Do Mutual Funds Keep Their Promises?*

Simona Abis and Anton Lines

Columbia University

First Draft: May 12, 2020

This Version: January 6, 2022

Abstract

Mutual fund prospectuses contain a wealth of qualitative information about fund strategies, yet a systematic analysis of this content is missing from the literature. We use machine learning to group together funds with similar strategy descriptions, and ask whether they act in accordance with the text. Despite weak legal recourse for investors, we find that mutual funds largely do keep their promises. We document a market-based disciplinary mechanism: when funds diverge from their group's core strategy, investors withdraw capital. Funds respond to these punitive outflows by reducing their divergence from the peer group average at a faster rate.

Keywords: Mutual Fund Strategies, Prospectuses, Market Discipline, Fund Flows, Unsupervised Machine Learning

*Correspondence: Simona Abis, e-mail: simona.abis@gsb.columbia.edu; Anton Lines, e-mail: anton.lines@gsb.columbia.edu. This paper was previously circulated under the title "Text-Based Mutual Fund Peer Groups". We are especially grateful to Svetlana Bryzgalova, Kent Daniel, Leonard Kostovetsky, Paul Tetlock, and conference/seminar participants at AFA 2021, Future of Financial Information Conference 2020, Columbia GSB, Minnesota 3M Seminar, London Business School, SFI, Toulouse, and City U HK for helpful comments. We also thank Luofeng Zhou, Tian Zhang, Ritt Keerati, and Gary Buranasampatanon for excellent research assistance.

1 Introduction

Considering the substantial search frictions that have been documented in the mutual fund industry (e.g. Sirri and Tufano (1998); Roussanov et al. (2021)), the *Principal Investment Strategy (PIS)* descriptions included in fund prospectuses have high potential value to investors. Mandated by the Securities and Exchange Commission, these descriptions are intended to "explain in general terms how the fund's adviser decides what securities to buy and sell ... [and] provide investors with essential information about the fund's investment approach and how the fund's portfolio would be managed." (SEC rule S7-10-97). But are PIS descriptions actually informative, and can their accuracy be relied upon? In public comments on the proposed rule adoption, investors expressed concerns about "boilerplate" disclosure.[1] And although SEC regulations prohibit false or misleading statements in official filings, investors have largely been unsuccessful in bringing litigation against fund advisers on this basis.[2] Regulatory enforcement actions by the SEC itself have been more successful, but have typically targeted the most flagrant errors, such as failure to disclose the use of leverage/derivatives, or deliberate inflation of past performance statistics.[3] In less extreme cases, the question of whether funds are adhering to their promised strategies is often open to interpretation and thus more difficult to enforce.

In this paper we ask whether the PIS descriptions of US active equity mutual funds are predictive of the funds' actual investment behavior, how investors react if a fund deviates from its "promised" strategy, and whether funds change their behavior in response to investor reactions. To enable this analysis, we obtain a comprehensive sample of prospectuses directly from EDGAR (the SEC's Electronic Data Gathering, Analysis, and Retrieval system) covering the period 2000-2017. Then, using a simple tool from unsupervised machine learning—the *k-means* algorithm—we distill the text into interpretable *strategy peer groups (SPGs)*: clusters of similar descriptions that represent distinct investment approaches.

Despite the legal challenges discussed above, we find that investors exert effective disciplinary pressure on "promise-breaker" funds, by withdrawing capital after months in which

[1] https://www.sec.gov/rules/final/33-7512r.htm

[2] In a landmark 2011 case, the US Supreme Court held that a fund adviser could not be held liable under SEC rule 10b-5 for false statements in its fund's prospectus because the adviser and the fund were separate legal entities (Janus Capital Group, Inc. v. First Derivative Traders, 564 U. S. 135 (2011)). Additionally, claims against fund managers that rely on sections 11(a) and 12(a)(2) of the Securities Act of 1933 have typically been dismissed by courts because mutual fund shares trade at net asset value (NAV), and therefore any depreciation in share value cannot be shown to be caused by the misleading statement (Geffen (2009)).

[3] See, for example, https://www.sec.gov/news/press-release/2012-2012-110htm; https://www.sec.gov/enforce/ia-5489-s.

funds' portfolio weights diverge from the core strategy of their peer group. Plausibly due to this endogenous response by investors, we find that most mutual funds *do* keep their promises. Funds' portfolio weights are on average closer to their own group's core strategy than that of other SPGs, and their average characteristics correspond (where measurable) to natural readings of the SPG's aggregated PIS text. Divergences from peer group averages mean-revert over time, and do so at a faster rate when high past divergences are coupled with investor outflows, suggesting that funds are responsive to market-based disciplinary forces. Lastly, we find evidence that investors make use of textual peer group comparisons when evaluating manager skill: SPG-adjusted returns have incremental explanatory value for fund flows beyond standard risk model alphas, and similar predictive power.

Our full analysis is structured as follows. As a preliminary exercise, we rule out the concern that strategy descriptions are mostly just legal boilerplate by documenting high cross-sectional variation in length, linguistic complexity, positive/negative sentiment, and frequencies of "litigious" and "uncertainty" words (Loughran and McDonald (2011)).

We then describe the construction of our strategy peer groups, using the *k-means* algorithm.[4] We start by encoding each fund's PIS description as a vector of relative word frequencies. Then, after randomly initializing k *centroid* vectors in the space spanned by the entire corpus of descriptions, each PIS is assigned to the closest centroid by Euclidean distance. Following this assignment, each centroid vector is recomputed as the mean vector of all documents assigned to it. The last two steps are repeated until convergence, resulting in k *clusters*, each with minimal distance between their constituents and the geometric center. To determine the number of clusters, we develop two criteria which we call *density* and *stability*. Intuitively, the density criterion is satisfied when new clusters contain enough funds and are sufficiently linguistically distinct, and the stability criterion is satisfied when most funds are classified into the same groups across consecutive ks.

According to these criteria, the universe of U.S. domestic active equity strategies can be optimally categorized into 17 SPGs, as shown in figure 1. Some strategies, such as *Large Cap*, *Mid Cap*, and *Small Cap*, correspond to the usual categories used by academics and industry consultants, but most appear to go beyond. Some are associated with firm characteristics

[4] A popular alternative for textual topic modeling is Blei et al. (2003)'s Latent Dirichlet Allocation (LDA), which estimates a posterior distribution over several topics for each document. Although this method yields similar results to k-means if we assign peer groups based on the most probable topic for each fund, we use k-means in our main specifications for two reasons: (i) philosophically, minimizing the distance to a single cluster center is more consistent with our interpretation of average portfolio weights as core strategies; and (ii) empirically, the topics generated by LDA have slightly less clear boundaries (e.g. two very similar topics for *quantitative*, while *small cap* and *large cap* strategies are merged into one topic; see appendix B).

(*Dividends*; *New Products & Services*; *Competitive Advantage*; *Price-Earnings Ratio*), some with investment philosophies (*Quantitative*; *Fundamental*; *Intrinsic Value*; *Long Term*; *Defensive*; *Tax*), some with secondary asset classes (*Fixed Income*; *Derivatives*), and some with international markets (*Foreign (ADR)*; *Foreign (Emerging Markets)*).[5]

To test whether SPGs capture actual fund behavior, we start by building a measure of strategy adherence (or, equivalently, strategy divergence). This approach has the advantages of quantifiability and objectivity but requires a simplifying assumption: that divergences from core strategies within each peer group are purely idiosyncratic. If this condition is satisfied, average idiosyncratic divergences will tend to zero as the number of funds increases, and the average portfolio weight vector within each SPG will capture its core strategy. Our measure of strategy divergence is therefore the (log-transformed) sum of squared differences between each fund's portfolio weight vector and the SPG-average weight vector.[6]

If funds are following their promised strategies, they should diverge *less* from their own peer group average than from a placebo strategy (i.e., some other peer group average). Indeed, across our full sample, we find that funds' portfolio weights are about 10% more similar to their own SPG average (at the 1% significance level), and this finding is robust to using alternative definitions of core strategies based on average returns. We also run *pairwise* fund-level regressions of fund similarity on dummies for membership in the same SPG, where similarity is measured by portfolio weight distances or differences between the first four return moments. We control for membership in alternative peer groups such as Daniel et al. (1997) or Fama and French (1993), as well as lagged dependent variables. These results indicate that funds generally follow their promised strategies, and that text-based peer groups capture a novel dimension of similarity.

Next, we examine investors' response to fund strategy divergence. We regress one-month-ahead net capital flows on our SPG divergence measure, controlling for performance and other fund-level attributes such as log total net assets, log fund age, and expense and turnover ratios. We also include funds' divergence from an alternative peer group average based on terciles of the Daniel et al. (1997) size, value, and momentum characteristics (which also nest the nine Morningstar equity style categories). The estimated coefficients on SPG divergence are negative and statistically significant (at either the 1% or 5% level, depending on the specification), indicating that net flows decline when funds diverge more from their core strategies. A one-standard deviation increase in divergence leads to lower flows by $450,000

[5] All funds in our sample hold an average of at least 80% of assets in U.S. common stock; however, what funds do with the remaining 20% is sometimes their most distinguishing feature.

[6] The log transformation is applied so that the variable is approximately normally distributed.

to $490,000 in the first month, and the effect persists for around twelve months. In terms of percentage flows, the cumulative annual effect is approximately 0.5%, which is a decrease of about 24% of the unconditional sample mean.[7] Importantly, we observe these responses only when computing divergence using the text-based peer group averages, and not when using the alternative DGTW/Morningstar peer groups, indicating that investors look beyond the classical size-value dimension when choosing among fund strategies.

The performance variables used as controls in the above regressions are also interesting in their own right. In addition to raw fund returns and alphas from standard risk models (CAPM, Fama-French-Carhart, and Fama and French (2015) five factors plus momentum), we also include peer-adjusted returns (i.e., raw returns minus the average return among funds in the same peer group) for both our SPGs and the alternative DGTW groups. Both peer-adjusted return variables predict future flows at the 1% significance level, independently from the standard alphas. Moreover, the effect of the SPG-adjusted returns is about twice that of the DGTW-adjusted returns, and about the same as CAPM alpha.

Do the investor responses documented above cause funds to change their behavior? To test this hypothesis, we regress funds' future SPG divergences on past divergences, net fund flows, and the interaction between them (along with the usual controls). Crucially, we should not expect *all* outflows to affect fund behavior, only those associated with high past divergences. Overall, we find that strategy divergence is mean-reverting at a rate of about 50% per year, confirming that funds anchor around the SPG core strategy. However, the main coefficient of interest is on the interaction term, which indicates that funds mean-revert at a faster rate following high past divergence combined with investor outflows. As an illustration of the magnitude, when time-t divergence is two standard deviations above its mean, then for each one standard deviation decline in flows, the mean reversion rate is increased by a factor of 1.16 in the next month and by 1.12 over the following year.

Naturally, since neither strategy divergence nor flows are exogenously determined, there is a concern that the correlations we document could be explained by alternative mechanisms. For example, funds that suffer losses after diverging from the peer group average may experience performance-related outflows. These funds may then meet redemptions by offloading the losing positions, which would result in lower future deviations. Another possibility is that funds may raise their management or marketing fees when they become more unique, which could lead to outflows. A third possibility is that when a fund family loses a resale

[7] While this magnitude may seem modest, bear in mind that we observe only equilibrium outcomes. Since funds mostly follow their promised strategies, it is plausible that they *anticipate* investor responses and pre-emptively adjust their behavior, which would dampen the observed investor response.

contract with a particular broker, it may anticipate outflows and attempt to compensate by increasing its funds' uniqueness to attract more attention elsewhere. However, our particular set of control variables allows us to rule out alternative explanations associated with fund performance or fees (as in the first two examples above), and the inclusion of time-varying fund family fixed effects eliminates explanations based on changing family-wide policies or distribution channels. In general, we are able to exclude alternatives stemming from the main drivers of flows identified in the prior literature.

Lastly, we supplement our quantitative core strategy measurement with a narrative approach. Specifically, we ask whether average stock characteristics for each SPG are consistent with common-sense readings of their aggregated PIS text. While this analysis is necessarily subjective, it provides a useful sanity check. In support of our previous conclusions, we find that (in most cases) funds' holdings align with their strategy descriptions. For example, *Dividend* funds hold stocks in older, larger firms with higher than average dividend yields, lower investment, and less excess cash; *New Products & Services* funds invest in firms from more innovative industries, with higher investment and R&D expenditure; *Competitive Advantage* funds invest in firms with higher profitability; and *Foreign (EM)* funds hold more ADRs and foreign-incorporated firms.

We are the first to provide a comprehensive analysis of the full "Principal Investment Strategies" (PIS) descriptions for active equity mutual funds. Abis (2020) uses the same underlying data, but only to identify quantitative funds. Our methodology allows us to uncover the complete strategy landscape without prior knowledge of what the strategies are.[8] Kostovetsky and Warner (2020) construct a textual measure of product differentiation, but do not explore the described strategies. Akey et al. (2021) use PIS text to determine the investment objectives of a different segment of the industry: ETFs and index funds. Other recent papers examine summary prospectuses: Krakow and Schäfer (2020) measure textual uniqueness within fund families as a proxy for disclosure informativeness; and Sheng et al. (2021) examine summary risk descriptions, finding that greater risk disclosure is negatively associated with future performance (but has no relation to fund flows).

Our paper contributes to the vast empirical literature on fund flows, which has previously examined responses to performance (e.g. Chevalier and Ellison (1997); Sirri and Tufano (1998); Barber et al. (2016); Berk and Van Binsbergen (2016)), fees (e.g. Barber et al. (2005); Ivković and Weisbenner (2009)), and marketing/product differentiation (e.g. Jain and Wu (2000); Cooper et al. (2005); Reuter and Zitzewitz (2006); Khorana and Servaes

[8] In a follow-up to our paper, Abis et al. (2021) develop and test a theoretical model of fund disclosure choice and investors' learning from this disclosure, relying on our SPGs to identify investment mandates.

(2012); Kostovetsky and Warner (2020)). We document a new driver of flows—strategy divergence. Our results are inconsistent with a pure product differentiation hypothesis, which would imply *higher* flows when funds are more unique compared to their peers, as in Kostovetsky and Warner (2020). However, despite the different conclusions, our findings are compatible with theirs since they use Morningstar categories to capture fund peer groups. One of the insights of our paper is that the true strategy landscape is more diverse; thus, what appears as a preference for fund uniqueness can be reinterpreted as diverse preferences for previously unmeasured strategies.

We also contribute to the literature on market discipline (see the survey in Flannery (1998)), which thus far has mostly focused on the banking sector. Moreover, since the response we document takes several months to accumulate, it is consistent with investors having limited attention (e.g., Barber and Odean (2007)) or bearing information/search costs (e.g. Hortacsu and Syverson (2004); Roussanov et al. (2020)). Finally, our analysis of the information content of mutual fund prospectuses highlights a potential alternative mechanism for alleviating investor search frictions, as opposed to welfare-reducing marketing expenses (Roussanov et al. (2021)).

Following this introduction, section 2 describes the data and construction of the strategy peer groups; section 3 presents our quantitative analysis of fund and investor behavior; section 4 presents the alternative narrative approach; and section 5 concludes.

2 Data and Clustering Methodology

2.1 Data

To construct our mapping of the mutual fund industry, we combine standard information about mutual fund characteristics, returns and holdings with a novel textual dataset of their "Principal Investment Strategy" descriptions, taken from mandatory disclosures to the SEC (prospectuses). Our combined sample runs from January 2000 to December 2017, covering $2,995$ unique funds and $320,750$ fund–month observations. Table 1 provides descriptive statistics for the final dataset.

Prospectuses: We obtain fund prospectuses from the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) of the SEC. The EDGAR system has been active since 1994, with a 2-year phase-in period. A number of subsequent reforms led to increasing standardization in disclosure formatting. In particular, as detailed in Release No. 33-7684, the SEC began to accept disclosures in HTML format in 2000. Prior to this year, the lack

of standardization renders textual analysis much less reliable, and coverage of funds is also much more sparse. We construct a comprehensive panel dataset of "Principal Investment Strategy" descriptions by fund–month, which we then merge to the traditional mutual fund datasets (see below). We are able to match 31,695 prospectuses to our funds of interest. Prospectuses may be published on any day of the year, and are often published less than once per quarter. Since any material change to the management of the fund must be reported to both the SEC and fund investors, for any month in which a prospectus is not available, we forward-fill that information using the latest available prospectus.

Fund characteristics and returns: We obtain fund characteristics and returns from the CRSP Survivorship-Bias-Free Mutual Fund dataset. We restrict the sample to equity funds and exclude international funds, sector funds, index funds, and underlying variable annuities. We account for incubation bias by excluding observations dated before the fund's first offer date (Evans (2010)). We also exclude funds with less than $5 million in Total Net Assets (TNA), following Kacperczyk et al. (2008). In the main sample, data are aggregated across all share classes of each fund for each time period. This is done by keeping the first offer date of the oldest share class, summing the TNA of all share classes, and averaging all other variables (e.g. fees, returns, turnover, etc), weighted by lagged TNA. Following Abis (2020), we identify each fund's share classes by constructing a comprehensive fund identifier using the CRSP Class Group identifier, the WFICN identifier in MFLinks, and fund names. This choice is particularly relevant for matching fund returns and characteristic to their holdings. Notably, the MFLinks table excludes many new funds in recent years (Shive and Yun (2013); Zhu (2020)). Finally, we exclude funds for which we have less than 12 months of observations.

Holdings: Fund foldings are obtained by combining the Thomson Reuters (formerly 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). The date of the switch is chosen to maximize coverage of active equity funds. We drop funds that hold fewer than 10 stocks, and funds that hold, on average, less than 80% of their assets (excluding cash) in common stock (Kacperczyk et al. (2008)). When observations are missing or only available quarterly, we forward-fill at a monthly frequency. Monthly holdings are already available for or 42% of the sample; 90% of the data is forward-filled for at most 1 quarter in total; and 99% is forward-filled for at most 2 quarters in total. Forward-filling is restricted to a maximum of 1 year.

Other: To construct some of the variables of interest we also use the CRSP monthly security database, Compustat, Fama–French factors, industry portfolios, and macroeconomic times series from FRED.

2.2 Principal Investment Strategies

The SEC requires all mutual fund families to publish quarterly prospectuses covering all of their funds. These prospectuses are divided into sections, each addressing a different regulatory question. In this paper we focus on a specific section: Principal Investment Strategies (PIS), corresponding to Item 9(c) of the N-1A mandatory disclosure form.[9] This item requires funds to disclose their main investing methodology, including the types of securities they tend to hold and the primary criteria used in selecting those securities. Funds provide narrative descriptions of their strategies, constrained only by the above requirements and the additional stipulation that they be written in "plain English" (rule 421(d) of the Securities Act).[10]

Figure 2 displays the cross-sectional distribution of PIS word counts, which show wide variation from as few as 20 to as many as 1500. This figure also indicates that the complete PIS sections are much longer than the excerpts provided by Morningstar (average word counts of 70 for Morningstar versus 306 in our sample), examined previously by Kostovetsky and Warner (2020).

We also observe significant variation in the textual sentiment and textual complexity of these sections. Panel 1 of figure 3 shows the distribution, across fund–month observations, of the Flesch-Kincaid grade level complexity measure (Kincaid et al. (1975)). Panel 2 displays the same distribution for the Flesch reading ease measure (Flesch (1948)). These measures are based on the relative number of total words to total sentences (average sentence length) and the relative number of total syllables to total words (average word syllable length) in a given text. They are calibrated to indicate, respectively, the number of years of schooling required to comprehend the text (panel 1) and a standardized readability index on a range of $[0, 100]$ (panel 2). The figure shows that there exists a large dispersion in the complexity (readability) of these sections, with most sections requiring between 5 to 25 years of schooling (with a mean of 13.7), or a reading ease score between 10 and 80.

Finally, we use the Loughran and McDonald dictionaries of *positive*, *negative*, *uncertainty* and *litigious* words to measure sentiment. These dictionaries are adapted to account for

[9] https://www.sec.gov/files/formn-1a.pdf
[10] https://www.sec.gov/rules/final/33-7497.txt

specific characteristics of financial language (Loughran and McDonald (2011)). Figure 4 shows the distribution of the frequency of *positive* (panel 1), *negative* (panel 2), *uncertainty* (panel 3) and *litigious* (panel 4) words for all pooled fund–month strategy descriptions. Here we also observe large cross-sectional variation in the sentiment of strategy sections.

These descriptive statistics suggest that the narrative descriptions in fund prospectuses may contain relevant and heterogeneous information about their strategies.

2.3 Strategy Peer Groups

We now describe our methodology for grouping funds into quantifiable and interpretable "Strategy Peer Groups" (SPGs) based on similarities in their PIS text.

Pre-processing: To convert the textual data to quantitative data, we use the "bag of words" approach. This procedure yields a list of all stemmed words and bi-grams (consecutive two-word combinations) for each document. We remove symbols, standard English stop-words (e.g. "is", "the", "and", etc), and a list of context-specific stop words.[11] We reduce each word to its root using the Porter stemmer algorithm (e.g. "company", "companies", ... = "compani").

The second step is to remove boilerplate language, which occurs in most prospectuses and is therefore less informative. This step reduces classification noise. To do this, we aggregate unique word stems found in any of the PIS sections into a corpus, then compute the frequency of all 4-grams (4 word combinations) in this corpus. For each document, we remove any 4-gram with a full-corpus frequency above the 0.1th percentile (this procedure removes 601 4-grams). We further remove single words and bi-grams that appear in more than 30% of PIS sections and in fewer than 5%. The remaining words and bi-grams are the "features" utilized in the clustering algorithm.

In the third step, we represent the entire corpus as a single matrix, whose columns are linguistic terms (words and bi-grams) and whose rows are the individual PIS sections (one for each fund-quarter). Each element in this matrix is the frequency of a particular feature in a particular PIS section, scaled by the number of sections in which that feature appears. This matrix is known as the term frequency–inverse document frequency, or $tfidf$, matrix.

[11] The full list of removed words, after stemming, is as follows: advis, alloc, asset, averag, bar, billion, chart, class, compani, describ, equiti, fmr, fund, iii, inform, invest, least, manag, may, might, money, mutual, net, nyse, object, page, portfolio, potenti, price, princip, prospectu, rang, reason, risk, russel, s&p, secur, sharehold, shown, state, stock, strategi, style, subadvis, total, unit, varieti, year.

Clustering: We use the *k-means* algorithm to group PIS sections by textual similarity. We choose this algorithm for its simplicity, and because it results in one cluster allocation per fund and explicitly maximizes differences between groups. This aligns with our measurement of core strategies as peer group average portfolio weights. By contrast, other popular models such as Latent Dirichlet Allocation (LDA) and Gaussian Mixture decompose each document into a weighted sum of topics.[12]

K-means takes as inputs: the $tfidf$ matrix, the desired number of clusters, and a tolerance parameter. The goal of the algorithm is to minimize the total Euclidean distance between the center of each cluster (called the "centroid") and all observations assigned to that cluster. In the context of textual analysis, the Euclidean distance is computed as follows:

$$\sqrt{\sum_{r=1}^{R} ||x_r - x_r^C||^2}, \tag{1}$$

where x_r is the $tfidf$ value for feature r in a specific document and x_r^C is the corresponding value in the cluster centroid. R is the total number of features.

Centroids are initialized randomly in the vector space of all documents, then updated using an iterative process. In each iteration, each document is assigned to the closest centroid (i.e., with the smallest Euclidean distance), after which the centroid is redefined as the mean $tfidf$ vector of all documents assigned to it in the previous iteration. This process continues until the Euclidean distance between cluster centroids in two consecutive iterations is smaller than the specified tolerance level. For a more detailed description of the k-means algorithm, see Appendix A.

The key hyperparameter is the number of clusters, k, where the optimal number varies depending on the true structure of the data. In order to find this optimum, we run the algorithm independently for consecutive ks from 10 to 20, then compare the obtained clusters according to two criteria:

- *Cluster stability.* For the approach to be robust it should not be very sensitive to the exact number of clusters chosen. Specifically, the algorithm should categorize the majority of PIS sections into the same group across consecutive ks. Funds belonging to the same cluster when $k = x$ should be homogeneously grouped into the same broader cluster when $k = x - 1$.

[12] It is possible to construct peer groups using the highest weighted topic for each fund. This procedure gives similar results to k-means clustering. See appendix B.

- *Cluster density.* Optimal clusters should be sufficiently populated and be sufficiently distinct from each other. When the optimal k is exceeded, the algorithm will split homogeneous groups into smaller categories with low density and a large overlap in identifying features, making them difficult to distinguish in terms of interpretability.

We choose the highest k such that both criteria remain satisfied. Appendix A.2 describes the quantitative implementation of this procedure in detail, while figure 5 provides a more intuitive illustration for $k = \{16, 17, 18\}$. The figure shows heat maps of the joint frequency of cluster assignments when increasing k from a smaller (rows) to a larger (columns) number. The number of PIS sections in each cluster combination is indicated by the color scale shown on the right side of the figure. To facilitate discussion, we have labeled all clusters by inspecting their word clouds and reading a random sample of prospectuses. Note that this labeling is not needed in order to assess the density and stability criteria, or indeed to generate any of the main results in the paper except for the narrative analysis in section 4.

The *stability* criterion is satisfied for all combinations of k displayed in figure 5. Both heat maps show a few high-density combinations and many more low-to-no-density ones. This indicates that the majority of observations are jointly classified into the same category across different specifications. Looking at the assigned labels, we observe that stability is preserved also from an interpretability perspective: most observations are assigned to clusters with an identical or semantically related label across different specifications.[13]

The *density* criterion remains satisfied when increasing k from 16 to 17 (panel 1): the *Competitive Advantage* cluster emerges, and is populated by observations formerly assigned to the *PE-Ratio*, *New Products & Services*, *Fundamental* and *Long Term* clusters. This new cluster has a high density and is characterized by a distinct word cloud. However, density is no longer satisfied when increasing k from 17 to 18 (panel 2): the new cluster, *Book Value*, accounts for only 0.8% of total observations.

Therefore, going forward, we utilize 17 strategy clusters as our main classification, since this number satisfies both criteria and provides the greatest interpretability.

Clusters as peer groups: The clusters identified through this methodology should be interpreted as groups of peer funds that emphasize similar aspects of their strategy in their PIS descriptions, the implicit assumption being that the number of words spent discussing a particular aspect of the strategy is proportional to its relevance for the fund. All 2,995

[13] In Appendix A, we formally show the similarity in identifying features by measuring the Euclidean distance between all pairs of cluster centroids.

funds of interest are assigned to a Strategy Peer Group (SPG), for every month they are active between January 2000 and December 2017. Although the SPGs are identified by a full distribution over words and bi-grams, the labels and key prominent features shown in figure 1 are a useful short-hand to indicate the characteristics that are most distinctive for a given peer group.

Figure 6 displays the number of funds assigned to each SPG in each month, along with their aggregate TNA. We observe that SPGs have different relative sizes which vary over time. Panel 1 of figure 7 shows the frequency of assignment to each SPG. The ten most common strategies in our sample are *Large Cap* (37,771 fund-month observations; 802 unique funds), *Fundamental* (35,655 observations; 708 funds), *New Products & Services* (23,991 observations; 464 funds), *Dividends* (21,050 observations; 386 funds), *Derivatives* (20,840 observations; 553 funds), *Competitive Advantage* (20,330 observations; 386 funds); *Defensive* (19,882 observations; 544 funds); *Quantitative* (18,466 observations; 390 funds), *Small Cap* (18,154 observations; 353 funds); and *Long Term* (17,186 observations; 316 funds).

Panel 2 of figure 7 shows that funds tend to be assigned to the same SPG over time. In fact, we observe that 1,087 funds are assigned to only one SPG throughout their lives. The vast majority of funds are assigned to a maximum of 5 SPGs. Only a very small number are assigned to more than 5 SPGs, which may be the result of estimation noise. Note that we do not impose any constraint on the cluster estimation that forces PIS sections from the same fund to be grouped into the same SPG. Hence, the stability of the assignment over time further confirms the robustness of our methodology.

The strategies identified by k-means clustering, shown in figure 1, are highly interpretable and relatively unambiguous. *Small Cap*, *Mid Cap*, and *Large Cap* correspond to generic stock selection within each of the three main buckets of market capitalization. The closest strategies to value and growth are *Intrinsic Value* and *Long Term*, respectively. Classification along these two axes is ubiquitous in the industry (e.g. Morningstar style boxes) and in academia. Four other strategies can be described as general investment philosophies: funds with *Quantitative* strategies claim to use proprietary stock-selection and factor models, and often say they rank stocks according to some objective criteria; funds with *Fundamental* strategies claim to rely more on qualitative research and traditional firm valuation; funds with *Defensive* strategies say they use cash-like securities to reduce risk in turbulent market conditions; and *Tax-Managed* strategies promise to minimize taxable distributions to shareholders. The next group of strategies focuses on specific company attributes that the funds look for when selecting stocks: funds with *Competitive Advantage* strategies claim to

look for companies with competitive advantages relative to their industry peers, and firms with strong balance sheets; funds with *New Products & Services* strategies claim to look for companies with innovative new product lines and new technologies; funds with *Dividends* strategies say they look for companies with high dividend yields; and funds with *PE-Ratio* strategies advertise the use a particular valuation method: price-earnings multiples. The last group of strategies concern the use of alternative asset classes, which are permitted to comprise at most 20% of the holdings of equity funds: *Derivatives*, *Fixed Income*, *Foreign (ADR)* and *Foreign (Emerging Markets)*.

3 Empirical Analysis

3.1 Deviations from Core Strategies

In this section, we construct general measures of funds' strategies that do not depend on any subjective interpretation of the text. Our underlying assumption is that the holdings of the average fund in each strategy peer group are representative of the group's core strategy—in other words, we assume that individual fund divergences are purely idiosyncratic and cancel out over many funds in a group.

The *core strategy* of each SPG is measured as the mean portfolio weight vector across all funds in the group, and an individual fund's divergence from the core strategy is measured by the distance of its holdings from this mean vector. For each month in our sample, we compute this distance relative to each fund's own assigned SPG, and relative to the average of other SPGs. If funds are following their promised strategies, we should observe smaller divergences from their own group average.

Formally, *SPG-Deviation* is defined as follows:

$$Deviation_{j,t}^{G_{j,t}} = \sum_{i=1}^{N_t^j} \left(w_{i,t}^j - \bar{w}_{i,t}^{G_{jt}} \right)^2 \tag{2}$$

for $G = [SPG,\ \widetilde{SPG}]$, where $SPG_{j,t}$ is the peer group to which fund j is assigned at time t, and $\widetilde{SPG}_{j,t}$ represents all groups other than $SPG_{j,t}$. $w_{i,t}^j$ is the weight on stock i in fund j's portfolio in month t, and $\bar{w}_{i,t}^{G_{j,t}}$ is the average weight on stock i at time t for all funds belonging to group $G_{j,t}$. We log-transform the deviation variables so that their distributions are approximately Gaussian.

For robustness, we also compute the SPG average return as an alternative measurement of the core strategy, and funds' absolute return deviation (tracking error) with respect to

this average return as a measure of strategy divergence. However, we expect this alternative to be much noisier than the divergence measure based on average portfolio weights, and thus for the results to be much weaker.

To test whether funds generally adhere to the strategies described in their prospectuses (which lead them to be categorized into a particular SPG) we run the following regression:

$$Deviation_{j,t} - \widetilde{Deviation}_{j,t} = \alpha + \gamma' X_{j,t} + \eta_t + \iota_f + \varepsilon_{j,t}. \tag{3}$$

The coefficient of interest is $\hat{\alpha}$, which estimates the difference between within-group *Deviation* and outside-group $\widetilde{Deviation}$ when all control variables are equal to their mean values (all controls are demeaned). The control variable vector, X, contains the log of total net assets (TNA), the log of fund age, the fund's expense and turnover ratios, monthly percentage flows, and monthly flow volatility. η_t are month fixed effects, and ι_f are fund family fixed effects. Standard errors are clustered by fund and month.

Table 2 reports the estimated $\hat{\alpha}$s in the third column, while the first two columns show the average deviations relative to the fund's own SPG (within) and relative to other SPGs (outside), respectively. Portfolio weight deviations are shown in the top row, and return absolute deviations are shown in the bottom row. Column 3 shows that both measures of distances are significantly lower *within* the assigned SPG than outside.[14] All results are significant at the 1% level. Note that the large t-statistics are due to the fact that the reported coefficients are regression intercepts. These results should be interpreted as differences in the mean of the dependent variable between funds in the same month and fund family, after controlling for fund-level characteristics.

Funds allocate their capital more similarly to the average fund in their own SPG, which also translates into greater similarity in raw returns. Thus we conclude that, on average, funds appear to be keeping their promises to investors. This interpretation does not depend on the average level of distance within any particular group, as long as funds with high in-group distance still have higher distance with respect to funds outside of the group. However, we acknowledge that our measure is limited by how well portfolio weights can capture all aspects of fund strategies. The $\hat{\alpha}$ coefficient could be biased towards zero if we miss a major axis of commonality for a particular strategy. One such example is the *Derivatives* SPG—we can observe the percentage of assets other than cash and common stock in the portfolio, but our data do not include derivative position values specifically, which should certainly be a significant factor for that SPG.

[14] Portfolio weight deviations are always negative due to the log transformation.

3.2 Pairwise Analysis

To allow for more strict control variables, we compute our deviation measure, as well as differences between the first four return moments, for each *pair* of funds in the full sample, every month.

$$Deviation_{i,j,t} = \alpha + \beta_1 SameSPG_{i,j,t} + \beta_2 SameDGTW_{i,j,t} + \beta_3 SameFF3_{i,j,t} + Deviation_{i,j,t-1} + \gamma X_{i,j,t} + \eta_t + \varepsilon_{i,j,t}; \quad (4)$$

$$MomentDiff_{i,j,t+(1,24)} = \alpha + \beta_1 SameSPG_{i,j,t} + \beta_2 SameDGTW_{i,j,t} + \beta_3 SameFF3_{i,j,t} + MomentDiff_{i,j,t-(24,1)} + \gamma X_{i,j,t} + \eta_t + \varepsilon_{i,j,t}. \quad (5)$$

$Deviation_{i,j,t}$ is defined as in equation 2, except the average portfolio weights $\bar{w}^G_{i,t}$ are replaced with the weights of another individual fund i. $MomentDiff_{i,j,t+(1,24)}$ is the absolute difference (between fund i and fund j, in month t) in each moment of the distribution of future 24-months rolling returns, for $Moment \in [Mean,\ StDev,\ Skewness,\ Kurtosis]$. The variables denoted by $SameCluster_{i,j,t}$ (for $Cluster = [SPG,\ DGTW,\ FF3]$) are indicator variables that take a value of 1 if funds i and j are assigned to the same peer group, and 0 otherwise. The three peer groups are, respectively, (i) our text-based strategy peer groups ($SPGs$); (ii) Daniel et al. (1997) holdings-based peer groups, constructed from terciles of market capitalization, book-to-market ratio, and past returns ($DGTW$); and (iii) peer groups constructed using funds' exposures to the Fama-French 3-factor model ($FF3$). In both regressions, we control for differences in the same fund-level control variables as in section 3.1 (represented here by $X_{i,j,t}$), as well as the *past* portfolio weight deviations and moment differences ($Deviation_{i,j,t-1}$ and $MomentDiff_{i,j,t-(24,1)}$, respectively). We also include month (η_t) fixed-effects, and cluster standard errors at the fund and month level.

The moment-similarity regressions have a predictive interpretation: a negative and significant β_1 indicates that belonging to the same SPG today correlates with lower differences (higher similarity) in *future* returns. The portfolio weight deviation regression is run contemporaneously, hence a negative and significant β_1 indicates that funds belonging to the same SPG also have more similar holdings. Since we control for alternative peer group co-membership, a positive and significant β_1 indicates that prospectuses have incremental

predictive power for holdings and return similarities relative to traditional factors or characteristic portfolios.

Tables 3 and 4 report the results. β_1 is always negative and statistically significant, even when controlling for alternative peer groups. The magnitudes of the coefficients on the $SameSPG$ indicator variable are between 6 and 10 times smaller than the coefficients on the $SameDGTW$ or $SameFF3$ indicator variables. This is not surprising. Indeed, the alternative peer groups are constructed directly from funds' holdings, which implies that they will mechanically have explanatory power for holdings similarities, and will likely predict similarities in returns. Instead, it is surprising that the text-based SPGs, which are constructed independently from funds' holdings or returns, have any incremental explanatory power at all. These results indicate the SPGs capture a novel dimension of similarity that is not subsumed by traditional classifications.

3.3 Investor Responses to Strategy Divergence

In the absence of stringent legal repercussions, what drives funds' general adherence to their strategy descriptions? Funds are known to change their names to take advantage of hot new investment styles (Cooper et al. (2005)) and for their exposure to risk factors to drift over time (Wermers (2012)), so the question is a pressing one. In this section, we examine the hypothesis that it is investors' exit decisions—their withdrawal of capital from the fund—that exert a disciplining force on fund behaviour.

To test this hypothesis, we use our measure of divergence from the core strategy of the fund's peer group (*SPG-Divergence*), constructed as per equation 2. To facilitate discussion of magnitudes, we standardize the measure to have mean zero and variance 1. We analyze the response of both percentage flows ($Flow(\%)$) and dollar flows ($Flow(\$)$), the latter in millions. Percentage flows are defined as follows:

$$Flow(\%)_{j,t} = \frac{TNA_{j,t} - TNA_{j,t-1} \times MRet_{j,t}}{TNA_{j,t-1}}; \tag{6}$$

where $MRet_{j,t}$ is the fund's gross monthly portfolio return from $t-1$ to t. Dollar flows are equal to percentage flows multiplied by $TNA_{j,t-1}$.

We then examine the relationship between fund flows and SPG divergence, controlling for fund characteristics, past performance, and fund family membership, which are already known to drive flows. Additionally, to check that any relationship between flows and divergence is driven specifically by the strategies described in fund prospectuses, and not simply

correlation with holdings-based peer groups such as Daniel et al. (1997) (henceforth, DGTW) or Morningstar style categories, we include the divergence from the core strategy of each fund's DGTW characteristics-based peer group. This measure, *DGTW-Divergence* is constructed in the same way as *SPG-Divergence* but with funds clustered according to terciles of the DGTW size, book-to-market, and past return characteristics. The 9 main Morningstar categories for U.S. Equity funds are nested within the 27 DGTW peer groups.[15]

We run variants of the following regression:

$$Flow_{j,t+1} = \alpha_{j \in f,t} + \beta' Divergence_{j,t} + \gamma' Performance_{j,t} + \lambda' X_{j,t} + \varepsilon_{j,t+1}, \tag{7}$$

where $Divergence_{j,t}$ is either (or both) of *SPG-Divergence* and *DGTW-Divergence*, measured for fund j in month t. Since capital flows are heavily influenced by distribution channels common to funds in the same family, and given the impact of overall market conditions (e.g. aggregate outflows during recessions) that may affect different families in different ways, all specifications include fund family×month fixed effects, represented by $\alpha_{j \in f,t}$. The vector $X_{j,t}$ contains the usual fund-level control variables (log TNA; fund age; expense ratio; turnover ratio) measured as of month t.

The vector $Performance_{j,t}$ includes the past twelve-month raw return, the past twelve-month return in excess of the strategy peer group average (*SPG-Adj return*), the DGTW selectivity measure (*DGTW-Adj return*), as well as past twelve-month alphas according to the Capital Asset Pricing Model (CAPM), the Fama-French-Carhart four-factor model (FFC), and the Fama-French five-factor model augmented with a momentum factor (FF6).[16] All returns and alpha are net of fees. The SPG-adjusted return is computed as follows:

$$Ret_{j,t}^{SPGAdj} = Ret_{j,t} - \frac{1}{M_{j,t}} \sum_{k=1}^{M_{j,t}} Ret_{k,t}, \tag{8}$$

[15] Although there are over a hundred Morningstar categories, they primarily reflect broad differences in asset classes. For the sample of funds we use in this paper—corresponding to Morningstar's "U.S. Equity" category group—the categories are "Large Value", "Large Blend", "Large Growth", "Mid-Cap Value", "Mid-Cap Blend", "Mid-Cap Growth", "Small Value", "Small Blend", "Small Growth", and "Leveraged Net Long". For a full list of Morningstar categories, see https://morningstardirect.morningstar.com/clientcomm/Morningstar_Categories_US_April_2016.pdf.

[16] All alphas are computed at a monthly frequency by subtracting monthly fund returns from a beta-adjusted factor return portfolio, where the betas are estimated using rolling 24-month regressions. We then aggregate the monthly alphas up to the annual frequency. Note that we do not include Berk and van Binsbergen (2015)'s value added measure in this regression as our goal is to understand investor responses. Value-added is a measure of fund manager skill in the presence of decreasing returns to scale, while investors only care about the returns they will experience on their own portfolios.

where $Ret_{j,t}$ is the annual return of fund j from month $t-12$ to month t, and $M_{j,t}$ is the number of funds in the same SPG as fund j during month t.

The null hypothesis in this regression is that *SPG-divergence* has no impact on fund flows. Referring to equation 7, we have:

- $H_0 : \beta = 0$ *(Investors either do not realize that funds are deviating from their promised strategies, or do not care, or have no outside options.)*

Using a two-sided test, we then have the following alternative hypotheses:

- $H_1 : \beta > 0$ *(Greater SPG-Divergence is associated with increased fund flows; investors prefer funds that are more unique relative to their peers.)*
- $H_2 : \beta < 0$ *(Greater SPG-Divergence is associated with decreased fund flows; investors prefer funds to adhere to their promised strategies, exemplified by peer group averages.)*

Estimated coefficients from equation 7 are reported in table 5. Columns 1-3 show specifications where the dependent variable is percentage flows, and columns 4-6 show specifications using dollar flows. The headline result is that across all specifications, funds that deviate more from the average portfolio weights of their Strategy Peer Groups experience *lower* flows on average in the subsequent month. Despite the standard errors being computed from only a few hundred degrees of freedom (due to two-way clustering, as recommended by Petersen (2009)), the estimated coefficients on *SPG-Divergence* are always significant at the 5% level, and often at the 1% level. The magnitudes are sensible for a single month effect: on average, a one-standard-deviation increase in *SPG-Divergence* leads to a decrease of between $454,000 and $493,000 in next-month net flows. We analyze the cumulative (multi-month) effect in the next section.

Notably, the negative relationship between strategy divergence and flows is *not* observed for the alternative DGTW peer groups, indicating that the effect is specifically due to the novel information on strategies contained in fund prospectuses. Therefore, we reject H_0 in favor of H_2: investors prefer funds to stick to their advertised strategies, and punish those that do not by withdrawing capital.

3.4 Dynamics of Investor Responses

Even though the first month after an increase in strategy divergence already shows a pronounced negative effect on fund flows, the full impact of investor responses may take several

months to accumulate. We would also like to verify whether the effect is permanent; i.e., that it does not reverse over time. As such, we re-estimate the regression in equation 7 with m-month-ahead flows ($Flow_{j,t+m}$, where $m \in [1, 2, 3, 6, 9, 12]$) as the dependent variable, focusing on the full specification with both *SPG-Divergence* and *DGTW-Divergence* (columns 3 and 6 in table 5).

The results are presented in table 6. Panel A shows specifications with percentage flows as the dependent variable and panel B shows specifications with dollar flows. In both cases, the negative effect of *SPG-Divergence* on flows persists over time, with the same sign but decaying magnitude, for up to twelve months after the initial divergence. The cumulative effect on dollar flows is the sum of the month-by-month coefficients: over the subsequent year, a one-standard-deviation increase in *SPG-Divergence* is associated with a decline of about $4.8 million. In terms of percentage flows, the compounded annual effect is a decrease of 0.47% for the average fund. This number is 24.3% of the unconditional sample mean.[17]

We note that this number likely understates the magnitude of investors' partial-equilibrium response to strategy divergence. If funds anticipate a response and pre-emptively reduce their divergence in the subsequent months, this would dampen the effect we are able to observe. We examine this possibility further in section 3.6.

As with the first-month results, we find no effect of *DGTW-Divergence* on flows in the subsequent year, affirming our earlier conclusion that investors evaluate funds' strategy adherence relative to the descriptions in their prospectuses rather than its correlation with holdings-based peer groups.

3.5 Fund Flows and Performance

Rational investors should allocate capital based only on active returns (i.e., total returns minus benchmark returns). Currently, the standard method of computing active returns is to subtract beta-adjusted market returns (CAPM) (Berk and Van Binsbergen (2016) and Barber et al. (2016)). However, if the strategies described in fund prospectuses matter to investors (for example, if they have preferences for characteristics beyond standard risk factors), peer-adjusted returns should also predict flows independently from risk model alphas.

To test this hypothesis, we examine the performance coefficients from regression equation 7, reported in panel B of table 5 and in table 6. Across all specifications, SPG-adjusted returns are positively related to one-month-ahead fund flows (significant at the 1% level), both in percentage and dollar terms, and this effect persists for at least 12 months (usually

[17] From table 1, the annual mean percentage flow is $(1 + 0.16/100)^{12} - 1 = 1.94\%$.

also at the 1% significance level). Controlling for CAPM alpha, Fama-French-Carhart (FFC) alpha, and six-factor alpha (Fama and French (2015) plus momentum) does not subsume the effect of SPG-adjusted returns. DGTW-adjusted returns (based on Daniel et al. (1997) benchmarks) are also positively related to future flows, but including them in the regressions does not attenuate the economic or statistical significance of the SPG-adjusted return coefficients. In fact, SPG-adjusted returns consistently have a stronger relationship with future flows than DGTW-adjusted returns. Despite the prominence of Morningstar size and value style categories in the industry, we find that investors rely more on text-based peer group comparisons.

Text-based Strategy Peer Groups appear to provide investors with novel, useful conditioning information when choosing how to allocate their capital. The magnitudes of the estimated coefficients are similar to those for CAPM alpha—indeed, in the case of percentage flows, they are even larger than the CAPM coefficients.[18] Our evidence indicates that, when allocating capital, investors pay attention to fund performance relative to other funds with similar strategy descriptions in addition to aggregate risk exposures. These findings paint a consistent picture of rational investors who balance the benefits of peer outperformance, which requires strategy divergence, against the costs of such divergence.

3.6 Funds' Response to Market Discipline

Having found that investors direct less capital to funds that diverge more from their core strategies, the natural follow-up question is whether funds change their behavior after investor exit. As discussed in section 3.4, a secondary question is whether funds anticipate negative responses from investors and reduce strategy divergence on their own.

To answer these questions, we estimate the dynamics of our standardized *SPG-Divergence* measure as a function of its own lags, as well as past net flows and the interaction between flows and lagged divergence. Specifically, for $m \in [1, 2, 3, 6, 9, 12]$, we estimate:

$$\begin{aligned} Divergence_{j,t+m} = \alpha_{j \in f,t} + \beta_1 Divergence_{j,t} + \beta_2 Flow_{j,t} + \beta_3 (Divergence_{j,t} \times Flow_{j,t}) + \\ \gamma' Performance_{j,t} + \lambda' X_{j,t} + \varepsilon_{j,t+1}. \end{aligned} \tag{9}$$

The regression includes the same fund attributes and performance controls as equation 7 (discussed in section 3.4), and the same fund family×month fixed effects. Standard errors

[18] Interestingly, in a departure from Berk and Van Binsbergen (2016) and Barber et al. (2016), but consistent with Jegadeesh and Mangipudi (2021), we find that the strongest overall predictor of future fund flows in our sample is Fama-French-Carhart (four factor) alpha.

are two-way clustered by fund and month. The coefficients of interest are β_1, which captures the "regular" autocovariance structure of *SPG-Divergence*, and β_3, which captures the incremental effect of fund outflows on the estimated autocovariance. Recall that we do not expect all outflows to be interpreted by the fund as investor disapproval; rather, assuming funds are responsive to investor exit decisions, they should only change their behavior when high outflows occur together with high strategy divergence.

Table 7 reports the estimated coefficients from equation 9. Since the coefficient on lagged divergence is estimated to be less than 1, we can infer that the variable exhibits mean reversion at a monthly rate of $1 - \beta_1 = 0.098$ (row 1, column 1). Over a year, the estimated mean-reversion rate is 0.51 (row 1, column 6). Therefore, independently of flows, funds portfolio weights are "anchored" to the core strategies of their peer groups. A plausible reason for this finding is that funds anticipate investor (or regulator) disapproval and do not stray from their prospectus strategy descriptions for too long.

However, the most important coefficient is on the interaction between past divergence and past flows (row 3). This coefficient (β_3) is estimated to be positive and significant (usually at the 1% or 5% significance levels) for up to at least 12 months. The positive sign indicates that, when net flows are negative, any level of *SPG-divergence* at time t is associated with lower divergence at time $t + m$; in other words, the rate of mean-reversion is increased. Of course, the estimated effect is symmetric: when net flows are positive, then the rate of mean-reversion is slower. Overall, funds' rate of reversion to their core strategies is significantly influenced by investor entry and exit decisions. These findings are indicative of funds' response to a market discipline effect.

The magnitude of the first-month β_3 coefficient of 0.001 (column 1, row 3) can be understood as follows. If time-t *SPG-Divergence* is two standard deviations above its mean, then a one-standard-deviation decline in flows (7.86%, as reported in table 1) increases the divergence mean-reversion speed by $0.001 \times 2 \times 7.86 = 0.0157$, which is a factor of approximately 16% of the regular mean-reversion rate discussed above. Over a yearly horizon, the estimated coefficient is 0.004. An important difference from equation 7 is that here the dependent variable is a "stock" variable rather than a "flow" variable. Thus, estimated effects are cumulative, and the same decline in flows results in an increase in mean-reversion of $0.004 \times 2 \times 7.86 = 0.0629$, or 12.3% of the regular annual mean-reversion rate.

3.7 Alternative Explanations

Since neither strategy divergence nor flows are exogenously determined, it is possible that the raw correlations we document could be explained by an alternative mechanism. One possibility is that the observed correlations are ultimately driven by fund performance. For example, suppose two otherwise similar funds diverge from their peer group average by taking a few different large positions. Then each fund's return will be driven by the idiosyncratic volatility of their unique bets. Suppose that one fund gets lucky and experiences high returns on its large holdings, whereas the other gets unlucky and experiences low returns. If the unlucky fund experiences performance-driven outflows and unwinds its large positions in the loser stocks in order to meet investor redemptions (while the lucky fund maintains or increases its large winner positions due to inflows), this could generate an interaction effect between flows and strategy divergence. A second possibility is that the correlations are driven by fees. For example, Kostovetsky and Warner (2020) report that new funds that are more unique charge higher fees. If funds also raise their fees when they become more unique over time, and if investors happen to react negatively to these fee increases, this could generate patterns similar to what we observe.

Explanations such as these are ruled out by our inclusion of an extensive list of performance variables, as well as funds' total expense ratios, in all of the regressions involving flows and strategy divergence.

A third possibility is that the correlations are driven by changing attributes of fund families or fund distribution channels. For instance, if a fund family stops selling shares via a particular broker, the clients of this broker may move their investments to another fund family that remained with the broker, resulting in gradual outflows from the first family. If the family anticipates these outflows and increases the uniqueness of its funds in an attempt to attract compensating attention through its other distribution channels, this could also generate a relationship between flows and strategy divergence.

This possibility is ruled out by our inclusion of fund family$\times$month fixed effects in all flow regressions. The identifying variation in these regressions comes from comparing two funds in the same fund family in the same month but with different levels of strategy divergence or flows. Overall, our empirical setting excludes alternative explanations stemming from the main drivers of flows identified in the prior literature.

4 Narrative Analysis

As a supplement to our quantitative findings on core strategy deviations, we conduct a narrative analysis of each strategy peer group (SPG), asking whether fund holdings line up with an intuitive reading of the strategy descriptions. This exercise necessarily involves a subjective component, and reasonable readers may disagree with some of our choices. We try to avoid overly detailed inferences and instead focus on broad observations that we believe are less likely to be controversial. This narrative exercise is intended to act as a sanity check on, and to provide context for, the objective similarity measures examined in section 3.1. The exercise also has independent descriptive value, as the active equity investment landscape in the United States has not previously been examined in this way.

For each fund, we compile fund attributes, performance statistics, risk-factor loadings, average industry composition, and average stock characteristics weighted by the percentage of assets invested in each stock. The fund attributes are log TNA, log age, expense ratio, and turnover ratio. The performance measures are rolling 24-month six-factor alphas (Fama and French (2015) plus momentum) and six-factor value-added (Berk and van Binsbergen (2015)). For risk-factor loadings, we again use the five Fama-French factors plus momentum. For industry composition, we use the Fama-French 48 industries. The stock characteristics are subdivided into five categories: (i) "traditional" characteristics (market beta, market capitalization, book-to-market ratio, past year's stock return, investment, and profitability); (ii) balance sheet variables (current assets, inventories, non-performing assets, PP&E, intangible assets, asset growth, cash and equivalents, current liabilities, deferred taxes, long-term debt, and leverage ratio); (iii) income statement variables (operating income, earnings growth, and R&D expenditures); (iv) market variables (dividend yield, equity issuance, equity repurchases, Amihud illiquidity, and firm age). Furthermore, we compute the following portfolio characteristics: percentage of holdings in ADRs (American Depository Receipts); percentage of holdings that are foreign incorporated; percentage of holdings in common stock; percentage of holdings in cash; and number of different stocks in the portfolio. All variables are normalized by subtracting the full-sample mean and dividing by the standard deviation.

For each of the above variables, and for each SPG, we run the following regression:

$$Y_{j,t} = \alpha + \beta I_{j,t}^{SPG_{j,t}} + \gamma X_{j,t} + \eta_t + \iota_{j \in f} + \varepsilon_{j,t}, \tag{10}$$

where $Y_{j,t}$ is the variable of interest for fund j at time t; η_t and $\iota_{j \in f}$ denote fixed effects for month t and membership of fund j in family f, respectively; $I_{j,t}^{SPG_{j,t}}$ is an indicator variable equal to 1 if fund j belongs to the SPG of interest at time t, and 0 otherwise; and $X_{j,t}$ is a vector of demeaned fund-level control variables (the natural logarithms of fund age and TNA, expense ratio, turnover ratio, net fund flows, and flow volatility; when $Y_{j,t}$ is one of these fund-level characteristics, it is excluded from the control vector). Because all independent variables are demeaned, α can be interpreted as the mean of the characteristic of interest when $I_{j,t}^{SPG_{j,t}} = 0$ (i.e. when fund j does not belong to the SPG of interest at time t). The β coefficient then indicates the average difference between the characteristic for the SPG of interest and all other SPGs (holding the other characteristics, family membership, and month constant).

In the discussion below, we highlight the most striking (and statistically significant) attributes of each SPG. The full breakdown for all of the above attributes is provided for the reader in tables 8 through 15.

Large Cap, Mid Cap, and Small Cap: The obviously relevant attributes for peer groups based on company size are SMB beta, average market capitalization, and average firm age. Accordingly, we find that funds in the *Large Cap* SPG have significantly lower SMB betas, and hold significantly older firms with larger average market cap. The SMB beta of the *Mid Cap* SPG is not significantly different from other funds, but average firm age and market cap are smaller. The *Small Cap* SPG has both a higher SMB beta (highest among all SPGs) and a much lower average firm age and market cap (lowest among all SPGs).

Intrinsic Value: Funds in this SPG invest in companies trading at a discount to perceived intrinsic value. This is the closest SPG to a typical "value" strategy, and as such is reflected in higher than average HML betas at the fund level, as well as higher than average book-to-market ratio at the stock level. The "discount to intrinsic value" attribute may also be reflected in the fact that these funds buy firms with higher share repurchases, potentially indicating undervaluation.

Long Term: Funds in this SPG claim to be searching for long term investment opportunities, or companies with long term growth potential. In the data, they have lower than average book-to-market ratios, both at the fund level and at the stock level (indicting a

growth style tilt), and also hold companies with higher intangibles (which includes capitalized R&D) and lower dividend yields (greater retained earnings, typically used to grow the firm).

Fixed Income: The *Fixed Income* SPG is distinguished by an emphasis on bonds as a secondary asset class. Consistent with this emphasis, we find funds in this SPG to have the lowest percentage of holdings in common stock, and the highest holdings in cash.

Derivatives: Funds in this SPG discuss the use of derivatives (particularly options and futures) and short selling to enhance their portfolio return. Unfortunately, we do not observe derivative positions in our dataset and are thus unable to validate this SPG using the narrative/descriptive approach.

Quantitative: Funds in this group are significantly younger and smaller, with higher turnover ratio and lower expense ratio. This confirms the findings of Abis (2020), who identifies quantitative funds using a different methodology (supervised machine learning). We also expect *Quantitative* funds to make greater use of trend-following strategies and hold stocks of more companies at a time, signifying their greater information-processing capabilities. Both of these expectations are validated by the data: funds in this SPG have higher momentum betas and hold stocks of more companies than other funds.

Fundamental: The *Fundamental* SPG is comprised of funds that engage in bottom-up, fundamental research on individual companies. As such, we would not expect any particular tilts in industry composition, risk factor loadings or stock characteristics. Funds in this peer group can be seen as generic traditional investment managers. Consistent with this intuition, we do not observe large differences between funds in this SPG and funds outside this SPG.

Defensive: In general, "defensive" strategies can mean either safer asset classes or defensive industries—the word cloud indicates that this SPG focuses on the former, specifically money market and other short term securities. The text generally permits managers to use these cash-like securities to reduce risk in adverse market conditions. In line with this objective, our results show that these funds hold a smaller fraction of their assets in common stock. The funds also have a lower than average market beta.

Tax: This SPG is characterized by tax-management strategies that explicitly aim to reduce the tax burden to the end investor through lower distributions. This strategy is difficult to assess using the simple characteristics we report in this paper, as it depends on the timing of sales relative to accrued capital gains. We note, however, that the *Tax* SPG holds much less of its assets in cash than other funds, which may correlate with lower taxable distributions.

Dividends: The "Dividends" group arguably has the most straightforward strategy and the clearest ex-ante expectations for investors: funds in this SPG should try to maximize dividend distributions. Accordingly, we find that these funds hold stocks with (by far) the highest dividend yield among all SPGs. The companies they buy are also typical high-dividend firms, in that they are larger and older, have lower investment and cash on the balance sheet (plausibly due to higher payout ratios). The industry composition of *Dividends* funds' portfolios is also tilted towards classical high-dividend industries such as utilities, telecoms, and banks.

PE Ratio: The use of the price-earnings ratio in security valuation is the common theme among funds in this SPG, who claim to look for companies whose prices are low relative to the prices implied by PE multiples for similar firms. This SPG also contains funds who use the PE ratio to identify "value" firms (i.e., those with low PE ratios). This is consistent with our finding that funds in this SPG have higher than average HML betas (though not higher book-to-market ratios at the stock level). Overall, our analysis for this SPG is limited.

New Products & Services: Funds in this SPG focus on companies with new or innovative product lines, and new or rapidly changing technologies. Consistent with this focus, funds in this SPG tend to avoid stable industries such as utilities, telecoms, and banks, while holding a significantly greater fraction of stocks from innovative industries such as technology and pharmaceuticals (drugs). Companies introducing new products and services should require higher levels of investment and R&D expenditure, and higher asset growth, all of which are confirmed in the data (along with a lower CMA beta, which also indicates high investment).

Competitive Advantage: According to the text, the two defining attributes of this SPG are companies with a sustainable competitive advantage and those with a strong balance sheet. Consistent with both of these attributes, we find that funds in this SPG tend to hold companies with higher than average profitability (highest among all SPGs) and companies with lower levels of debt (lowest among all SPGs) and lower leverage.

Foreign: The two *Foreign* SPGs—*Emerging Markets (EM)* and *American Depository Receipts (ADR)*—represent quite different strategies. While the former is explicitly focused on buying stocks of foreign companies, usually in emerging markets, the latter strategy appears to simply indicate a larger investment opportunity set, where funds are permitted (but not required) to hold ADRs. The characteristics data are consistent with this interpretation: *Foreign (EM)* funds hold a higher fraction of both ADRs and foreign-incorporated firms in their portfolios, compared to all other funds. However, there are no significant differences for *Foreign (ADR)* funds.

While we are not able to measure every relevant characteristic associated with the various strategies—derivative and bond positions, for example—we do not find any inconsistencies among the large set of characteristics that we *are* able to measure, and in most cases the measured characteristics directly affirm the text. Overall, we conclude that fund characteristics are broadly consistent with the investment approaches described in the text.

5 Conclusion

In this paper we ask whether U.S. active equity mutual funds actually follow the strategies described in the "Principal Investment Strategies" sections of their prospectuses, and how investors respond if they diverge from these strategies. We use an unsupervised machine learning algorithm, *k-means*, to group together funds with similar strategy descriptions, and we develop new criteria to determine the optimal number of groups. Each of the resulting 17 *strategy peer groups (SPGs)* capture distinct and interpretable approaches to investment, most of which go beyond the standard size-value axis widely used to measure fund styles.

We use the average portfolio weights among funds in each SPG to capture the core strategy of the group. To answer the question of whether funds are following their promised strategies, we compute fund-level divergences from the core strategy of each SPG and find that funds are on average closer to the core strategies of their own peer groups than to others. Using pairwise regressions, we also confirm that any two funds tend to be more similar in terms of both portfolio weights and returns if they are in the same SPG, controlling for past similarity, differences in fund attributes, and whether their holdings are in the same terciles of size, value, and momentum. As a sanity check on these quantitative results, we also conduct a narrative analysis of the 17 strategies and confirm that average characteristics

of the funds in each SPG generally line up with common-sense readings of their aggregate PIS text.

We then document a market discipline effect that operates via investors' net capital flows. Controlling for performance and other fund characteristics, flows are lower for funds that diverge more from the core strategy of their SPG. The investor response is persistent, lasting up to twelve months after the initial divergence. Investors also chase SPG-adjusted returns independently of standard factor model alphas, with similar levels of predictive power. Finally, we show that funds are responsive to the disciplinary effect of fund flows. Strategy divergence mean-reverts over time, and does so at a faster rate following months in which funds diverged more from their core strategies *and* experienced outflows.

Our results uncover a diverse fund strategy landscape that was previously unknown from traditional analyses of holdings and returns. We demonstrate the importance of these strategies to investors, and show that they effectively discipline funds despite lacking access to reliable legal enforcement channels.

References

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

Abis, S., A. M. Buffa, A. Javadekar, and A. Lines (2021). Learning from prospectuses. *Working Paper*.

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

Barber, B. and T. Odean (2007). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies 21*, 785–818.

Barber, B. M., X. Huang, and T. Odean (2016). Which factors matter to investors? evidence from mutual fund flows. *The Review of Financial Studies 29*(10), 2600–2642.

Barber, B. M., T. Odean, and L. Zheng (2005). Out of sight, out of mind: The effects of expenses on mutual fund flows. *The Journal of Business 78*(6), 2095–2120.

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

Berk, J. B. and J. H. Van Binsbergen (2016). Assessing asset pricing models using revealed preference. *Journal of Financial Economics 119*(1), 1–23.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *the Journal of machine Learning research 3*, 993–1022.

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

Cooper, M. J., H. Gulen, and P. R. Rau (2005). Changing names with style: Mutual fund name changes and their effects on fund flows. *The Journal of Finance 60*(6), 2825–2858.

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.

deHaan, E., Y. Song, C. Xie, and C. Zhu (2020). Disclosure obfuscation in mutual funds.

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

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

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

Flannery, M. J. (1998). Using market information in prudential bank supervision: A review of the us empirical evidence. *Journal of Money, Credit and Banking*, 273–305.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology 32*(3), 221–233.

Geffen, D. M. (2009). A shaky future for securities act claims against mutual funds. *Securities Regulation Law Journal 37*, 20.

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

Hortacsu, A. and C. Syverson (2004). Product differentiation, search costs, and competition in the mutual fund industry: A case study of sp 500 index funds. *The Quarterly Journal of Economics 119*(2), 403–456.

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

Ivković, Z. and S. Weisbenner (2009). Individual investor mutual fund flows. *Journal of Financial Economics 92*(2), 223–237.

Jain, P. C. and J. S. Wu (2000). Truth in mutual fund advertising: Evidence on future performance and fund flows. *The journal of finance 55*(2), 937–958.

Jegadeesh, N. and C. S. Mangipudi (2021). What do fund flows reveal about asset pricing models and investor sophistication? *The Review of Financial Studies 34*(1), 108–148.

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

Khorana, A. and H. Servaes (2012). What drives market share in the mutual fund industry? *Review of Finance 16*(1), 81–113.

Kincaid, P. J., R. P. Fishburne, R. L. J. Rogers, and B. S. Chissom (1975). Derivation od new readability formulas (automated readability index, fox count and flesch reading ease formula) for navy enlisted personel. *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. *Swiss Finance Institute Research Paper* (20-54).

Loughran, T. and B. McDonald (2011). When is a liability not a liability? textual analysis dictionaries and 10-ks. *Journal of FInance 66*(1), 35–65.

Loughran, T. and B. McDonald (2020). Textual analysis in finance. *Annual Review of Financial Economics 12*, 357–375.

Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of financial studies 22*(1), 435–480.

Reuter, J. and E. Zitzewitz (2006). Do ads influence editors? advertising and bias in the financial media. *The Quarterly Journal of Economics 121*(1), 197–227.

Roussanov, N., H. Ruan, and Y. Wei (2021). Marketing mutual funds. *The Review of Financial Studies 34*(6), 3045–3094.

Roussanov, N. L., H. Ruan, and Y. M. Wei (2020). Mutual fund flows and performance in (imperfectly) rational markets? *Working Paper*.

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.

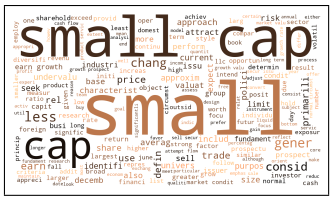
Sirri, E. R. and P. Tufano (1998). Costly search and mutual fund flows. *The journal of finance 53*(5), 1589–1622.

Wermers, R. (2012). Matter of style: The causes and consequences of style drift in institutional portfolios. *Working paper*.

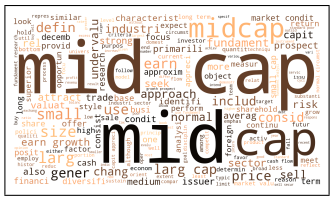
Zhu, Q. (2020). The missing new funds. *Management Science*.

Figure 1: **Strategy Peer Groups:** Word clouds for all estimated SPGs using the k-means algorithm with $k = 17$ clusters. These represent the frequency of features (words and bigrams) in the strategy sections within each SPG. Word sizes are proportional to frequency.

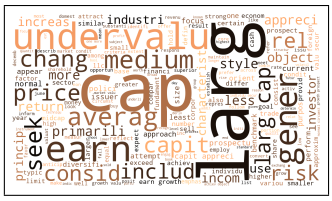
(a) Small Cap



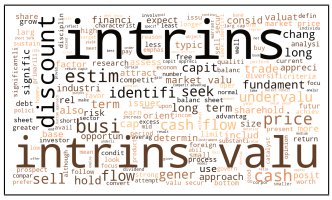
(b) Mid Cap



(c) Large Cap



(d) Intrinsic Value



(e) Long Term

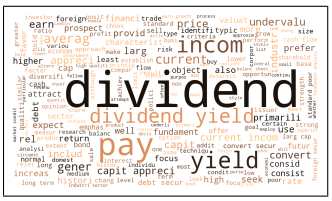
(f) Quantitative

(g) Fundamental

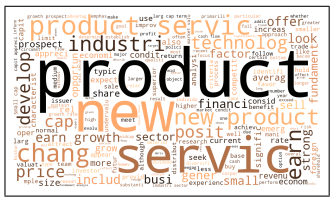
(h) Defensive

(i) Tax

(j) Dividends

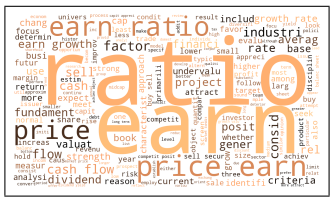


(k) Products & Services



(l) Competitive Advantage

(m) PE-Ratio



(n) Fixed Income

(o) Derivatives

(p) Foreign (EM)

(q) Foreign (ADR)

Figure 2: **Distribution of Word Counts for Strategy Descriptions:** Pooled distribution of the number of words across all fund–month observations.

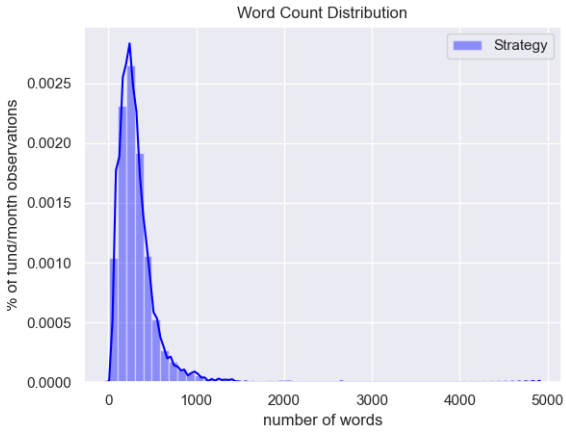


Figure 3: **Distribution of Complexity Measures for Strategy Descriptions:** Panel 1 displays the pooled distribution of the Flesch-Kincaid grade level complexity measure across all fund–month observations, for Strategy sections. This measure indicates the number of years of schooling required in order to comprehend each section. Panel 2 displays the same distribution for the Flesch-Kincaid reading ease measure. This measure indicates, on a scale of $[1, 100]$, how easily a section can be read (a higher score indicates lower complexity). Both measures are based on the relative number of total words to total sentences (average sentence length) and the relative number of total syllables to total words (average word length) contained in each section.

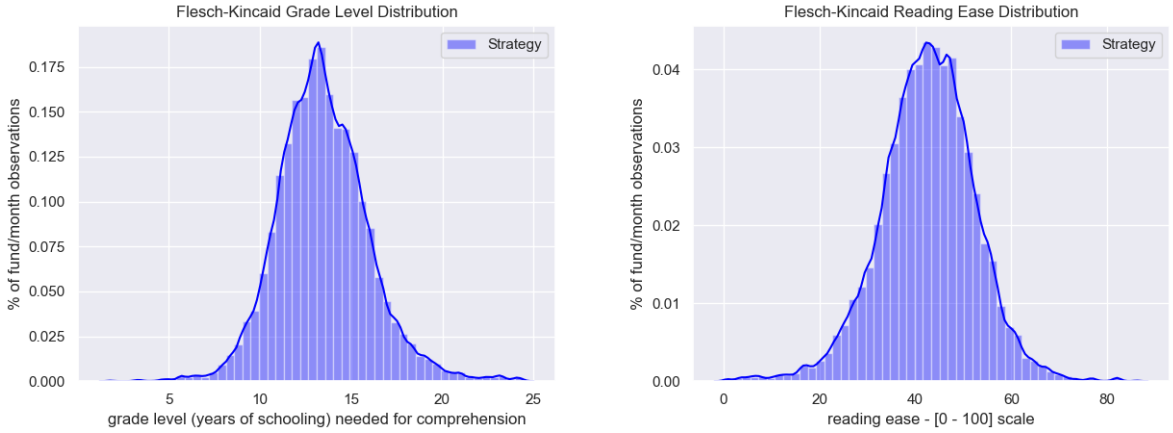


Figure 4: **Frequency of Positive, Negative, Uncertainty and Litigious Words in Strategy Descriptions:** Panel 1 displays the pooled distribution of the frequency of *positive* words for all fund-month observations. Panel 2 display the same distribution for the frequency of *negative* words; panel 3 for the frequency of *uncertainty* words; and panel 4 for the frequency of *litigious* words. The frequency of words per section is obtained by computing the percentage of the total number of words in each section that appears in the Loughran and McDonald *positive*, *negative*, *uncertainty* or *litigious* dictionaries. These dictionaries are adapted to account for specific characteristics of financial language (Loughran and McDonald (2011)).

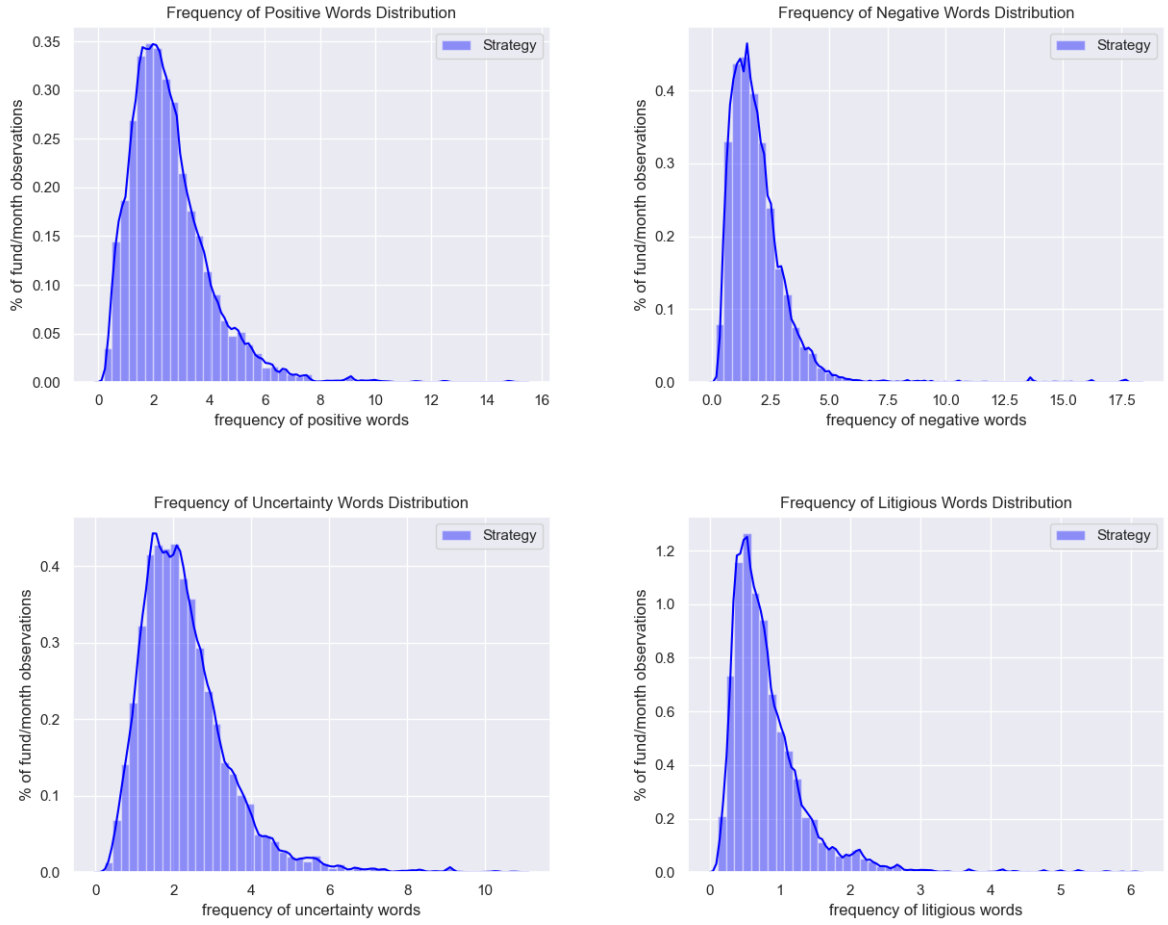


Figure 5: **K-Means Cluster Assignment:** Each square in each heatmap represents the number of strategy sections assigned to a specific cluster combination: applying the k-means algorithm with a lower (rows) versus higher (columns) number of clusters. Darker colors indicate a higher number of sections classified in that specific combination, according to the color map on the right hand side. Panel 1 shows the cross-allocation of strategy sections when going from 16 (rows) to 17 (columns) clusters. Panel 2 shows the cross-allocation of strategy sections when going from 17 (rows) to 18 (columns) clusters.

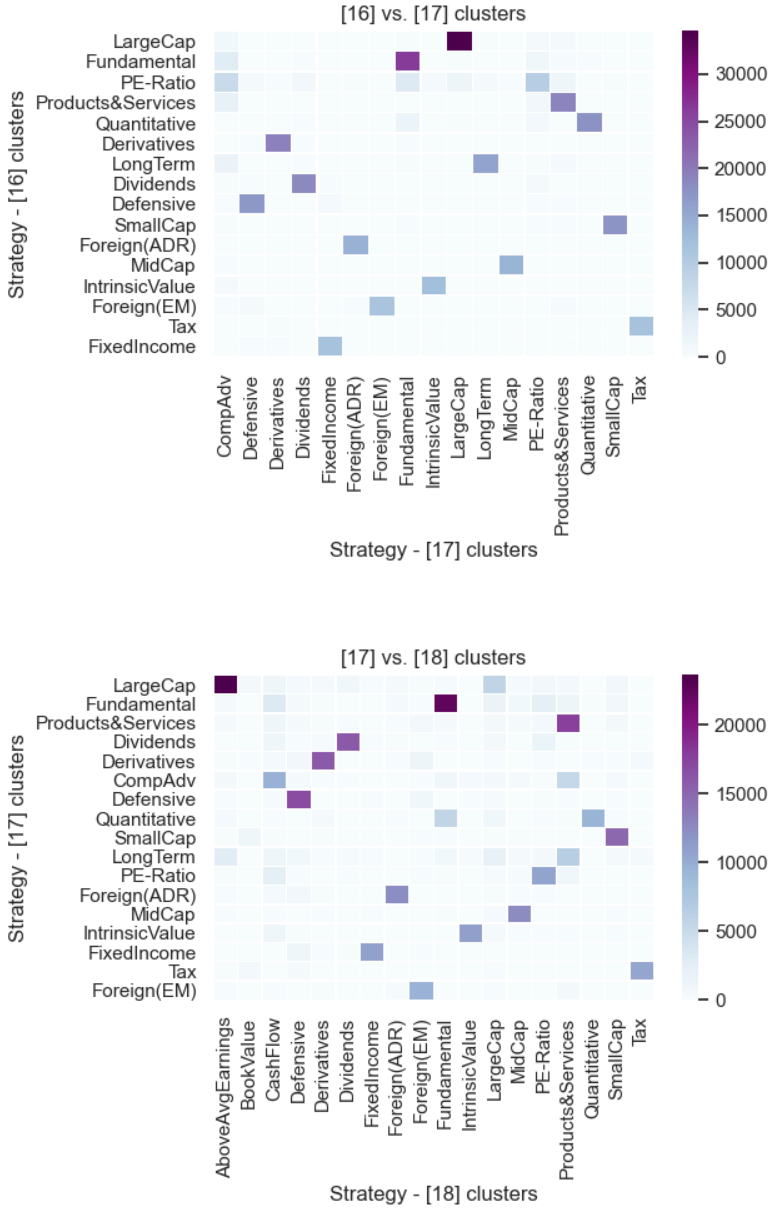


Figure 6: **TNA by SPG Over Time:** Panel 1 displays the cumulative TNA managed by funds in different SPGs between January 2000 and December 2017. Panel 2 displays the number of funds assigned to the different SPGs between January 2000 and December 2017.

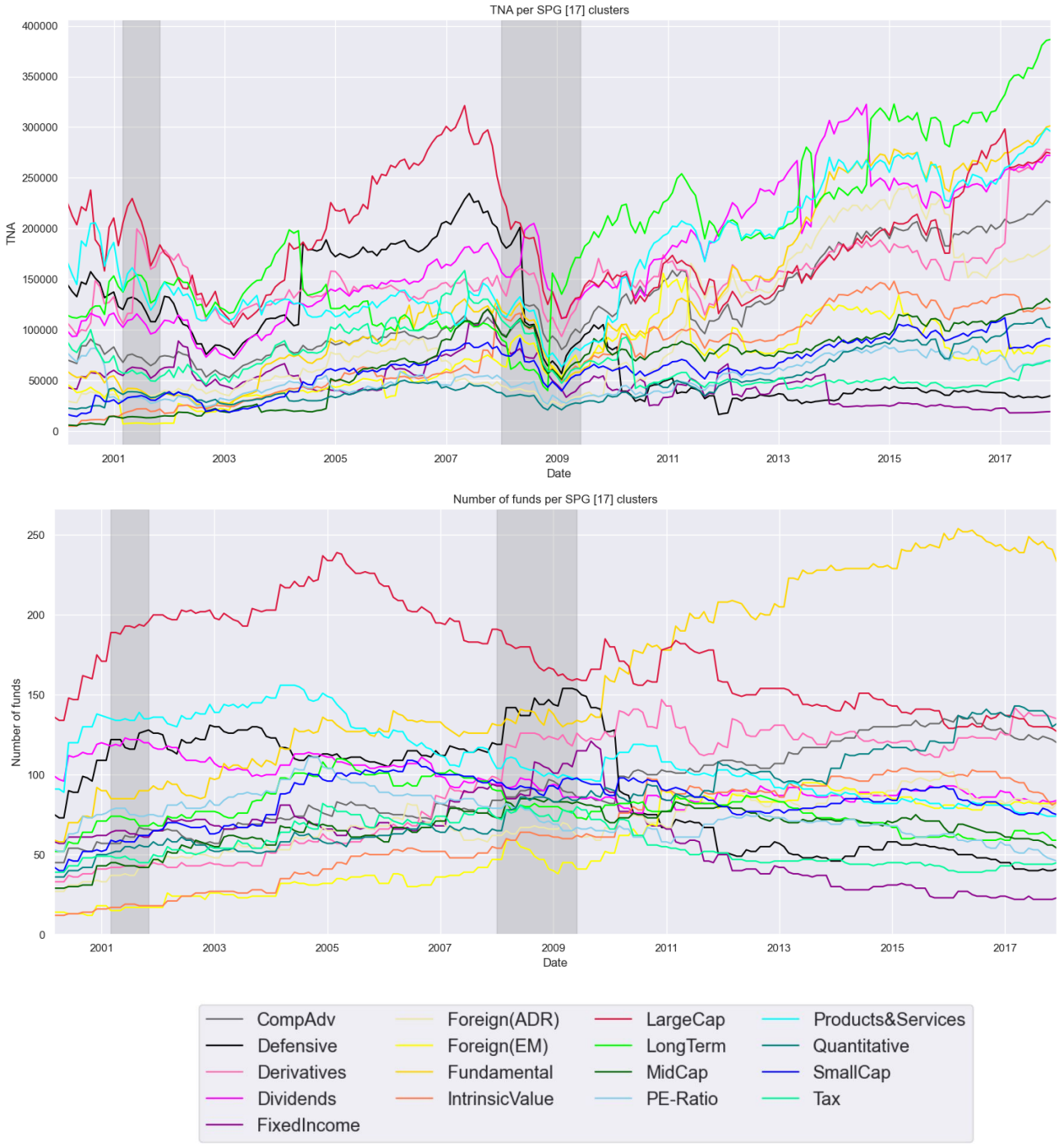


Figure 7: **SPG Assignments Per Fund:** Panel 1 shows the number of fund-month observations assigned to each of the SPGs using the k-means algorithm with 17 clusters. The number next to each bar indicates the unique number of funds assigned each SPG at some point in their lives. Panel 2 shows the number of different SPG assignments that unique funds receive throughout their lives.

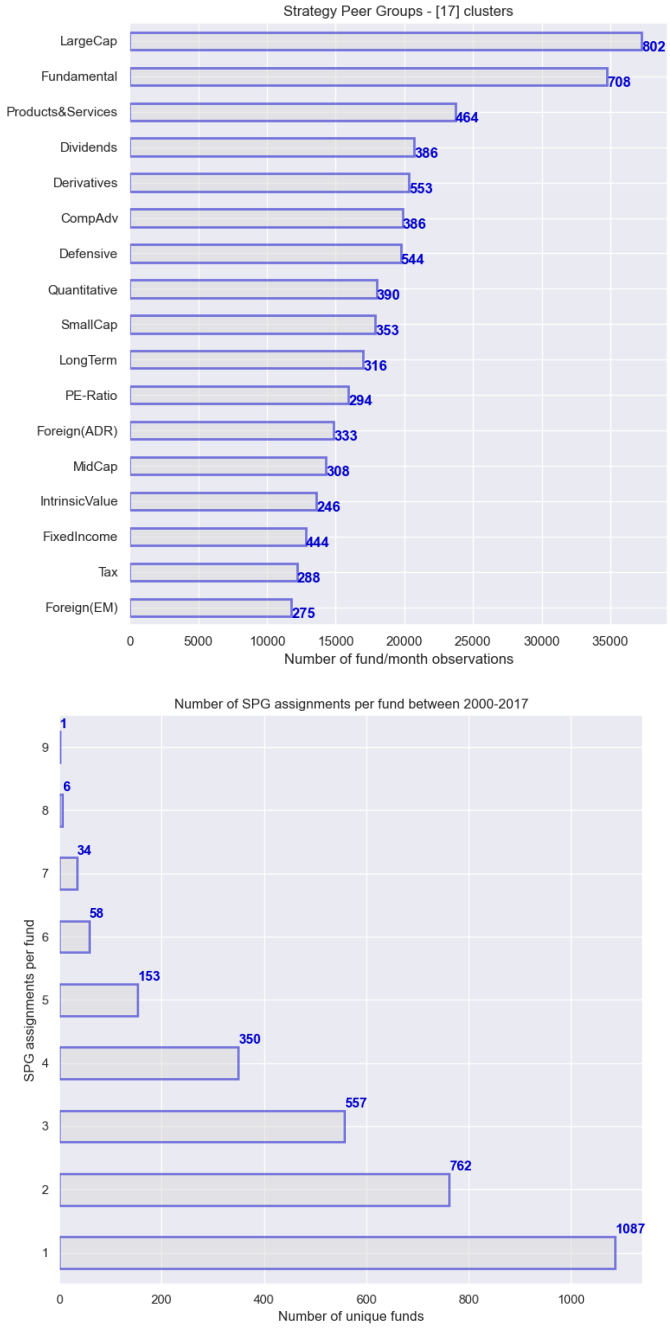


Table 1: **Summary Statistics:** Variables: Total Net Assets (TNA) in millions of dollars; fund age in months; fund expense, ratio winsorized at the 0.1% level; fund turnover ratio, winsorized at the 0.1% level; monthly net fund flows as a percentage of TNA; flow volatility; and fund cash holdings as a percentage of TNA. For each variable, the table displays the number of available observations (*count*), the mean (*mean*), the standard deviation (*sd*), the minimum (*min*) and maximum (*max*) values, and the 25th ($p25$), 50th ($p50$) and 75th ($p75$) percentiles.

	count	mean	sd	min	p25	p50	p75	max
TNA (M$)	322871	1203	4375	5	68	245	904	177462
Age (months)	322859	178	157	0.00	78	138	220	1121
ExpenseRatio	321377	1.22	0.42	0.11	0.98	1.18	1.42	4.43
TurnoverRatio	313340	80.42	76.07	1.00	33	62	102	815
Flow (%)	322608	0.16	7.86	-45.61	-1.49	-0.44	0.75	143.15
FlowVol	315638	2022	5065	4.55	154	532	1765	79135
Cash (%)	300429	3.05	4.66	-20.87	0.45	1.87	4.10	51.24

Table 2: **Divergence from Core Strategies:** This table reports average divergence of each fund from the core strategy of its own peer group (column 1), average placebo divergence from the average of other peer groups (column 2), and the difference between these two estimates (column 3) (see section 3.1). The first row shows results for portfolio weight divergence, computed as the log sum of squared differences between the fund's portfolio weights and the average weights for funds in the same Strategy Peer Group. The second row shows results for return divergence, computed as the absolute difference in returns between each fund and the average of its peer group. The differences in column 3 are estimated controlling for fund age, log assets, expense ratio, turnover ratio, monthly fund flows, and monthly fund flow volatility, as well as month and fund family fixed effects. All non-binary independent variables are demeaned. Standard errors are two-way clustered by fund and month.

	(1) Within SPG	(2) Outside SPG	(3) Difference
Portfolio Weight Divergence	-4.766***	-4.330***	-0.436***
	(-443.55)	(-351.64)	(-52.33)
Return Divergence	1.393***	1.442***	-0.0484***
	(201.55)	(215.92)	(-17.35)
Controls	Yes	Yes	Yes
FE	Family+Month	Family+Month	Family+Month
Cluster	Fund+Month	Fund+Month	Fund+Month
Obs	288012	288012	288012

t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01

Table 3: **Pairwise Strategy Divergence:** This table reports the results of regressing pairwise portfolio weight divergence on a dummy variable for whether the two funds are grouped into the same SPG (see section 3.2). The regressions control for past differences in fund-level characteristics (log assets, log age, expense ratio, turnover ratio, net flows, net flow volatility), past portfolio weight divergence, and dummies for co-membership in alternative peer groups, based on holdings characteristics (Daniel et al. (1997)) and risk factor loadings (Fama and French (1993)). All specifications include month fixed effects, and standard errors are clustered by fund and month.

	Portfolio Weight Divergence			
Same SPG	-0.0188***	-0.0172***	-0.0175***	-0.0161***
	(-12.20)	(-11.59)	(-11.79)	(-11.20)
Same DGTW		-0.174***		-0.159***
		(-22.66)		(-24.09)
Same FF3			-0.175***	-0.159***
			(-17.22)	(-17.04)
Controls	Yes	Yes	Yes	Yes
FixedEffects	Month	Month	Month	Month
Clustering	Fund+Month	Fund+Month	Fund+Month	Fund+Month
R2	0.0354	0.0739	0.0729	0.105
Observations	159681454	159681454	159681454	159681454

t statistics in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Table 4: **Pairwise Return Moment Distances:** This table reports the results of regressing pairwise absolute return moment differences (mean, standard deviation, skewness, and kurtosis) on a dummy variable for whether the two funds are grouped into the same SPG (see section 3.2). The return moments are estimated using *future* data from the subsequent 24 months. The regressions control for past return moment differences (estimated over the previous 24 months) as well as differences in fund-level characteristics (log assets, log age, expense ratio, turnover ratio, net flows, net flow volatility) and dummies for co-membership in alternative peer groups, based on holdings characteristics (Daniel et al. (1997)) and risk factor loadings (Fama and French (1993)). All specifications include month fixed effects, and standard errors are clustered by fund and month.

	Ret. Mean	Ret. Std.	Ret. Skew.	Ret. Kurt.
Same SPG	-0.00731***	-0.0431***	-0.00617***	-0.00821***
	(-4.35)	(-8.39)	(-4.06)	(-3.19)
Same DGTW	-0.0481***	-0.216***	-0.0425***	-0.0692***
	(-9.84)	(-19.39)	(-6.86)	(-8.56)
Same FF3	-0.0414***	-0.197***	-0.0329***	-0.0584***
	(-8.69)	(-18.31)	(-6.23)	(-9.76)
Controls	Yes	Yes	Yes	Yes
FixedEffects	Month	Month	Month	Month
Clustering	Fund+Month	Fund+Month	Fund+Month	Fund+Month
R2	0.0870	0.0953	0.179	0.150
Observations	87938298	87938298	87938298	87935610

t statistics in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Table 5: **Fund Flows, Strategy Divergence, and Performance:** This table reports estimated coefficients from regressions of one-month-ahead net fund flows on divergence from the fund's core strategy and various measures of past performance (see section 3.3). Strategy divergence comes in two flavors: SPG-divergence is the log sum of squared differences between the fund's portfolio weights and the average weights for funds in the same Strategy Peer Group; DGTW-divergence is constructed in the same way but with respect to average portfolio weights among funds whose holdings are in the same Daniel et al. (1997) size, book-to-market, and past return terciles. The dependent variable is either percentage flows (columns 1-3) or dollar flows in millions (columns 4-6). Panel A reports the coefficients on funds' strategy divergences and the control variables (log TNA, log age, expense ratio, turnover ratio), while panel B reports the coefficients on peer-adjusted returns for both the SPG and DGTW peer groups, as well as standard factor model alphas. Coefficients in both panels are from the same regressions. All specifications include fund-family×month fixed effects, and standard errors are two-way clustered by fund and month.

Panel A: Divergence from Core Strategies

	% Flow (t+1)			$ Flow (t+1)		
	(1)	(2)	(3)	(4)	(5)	(6)
Divergence (SPG)	-0.037**		-0.052***	-0.493**		-0.454**
	(-2.12)		(-2.84)	(-2.45)		(-2.20)
Divergence (DGTW)		0.000	0.028		-0.304	-0.058
		(0.02)	(1.64)		(-1.59)	(-0.30)
Log TNA	0.021	0.022	0.021	-2.223***	-2.214***	-2.218***
	(1.47)	(1.60)	(1.52)	(-7.84)	(-7.82)	(-7.82)
Log Age	-0.529***	-0.529***	-0.533***	-4.048***	-4.040***	-4.069***
	(-16.03)	(-16.14)	(-16.20)	(-9.06)	(-9.03)	(-9.09)
Expense Ratio	-0.304***	-0.263***	-0.291***	0.609	0.839	0.684
	(-3.60)	(-3.13)	(-3.48)	(0.77)	(1.05)	(0.85)
Turnover Ratio	-0.001*	-0.001*	-0.001*	-0.001	-0.000	-0.001
	(-1.89)	(-1.75)	(-1.91)	(-0.18)	(-0.12)	(-0.20)
R2	0.333	0.333	0.334	0.286	0.285	0.286
Obs	239,661	239,680	239,661	239,661	239,680	239,661

Continued on following page ...

... Continued from previous page

Panel B: Performance

	% Flow (t+1)			$ Flow (t+1)		
	(1)	(2)	(3)	(4)	(5)	(6)
SPG-Adj Return	0.042***		0.036***	0.197***		0.175***
	(6.52)		(5.54)	(3.32)		(2.93)
DGTW-Adj Return		0.023***	0.020***		0.090***	0.077**
		(6.36)	(5.63)		(2.93)	(2.49)
Raw Return	0.013**	0.030***	0.007	0.055	0.143***	0.031
	(2.03)	(5.44)	(1.03)	(1.10)	(3.49)	(0.59)
CAPM Alpha	0.030***	0.042***	0.035***	0.309***	0.360***	0.326***
	(6.05)	(8.41)	(7.00)	(6.56)	(7.94)	(6.91)
FFC Alpha	0.074***	0.068***	0.066***	0.587***	0.562***	0.557***
	(8.02)	(7.42)	(7.24)	(7.19)	(6.86)	(6.75)
FF6 Alpha	0.023***	0.022***	0.023***	0.054	0.051	0.053
	(2.98)	(2.87)	(2.96)	(0.78)	(0.73)	(0.76)
R2	0.333	0.333	0.334	0.286	0.285	0.286
Obs	239,661	239,680	239,661	239,661	239,680	239,661

t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01

Table 6: **Dynamics of Fund Flows:** This table extends the regressions in table 5 to the case of m-month-ahead fund flows, where $m \in [1, 2, 3, 6, 9, 12]$ (see section 3.4). Panel A reports the results for percentage flows, and panel B for dollar flows. Included in the regression but omitted for brevity are the same control variables (log TNA, log age, expense ratio, turnover ratio). All regressions include fund-family×month fixed effects, and standard errors are two-way clustered by fund and month.

Panel A: Percentage Flows

	% Flow					
	(t+1)	(t+2)	(t+3)	(t+6)	(t+9)	(t+12)
Divergence (SPG)	-0.052***	-0.047**	-0.038**	-0.038**	-0.038**	-0.028
	(-2.84)	(-2.59)	(-2.07)	(-2.00)	(-2.00)	(-1.45)
Divergence (DGTW)	0.028	0.027	0.016	0.024	0.031*	0.024
	(1.64)	(1.62)	(0.97)	(1.38)	(1.79)	(1.36)
SPG-Adj Return	0.036***	0.028***	0.025***	0.015**	0.012*	0.017***
	(5.54)	(4.14)	(3.75)	(2.26)	(1.91)	(2.69)
DGTW-Adj Return	0.020***	0.021***	0.021***	0.020***	0.016***	0.012***
	(5.63)	(5.85)	(5.61)	(5.64)	(5.07)	(4.15)
CAPM Alpha	0.035***	0.032***	0.030***	0.020***	0.014**	0.009
	(7.00)	(6.32)	(5.50)	(3.77)	(2.48)	(1.56)
FFC Alpha	0.066***	0.067***	0.065***	0.070***	0.070***	0.061***
	(7.24)	(7.22)	(6.97)	(7.31)	(7.47)	(6.78)
FF6 Alpha	0.023***	0.023***	0.025***	0.022***	0.020**	0.022***
	(2.96)	(3.00)	(3.31)	(2.81)	(2.57)	(2.75)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R2	0.334	0.330	0.326	0.316	0.310	0.305
Obs	239,661	236,902	234,107	226,712	219,662	212,633

Continued on following page ...

... Continued from previous page

Panel B: Dollar Flows

	$ Flow					
	(t+1)	(t+2)	(t+3)	(t+6)	(t+9)	(t+12)
Divergence (SPG)	-0.454**	-0.432**	-0.356*	-0.391*	-0.388*	-0.410*
	(-2.20)	(-2.09)	(-1.72)	(-1.89)	(-1.84)	(-1.88)
Divergence (DGTW)	-0.058	-0.061	-0.184	-0.166	-0.174	-0.207
	(-0.30)	(-0.32)	(-0.93)	(-0.83)	(-0.88)	(-1.03)
SPG-Adj Return	0.175***	0.152**	0.152**	0.118*	0.131**	0.193***
	(2.93)	(2.44)	(2.41)	(1.88)	(2.12)	(3.09)
DGTW-Adj Return	0.077**	0.076**	0.073**	0.052*	0.028	0.014
	(2.49)	(2.44)	(2.32)	(1.69)	(0.94)	(0.47)
CAPM Alpha	0.326***	0.304***	0.278***	0.199***	0.143**	0.087
	(6.91)	(6.40)	(5.65)	(3.90)	(2.58)	(1.52)
FFC Alpha	0.557***	0.567***	0.559***	0.578***	0.549***	0.489***
	(6.75)	(6.74)	(6.55)	(6.59)	(6.22)	(5.81)
FF6 Alpha	0.053	0.062	0.074	0.076	0.073	0.091
	(0.76)	(0.90)	(1.07)	(1.06)	(0.99)	(1.21)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R2	0.286	0.286	0.286	0.285	0.286	0.288
Obs	239,661	236,904	234,110	226,714	219,664	212,635

t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01

Table 7: **Dynamics of Strategy Divergence:** This table reports estimated coefficients from regressions of SPG-divergence in month $t+m$, where $m \in [1, 2, 3, 6, 9, 12]$, on SPG divergence in month t, percentage fund flows, and the interaction between them (see section 3.6). Also included in the regressions are the usual performance variables and fund-level control variables (log TNA, log age, expense ratio, turnover ratio) as of month t. All regressions include fund-family×month fixed effects, and standard errors are two-way clustered by fund and month.

	Divergence (SPG)					
	(t+1)	(t+2)	(t+3)	(t+6)	(t+9)	(t+12)
Divergence (SPG) (t)	0.902***	0.822***	0.751***	0.622***	0.544***	0.490***
	(157.57)	(123.05)	(111.10)	(65.60)	(49.23)	(41.56)
% Flow	-0.001***	-0.001***	-0.001**	-0.001	-0.002**	-0.001
	(-3.21)	(-2.72)	(-2.00)	(-1.35)	(-2.01)	(-1.18)
Flow × *Divergence*	0.001**	0.001*	0.002**	0.003**	0.004***	0.004***
	(2.55)	(1.79)	(2.19)	(2.35)	(3.37)	(2.83)
SPG-Adj Return	0.000	0.000	0.000	0.001	0.001	0.002
	(0.34)	(0.32)	(0.52)	(0.52)	(0.40)	(0.95)
DGTW-Adj Return	-0.000	-0.000	-0.001	-0.001	-0.000	-0.000
	(-0.41)	(-0.47)	(-0.99)	(-0.76)	(-0.41)	(-0.37)
CAPM Alpha	0.001	0.001*	0.002**	0.003**	0.003**	0.003**
	(1.53)	(1.95)	(2.13)	(2.32)	(2.46)	(2.14)
FFC Alpha	-0.000	-0.001	-0.002	-0.004*	-0.005**	-0.005*
	(-0.41)	(-0.97)	(-1.56)	(-1.83)	(-2.10)	(-1.89)
FF6 Alpha	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(-1.53)	(-1.13)	(-0.66)	(-0.66)	(-0.50)	(-0.69)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R2	0.881	0.793	0.720	0.610	0.554	0.521
Obs	238,567	235,308	232,095	224,240	216,871	209,589

t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01

Table 8: **Fund Attributes by SPG:** This table reports *differences* between average fund attributes within a particular Strategy Peer Group (SPG) and the average across all other SPGs. Coefficients are estimated in separate regressions of the dependent vars (log of fund age, log of TNA, expense ratio, turnover ratio, post-fee FF6 Alpha, and post-fee FF6 Value Added) on dummies for each SPG (see section 4). Also shown are the *overall* averages across the full sample. We include month and fund family fixed effects. Standard errors are two-way clustered by fund and month.

	ln(TNA)	ln(Age)	Expenses	Turnover	Alpha	ValAdd
Large Cap	-0.0268	0.0382	-0.0255**	-4.658**	0.0174	5.930
	(-0.69)	(1.52)	(-2.19)	(-2.33)	(1.65)	(0.42)
Fundamental	0.0151	-0.0400	0.00132	2.480	-0.00580	6.635
	(0.33)	(-1.47)	(0.12)	(0.97)	(-0.64)	(0.64)
Products & Services	0.0876	0.0365	0.0586***	8.583**	0.0208	19.88
	(1.61)	(1.17)	(4.09)	(2.57)	(1.07)	(0.92)
Dividends	0.0447	0.0279	-0.0676***	-10.60***	-0.0556**	-67.09**
	(0.68)	(0.69)	(-4.27)	(-3.90)	(-1.99)	(-2.08)
Derivatives	-0.0441	-0.0261	-0.0346**	13.94***	-0.00723	-7.185
	(-0.76)	(-0.66)	(-2.42)	(4.25)	(-0.60)	(-0.29)
Comp. Advantage	0.0784	-0.00135	0.0377**	-0.761	0.0288	9.328
	(1.07)	(-0.04)	(2.09)	(-0.24)	(1.60)	(0.49)
Defensive	0.123**	-0.0464	-0.0289**	-3.678	-0.00308	-4.425
	(2.21)	(-1.32)	(-2.13)	(-1.09)	(-0.20)	(-0.19)
Quantitative	-0.251***	-0.116***	-0.0713***	27.24***	-0.0272	-38.34**
	(-3.75)	(-3.20)	(-3.90)	(6.11)	(-1.45)	(-2.54)
Small Cap	0.00894	-0.104***	0.0725***	1.201	0.0207	10.20
	(0.13)	(-3.05)	(3.88)	(0.37)	(1.18)	(0.73)
Long Term	0.103	0.102***	0.0365*	-11.01***	0.00623	27.12
	(1.61)	(2.67)	(1.90)	(-4.18)	(0.38)	(1.12)
PE-Ratio	-0.0275	0.0813**	0.0126	-2.328	-0.0198	8.075
	(-0.36)	(2.04)	(0.71)	(-0.65)	(-1.16)	(0.44)
Foreign (ADR)	-0.0438	-0.0406	-0.0254	-1.840	-0.00304	-10.23
	(-0.69)	(-1.04)	(-1.46)	(-0.59)	(-0.20)	(-0.51)
Mid Cap	-0.104	0.0168	0.0276	4.676	0.00907	9.822
	(-1.27)	(0.42)	(1.37)	(1.36)	(0.35)	(0.41)
Intrinsic Value	0.0239	0.000798	0.0399**	-18.96***	0.0112	16.88
	(0.27)	(0.01)	(2.44)	(-7.20)	(0.60)	(0.77)
Fixed Income	0.0841	0.0612*	0.00454	3.322	-0.00414	-5.040
	(1.42)	(1.66)	(0.26)	(0.88)	(-0.23)	(-0.20)
Tax	-0.105	-0.0154	-0.00645	-11.38**	0.000438	10.64
	(-1.11)	(-0.33)	(-0.29)	(-2.59)	(0.03)	(0.50)
Foreign (EM)	-0.0168	0.0607	-0.00695	-7.608**	-0.000198	14.43
	(-0.17)	(1.26)	(-0.32)	(-2.44)	(-0.01)	(0.60)
Overall	5.545***	4.898***	1.224***	81.29***	-0.0945***	-73.24***
	(325.90)	(472.54)	(278.75)	(99.19)	(-73.97)	(-35.53)
Obs	288,012	288,012	288,012	288,012	268,509	268,509

t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01

Table 9: **Risk Factor Exposures by SPG:** This table reports *differences* between funds' loadings on the Fama and French (2015) five factors plus momentum within a particular Strategy Peer Group (SPG) and the average across all other SPGs. Coefficients are estimated in separate regressions of fund loadings (estimated from prior 12 months' daily returns) on dummies for each SPG (see section 4). Also shown are the *overall* averages across the full sample. We control for (demeaned) log age, log assets, expense and turnover ratio, fund flows, flow volatility, as well as month fund and family fixed effects. Standard errors are two-way clustered by fund and month.

	Market beta	SMB beta	HML beta	MOM beta	RMW beta	CMA beta
LargeCap	-0.00302	-0.0357**	-0.00561	0.00182	-0.00575	-0.00825
	(-0.65)	(-2.57)	(-0.64)	(0.46)	(-0.89)	(-1.28)
Fundmntl	0.00153	0.0309*	0.0101	0.00476	-0.0130*	-0.00954
	(0.32)	(1.90)	(1.01)	(1.08)	(-1.71)	(-1.29)
ProdServ	0.0194***	0.0601***	-0.0653***	0.0298***	-0.0764***	-0.0391***
	(2.77)	(3.18)	(-6.34)	(5.16)	(-7.20)	(-4.67)
Dividends	-0.0203**	-0.179***	0.0863***	-0.0420***	0.105***	0.106***
	(-2.54)	(-11.13)	(7.95)	(-7.38)	(11.05)	(10.44)
Deriv	-0.0305***	-0.0472***	0.00147	-0.00753	0.00763	0.0110
	(-3.65)	(-2.75)	(0.15)	(-1.35)	(0.86)	(1.34)
CompAdv	0.00610	-0.0171	-0.0985***	0.0170**	-0.0604***	-0.0622***
	(0.92)	(-0.74)	(-7.67)	(2.53)	(-5.31)	(-5.18)
Defensive	-0.0119*	-0.0373**	0.00344	-0.00717	0.00265	0.0115
	(-1.90)	(-2.57)	(0.36)	(-1.35)	(0.30)	(1.28)
Quantit	0.00603	-0.0385	0.0361***	0.0158***	0.0790***	0.0238***
	(0.88)	(-1.65)	(3.06)	(2.76)	(9.18)	(2.71)
SmallCap	0.0205***	0.372***	0.0402***	0.00406	-0.00438	-0.0135**
	(2.91)	(19.48)	(3.22)	(0.79)	(-0.43)	(-1.98)
LongTerm	0.00944	-0.0176	-0.0375***	0.000236	-0.0144	-0.0373***
	(1.15)	(-0.82)	(-2.75)	(0.04)	(-1.29)	(-3.43)
PE-Ratio	0.00556	0.00966	0.0423***	-0.00407	0.0364***	0.0206**
	(0.75)	(0.38)	(3.23)	(-0.53)	(3.23)	(2.11)
Foreign(ADR)	0.00985	-0.00703	-0.0145	0.0129**	-0.0130	-0.0271**
	(1.24)	(-0.30)	(-1.03)	(2.37)	(-1.09)	(-1.97)
MidCap	0.00935	-0.00812	-0.0272**	0.00937	-0.0136	0.0190*
	(1.30)	(-0.47)	(-2.10)	(1.61)	(-1.27)	(1.88)
IntrValue	-0.00306	-0.0716***	0.0675***	-0.0485***	0.0260**	0.0360***
	(-0.34)	(-3.14)	(5.34)	(-7.40)	(2.17)	(3.09)
FixedInc	-0.0154	-0.0125	0.00391	-0.0202***	-0.0160	0.0115
	(-1.64)	(-0.72)	(0.37)	(-3.60)	(-1.50)	(1.14)
Tax	0.00110	-0.0615**	-0.0139	0.00367	0.0135	0.00345
	(0.17)	(-2.14)	(-1.13)	(0.67)	(1.19)	(0.36)
Foreign(EM)	-0.00647	0.0294	-0.0160	0.0106	-0.0335***	-0.0351***
	(-0.79)	(1.09)	(-1.05)	(1.55)	(-3.00)	(-2.82)
Overall	0.980***	0.222***	-0.00637*	0.0307***	-0.0542***	-0.0319***
	(970.75)	(34.85)	(-1.85)	(21.39)	(-20.49)	(-13.29)
Obs	285,996	285,996	285,996	285,996	285,996	285,996

t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01

Table 10: **Industry Exposures by SPG:** This table displays *differences* in average industry exposures between funds in each Strategy Peer Group (SPG) and the average across all other SPGs. These differences are estimated in separate regressions of fund-level industry exposures on dummy variables for each SPG (see section 4), controlling for fund age, log assets, expense ratio, turnover ratio, monthly fund flows, monthly fund flow volatility, and month as well as fund family fixed effects. All non-binary independent variables are demeaned. Standard errors are two-way clustered by fund and month. Industries are the Fama-French 48 industries—a selected sample is displayed for brevity.

	Util	Telcm	Hshld	Oil	Banks	Tech	Drugs
LargeCap	-0.23**	-0.16	0.05	-0.28	-0.05	0.00	-0.05
	(-2.29)	(-1.07)	(0.98)	(-1.12)	(-0.16)	(1.38)	(-0.20)
Fundamental	0.22	-0.34**	-0.08	0.37	0.57	0.00	-0.18
	(1.58)	(-2.54)	(-1.47)	(1.12)	(1.64)	(0.81)	(-0.64)
ProductsServices	-0.68***	-0.46**	-0.08	-0.36	-2.77***	0.01***	0.99**
	(-6.04)	(-2.04)	(-1.38)	(-0.96)	(-6.33)	(3.43)	(2.51)
Dividends	1.71***	1.83***	0.11*	0.31	4.09***	-0.02***	-0.27
	(8.89)	(7.95)	(1.79)	(1.02)	(7.44)	(-5.91)	(-0.83)
Derivatives	0.14	0.37**	-0.00	-0.02	0.04	0.00	0.01
	(0.96)	(2.16)	(-0.01)	(-0.06)	(0.10)	(0.73)	(0.03)
CompAdvantage	-0.87***	-0.61***	0.35***	-1.33***	-1.18***	0.01*	1.15***
	(-7.75)	(-3.58)	(3.94)	(-3.77)	(-2.86)	(1.86)	(2.78)
Defensive	0.03	0.16	-0.02	0.07	0.87*	-0.00	-0.56*
	(0.23)	(0.84)	(-0.33)	(0.21)	(1.81)	(-0.53)	(-1.85)
Quantitative	0.25	0.41**	-0.09*	0.63*	1.19***	0.00	-0.37
	(1.63)	(2.43)	(-1.67)	(1.74)	(2.63)	(0.08)	(-1.19)
SmallCap	-0.24**	-1.53***	-0.12*	2.62***	-1.41***	-0.01***	0.32
	(-2.01)	(-9.00)	(-1.79)	(4.37)	(-3.11)	(-2.69)	(0.61)
LongTerm	-0.55***	-0.19	0.10	-0.52	-1.35***	0.00	0.93*
	(-2.69)	(-0.76)	(1.03)	(-1.14)	(-2.66)	(1.08)	(1.81)
PERatio	0.39**	-0.13	0.03	0.26	0.79	-0.00	-0.99***
	(2.49)	(-0.65)	(0.26)	(0.56)	(1.36)	(-1.25)	(-2.91)
Foreign(ADR)	-0.51**	-0.28	-0.19**	0.54	0.20	0.00	0.38
	(-2.07)	(-1.12)	(-2.46)	(1.18)	(0.40)	(0.10)	(1.13)
MidCap	0.43**	-0.39**	-0.07	-1.83***	-2.75***	-0.00	-1.57***
	(2.29)	(-2.07)	(-0.91)	(-4.61)	(-5.99)	(-1.20)	(-4.42)
IntrinsicValue	0.14	1.87***	0.02	-1.57***	1.91***	-0.00	-0.57
	(0.77)	(3.52)	(0.18)	(-3.25)	(3.41)	(-0.01)	(-1.63)
FixedIncome	0.19	0.14	-0.06	0.13	0.21	0.00	0.49
	(1.34)	(0.54)	(-0.78)	(0.35)	(0.44)	(0.03)	(0.99)
Tax	-0.41*	0.28	0.12	1.28**	0.37	0.00	0.66
	(-1.95)	(1.07)	(1.24)	(2.14)	(0.73)	(0.91)	(1.29)
Foreign(EM)	-0.17	-0.08	-0.13	-0.60	-0.54	0.00	-0.55*
	(-0.80)	(-0.29)	(-1.43)	(-1.01)	(-1.04)	(1.16)	(-1.70)
Obs	288,012	288,012	288,012	288,012	288,012	288,012	288,012

t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01

Table 11: **Traditional Stock Characteristics by SPG:** This table reports *differences* between various average stock characteristics within a particular Strategy Peer Group (SPG) and the average across all other SPGs. The differences are estimated in separate regressions of fund-level average stock chracteristics on dummy variables for each SPG (see section 4), controlling for fund age, log assets, expense ratio, turnover ratio, monthly fund flows, monthly fund flow volatility, and month as well as fund family fixed effects. All non-binary independent variables are demeaned. Standard errors are two-way clustered by fund and month.

	MarketBeta	MarketCap	BookToMkt	Momentum	Investment	Profitability
LargeCap	-0.942**	7.372**	-0.0126	-0.935	-0.492	0.552
	(-2.03)	(2.47)	(-0.13)	(-1.03)	(-0.75)	(0.60)
Fundmntl	0.618	-4.409	0.113	1.562*	-0.296	-0.715
	(1.17)	(-1.36)	(1.09)	(1.68)	(-0.55)	(-0.73)
ProdServ	0.843	-14.96***	-0.337***	6.930***	2.542***	4.726***
	(1.09)	(-3.82)	(-3.01)	(4.27)	(2.73)	(4.17)
Dividends	-2.774***	38.89***	0.763***	-10.61***	-2.195***	-7.378***
	(-4.28)	(10.74)	(5.60)	(-6.48)	(-3.80)	(-7.89)
Deriv	0.796	7.771**	0.118	-1.397	0.502	-1.820
	(0.99)	(2.08)	(1.02)	(-1.13)	(0.63)	(-1.34)
CompAdv	-0.0285	6.995	-0.817***	3.233**	-0.420	7.258***
	(-0.04)	(1.40)	(-5.50)	(2.31)	(-0.53)	(5.79)
Defensive	-0.825	5.265	0.0531	-2.519**	-0.530	0.341
	(-1.25)	(1.64)	(0.56)	(-2.27)	(-0.87)	(0.32)
Quantit	-0.247	9.875**	0.0675	1.021	-2.325***	-0.447
	(-0.38)	(2.22)	(0.56)	(0.73)	(-4.68)	(-0.41)
SmallCap	2.153**	-68.46***	0.350*	5.071***	5.652***	-2.402
	(2.41)	(-18.03)	(1.76)	(3.36)	(4.45)	(-1.46)
LongTerm	0.626	5.138	-0.408***	0.562	0.912	2.421
	(0.72)	(1.11)	(-2.63)	(0.38)	(0.77)	(1.56)
PERatio	-0.236	-2.696	0.0336	-1.647	-0.569	-1.198
	(-0.27)	(-0.51)	(0.26)	(-1.02)	(-0.62)	(-0.84)
Foreign(ADR)	1.428*	1.856	-0.0654	3.049**	0.460	1.518
	(1.71)	(0.39)	(-0.34)	(2.23)	(0.70)	(1.10)
MidCap	0.249	-12.88***	-0.156	0.335	-1.184**	3.009**
	(0.35)	(-3.40)	(-0.69)	(0.29)	(-2.30)	(2.45)
IntrValue	-1.332	13.17***	0.287**	-7.736***	-2.803***	-3.813***
	(-1.26)	(2.81)	(2.04)	(-6.64)	(-3.58)	(-2.94)
FixedInc	0.114	2.770	0.101	-2.037*	-0.132	-3.458**
	(0.13)	(0.70)	(0.89)	(-1.69)	(-0.15)	(-2.39)
Tax	1.240	18.27***	0.0109	1.271	-0.348	-1.022
	(1.40)	(3.08)	(0.06)	(0.72)	(-0.30)	(-0.64)
Foreign(EM)	-1.173	-10.12*	-0.204	1.690	0.466	0.610
	(-1.27)	(-1.93)	(-1.62)	(1.35)	(0.83)	(0.41)
Obs	288,012	288,012	288,012	288,012	288,012	288,012

t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01

Table 12: **Balance Sheet Variables by SPG:** This table reports *differences* between average balance sheet characteristics of fund holdings within a particular Strategy Peer Group (SPG) and the average across all other SPGs. These differences are estimated in separate regressions of the dependent variables on dummies for each SPG (see section 4), controlling for fund age, log assets, expense ratio, turnover ratio, monthly fund flows, monthly fund flow volatility, and month as well as fund family fixed effects. All non-binary independent variables are demeaned. Standard errors are two-way clustered by fund and month.

	CurrAsts	Invent.	NonPrfAst	PP&E	Intang.	AstGrw	Cash
LargeCap	-0.05	0.47	-0.46	-1.42	0.78	-0.24	-0.23
	(-0.03)	(0.59)	(-1.40)	(-1.33)	(0.60)	(-0.52)	(-0.18)
Fundmntl	-0.72	0.26	1.10**	1.03	-1.75	0.66	-0.79
	(-0.49)	(0.42)	(2.53)	(0.89)	(-1.27)	(1.12)	(-0.55)
ProdServ	13.88***	-0.95	-0.96**	-1.27	5.25***	4.15***	11.83***
	(7.54)	(-1.27)	(-2.28)	(-0.84)	(3.48)	(4.85)	(6.07)
Dividends	-20.09***	-2.83***	0.98**	1.72	-4.54***	-6.18***	-14.61***
	(-12.05)	(-3.30)	(2.53)	(1.33)	(-3.24)	(-7.37)	(-9.32)
Deriv	-2.20	-0.24	-0.42	0.77	-2.87*	-1.14**	-0.72
	(-1.35)	(-0.30)	(-0.79)	(0.67)	(-1.73)	(-1.99)	(-0.40)
CompAdv	8.99***	-0.03	-1.94***	-6.17***	9.09***	2.22***	8.13***
	(4.43)	(-0.03)	(-3.61)	(-3.82)	(4.53)	(3.01)	(3.59)
Defensive	-2.00	0.27	1.44*	-2.50*	0.21	-0.86	-1.80
	(-1.23)	(0.25)	(1.86)	(-1.76)	(0.14)	(-1.33)	(-1.11)
Quantit	-5.96***	2.74***	-0.03	0.44	-8.78***	-2.63***	-5.30***
	(-3.53)	(3.59)	(-0.06)	(0.37)	(-4.94)	(-4.34)	(-3.38)
SmallCap	13.13***	2.24**	1.77***	6.30***	-11.08***	3.15***	9.17***
	(6.27)	(1.99)	(2.92)	(3.62)	(-5.93)	(3.19)	(3.85)
LongTerm	4.57**	1.24	-1.61***	-2.40	4.83**	1.21	4.83**
	(2.04)	(1.13)	(-3.35)	(-1.31)	(2.40)	(1.49)	(2.04)
PERatio	-4.20*	2.80**	0.92*	2.58	-2.46	-0.56	-5.80***
	(-1.94)	(2.48)	(1.69)	(1.38)	(-1.35)	(-0.63)	(-3.04)
Foreign(ADR)	1.51	-0.55	-0.22	-0.08	0.18	2.23***	2.40
	(0.73)	(-0.67)	(-0.36)	(-0.05)	(0.10)	(2.97)	(1.22)
MidCap	1.16	0.10	-2.35***	3.32*	13.10***	-0.27	-5.31***
	(0.64)	(0.11)	(-4.78)	(1.94)	(6.06)	(-0.37)	(-3.11)
IntrValue	-9.28***	-2.65**	1.24	-4.22**	5.47**	-3.82***	-5.77***
	(-4.61)	(-2.34)	(1.50)	(-2.08)	(2.10)	(-5.41)	(-2.63)
FixedInc	-2.16	-1.80*	1.11*	1.20	-3.80**	-0.41	-0.58
	(-1.12)	(-1.74)	(1.76)	(0.81)	(-2.19)	(-0.47)	(-0.30)
Tax	-4.10*	-1.27	-0.64	1.34	-4.68**	0.31	-1.24
	(-1.88)	(-1.22)	(-1.41)	(0.61)	(-2.38)	(0.36)	(-0.56)
Foreign(EM)	3.09	-1.90	-0.15	-0.11	2.36	1.39**	3.53*
	(1.37)	(-1.58)	(-0.22)	(-0.06)	(0.97)	(1.98)	(1.69)
Obs	288,012	288,012	288,012	288,012	288,012	288,012	288,012

t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01

Table 13: **Additional Balance Sheet and Income Statement Variables by SPG:** This table reports *differences* between average characteristics of fund holdings within a particular Strategy Peer Group (SPG) and the average across all other SPGs. These differences are estimated in separate regressions of the dependent variables on dummies for each SPG (see section 4), controlling for fund age, log assets, expense ratio, turnover ratio, monthly fund flows, monthly fund flow volatility, and month as well as fund family fixed effects. All non-binary independent variables are demeaned. Standard errors are two-way clustered by fund and month.

	CurrLiab	Levrg	DefTax	LTDebt	OperInc	EarnGrw	R&D
LargeCap	0.97	0.02	2.02	-2.76***	-2.22	-0.17	0.04
	(1.14)	(0.06)	(1.21)	(-3.16)	(-1.25)	(-0.80)	(0.03)
Fundmntl	-1.44*	-0.65*	-1.32	-0.10	1.60*	0.07	-1.91*
	(-1.68)	(-1.75)	(-0.72)	(-0.10)	(1.74)	(0.31)	(-1.79)
ProdServ	4.69***	-2.33***	-7.74***	-2.71**	-2.20	-0.13	6.07**
	(4.62)	(-5.01)	(-3.80)	(-2.34)	(-1.54)	(-0.38)	(2.52)
Dividends	-5.69***	2.82***	19.86***	2.89***	2.92***	0.28	-4.13***
	(-5.78)	(6.16)	(9.05)	(2.62)	(4.17)	(1.08)	(-3.70)
Deriv	-0.22	0.43	3.52*	-0.09	-2.94	0.14	-1.34
	(-0.22)	(0.91)	(1.69)	(-0.10)	(-0.83)	(0.62)	(-1.20)
CompAdv	4.39***	-1.24**	1.08	-5.62***	-0.18	0.25	-0.59
	(4.43)	(-2.31)	(0.40)	(-4.40)	(-0.11)	(0.63)	(-0.35)
Defensive	-0.50	0.88**	-3.04	-0.77	-0.48	0.11	-0.04
	(-0.50)	(2.11)	(-1.28)	(-0.69)	(-0.36)	(0.31)	(-0.02)
Quantit	0.99	2.15***	7.89***	0.67	4.33***	0.19	-3.40***
	(0.90)	(4.70)	(3.61)	(0.64)	(3.46)	(0.72)	(-3.48)
SmallCap	-2.31*	-2.58***	-19.50***	3.13**	-4.85**	-0.74	7.54**
	(-1.93)	(-4.79)	(-9.93)	(2.05)	(-2.17)	(-1.64)	(2.48)
LongTerm	1.65	-0.09	2.44	-0.59	0.19	0.04	2.51
	(1.29)	(-0.16)	(0.95)	(-0.31)	(0.21)	(0.11)	(0.97)
PERatio	-2.02	-0.05	-0.46	-0.34	1.50	-0.22	-0.76
	(-1.41)	(-0.08)	(-0.17)	(-0.28)	(1.64)	(-0.67)	(-0.38)
Foreign(ADR)	-0.08	-1.09**	2.00	-0.97	0.08	0.18	0.99
	(-0.06)	(-2.05)	(0.69)	(-0.71)	(0.10)	(0.62)	(0.49)
MidCap	3.05**	-0.17	-10.15***	5.48***	1.59**	0.21	-2.10**
	(2.45)	(-0.33)	(-4.57)	(4.34)	(2.04)	(0.51)	(-2.24)
IntrValue	-2.48*	3.05***	4.55*	4.72**	4.98	-0.75	-4.99***
	(-1.89)	(3.43)	(1.72)	(2.44)	(1.56)	(-1.16)	(-2.98)
FixedInc	-1.46	0.29	-3.03	3.33*	-0.16	-0.46	0.98
	(-1.23)	(0.60)	(-1.32)	(1.88)	(-0.15)	(-1.45)	(0.49)
Tax	-1.07	0.53	8.05**	-2.32	0.30	0.51	1.05
	(-0.77)	(0.90)	(2.48)	(-1.59)	(0.16)	(1.49)	(0.35)
Foreign(EM)	-0.13	-0.37	-6.30**	1.58	-1.56	0.65	-0.68
	(-0.09)	(-0.69)	(-2.14)	(1.18)	(-1.32)	(1.13)	(-0.71)
Obs	288,012	288,012	288,012	288,012	288,012	288,012	288,012

t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01

Table 14: **Security Market Characteristics by SPG:** This table reports *differences* between average market characteristics of securities held by funds within a particular Strategy Peer Group (SPG) and the average across all other SPGs. These differences are estimated in separate regressions of the dependent variables on dummies for each SPG (see section 4), controlling for fund age, log assets, expense ratio, turnover ratio, monthly fund flows, monthly fund flow volatility, and month as well as fund family fixed effects. All non-binary independent variables are demeaned. Standard errors are two-way clustered by fund and month.

	DivYield	Issuance	Repurchase	Illiquid	FirmAge
LargeCap	-0.23*	0.03	2.36**	0.00	5.49**
	(-1.74)	(0.07)	(2.29)	(0.33)	(2.05)
Fundmntl	-0.16	0.17	-1.75	-0.00	-5.21
	(-1.14)	(0.43)	(-1.39)	(-0.33)	(-1.64)
ProdServ	-0.74***	-1.49***	-4.91***	-0.00*	-23.29***
	(-4.03)	(-3.63)	(-3.59)	(-1.92)	(-6.36)
Dividends	3.48***	3.01***	5.07***	0.01**	61.97***
	(10.00)	(6.35)	(4.07)	(2.40)	(15.40)
Deriv	0.36	0.10	1.90	-0.00	8.23**
	(1.52)	(0.20)	(1.29)	(-1.28)	(2.16)
CompAdv	-0.83***	-1.32**	-3.60**	-0.01**	-11.85***
	(-4.34)	(-2.12)	(-2.08)	(-2.48)	(-2.72)
Defensive	0.27*	0.55	1.93*	0.00	2.11
	(1.70)	(1.23)	(1.69)	(0.39)	(0.70)
Quantit	0.12	2.23***	11.79***	0.01**	17.54***
	(0.69)	(3.43)	(6.83)	(2.14)	(4.19)
SmallCap	-1.06***	-2.82***	-19.07***	-0.01**	-53.29***
	(-6.26)	(-4.53)	(-11.80)	(-1.98)	(-16.02)
LongTerm	-0.66**	0.71	-1.07	-0.00	-3.30
	(-2.35)	(1.32)	(-0.71)	(-0.35)	(-0.74)
PERatio	0.01	-0.05	-0.21	0.00	9.79*
	(0.07)	(-0.09)	(-0.12)	(0.77)	(1.96)
Foreign(ADR)	-0.46**	0.24	1.14	0.00	-4.14
	(-1.97)	(0.40)	(0.68)	(0.69)	(-0.80)
MidCap	-0.58**	-1.42**	0.18	-0.00	-18.66***
	(-2.33)	(-2.08)	(0.12)	(-0.68)	(-4.78)
IntrValue	0.49*	0.18	10.80***	-0.00	15.32***
	(1.94)	(0.24)	(5.00)	(-0.42)	(3.08)
FixedInc	0.36	-0.35	0.51	0.01	3.34
	(1.36)	(-0.50)	(0.38)	(1.47)	(0.84)
Tax	0.11	1.79***	4.87**	-0.00	17.78***
	(0.45)	(2.65)	(2.32)	(-1.42)	(3.22)
Foreign(EM)	-0.50*	-1.61**	-6.02***	-0.00	-20.42***
	(-1.93)	(-2.32)	(-3.02)	(-1.19)	(-4.12)
Obs	288,012	288,012	288,012	288,012	288,012

t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01

Table 15: **Holdings Characteristics by SPG:** This table reports *differences* between average fund holdings characteristics for a particular Strategy Peer Group (SPG) and the average across all other SPGs. These differences are estimated in separate regressions of the dependent variables on dummies for each SPG (see section 4), controlling for fund age, log assets, expense ratio, turnover ratio, monthly fund flows, monthly fund flow volatility, and month as well as fund family fixed effects. All non-binary independent variables are demeaned. Standard errors are two-way clustered by fund and month.

	ADR%	ForeignInc	NumStocks	Cash%	CommStk%
LargeCap	-0.02	0.06	-2.05	2.45	4.48*
	(-0.38)	(0.34)	(-0.66)	(1.00)	(1.71)
Fundmntl	-0.09	0.06	-2.61	0.64	2.05
	(-1.15)	(0.24)	(-0.98)	(0.28)	(0.72)
ProdServ	-0.00	0.15	-1.42	-2.34	7.48**
	(-0.05)	(0.49)	(-0.71)	(-0.82)	(2.24)
Dividends	0.95***	0.37	-6.47*	-4.02	-23.36***
	(6.40)	(1.39)	(-1.75)	(-1.30)	(-4.80)
Deriv	0.02	0.18	3.52	2.23	-5.59
	(0.23)	(0.67)	(0.99)	(0.55)	(-1.09)
CompAdv	-0.03	0.13	-13.55***	0.98	8.85*
	(-0.31)	(0.47)	(-5.61)	(0.22)	(1.81)
Defensive	0.11	0.39*	-3.48*	3.40	-9.22**
	(1.17)	(1.68)	(-1.67)	(1.00)	(-2.36)
Quantit	-0.65***	-1.81***	18.93***	-15.44***	25.16***
	(-7.95)	(-7.54)	(4.89)	(-4.23)	(6.67)
SmallCap	-0.36***	-0.88***	40.14***	3.45	-3.90
	(-4.26)	(-2.91)	(4.79)	(1.15)	(-1.12)
LongTerm	0.02	-0.22	-11.17*	-1.33	7.61*
	(0.20)	(-0.72)	(-1.91)	(-0.33)	(1.66)
PERatio	-0.22**	-0.34	-4.50	-1.03	7.39*
	(-2.08)	(-1.03)	(-1.30)	(-0.24)	(1.73)
Foreign(ADR)	0.03	0.46	-4.18**	-3.89	4.31
	(0.18)	(1.45)	(-2.02)	(-1.32)	(1.09)
MidCap	-0.20	0.51	-6.47	-3.47	6.19*
	(-1.32)	(1.49)	(-1.54)	(-1.37)	(1.86)
IntrValue	0.10	0.50	-13.87***	3.66	-12.79**
	(0.67)	(1.62)	(-7.47)	(0.84)	(-2.01)
FixedInc	-0.09	-0.03	-1.43	21.33***	-26.77***
	(-0.87)	(-0.10)	(-0.65)	(4.40)	(-4.04)
Tax	0.12	-0.29	15.03	-12.45***	11.11**
	(0.84)	(-0.94)	(1.62)	(-3.34)	(2.54)
Foreign(EM)	0.46***	1.08***	-8.30**	7.60*	-16.74**
	(3.30)	(2.60)	(-2.13)	(1.81)	(-2.01)
Obs	288,012	288,012	288,012	275,825	275,825

t statistics in parentheses; * p<0.10, ** p<0.05, *** p<0.01

A K-means

A.1 The Algorithm

The K-Means algorithm takes as inputs the $tfidf$ matrix, the number of desired clusters (k) and a tolerance threshold (τ). The algorithm is initialized by choosing k points in the vector space (centroids). Points are deliberately chosen to be far from each other in order to minimixe the likelihood of converging to a local minimum. Each chosen point represents a features vector of same length as the number of chosen features $(10,000)$, whose elements exist in $[0,1]$. Then the following steps are repeated until the pre-defined tolerance level is reached:

1. Calculate the euclidian distance between the vectors representing each document (rows of the tfidf matrix) and each of the k centroid vectors as follows:

$$\sum_{r=1}^{R} ||x_r - x_r^C||^2 \tag{11}$$

   where x_r is the frequency assigned to feature r in a specific document and x_r^C is the frequency assigend to feature r in a cluster's centroid. R is the total number of features.

2. Assign each document to the closest centroid (form clusters)

3. Generate new centroids (features vectors) by taking the item-by-item average of the feature vectors of all documents assigned to the same cluster

4. Calculate the euclidian distance between the centroids at iteration n and those at iteration $n+1$.

   - If the largest distance is greater than the tolerance level τ, repeat all steps
   - Otherwise exit the loop and return the formed clusters (convergence)

We ran the above algorithm with different specifications for the user defined parameters (k and τ). All runs are independent (use different seeds). Despite the possibility of K-means reaching a local optimum, in our setting, the procedure is robust to changes in initial parameters (see discussion in Section A.2).

In the main specification we use the default value for τ in Python's scikit-learn implementation: 0.0001. We use 17 clusters for the determination of the Strategy Peer Groups (SPG).

A.2 Optimal Number of Clusters

In order to choose the correct number of clusters we perform independent runs of the K-means algorithm for $K = [10, 20]$. We then compare the categorization between each consecutive [K] and [K+1] optimal solutions. We base the choice of the optimal number of clusters on two criteria which we label density and stability.

Observations Cross-Classification: Stability Define the crosstab matrix as the number of observations falling in cluster i under [K] and cluster j under [K+1]

$$CrossTab_{(i,j)} = \#\ clustered\ as\ i\ under\ [K]\ and\ j\ under\ [K+1],$$

If we treat [K+1] as the ground truth, and [K] as the predicted value, for any combination (i, j), the denominator of its precision is the sum for all j given i, and the denominator of its recall is the sum for all i given j. Formally, define precision and recall as:

$$Precision_{(i,j)} = \frac{CrossTab_{(i,j)}}{\sum_{i=1}^{K} CrossTab_{(i,j)}}$$

$$Recall_{(i,j)} = \frac{CrossTab_{(i,j)}}{\sum_{j=1}^{K+1} CrossTab_{(i,j)}}$$

Intuitively, if $Precision_{(i,j)}$ is large, it means that observations classified as i under [K], are very likely to be classified as j under [K+1], meaning that i under [K] is likely to be a subset of j under [K+1]. Similarly, if $Recall_{(i,j)}$ is large, it means that observations classified as j under [K+1], are very likely to be classified as i under [K], meaning that j under [K+1] is likely to be a subset of i under [K].

We finally combine the two above criteria into an $Fscore$ matrix indicating the harmonic mean of precision and recall. Due to the characteristic of the harmonic mean, if $Fscore_{(i,j)}$ is large, $Precision_{(i,j)}$ and $Recall_{(i,j)}$ are both expected to be large, and cluster i under [K] is likely to be in line with cluster j under [K+1].

$$Fscore_{(i,j)} = 2 \cdot \frac{Precision_{(i,j)} \cdot Recall_{(i,j)}}{Precision_{(i,j)} + Recall_{(i,j)}}$$

We currently use a threshold of 0.5 to tell whether the F1 score is large enough:

$$\widehat{Fscore}_{(i,j)} = Fscore_{(i,j)} > 0.5$$

Note that a high F1 score in this context indicates a high stability over the independent optimal allocations found using the K-means algorithm with [K] and [K+1] clusters (i.e. these are unlikely to be local minima).

Euclidean Distance: Density Define the Eucledian distance between the centroids of any pair of clusters under [K] and [K+1] as:

$$Dist_{(i,j)} = ||C_i^K - C_j^{K+1}||_2$$

where C_i^K indicates the centroid of tfidf vector of cluster i under [K], and C_j^{K+1} indicates the centroid of tfidf vector of cluster j under [K+1].

If the distance between two centroids is very small, the underlying clusters are likely to be very similar in meaning. We currently, use the threshold 0.2 to tell whether the distance is small enough:

$$\widehat{Dist}_{(i,j)} = Dist(i,j) < 0.2$$

Optimal Choice: Stability and Density For row i in a criterion matrix ($\widehat{Fscore}_{(i,j)}$ or $\widehat{Dist}_{(i,j)}$), if the sum of that row is 0 (i.e. it does not include '1s'), cluster i is likely to be a new cluster; if the sum of that row is 1 (i.e. the row includes only 1 '1' in column j), then cluster i under [K] is likely to be in line with cluster j under [K+1]; if the sum exceeds 1 (i.e. the row includes more than 1 '1'), cluster i is likely to split into multiple fractions under [K+1], and the new clusters under [K+1] are columns whose values are 1. A similar reasoning applies to column j. Note that matched clusters are those for which the corresponding column/row in the criteria matrices equals to 1; if the column's sum equals 0, the column cluster is a new cluster; if row sum is greater than 1, the corresponding rows are new clusters.

In essence, stability indicates that most clusters could be matched across the two independent runs [K] and [K+1]. Density indicates that any unmatched cluster is non-trivial. We choose the optimal K (K^*) such that both criteria are satisfied.

B Alternative Clustering: LDA

Latent Dirichlet Allocation (Blei et al. (2003)) is commonly used for textual topic modeling in the finance and economics literatures (see the recent survey in Loughran and McDonald

(2020)). The reader may wonder why we use k-means clustering insead of the more common LDA. In practice, the results of the two methodologies are fairly similar, as can be seen in figure B1. This figure shows the topic word clouds resulting from an application of LDA with prior topic probability $Dir(1/k)$ and prior word probability $Dir(1/N)$, where k is the number of topics and N is the number of terms (words and bi-grams) in the corpus. For comparison with our *k-means* results, we choose $k = 17$.

Despite the similarities, k-means is more suited to our particular application because it explicitly generates maximum differences between groups, whereas LDA seeks to maximize the likelihood of a generative model, even if the resulting topics are fairly similar. Figure B1 illustrates this problem with LDA: there are two "Quantitative" topics, two "Fundamental" topics, and two "PE-Ratio" topics, with quite similar word distributions. On the other hand, topics that are separated by k-means are sometimes merged by LDA: e.g. "Tax" and "Dividends"; and "Small Cap" and "Large Cap".

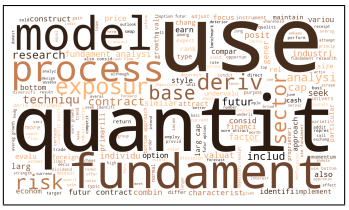
Another issue with LDA is that it is much less stable than k-means. The example in figure B1 would not be reproducible by re-running the algorithm, whereas k-means, even with different random seeds, delivers similar clusters every time it is run on the same data.

Finally, one advantage of LDA, in principle, is that it estimates a distribution over topics for each document. In the context of fund strategies, it would be useful to have the ability to assign multiple strategies to the same fund. In practice, however, due to the short length of the PIS descriptions (about 300 words, on average), over 90% of funds have virtually all of their weight in a single topic. The relatively short documents may also be the reason why LDA is less stable in our setting.

Figure B1: **LDA Topics:** Word clouds for $k = 17$ topics estimated by applying Latent Dirichlet Allocation to fund strategy descriptions. The word clouds represent the distribution of features (words and bi-grams) for each topic. Word sizes are proportional to frequency.

(a) Derivatives/Short Selling

(b) Quantitative 1



(c) Quantitative 2

(d) PE-Ratio 1

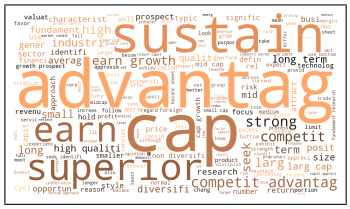
(e) PE-Ratio 2

(f) Free Cash Flow

(g) Foreign (EM)

(h) Foreign (ADR)

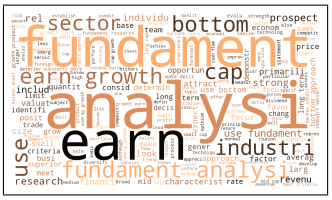
(i) Competitive Advantage



(j) Intrinsic Value

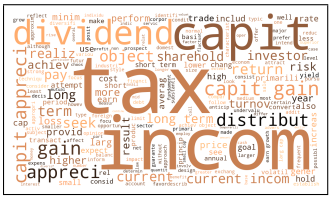
(k) Defensive

(l) Fundamental 1



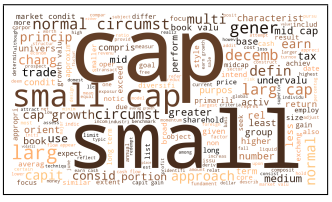
(m) Fundamental 2

(n) Tax/Dividends



(o) New Products

(p) Small Cap/Large Cap



(q) Fixed Income