}_{j,t} = R_{j,t} - \hat{\beta}'_{j,t-1} V_t. \tag{40}$$

4.2.5 Fund Flows

In order to run flow-performance sensitivity regressions for each fund-month, we construct fund flows as the growth in TNA not attributable to returns:

$$Flow_{j,t} = \frac{TNA_t - TNA_{t-1}(1 + r_{j,t})}{TNA_{t-1}}. \tag{41}$$

We also construct annual flows as:

$$Flow_{j,A} = \frac{TNA_t - TNA_{t-12}(1 + r_{j,A})}{TNA_{t-12}}, \tag{42}$$

where $r_{j,A}$ are annualized fund returns.

4.3 Test of Prediction 1: Prospectus Choice

Proposition 1 shows that managerial incentive to disclose more information is positively related to the degree of specialization of a fund's strategy. In order to test this prediction we run the following regressions:

$$Boiler_{j,t} = a + bSpecialization_{j,t} + \gamma' X_{j,t} + \iota_t + \zeta_{SPG} + \varepsilon_{j,t}, \tag{43}$$

where $Boiler_{j,t}$ measures the percentage of boilerplate in the prospectus of fund j at time t (see subsection 4.2.2) and $Specialization_{j,t}$ is as defined in subsection 4.2.3. $X_{j,t}$ is a vector of demeaned fund level control variables: the log of fund size and age, the turnover and expense ratios, net fund flows and flow volatility, and style control variables constructed as the average loading of the stocks held by fund j at time t on the Fama and French (2015) size, value, momentum, investment and profitability factors. The same control variables are used in all following specifications. We further add month and mandate fixed effects, and cluster standard errors at the month and mandate level. Fund mandates (SPGs) are constructed as described in section 4.2.

Table II displays the results of these regressions. All specifications show that higher strategy specialization corresponds to a significantly lower level of boilerplate in the strategy descriptions, confirming the first prediction of the model. The results persist when adding SPG fixed effects/clustering, and are also robust to measuring boilerplate using an indicator variable that takes a value of one for above-median boilerplate ($D50\_Boiler$; columns 3 and 4).

4.4 Test of Prediction 2: Fund Size

The second prediction of the model is that, within each investment mandate, funds with generic prospectuses should be larger. We test this prediction by running the following regression:

$$ln(TNA_{j,t}) = a + bBoiler_{j,t} + \gamma' X_{j,t} + \iota_t + \zeta_{SPG} + \varepsilon_{j,t} \tag{44}$$

where $ln(TNA_{j,t})$ is the natural logarithm of the size of fund j at time t. We log-transform this variable due to its high skewness.

Table III reports the results. The positive and significant $\hat{b}$ coefficients indicate that indeed fund size grows with the percentage of boilerplate in funds' prospectuses.

4.5 Test of Prediction 3: Fund Flow Volatility

The other side of this trade-off, which incentivizes managers with specialized strategies to disclose more detailed prospectuses, is the reduction in flow volatility. Hence, we show that within each mandate, funds that disclose more detailed prospectuses also experience a lower flow volatility. We do so by running the following regressions:

$$FlowVol_{j,t} = a + bBoiler_{j,t} + \gamma' X_{j,t} + \iota_t + \varepsilon_{j,t} | SPG, \tag{45}$$

where $FlowVol_{j,t}$ is the 12-months rolling net flows volatility of fund j at time t. The regression is conditional on observations belonging to each SPG of interest (we run a separate regression for each SPG). The coefficient of interest is $\hat{b}$. A positive and significant $\hat{b}$ would indicate that funds with a higher percentage of boilerplate in their strategy descriptions experience higher flow volatility.

Table IV reports the results. We observe that for 12 of the SPGs, a higher boilerplate corresponds to a significantly higher net flow volatility. Results are not significant for two of the SPGs, while they go in the opposite direction for the *Derivatives* and *Long Term* SPGs. Overall, the results are consistent with the model's prediction.

4.6 Test of Prediction 4: Flow-Performance Sensitivity

The fourth prediction of the model is that flow-performance sensitivity should be increasing in the level of detail of fund prospectuses. To test this prediction, we run the following flow-performance sensitivity regression:

$$Flow_{j,t+1} = a + b_0 \hat{\alpha}_{j,t} + b_1 [\hat{\alpha}_{j,t} \times Boiler_{j,t}] + \gamma' X_{j,t} + \iota_t + \zeta_{SPG} + \varepsilon_{j,t}, \tag{46}$$

where $Boiler_{j,t}$ is the boilerplate measure for the PIS description of fund j in month t, constructed as described in section 4.2.2. $\hat{\alpha}_{j,t}$ is fund j's active return at time t, as explained in section 4.2.4.

$X_{j,t}$ are demeaned fund-specific control variables, ι_t are month fixed-effects and ζ_{SPG} are SPG fixed effects. Standard errors are clustered at the fund and SPG level.

The coefficient of interest is $\hat{b}_1$, which indicates by how much the flow-performance sensitivity of fund j in month t changes as the boilerplate measure increases. We use both the continuous measure of boilerplate described in section 4.2, as well as an indicator variable equal to 1 when boilerplate is above its unconditional median, zero otherwise ($D50\_Boiler$). In the latter specification, $\hat{b}_0$ represents the flow-performance sensitivity for low boilerplate funds, while $\hat{b}_0 + \hat{b}_1$ represents flow-performance sensitivity for high boilerplate funds. Hence, b_1 is the incremental effect.

Table 4 reports the estimated coefficients for these regressions. We observe that $\hat{b}_1$ is always negative and significant, indicating that funds with higher boilerplate experience a lower flow-performance sensitivity, confirming the prediction of the model.

5 Conclusion

In this paper we study the interplay between investor learning and fund managers' incentive to disclose information through prospectuses. We show both theoretically and empirically that, depending on the degree of specialization of a manager's strategy, her incentive to disclose information through prospectuses changes. Greater information disclosure provides a clear signal about the fund's strategy to investors, and therefore about the appropriate benchmark. In response to this signal, investors learn faster about the fund's skill and so exhibit a higher flow-performance sensitivity. Less disclosure, on the other hand, causes investors to make more mistakes in detemining the correct benchmark for the fund, which leads to higher net flows volatility but also greater fund size. In equilibrium, managers with more specialized strategies have a greater incentive to disclose information, while managers with more standardized strategies have an incentive to obscure information by writing generic prospectuses. We confirm these predictions empirically, using Principal Investment Strategy descriptions in fund prospectuses.

APPENDIX

A Proofs

Lemma A.1 (Truncated Normal Distribution). *Let X be normally distributed with mean μ_X and variance σ_X^2, $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$. Define $Y \equiv (X|\underline{x} < X < \overline{x})$. It follows that Y has a truncated normal distribution with density equal to*

$$h_Y(y) = \begin{cases} \sigma_X^{-1} \frac{\phi((y-\mu_X)/\sigma_X)}{\Phi(\overline{z})-\Phi(\underline{z})} & \text{if } \underline{x} < y < \overline{x} \\ 0 & \text{otherwise} \end{cases}, \tag{A.1}$$

and first and second moments equal to

$$\mathbb{E}[Y] = \mu_X + \sigma_X \frac{\phi(\underline{z}) - \phi(\overline{z})}{\Phi(\overline{z}) - \Phi(\underline{z})}, \tag{A.2}$$

$$\mathbb{V}\text{ar}[Y] = \sigma_X^2 \left(1 + \frac{\underline{z}\phi(\underline{z}) - \overline{z}\phi(\overline{z})}{\Phi(\overline{z}) - \Phi(\underline{z})} - \left(\frac{\phi(\underline{z}) - \phi(\overline{z})}{\Phi(\overline{z}) - \Phi(\underline{z})} \right)^2 \right), \tag{A.3}$$

where $\underline{z} \equiv (\underline{x} - \mu_X)/\sigma_X$, $\overline{z} \equiv (\overline{x} - \mu_X)/\sigma_X$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and the cumulative distribution function of a standard normal distribution, respectively.

Proof. See Johnson et al. (1994). □

Proof of Proposition 1. A fund manager chooses a generic prospectus over a detailed prospectus when (28) is larger than (27). Let $\Psi(\gamma_i)$ denote the difference between (28) and (27), multiplied by $2c/f$:

$$\Psi(\gamma_i) = \frac{2c}{f}\mathbb{E}[\lambda_g^\alpha F_2]\gamma_i - \mathbb{V}\text{ar}[\lambda_g^\alpha(\varepsilon_i + F_2(\gamma_i - \lambda^b\eta_i))] + \frac{\sigma_\alpha^4\sigma_\varepsilon^2}{(\sigma_\alpha^2 + \sigma_\varepsilon^2)^2}, \tag{A.4}$$

where, for brevity, we omitted the subscript 0 and superscript $\mathcal{M}_i$ from the moments' operators. The variance term in (A.4) can be written as

$$\begin{aligned} \mathbb{V}\text{ar}[\lambda_g^\alpha(\varepsilon_i + F_2(\gamma_i - \lambda^b\eta_i))] &= \mathbb{V}\text{ar}[\lambda_g^\alpha F_2]\gamma_i^2 + \mathbb{V}\text{ar}[\lambda_g^\alpha(\varepsilon_i - \lambda^b\eta_i F_2)] + 2\mathbb{C}\text{ov}[\lambda_g^\alpha(\varepsilon_i - \lambda^b\eta_i F_2), \lambda_g^\alpha F_2]\gamma_i \\ &= \mathbb{V}\text{ar}[\lambda_g^\alpha F_2]\gamma_i^2 + \sigma_\varepsilon^2\mathbb{E}[(\lambda_g^\alpha)^2] + (\lambda^b\sigma_\eta)^2\mathbb{E}[(\lambda_g^\alpha F_2)^2], \end{aligned} \tag{A.5}$$

where the second equality follows from

$$\begin{aligned}
\mathbb{V}\text{ar}[\lambda_g^\alpha(\varepsilon_i - \lambda^b \eta_i F_2)] &= \mathbb{V}\text{ar}[\lambda_g^\alpha \varepsilon_i] + (\lambda^b)^2 \mathbb{V}\text{ar}[\eta_i \lambda_g^\alpha F_2] - 2\lambda^b \mathbb{C}\text{ov}[\lambda_g^\alpha \varepsilon_i, \eta_i \lambda_g^\alpha F_2] \\
&= \sigma_\varepsilon^2 \mathbb{E}[(\lambda_g^\alpha)^2] + (\lambda^b \sigma_\eta)^2 \mathbb{E}[(\lambda_g^\alpha F_2)^2] - 2\lambda^b \Big(\mathbb{E}[\varepsilon_i \eta_i (\lambda_g^\alpha)^2 F_2] - \mathbb{E}[\lambda_g^\alpha \varepsilon_i] \mathbb{E}[\eta_i \lambda_g^\alpha F_2] \Big) \\
&= \sigma_\varepsilon^2 \mathbb{E}[(\lambda_g^\alpha)^2] + (\lambda^b \sigma_\eta)^2 \mathbb{E}[(\lambda_g^\alpha F_2)^2] - 2\lambda^b \mathbb{E}[\varepsilon_i] \mathbb{E}[\eta_i] \mathbb{C}\text{ov}[\lambda_g^\alpha, \lambda_g^\alpha F_2] \\
&= \sigma_\varepsilon^2 \mathbb{E}[(\lambda_g^\alpha)^2] + (\lambda^b \sigma_\eta)^2 \mathbb{E}[(\lambda_g^\alpha F_2)^2] \qquad \text{(A.6)}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{C}\text{ov}[\lambda_g^\alpha(\varepsilon_i - \lambda^b \eta_i F_2), \lambda_g^\alpha F_2] &= \mathbb{C}\text{ov}[\lambda_g^\alpha \varepsilon_i, \lambda_g^\alpha F_2] - \lambda^b \mathbb{C}\text{ov}[\eta_i \lambda_g^\alpha F_2, \lambda_g^\alpha F_2] \\
&= \Big(\mathbb{E}[\varepsilon_i (\lambda_g^\alpha)^2 F_2] - \mathbb{E}[\lambda_g^\alpha \varepsilon_i] \mathbb{E}[\lambda_g^\alpha F_2] \Big) - \Big(\mathbb{E}[\eta_i (\lambda_g^\alpha F_2)^2] - \mathbb{E}[\eta_i \lambda_g^\alpha F_2] \mathbb{E}[\lambda_g^\alpha F_2] \Big) \\
&= \mathbb{E}[\varepsilon_i] \mathbb{C}\text{ov}[\lambda_g^\alpha, \lambda_g^\alpha F_2] - \mathbb{E}[\eta_i] \mathbb{V}\text{ar}[\lambda_g^\alpha F_2] \\
&= 0, \qquad \text{(A.7)}
\end{aligned}$$

given that $\mathbb{V}\text{ar}[XY] = \mathbb{E}[X^2]\mathbb{E}[Y^2] - \mathbb{E}[X]^2 \mathbb{E}[Y]^2$ when X and Y are independent, and that $\mathbb{E}[\varepsilon_i] = \mathbb{E}[\eta_i] = 0$. Therefore, $\Psi(\gamma_i)$ is a quadratic function of γ_i,

$$\Psi(\gamma_i) = a_0 + 2a_1\gamma_i + a_2\gamma_i^2, \qquad \text{(A.8)}$$

with coefficients

$$\begin{aligned}
a_0 &= \sigma_\varepsilon^2 \left((\lambda_d^\alpha)^2 - \mathbb{E}[(\lambda_g^\alpha)^2] \right) - (\lambda^b \sigma_\eta)^2 \mathbb{E}[(\lambda_g^\alpha F_2)^2], \qquad \text{(A.9)} \\
a_1 &= (c/f)\mathbb{E}[\lambda_g^\alpha F_2], \qquad \text{(A.10)} \\
a_2 &= -\mathbb{V}\text{ar}[\lambda_g^\alpha F_2]. \qquad \text{(A.11)}
\end{aligned}$$

The equation $\Psi(\gamma_i) = 0$ admits two distinct (real) solutions $(\underline{\gamma}, \overline{\gamma})$ if its discriminant $(a_1^2 - a_2 a_0)$ is strictly positive. By dividing the discriminant by $\mathbb{V}\text{ar}[\lambda_g^\alpha F_2]$, we obtain the existence condition in (30). Since $\partial^2\Psi(\gamma_i)/\partial\gamma_i^2 = -2\mathbb{V}\text{ar}[\lambda_g^\alpha F_2] < 0$, $\Psi(\gamma_i)$ is concave, implying that $\Psi(\gamma_i) > 0$ (i.e., a generic prospectus is optimal) for $\underline{\gamma} < \gamma_i < \overline{\gamma}$, and negative (i.e., a detailed prospectus is optimal) otherwise. If the discriminant $(a_1^2 - a_2 a_0)$ is negative, then $\Psi(\gamma_i)$ is always negative, thus making a detailed prospectus optimal for any γ_i. Given a measure one of fund managers, the fraction of managers adopting a generic prospectus, m, is equal to $\mathbb{P}(\underline{\gamma} < \gamma_i < \overline{\gamma}) = \Phi(\overline{\gamma}/\sigma_\gamma) - \Phi(\underline{\gamma}/\sigma_\gamma)$.

In the limit cases $\sigma_\eta = 0$ and $\sigma_\eta \to \infty$, $\lambda^b \sigma_\eta = 0$, which makes a_0 always positive. This is because $\lambda_d^\alpha > \lambda_g^\alpha(F_2)$ for any $F_2 \neq 0$, and $\lambda_d^\alpha = \lambda_g^\alpha(F_2)$ for $F_2 = 0$, thus making $(\lambda_d^\alpha)^2 > \mathbb{E}[(\lambda_g^\alpha)^2]$. Since $a_0 > 0$, it follows that $\overline{\gamma} > 0$ and $\underline{\gamma} < 0$. □

Proof of Proposition 2. Given the equilibrium fund size in (13), the expected fund size associated with detailed and generic prospectuses are equal to

$$\mathbb{E}[q_{i2}(d)] = \frac{\mathbb{E}[\hat{\alpha}_2(d)] - f}{c} = \frac{\bar{\alpha} - f}{c}, \tag{A.12}$$

$$\mathbb{E}[q_{i2}(g)] = \frac{\mathbb{E}[\hat{\alpha}_2(g)] - f}{c} = \frac{\bar{\alpha} - f}{c} + \frac{1}{c}\mathbb{E}[\lambda_g^\alpha F_2]\gamma_i, \tag{A.13}$$

respectively. Since $\mathbb{E}[q_{i2}(d)]$ does not depend on the fund strategy γ_i, the cross-sectional average is equal to $(\bar{\alpha} - f)/c$. Since, instead, $\mathbb{E}[q_{i2}(g)]$ depends on the fund strategy γ_i, the cross-sectional average is equal to

$$\frac{\bar{\alpha} - f}{c} + \frac{1}{c}\mathbb{E}[\lambda_g^\alpha F_2]\mathbb{E}[\gamma_i | \underline{\gamma} < \gamma_i < \overline{\gamma}], \tag{A.14}$$

where the last term is the mean of a truncated normal distribution. By the results in Lemma A.1,

$$\mathbb{E}[\gamma_i | \underline{\gamma} < \gamma_i < \overline{\gamma}] = \sigma_\gamma \frac{\phi(\underline{\gamma}/\sigma_\gamma) - \phi(\overline{\gamma}/\sigma_\gamma)}{\Phi(\overline{\gamma}/\sigma_\gamma) - \Phi(\underline{\gamma}/\sigma_\gamma)} \tag{A.15}$$

and (32) obtains.

Since the function $\Lambda(F_2) = \lambda_g^\alpha F_2 = \sigma_\alpha^2 F_2/(\sigma_\alpha^2 + \sigma_\varepsilon^2 + [\sigma_b^2(1 - \lambda^b) + \sigma_\gamma^2]F_2^2)$ is symmetric around 0 (i.e., $\Lambda(-F_2) = -\Lambda(F_2)$), it follows that $\mathbb{E}[\lambda_g^\alpha F_2]$ is positive if $\mu_F > 0$, and is negative if $\mu_F < 0$. This is because the probability density function of the factor return F_2, $h_{F_2}(\cdot)$, is such that $h_{F_2}(x) > h_{F_2}(-x)$ for $x > 0$ if $\mu_F > 0$, and $h_{F_2}(x) < h_{F_2}(-x)$ for $x > 0$ if $\mu_F < 0$. Moreover, when $\mathbb{E}[\lambda_g^\alpha F_2] > 0$, the linear coefficient a_1 in (A.8) is positive, implying that $|\overline{\gamma}| > |\underline{\gamma}|$, if the existence condition in (30) is satisfied. When, instead, $\mathbb{E}[\lambda_g^\alpha F_2] < 0$, the opposite holds true: the linear coefficient a_1 in (A.8) is negative and $|\overline{\gamma}| < |\underline{\gamma}|$. Therefore, it follows that $\phi(\underline{\gamma}/\sigma_\gamma) - \phi(\overline{\gamma}/\sigma_\gamma)$ has the same sign as $\mathbb{E}[\lambda_g^\alpha F_2]$, thus making the cross-sectional average of the expected fund size associated with a generic prospectus always larger than that associated with a detailed prospectus.

While the cross-sectional dispersion (variance) of $\mathbb{E}[q_{i2}(d)]$ is equal to zero, that of $\mathbb{E}[q_{i2}(g)]$ is positive and equal to

$$\frac{\sigma_\gamma^2}{c^2}\mathbb{E}[\lambda_g^\alpha F_2]^2 \left(1 + \frac{(\underline{\gamma}/\sigma_\gamma)\phi(\underline{\gamma}/\sigma_\gamma) - (\overline{\gamma}/\sigma_\gamma)\phi(\overline{\gamma}/\sigma_\gamma)}{\Phi(\overline{\gamma}/\sigma_\gamma) - \Phi(\underline{\gamma}/\sigma_\gamma)} \left(\frac{\phi(\underline{\gamma}/\sigma_\gamma) - \phi(\overline{\gamma}/\sigma_\gamma)}{\Phi(\overline{\gamma}/\sigma_\gamma) - \Phi(\underline{\gamma}/\sigma_\gamma)}\right)^2\right), \tag{A.16}$$

where the last term corresponds to $\mathbb{V}\text{ar}[\gamma_i | \underline{\gamma} < \gamma_i < \overline{\gamma}]$ and is obtained by Lemma A.1. □

Proof of Proposition 3. Fun size at time 1 and time 2 are equal to $(\bar{\alpha}-f)/c$ and $(\hat{\alpha}_2(p)-f)/c$, respectively. Therefore, fund flows at time 2 associated with a fund prospectus p, $q_{i2}(p)/q_{i1}-1$, are equal to

$$flow_{i2}(p) = \frac{\hat{\alpha}_2(p)-\bar{\alpha}}{\bar{\alpha}-f}. \tag{A.17}$$

Fund flow volatility is therefore equal to

$$\begin{aligned}\sqrt{\mathbb{V}\mathrm{ar}[flow_{i2}(p)]} &= \frac{1}{\bar{\alpha}-f}\sqrt{\mathbb{V}\mathrm{ar}[\hat{\alpha}_2(p)]}\\ &= \frac{1}{\bar{\alpha}-f}\sqrt{\mathbb{V}\mathrm{ar}[\lambda_p^\alpha(\varepsilon_i+\mathbb{1}_{\{p=g\}}F_2(\gamma_i-\lambda^b\eta_i))]},\end{aligned} \tag{A.18}$$

where the second equality follows from (26). When the prospectus is detailed ($p=d$), the fund flow volatility and its cross-sectional average (since $\sqrt{\mathbb{V}\mathrm{ar}[flow_{i2}(d)]}$ does not depend on γ_i) are both equal to

$$\sqrt{\mathbb{V}\mathrm{ar}[flow_{i2}(d)]} = \frac{1}{\bar{\alpha}-f}\sqrt{\mathbb{V}\mathrm{ar}[\lambda_d^\alpha\varepsilon_i]} = \frac{\lambda_d^\alpha\sigma_\varepsilon}{\bar{\alpha}-f}. \tag{A.19}$$

When the prospectus is generic ($p=g$), the fund flow volatility is equal to

$$\begin{aligned}\sqrt{\mathbb{V}\mathrm{ar}[flow_{i2}(g)]} &= \frac{1}{\bar{\alpha}-f}\sqrt{\mathbb{V}\mathrm{ar}[\lambda_g^\alpha(\varepsilon_i+F_2(\gamma_i-\lambda^b\eta_i))]}\\ &= \frac{1}{\bar{\alpha}-f}\sqrt{\sigma_\varepsilon^2\mathbb{E}[(\lambda_g^\alpha)^2]+(\lambda^b\sigma_\eta)^2\mathbb{E}[(\lambda_g^\alpha F_2)^2]+\mathbb{V}\mathrm{ar}[\lambda_g^\alpha F_2]\gamma_i^2},\end{aligned} \tag{A.20}$$

where the second equality follows from (A.5). Since $\sqrt{\mathbb{V}\mathrm{ar}[flow_{i2}(g)]}$ depends on γ_i, the cross-sectional average is obtained by integrating $\sqrt{\mathbb{V}\mathrm{ar}[flow_{i2}(g)]}$ over γ_i in the interval $[\underline{\gamma},\overline{\gamma}]$, thus obtaining (33). $\square$

Proof of Proposition 4. Fund flows at time 2 associated with a fund prospectus p are as in (A.17), where, based on (19), $\hat{\alpha}_2(p)=(1-\lambda_p^\alpha)\bar{\alpha}+\lambda_p^\alpha\hat{r}_{i2}^A$. It follows that the flow-performance sensitivity $fps_{i2}(p)\equiv\partial flow_{i2}(p)/\partial\hat{r}_{i2}^A$ is equal to

$$\frac{\lambda_p^\alpha}{\bar{\alpha}-f}. \tag{A.21}$$

Since $\lambda_d^\alpha \geqslant \lambda_g^\alpha$ for any realization of the factor return F_2, the flow-performance sensitivity is always larger for funds adopting detailed prospectuses. Moreover, since λ_g^α decreases with F_2, σ_γ, and

σ_b (while λ_g^α is independent of them), the difference in flow-performance sensitivity across funds adopting detailed and generic prospectuses increases in those quantities. $\square$

B Boilerplate Measure Examples

In order to better understand the measure, consider the two examples below. They have been chosen to be of same length (68 words), which is much below the median PIS size. One of them is ranked in the top 10% of the boilerplate distribution while the other belongs to the bottom 10%. In both examples, the bold words indicate the *boilerplate vector* while all other words have been highlighted in red.

High boilerplate example: *Crsp_cl_grp: 1000270 Date: 2012-09-30*

Full text before pre-processing:

> Under normal circumstances, the Fund invests at least 80% of its net assets, plus any borrowings for investment purposes, in dividend-paying equity securities. The Fund usually invests in equity securities of companies with large market capitalizations (those included in the S&P 500 Index, which as of December 31, 2011 ranged between $1.57 billion and $401.25 billion), but may also invest in equity securities of companies with medium market capitalizations (those included in the Russell Midcap$$ Index, which as of December 31, 2011 ranged between $117 million and $20.51 billion). The Fund invests in value equity securities, an investment strategy that emphasizes buying equity securities that appear to be undervalued. The Fund will also invest in real estate investment trusts and securities of foreign issuers.

Pre-processed text with boilerplate based on 0.1% *measure in bold and rest in red:*

Boilerplate top 0.1% : 58.82%

> **normal circumst fund invest least net asset plu borrow invest purpos dividend pay equiti secur** fund usual **invest equiti secur compani larg market capit** includ decemb rang billion billion **may also invest equiti secur compani medium market capit** includ russel midcap decemb **rang million billion fund** invest valu equiti secur invest strategi emphas buy equiti secur appear undervalu fund also **invest real estat invest trust** secur foreign issuer

Low boilerplate example: *Crsp_cl_grp: 2000410 Date: 2003-12-31*

Full text before pre-processing:

The Fund invests in the common stocks that comprise the Standard & Poor's Composite Stock Price Index ("S&P 100"), in an effort to provide investment results that correspond to or exceed the aggregate price and dividend performance of the S&P 100. The S&P 100 is a market capitalization-weighted index of 100 common stocks from a broad range of industries. Under normal conditions, at least 80% of the Fund's assets will be invested to correspond as closely as possible to the relative weighting of the S&P 100 in order to attempt to achieve a high degree of correlation between the performance of the Fund's portfolio and that of the S&P 100. The remaining 20% of the Fund's assets will, under normal conditions, also be invested in stocks that are included in the S&P 100, but the Fund's position in such stocks may be greater (overweighted) or less (underweighted) compared to such stocks' weightings in the S&P 100.

Pre-processed text with boilerplate based on 1% *measure in bold and rest in red:*

Boilerplate top 0.1% : 5.88%

fund invest common stock compris standard poor composit stock price effort provid invest result correspond exceed aggreg price dividend perform market capit weight common stock broad rang industri normal condit least fund asset invest correspond close possibl rel weight order attempt achiev high degre correl perform fund portfolio remain fund asset normal condit also invest stock includ fund posit stock may greater overweight less underweight compar stock weight

References

Abis, S. (2020). Man vs. machine: Quantitative and discretionary equity management. Working paper.

Abis, S. and A. Lines (2020). Text-based mutual fund peer groups. Working paper.

Akey, P., A. Z. Robertson, and M. Simutin (2021). Closet active management of passive funds. Working paper.

Baks, K. P., A. Metrick, and J. Wachter (2001). Should investors avoid all actively managed mutual funds? A study in bayesian performance evaluation. *Journal of Finance 56*(1), 45–85.

Berk, J. and R. Green (2004). Mutual fund flows and performance in rational markets. *Journal of Political Economy 112*(6), 1269–1295.

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

Brown, D. P. and Y. Wu (2016). Mutual fund flows and cross-fund learning within families. *Journal of Finance 71*(1), 383–424.

Buffa, A. M. and A. Javadekar (2020a). The allocation of talent across mutual fund strategies. Working paper.

Buffa, A. M. and A. Javadekar (2020b). Decomposing fund activeness. Working paper.

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

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

Choi, D., B. Kahraman, and A. Mukherjee (2016). Learning about mutual fund managers. *Journal of Finance 71*(6), 2809–2860.

Daniel, K., M. Grinblatt, S. Titman, and R. Wermers (1997). Measuring mutual fund performance with characteristic-based benchmarks. *Journal of Finance 52*(3), 1035 – 1058.

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

Fama, E. F. and K. R. French (2010). Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance 65*(5), 1915–1947.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Franzoni, F. and M. C. Schmalz (2017). Fund flows and market states. *Review of Financial Studies 30*(8), 2621–2673.

Gârleanu, N. and L. H. Pedersen (2018). Efficiently inefficient markets for assets and asset management. *Journal of Finance 73*(4), 1663–1712.

Harvey, C. R. and Y. Liu (2019). Cross-sectional alpha dispersion and performance evaluation. *Journal of Financial Economics 134*(2), 273–296.

Hillert, A., A. Niessen-Ruenzi, and S. Ruenzi (2016). Mutual fund shareholder letters: flows, performance, and managerial behavior. Working paper.

Huang, J. C., K. D. Wei, and H. Yan (2012). Investor learning and mutual fund flows. Working paper.

Hwang, B.-H. and H. H. Kim (2017). It pays to write well. *Journal of Financial Economics 124*(2), 373–394.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1994). *Continuous Univariate Distributions*. John Wiley and Sons.

Jones, C. S. and J. Shanken (2005). Mutual fund performance with learning across funds. *Journal of Financial Economics 78*(3), 507–552.

Kacperczyk, M., C. Sialm, and L. Zheng (2008). Unobserved actions of mutual funds. *Review of Financial Studies 21*(6), 2379–2416.

Kacperczyk, M., S. Van Nieuwerburgh, and L. Veldkamp (2016). A rational theory of mutual funds' attention allocation. *Econometrica 84*(2), 571–626.

Kincaid, P. J., R. P. Fishburne, R. L. J. Rogers, and B. S. Chissom (1975). Derivation of new readability formulas (automated readability index, fox count and flesh reading ease formula) for navy enlisted personal. *Institute for Simulation and Training 56*.

Kostovetsky, L. and J. B. Warner (2020). Measuring innovation and product differentiation: Evidence from mutual funds. *Journal of Finance 75*(2), 779–823.

Krakow, N. J. and T. Schäfer (2020). Mutual funds and risk disclosure: Information content of fund prospectuses. Working Paper 20-54.

Pástor, L. and R. F. Stambaugh (2002). Investing in equity mutual funds. *Journal of Financial Economics 63*(3), 351–380.

SEC (2016). How to read a mutual fund prospectus. Investor Alerts and Bulletins.

Sheng, J., N. Xu, and L. Zheng (2021). Do mutual funds walk the talk? A textual analysis of risk disclosure by mutual funds. Working paper.

Shive, S. and H. Yun (2013). Are mutual funds sitting ducks? *Journal of Financial Economics 107*(1), 220–237.

Sirri, E. R. and P. Tufano (1998). Costly search and mutual fund flows. *Journal of Finance 53*(5), 1589–1622.

Starks, L. T. and S. Y. Sun (2016). Economic policy uncertainty, learning and incentives: Theory and evidence on mutual funds. Working paper.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance 62*(3), 1139–1168.

Tucker, A. M. and Y. Xia (2020). The promise & perils of plain english mutual fund disclosures. *Working Paper*.

Tucker, A. M., Y. Xia, and S. Smelcer (2020). It ain't just what funds disclose (it's the way that they do it). Working paper.

Zhang, A. L. (2020). Uncovering mutual fund private information with machine learning. Working paper.

Zhu, Q. (2020). The missing new funds. *Management Science*.

Figure 1: Equilibrium

In this figure we plot a fund manager's expected utility $\mathbb{E}[v_i(q_{i2} \cdot f)]$ (top-left panel), the expected fund size $\mathbb{E}[q_{i2}]$ (top-right panel), and the fund flow volatility $\mathbb{V}\mathrm{ar}[q_{i2}/q_{i1} - 1]$ (bottom-left panel) at time 2, as a function of the fund strategy γ_i. The bottom-right panel plots the flow-performance sensitivity $\lambda_p^\alpha/(\bar{\alpha} - f)$, as a function of the factor return F_2. Parameter values are: $\mu_F = 0.05$, $\sigma_F = 0.1$, $\sigma_\varepsilon = 0.15$, $\bar{\alpha} = 0.02$, $\bar{b} = 1$, $\sigma_\alpha = \sigma_b = \sigma_\eta = \sigma_\gamma = 0.25$, $f = 0.01$, $c = 0.001$.

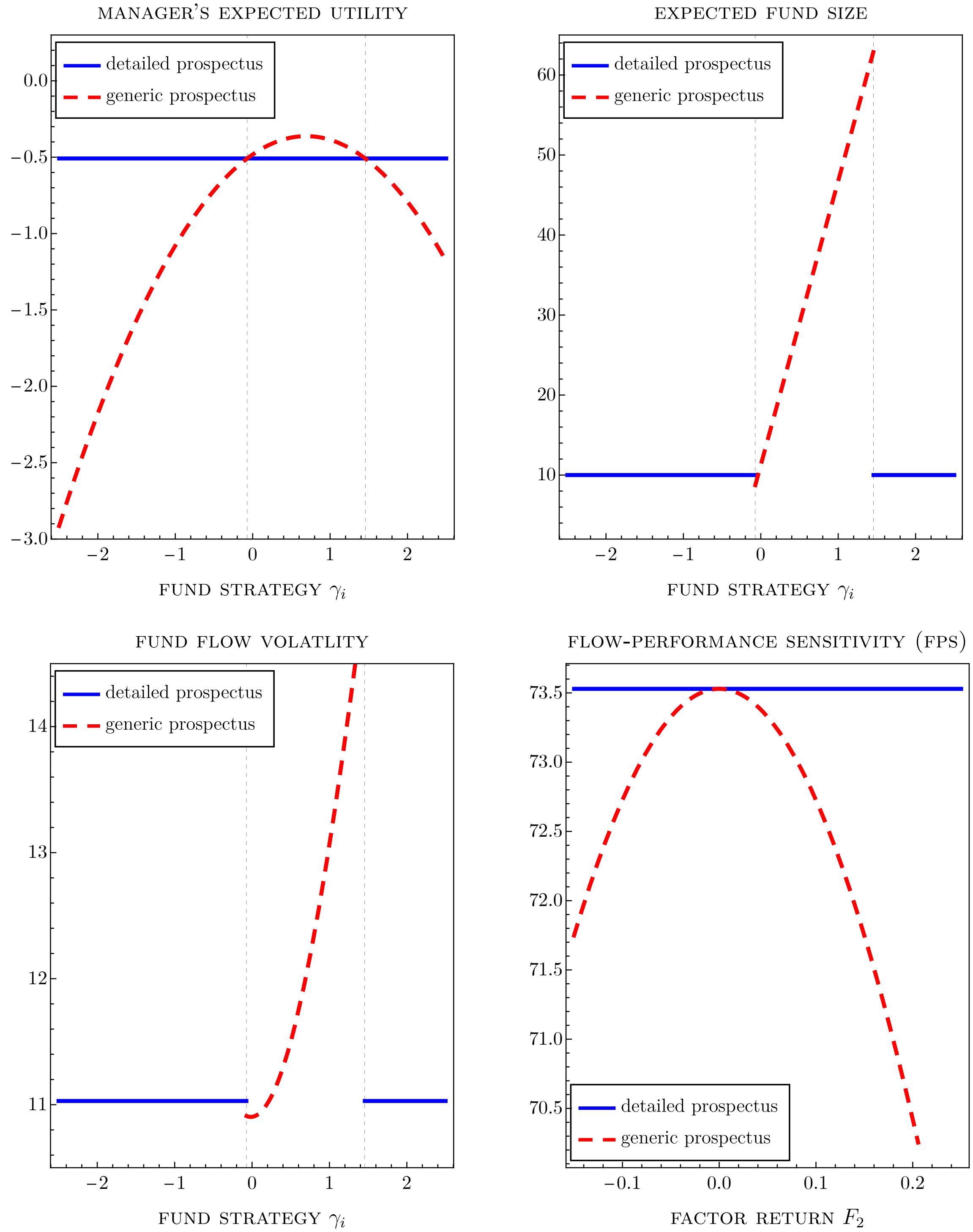

Figure 2: Boilerplate Phrases

This figure contains a list of the top 100 boilerplate phrases, ordered by frequency.

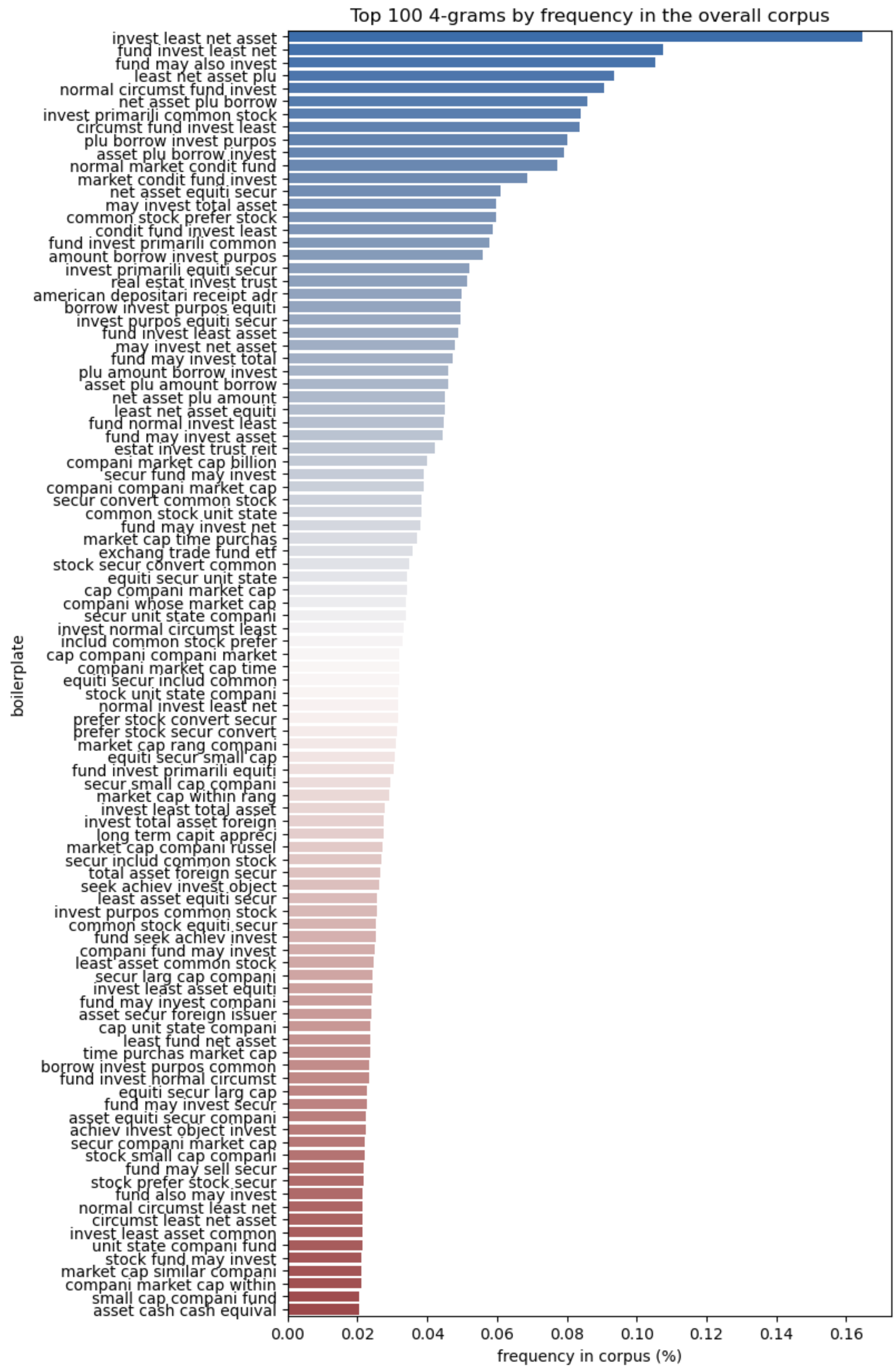

Table I: Boilerplate and Text Length/Complexity/Financial Content

This table shows the association between the percentage of boilerplate phrases in each strategy description, and the length, complexity, and financial content of those descriptions. Dependent variables are as follows. Column 1: Flesch-Kincaid grade-level complexity score; column 2: length of the description in words; columns 3-6: percentage of financial terms present in the description, as defined in equation (37). Column 3 displays results for the un-weighted ratio of financial terms to total terms, while columns 4-6 weight each term by the inverse of their frequency among other PIS descriptions, where a indicates how heavily the more common terms are down-weighted. Primary independent variable: fund-month percentage of boilerplate, as defined in equation (35). Fund level control variables: fund age, size, expense ratio, turnover ratio, net flow, flow volatility, and style, measured as the weighted average of the beta of stocks held with respect to the Fama-French 5 factors and momentum. All regressions include month and strategy fixed effects. The regression only includes fund-month combinations for which a prospectus is available (no forward filling). t-statistics are in parenthesis. ***, **, * indicate significance at less than 1%, 5%, and 10%, respectively.

	Complexity	Length	Financial Content			
			$a = 0$	$a = 1$	$a = 2$	$a = 4$
	(1)	(2)	(3)	(4)	(5)	(6)
Boiler %	-2.874***	-353.7***	-0.000351	-0.161*	-0.312***	-0.583***
	(-5.39)	(-5.70)	(-0.01)	(-2.09)	(-3.24)	(-4.57)
Ln(TNA)	-0.0928**	6.366***	-0.00382	-0.00459	-0.00548	-0.00756
	(-2.60)	(2.92)	(-1.40)	(-1.50)	(-1.57)	(-1.69)
Ln(Age)	-0.0798	-24.61***	-0.00587	-0.00765	-0.00942*	-0.0127*
	(-1.09)	(-6.87)	(-1.37)	(-1.62)	(-1.75)	(-1.82)
Expenses	11.33	920.8	0.384	0.730	0.981	1.247
	(0.48)	(0.83)	(0.30)	(0.54)	(0.66)	(0.67)
Turnover	-0.0767	2.131	-0.0171***	-0.0149**	-0.0132*	-0.0104
	(-0.79)	(0.71)	(-3.23)	(-2.40)	(-1.76)	(-1.01)
Flow Volatility	0.0563	-8.937	0.00984	0.0123	0.0145*	0.0177*
	(0.50)	(-1.37)	(1.47)	(1.68)	(1.80)	(1.87)
Flow	-0.00825*	0.150	-0.000617	-0.000685	-0.000743	-0.000832
	(-2.10)	(0.35)	(-1.62)	(-1.65)	(-1.64)	(-1.59)
Style (HML)	0.0400	7.844**	-0.00117	0.00152	0.00430	0.00987
	(0.59)	(2.52)	(-0.18)	(0.24)	(0.65)	(1.29)
Style (SMB)	-0.263***	3.296	-0.0262**	-0.0284***	-0.0313***	-0.0380***
	(-3.53)	(0.62)	(-2.69)	(-2.94)	(-3.21)	(-3.57)
Style (MOM)	-0.218*	-7.075	-0.00713	-0.00890	-0.0109	-0.0151
	(-1.76)	(-1.40)	(-0.93)	(-1.02)	(-1.08)	(-1.15)
Style (RMW)	0.0329	1.538	0.00946**	0.00934**	0.00935*	0.00969
	(0.64)	(0.43)	(2.35)	(2.14)	(1.96)	(1.71)
Style (CMA)	0.0366	-0.655	-0.00229	-0.00121	-0.000198	0.00150
	(0.81)	(-0.26)	(-0.77)	(-0.36)	(-0.05)	(0.31)
Intercept	15.28***	453.7***	1.110***	1.163***	1.215***	1.315***
	(30.85)	(15.27)	(39.17)	(39.83)	(37.75)	(31.17)
SPG+Month FE	Y	Y	Y	Y	Y	Y
SPG+Month Cluster	Y	Y	Y	Y	Y	Y
R2	0.07	0.23	0.04	0.05	0.06	0.09
Obs	24,847	24,847	24,847	24,847	24,847	24,847

Table II: Prospectus Choice

This table documents the relationship between fund strategy specialization and the percentage of boilerplate phrases in prospectus strategy descriptions. Dependent variables are as follows. Columns 1-2: fund-month percentage of boilerplate, as defined in equation (35); columns 3-4: an indicator variable that takes a value of 1 when boilerplate is above its unconditional mean. Primary independent variable: strategy specialization, as measured by holdings dispersion with respect to other funds in the same Strategy Peer Group (SPG). Holdings dispersion is defined as the logged sum of squared differences between each fund's portfolio weight vector and the average portfolio weight vector for the entire peer group. Fund level control variables: fund age, size, expense ratio, turnover ratio, net flow, flow volatility, and style, measured as the weighted average of the beta of stocks held with respect to the Fama-French 5 factors and momentum. All regressions include month fixed effects. Columns 2 and 4 also include SPG fixed effects. t-statistics are in parenthesis. ***, **, * indicate significance at less than 1%, 5%, and 10%, respectively.

	Boiler %		D50_Boiler	
	(1)	(2)	(3)	(4)
Ln(Dispersion)	-0.00712***	-0.0143***	-0.0329***	-0.0583***
	(-19.65)	(-7.94)	(-22.86)	(-9.25)
Ln(TNA)	0.00274***	0.000574	0.00945***	0.00128
	(15.72)	(0.62)	(12.42)	(0.26)
Ln(Age)	-0.00945***	-0.00415*	-0.0333***	-0.0127
	(-27.96)	(-1.87)	(-27.05)	(-1.52)
Expenses	-0.935***	-1.414	-3.169***	-5.488*
	(-15.75)	(-1.72)	(-11.17)	(-1.78)
Turnover	0.00436***	0.00292	0.0104***	0.00284
	(10.61)	(1.02)	(5.63)	(0.28)
Flow	-0.000122*	-0.0000700	-0.00109***	-0.000886
	(-1.74)	(-0.48)	(-3.33)	(-1.53)
Flow Volatility	0.00293***	0.00687	0.0248***	0.0386
	(2.99)	(1.33)	(6.67)	(1.57)
Style (HML)	0.00332***	0.00359	0.0135***	0.0169
	(4.68)	(1.30)	(5.33)	(1.25)
Style (SMB)	0.0175***	0.0150***	0.0762***	0.0650***
	(23.06)	(3.15)	(24.98)	(4.07)
Style (MOM)	0.0116***	0.00870	0.0385***	0.0289*
	(9.19)	(1.74)	(7.64)	(1.82)
Style (RMW)	0.000409	0.000407	-0.00103	-0.000317
	(0.74)	(0.26)	(-0.44)	(-0.05)
Style (CMA)	-0.00161***	-0.000468	-0.00385*	0.00178
	(-3.05)	(-0.26)	(-1.67)	(0.22)
Intercept	0.209***	0.167***	0.460***	0.316***
	(82.57)	(14.15)	(39.40)	(6.59)
Month FE	Y		Y	
SPG+Month FE		Y		Y
Month Cluster	Y		Y	
SPG+Month Cluster		Y		Y
R2	0.04	0.16	0.04	0.11
Obs	259,017	259,017	259,017	259,017

Table III: Fund Size

This table documents the relationship between boilerplate percentage and fund size. Dependent variable: the natural logarithm of assets under management. Primary independent variables are as follows. Columns 1-2: fund-month percentage of boilerplate, as defined in equation (35); columns 3-4: an indicator variable that takes a value of 1 when boilerplate is above its unconditional mean. Fund level control variables: fund age, size, expense ratio, turnover ratio, net flow, flow volatility, and style, measured as the weighted average of the beta of stocks held with respect to the Fama-French 5 factors and momentum. All regressions include month fixed effects. Columns 2 and 4 also include Strategy Peer Group (SPG) fixed effects. t-statistics are in parenthesis. ***, **, * indicate significance at less than 1%, 5%, and 10%, respectively.

	Ln(TNA)			
	(1)	(2)	(3)	(4)
Boiler %	0.325***	0.237*		
	(16.02)	(1.80)		
D50_B			0.0624***	0.0435
			(12.46)	(1.25)
D_Middle_B_01			-0.0485***	-0.0427
			(-10.14)	(-1.12)
Ln(Age)	0.713***	0.704***	0.712***	0.704***
	(226.68)	(22.92)	(219.34)	(22.77)
Expenses	-105.8***	-107.6***	-106.2***	-108.0***
	(-30.68)	(-16.77)	(-30.77)	(-16.78)
Turnover	-0.214***	-0.204***	-0.215***	-0.205***
	(-33.52)	(-4.52)	(-33.21)	(-4.54)
Flow	0.0165***	0.0166***	0.0165***	0.0167***
	(19.27)	(8.44)	(19.35)	(8.36)
Flow Volatility	2.255***	2.223***	2.251***	2.219***
	(116.17)	(24.74)	(115.39)	(24.44)
Style (HML)	0.0239***	0.0425**	0.0237***	0.0423**
	(2.70)	(2.17)	(2.67)	(2.17)
Style (SMB)	0.0873***	0.0892***	0.0876***	0.0891***
	(14.90)	(3.62)	(15.21)	(3.55)
Style (MOM)	0.0149	0.0205	0.0166	0.0217
	(0.93)	(0.54)	(1.03)	(0.58)
Style (RMW)	-0.0353***	-0.0293*	-0.0344***	-0.0286*
	(-5.25)	(-1.82)	(-5.15)	(-1.80)
Style (CMA)	0.0157***	0.0203	0.0163***	0.0209
	(2.84)	(1.02)	(2.95)	(1.04)
Intercept	2.993***	3.077***	3.067***	3.137***
	(64.56)	(16.85)	(64.32)	(16.90)
Month FE	Y		Y	
SPG+Month FE		Y		Y
Month Cluster	Y		Y	
SPG+Month Cluster		Y		Y
R2	0.49	0.49	0.49	0.49
Obs	259,052	259,052	258,209	258,209

Table IV: Flow Volatility

This table documents the relationship between boilerplate percentage and fund flow volatility. Dependent variable: 12-months rolling volatility in net fund flows. Primary independent variables are as follows. Column 1: fund-month measure of boilerplate, as defined in equation (35); column 2: indicator variable taking value of 1 when boilerplate is above its unconditional mean. Fund level control variables: age, size, expense ratio, turnover ratio, net funds flows growth, volatility of net fund flows and style measured as the weighted average of the beta of stocks held with respect to the Fama-French 5 factors and momentum. All regressions include month fixed effects and clustering. Regressions are conditioned on observations belonging to each SPG. Only the coefficients of interest are reported. *t*-statistics are in parenthesis. ***, **, * indicate significance at less than 1%, 5%, and 10%, respectively.

	Flow Volatility	
Boilerplate measure	Boiler %	D50_Boiler
	(1)	(2)
LargeCap	0.0747***	0.0363***
	(8.44)	(14.05)
MidCap	-0.0653***	-0.00962***
	(-4.09)	(-2.66)
SmallCap	0.0758***	0.0155***
	(5.96)	(4.35)
Fundamental	0.106***	0.0337***
	(8.21)	(8.85)
Quantitative	0.138***	0.0490***
	(4.80)	(9.96)
LongTerm	-0.305***	-0.0371***
	(-10.78)	(-8.27)
PE-Ratio	0.0505***	-0.00380
	(3.37)	(-1.07)
Intrinsic Value	-0.0338	0.000753
	(-1.12)	(0.14)
Defensive	0.0317*	-0.000279
	(1.72)	(-0.08)
Dividends	0.169***	0.0364***
	(7.75)	(7.26)
Derivatives	-0.215***	-0.0658***
	(-6.80)	(-10.89)
Competitive Advantage	0.0952***	0.0309***
	(3.90)	(5.89)
Products & Services	0.00952	-0.0109***
	(0.42)	(-3.24)
Tax	0.273***	0.0672***
	(9.23)	(10.47)
Fixed Income	-0.00257	0.00262
	(-0.09)	(0.52)
Foreign (ADR)	0.101***	0.0425***
	(3.81)	(5.99)
Foreign (EM)	-0.101***	-0.00507
	(-3.48)	(-0.57)
Obs	10,443	10,443

Table V: Flow-performance Sensitivity

This table documents the relationship between boilerplate percentage and fund flow-performance sensitivity. Dependent variable: net fund flows. Primary independent variables: lagged active returns and its interaction with fund-month percentage of boilerplate, as defined in equation (35), (column 1-2), or with an indicator variable that takes a value of 1 when boilerplate is above its unconditional mean (columns 3-4). Fund level control variables: age, size, expense ratio, turnover ratio, net funds flows growth, volatility of net fund flows and style measured as the weighted average of the beta of stocks held with respect to the Fama-French 5 factors and momentum. All regressions include month fixed effects. Columns 2 and 4 also include SPG fixed effects. *t*-statistics are in parenthesis. ***, **, * indicate significance at less than 1%, 5%, and 10%, respectively.

	Net Flows (t+1)			
	(1)	(2)	(3)	(4)
Active Returns	0.234***	0.234***	0.229***	0.229***
	(14.58)	(16.40)	(17.73)	(17.91)
Active Returns × Boiler	-0.105*	-0.104*		
	(-1.68)	(-1.88)		
Active Returns × D50_Boiler			-0.0304**	-0.0304**
			(-2.26)	(-2.42)
Ln(TNA)	0.0133*	0.0138	0.0134*	0.0144
	(1.91)	(1.23)	(1.88)	(1.29)
Ln(Age)	-0.373***	-0.374***	-0.381***	-0.382***
	(-25.46)	(-15.00)	(-25.93)	(-15.84)
Expenses	-17.70***	-18.30***	-15.77***	-15.94***
	(-5.05)	(-3.83)	(-4.51)	(-3.18)
Turnover	-0.113***	-0.0951***	-0.106***	-0.0908**
	(-6.38)	(-3.39)	(-5.79)	(-2.87)
Flow	0.376***	0.375***	0.376***	0.375***
	(57.97)	(37.16)	(57.65)	(38.18)
Flow Volatility	-0.0657**	-0.0661*	-0.0675**	-0.0675*
	(-2.17)	(-2.04)	(-2.14)	(-1.92)
Style (HML)	0.0769**	0.0761**		
	(2.52)	(2.35)		
Style (SMB)	0.0626***	0.0687**		
	(2.95)	(2.41)		
Style (MOM)	0.263***	0.273***		
	(6.36)	(4.91)		
Style (RMW)	0.0670***	0.0681***		
	(4.15)	(3.03)		
Style (CMA)	0.0314	0.0296		
	(1.57)	(1.16)		
Intercept	1.951***	1.948***	1.970***	1.961***
	(22.60)	(14.62)	(22.83)	(15.22)
Month FE	Y		Y	
SPG+Month FE		Y		Y
Month Cluster	Y		Y	
SPG+Month Cluster		Y		Y
R2	0.18	0.18	0.18	0.18
Obs	264,333	264,333	269,890	269,890