What do hedge funds say?*

Juha Joenväärä† Jari Karppinen‡ Cristian Tiu§

March 22, 2019

Abstract

We investigate the information content of hedge fund strategy descriptions by testing two theories on writing sophistication. The first theory predicts that more sophisticated writing is positively linked to manager's talent, while the second associates sophistication with deceptive behavior. We find that in support of the first theory, funds with lexically diverse strategies outperform, are more likely to survive, take less financial risk and encounter fewer legal problems. In support to the second theory, we find that the syntactic complexity of strategy descriptions predicts more legal problems for the fund, while being associated weakly (or insignificantly) with outperformance and survival. Our mixed findings are consistent with an equilibrium in which talented managers write lexically diverse descriptions while less talented mangers gamble an entry into the industry, but their attempts to signal talent result in syntactically cumbersome and difficult to understand strategy descriptions.

*We are grateful for comments by Steve Dimmock and Veljko Fotak and we thank Kathleen Sheehan of Educational Test Services for her help and assistance with various measures of text sophistication. All errors are ours.

†Department of Economics, Accounting and Finance, University of Oulu, Oulu, Finland. Email: juha.joenvaara@oulu.fi

‡Department of Economics, Accounting and Finance, University of Oulu, Finland. Email: jari.karppinen@oulu.fi

§238 Jacobs Hall, Department of Finance, University at Buffalo, Buffalo, NY 14260. Email: ctiu@buffalo.edu

1 Introduction

Hedge funds are investment vehicles famed for their investment prowess, but sometimes also found at the center of financial scandals. Because they are secretive, investors seek any information that will help separating the good managers from those who are deceptive. By and large, financial economics literature has been using returns-based information to assess hedge fund quality. However, at the same time this information seems problematic. For example, Goetzmann, Ingersoll, Spiegel, and Welch (2007) assess that hedge fund manipulate returns and Kosowski, Naik, and Teo (2007) warn that returns are not normally distributed, making performance test statistics more difficult to construct. Bollen, Joenväärä, and Kauppila (2018) in fact show that most returns-based hedge fund performance predictors proposed by the literature fail to identify outperforming hedge funds in the recent period. These problems highlight the importance of seeking other fund characteristics than returns that may be useful in evaluate hedge funds. One such example, overlooked by financial economists, is given by hedge fund managers' writings.

It goes without saying that managers' writing samples reveal information about their authors and by extension, analyzing them may help a potential investor infer the quality of a fund. In this study we analyze the information content of hedge fund strategy descriptions. While other writing samples authored by hedge fund managers, such as investor letter or tweets, may contain more specific information about a fund, their availability is limited. By contrast, strategy descriptions are widely available in commercial databases and therefore their systematic study is possible.

In order to analyze the information content of strategy descriptions we focus exclusively on their level of sophistication and rely on two theories linking writing sophistication and fund quality. These theories offer mixed suggestions regarding the potential relationship between managers' text sophistication and the quality of a hedge fund. The first theory suggests that sophistication predicts fund quality. For example, Li, Zhang, and Zhao (2011) show that managers who attended universities with higher SAT averages (and who presumably write in a more sophisticated way) have higher raw and risk adjusted returns, take less risk and attract more inflows. From the point of view of this study, sophistication therefore predicts higher fund quality.

The second theory links writing sophistication to deceptive behavior. For example, just as the English expression "fine print" does not represent a beautiful text but rather an obscure, incomprehensible and misleading reference, studies in the psychology literature, such as Moffitt and Burns (2009) or Vrij, Granhag, and Porter (2010) show that the use of sophisticated vocabulary is associated with the intention to deceive and with lying.

In order to measure text sophistication we rely on two standard measures which can be computed through natural language processing tools. The systematic and automatic extraction of information using such tools was pioneered in financial economics research by Tetlock (2007), Loughran and McDonald (2011), and Loughran and McDonald (2014).[1] The first measure is lexical diversity (more

[1] A survey on the use of these techniques in the financial economics is Loughran and McDonald (2016), and another survey covering the use of textual analysis in the accounting literature is Fisher, Garnsey, and Hughes (2016). The information volunteered by hedge funds in writing, however, is relevant and consequential only as long as investors take it into consideration. To this point, Joenväärä and Tiu (2017) show that investors consider at least the hedge fund name – the most rudimentary type of text created by a hedge fund manager – when making investment decisions.

precisely we use the Shannon (1948) diversity index) – or simply put the degree to which a text contains a wide variety of distinct words. The second measure is syntactic complexity, calculated using a tool developed by Educational Testing Services and described in more detail in Sheehan (2015) – essentially, how complex the sentences are within a given text.

The two theories outlined above offer conflicting advice to a potential hedge fund investor: on the one hand, more sophisticated strategy descriptions appear to predict hedge fund quality, while at the same time funds with more complex strategy descriptions are expected to engage in deceptive behavior. This ambiguity suggests the possibility of a pooling equilibrium, such as in Stein (2005), in which bad hedge funds attempt to signal to investors that they are of high quality by posting sophisticated strategy descriptions.

However, our tests of the two theories outlined above reveal that one of the measures of text sophistication, namely, lexical diversity, is unambiguously associated with fund quality. More precisely, it is possible to separate the good funds from the bad using particularities of their strategy descriptions. For example, in terms of performance, the funds whose strategy descriptions are in the more lexically diverse decile outperform, in terms of Fung and Hsieh alphas, the funds in the lowest lexically diverse decile by 3.56% per year. In terms of appraisal ratios, the decile difference is 1.00 per year (in favor of the funds with the most lexically diverse strategies). The same differences, also statistically significant, are observed for a battery of measures of performance, including appraisal ratio and manipulation-proof performance measures. Funds posting strategies with high lexical diversity are also less likely to become extinct than funds with funds whose strategy descriptions are less lexically diverse. Furthermore, the funds with more lexically

diverse strategy descriptions were subjected to fewer regulatory actions and were associated with fewer civil/criminal legal problems than those funds whose strategy descriptions are less lexically diverse. It therefore appears that when a fund seeks to deceive its investors with a more sophisticated strategy description, that sophistication does not materialize as higher lexical diversity.

Our tests further reveal that the second measure of text sophistication we consider, namely, syntactic complexity, is associated with a more deceptive nature of the hedge fund. In particular, we find strong evidence that funds whose strategy descriptions are more syntactically complex experience more regulatory actions and report more civil/criminal legal problems than the funds with less syntactically complex strategy descriptions. While we find evidence that syntactic complexity is positively related to performance and survival, we also find that these relationships are weak, and they are mostly rendered insignificant by controlling for the other measure of writing sophistication, lexical diversity.

In order to estimate whether investors recognize those measures of text sophistication predicting fund quality, we analyze the response of flows to measures of text sophistication. We find that lexical diversity strongly predicts inflows, while the relationship between flows and syntactic complexity is not statistically or economically significant. Higher lexical diversity is also associated with higher fund leverage. This result is reassuring in that both investors and creditors prefer those hedge funds whose sophisticated writing indicates abnormal performance and integrity.

These findings suggest that quality managers use lexically diverse vocabulary to outline their strategies, and investors recognize their quality. Managers who are less talented, on the other hand, attempt to enter the industry risking that

their lack of talent is revealed. These latter funds attempt to deceive investors by using strategy descriptions that are syntactically complex - a form of able writing - but at the same time less clear. If they are lucky, these less talented funds perform well enough to warrant them an apparition in a database and a small degree of outperformance - much like the window dressers of Agarwal, Gay, and Ling (2014) - but at the other hand their tendency to deceive investors with strategy descriptions eventually expands to other deceptive behavior and eventually generates legal problems for the fund.

The paper is organized as follows. Section 2 motivates the measures for text characteristics and develops hypotheses. Section 3 describes the data used in this study. Section 4 examines the test sophistication of strategy descriptions, fund performance and financial risk. Section 5 examines the relation between disciplinary disclosures and strategy description sophistication. Section 6 analyzes the response of fund flows to text sophistication. Section 7 discusses robustness checks. Section 8 concludes.

2 Hypothesis development

How are the text characteristics of a hedge fund's strategy description linked with the quality of the fund?

The financial economics as well as the psychology literature separately offer some guidance, and their prescriptions are mixed. To start with, Hwang and Kim (2017) show that disclosure documents that are more difficult to read (or equivalently, more sophisticated) cause a firm to trade at a discount relative to its fundamental value. By contrast, Li et al. (2011) show that managers who attended

universities whose students have higher SAT average scores (and who presumably write in a more sophisticated way employing richer vocabularies) have higher raw and risk adjusted returns, take less risk and attract more inflows. Finally, studies in the psychology literature, such as Moffitt and Burns (2009) or Vrij et al. (2010) show that the use of sophisticated vocabulary is associated with the intention to deceive and with lying.

Inspired by evidence that managers with higher SAT scores outperform (and sophisticated writing is correlated with a high standardized test score), we hypothesize that hegde funds having more sophisticated strategy descriptions outperform. As the psychology literature shows, however, a sophisticated strategy description may also signal deceptive behavior, and in turn this sends an undesired signal to potential investors. We hypothesize that funds with true investment talent will specialize their type of writing sophistication, and this specialized text characteristic will be in turn associated not only with outperformance, but also with less deceptive behavior.

Funds with lower quality will attempt to mimic the good funds (as in Stein (2005)) by writing strategy descriptions that are also sophisticated. Because these funds behave in ways similar to the window-dressers of Agarwal et al. (2014), when signaling that they too are quality funds, they will use writing sophistication that has the potential to allow for opaque, difficult to understand texts. In their attempt to dupe investors, these funds will use their texts high in that measure of sophistication that is closest to deceptive behavior. This will be a type of sophistication not readily present in what the honest managers of good funds write.

We thus predict that there is a measure of text sophistication such that funds

with more sophisticated strategies, as described by that measure, outperform and engage in less deceptive behavior, having therefore fewer legal problems. We also predict that another measure of sophistication will in turn predict deceptive behavior, and consequently legal problems for the fund.

As mentioned, the measure of sophistication predicting quality is lexical diversity, while the measure of performance predicting deceptive behavior and more legal problems is syntactic complexity. We continue by describing our measures of writing sophistication.

2.1 Measures of writing sophistication

In order to assess the quality of writing samples put forth by hedge funds we use texts in which hedge funds invested energy, and that are not treated lightly: description of hedge funds' investment strategies. While hedge funds contributed to the corpus of written English on an ongoing basis - with their managers writing op-ed pieces, commenting on financial news, writing on various current issues, publishing books or offering television interviews, strategy descriptions are well-thought, reflect the vision of the entire fund rather than a manager's personal opinion and will likely be accessed by all the potential investors. From that point of view, strategy descriptions are the perfect English text indubitably and closely related to a fund.

We therefore collect hedge funds' strategy description texts from commercial hedge fund databases. For each strategy description we then compute text-based measures of writing sophistication.

Before we calculate particular measures of writing sophistication, however, we

point out that all these measures are correlated with text length, as there is more opportunity to exhibit sophisticated writing in a longer text. We therefore calculate the logarithm of *text length* in characters and use this measure the simplest proxy for text sophistication. This is similar to measures of complexity first proposed by Loughran and McDonald (2014), who argue that 10-K document file size provides a simple proxy for how easily readable the entire filing is. Just like in their case, in our study text length captures roughly the quantity of information provided in the written strategy description of a fund.[2]

Simply measuring length ignores, however, the quality of writing in those texts. To better capture this quality, one suggestion made by the literature on semantic content is to use dictionary-based word-count indices. However, such method depends on the choice of the word list, and choosing specific words to the finance profession may lead to endogeneity problems is estimating whether a text is sophisticated. To avoid such problems, we use measures that quantitatively characterize the sophistication of text that are well-established in literature. The first measure we employ in our study *lexical diversity* which represents the diversity of the vocabulary appeared in a text. Higher lexical diversity means the text features richer vocabulary with more synonyms use and less repetition of words. There are several approaches to measuring lexical diversity documented in the literature (for example, see Tweedie and Baayen (1998) and Jarvis (2013)).

Typical lexical diversity measures, such as the type-token ratio or its variants are computed by first applying standard textual analysis preprocessing steps (see, e.g., Buehlmaier and Whited (2018)), in which all non-alphanumeric characters

[2]Text length is always positive, hence its distribution has a right skew, and taking the log transformation is makes the variable more normally distributed.

are removed, all letters are converted to lowercase, English stop words removed, and texts stemmed. As a result, we obtain a list of tokens representing the words appearing in the text, and their frequencies. For example, Carroll's corrected type-token ratio (CTTR) is calculated as

$$\frac{V}{\sqrt{2 \times N}}$$

where N is the total number of tokens in the text and V is the number of distinct tokens.

For our analysis, we specifically use the Shannon diversity index[3] proposed by (Shannon, 1948) as a measure of lexical diversity. Shannon diversity index is given by

$$H' = -\sum_{i=1}^{V} p_i \ln p_i$$

where p_i is the proportion of ith word in the text containing V distinct words. The results of our analysis remain are similar when we use CTTR or the Shannon diversity index. The justification for using the Shannon entropy in our reported results is simply that the concept has been applied to financial economics in other contexts, for example to measure the diversification of a portfolio as in Meucci (2009). As we mentioned at the outset, our measure of lexical diversity is correlated with the text length (see Fergadiotis, Wright, and Green (2015)), and consequently in our tests we also simultaneously control for the latter.

As a second measure of text sophistication we consider the *syntactic complexity* of the text. Syntactic complexity is calculated using the the TextEvaluator® Score

[3]Shannon originally introduced the notion of information entropy to measure the degree of uncertainty (or unpredictability) in a message. The information entropy of a message can be understood as the amount of information the message contains.

(Sheehan, 2015). According to Napolitano, Sheehan, and Mundkowsky (2015) and Sheehan (2016), syntactic complexity encapsulates all information regarding how complex the sentences are within a text. It relies on information from syntactic parse trees, and part-of-speech tags, as well as basic measures such as the number of extremely long sentences and the size of the longest paragraph, and an automated version of the word "depth" measure introduced by Yngve (1960). This last feature, called average maximum Yngve depth, is designed to capture variation in the memory load imposed by sentences with varying syntactic structures. It is estimated by first using a syntactic parser to assign a depth classification to each word in a text, then determining the maximum depth represented within each sentence, and then averaging resulting sentence-level estimates to obtain a passage-level estimate. As a tool proposed by Educational Text Services to assess text sophistication, syntactic complexity certainly lends itself to identifying hedge funds whose managers wrote such that their standard test scores - and therefore the quality of the schools they graduated from - are high.

We continue by providing some examples of text sophistication calculations for two funds of different quality.

2.2 Hedge fund investment strategy descriptions

This section shows some examples of strategy description as reported to hedge fund databases. In the first example, the strategy description has a clear syntax structure, but the lexical diversity is high. The authors write clearly, but show great ability to use synonyms and make their text highly readable. In the second, syntax structure is nearly needlessly complicated, making the strategy difficult to

understand.

2.2.1 Sophistication and quality

The following paragraph contains the strategy description of CNH Diversified Opportunities Master Account LP managed by CNH Partners LLC, a fund whose principals are familiar to the reader as financial economists:

Led by principals Mark Mitchell, PhD, and Todd Pulvino, PhD, Diversified Opportunities (the "Fund") is an opportunistic event-driven hedge fund targeting market neutrality. The Fund focuses on liquidity-providing investments across a broad range of global corporate securities using proprietary quantitative screens and a fundamental research approach. The strategy is designed to capture systematic market as well as idiosyncratic security pricing anomalies related to mergers/acquisitions, credit/distressed events, changes in corporate capital structures and other arbitrage opportunities. Strategic Advantages: Principals Mitchell and Pulvino have applied a disciplined approach to managing arbitrage strategies since 2001. Frequent experience taking activist stance through lawsuits and serving on creditor committees. Historical proprietary databases inform investment thesis: -Merger arbitrage database tracking over 15,000 deals since 1962. -Convertible arbitrage database tracking over 3,000 issues since 1985 - Other proprietary databases of corporate spin-offs, high yield bonds, dual-class securities. Diversified approach allows fund to migrate toward most attractive dislocations and to withstand short-term pricing fluctuations. Market dislocation of 2008 created an historically attractive opportunity set across the Fund's underlying strategies. Investment Style: Quantitative tools are used to synthesize data, evaluate trading strategies, screen investment opportunities. Fundamental research and security selection is used to identify the most promising investments. Activist strategies are used with corporate management, including serving on creditor committees and actively participating

in balance sheet restructurings.

To illustrate our measures of sophistication, this strategy description has a text length above median at $\ln(1744) \approx 7.46$ and a syntactic complexity below median at 54 (we will discuss summary statistics in the next section). A lower syntactic complexity means that the text is easier to read. The lexical diversity of the text measures at 4.48. In terms of lexical diversity the fund thus belongs to the highest decile in the sample, a feature easily noted as we note diverse terms as well a lack of term repetitions.[4]

Our hypothesis states that having such a well written strategy description is associated with a higher quality for the fund, both in terms of performance as well as in the absence of legal problems.

2.2.2 Sophistication and deception

We continue by providing another sophisticated strategy example, but this kind with a different sophistication. The following paragraph contains the strategy description of Fairfield Sentry Ltd fund managed by Fairfield Greenwich Group. This fund came to the attention of general public as a feeder to Bernard Madoff's

[4]To calculate this measure, which captures word variety, we note that the text has 104 distinct tokens: fund (5), invest (5), strategi (5), opportun (4), corpor (4), secur (4), use (4), arbitrag (4), databas (4), market (3), proprietari (3), approach (3), sinc (3), princip (2), mitchel (2), phd (2), pulvino (2), diversifi (2), across (2), quantit (2), screen (2), fundament (2), research (2), price (2), merger (2), manag (2), activist (2), serv (2), creditor (2), committe (2), histor (2), track (2), attract (2), disloc (2), led (1), mark (1), todd (1), opportunist (1), event-driven (1), hedg (1), target (1), neutral (1), focus (1), liquidity-provid (1), broad (1), rang (1), global (1), design (1), captur (1), systemat (1), well (1), idiosyncrat (1), anomali (1), relat (1), acquisit (1), credit (1), distress (1), event (1), chang (1), capit (1), structur (1), strateg (1), advantag (1), appli (1), disciplin (1), frequent (1), experi (1), take (1), stanc (1), lawsuit (1), inform (1), thesi (1), deal (1), convert (1), issu (1), spin-off (1), high (1), yield (1), bond (1), dual-class (1), allow (1), migrat (1), toward (1), withstand (1), short-term (1), fluctuat (1), creat (1), set (1), under (1), style (1), tool (1), synthes (1), data (1), evalu (1), trade (1), select (1), identifi (1), promis (1), includ (1), activ (1), particip (1), balanc (1), sheet (1), restructur (1).

fund.

The Fund seeks to obtain capital appreciation of its assets principally through the utilization of a non-traditional options trading strategy described as 'split strike conversion', to which the Fund allocates the predominant portion of its assets. The investment strategy has defined risk and reward parameters. The establishment of a typical position entails (i) the purchase of a group or basket of equity securities that are intended to highly correlate to the S&P 100 Index, (ii) the purchase of out-of-the-money S&P 100 Index put options with a notional value that approximately equals the market value of the basket of equity securities and (iii) the sale of out-of-the- money S&P 100 Index call options with a notional value that approximately equals the market value of the basket of equity securities. The basket typically consists of between 40 to 50 stocks in the S&P 100 Index. The primary purpose of the long put options is to limit the market risk of the stock basket at the strike price of the long puts. The primary purpose of the short call options is to largely finance the cost of the put hedge and to increase the stand-still rate of return. The 'split strike conversion' strategy is implemented by Bernard L. Madoff Investment Securities LLC (BLM), a broker-dealer registered with the Securities and Exchange Commission, through accounts maintained by the Fund at that firm. The services of BLM and its personnel are essential to the continued operation of the Fund, and its profitability, if any. The Investment Manager, in its sole and exclusive discretion, may allocate a portion of the Fund's assets (never to exceed, in the aggregate, 5% of the Fund's Net Asset Value, measured at the time of investment) to alternative investment opportunities other than its 'split strike conversion' investments.

For the Fairfield Sentry Ltd fund the text length is $\ln(1825) \approx 7.51$. The lexical diversity is 4.39.[5] Both measures are above median when compared to the full

[5]The text has 100 distinct tokens (frequency shown in parentheses): fund (6), invest (6), option (5), basket (5), secur (5), valu (5), asset (4), strike (4), s (4), p (4), index (4), put (4), strategi (3), split (3), convers (3), equiti (3), market (3), alloc (2), portion (2), risk (2), typic

sample. However, the syntactic complexity of this text computed by TextEvaluator is notably high at 84. This is not surprising: the text is difficult to read, with its long sentences whose flow is interrupted by insertions of definitions of terms that are first used and then defined. This is the hallmark of a fund using a complex strategy description but in reality obfuscating investor understanding rather than clarifying it.

3 Data

This paper uses a comprehensive hedge fund dataset that is constructed by merging several commercial databases. Using comprehensive data mitigates the potential for a myriad of data mining biases, and therefore we use a wide cross-section of funds spanning a long study period. One particular and novel aspect of our database is yearly snapshots starting from 2007. These snapshots allow us to conduct out-of-sample tests entirely free from well-known biases such as the look-ahead and backfill bias.

(2), purchas (2), notion (2), approxim (2), equal (2), call (2), stock (2), primari (2), purpos (2), long (2), blm (2), seek (1), obtain (1), capit (1), appreci (1), princip (1), util (1), non-tradit (1), trade (1), describ (1), predomin (1), defin (1), reward (1), paramet (1), establish (1), posit (1), entail (1), group (1), intend (1), high (1), correl (1), ii (1), out-of-the-money (1), iii (1), sale (1), out-of-th (1), money (1), consist (1), limit (1), price (1), short (1), larg (1), financ (1), cost (1), hedg (1), increas (1), stand-stil (1), rate (1), return (1), implement (1), bernard (1), l (1), madoff (1), llc (1), broker-deal (1), regist (1), exchang (1), commiss (1), account (1), maintain (1), firm (1), servic (1), personnel (1), essenti (1), continu (1), oper (1), profit (1), manag (1), sole (1), exclus (1), discret (1), may (1), never (1), exceed (1), aggreg (1), net (1), measur (1), time (1), altern (1), opportun (1) .

3.1 Commercial hedge fund databases

We apply the aggregation procedure of Joenväärä, Kosowski, and Tolonen (2016) to construct our hedge fund sample. The final data set combines BarclayHedge, EurekaHedge, eVestment and Hedge Fund Research databases, yielding a total of 21,379 funds with 1,449,207 monthly time series observations of returns and assets under management (AUM), and covering the period from January 1994 to December 2016. From monthly series, the returns and AUMs are aggregated to 121,643 annual observations. For each fund, we collect strategy description texts as well as variables related to compensation structure and share restrictions. We exclude texts that do not contain at least one meaningful sentence in English describing the investment strategy. Although we have access to Lipper TASS, Morningstar and Preqin hedge fund databases, we do not use them since they do not contain textual information on hedge fund strategies.[6] The time-varying and time-invariant fund characteristics based on annual data are shown and defined in Table 1 (Panels A and B).

We classify the funds based on investment styles provided by the database vendors into four broad strategies: *directional*, *relative value*, *security selection*, and *multiprocess* traders as in Agarwal, Daniel, and Naik (2009), who in turn motivate their classification by the work of Fung and Hsieh (1997) and Brown and Goetzmann (2003). Strategy classifications are important and given that hedge funds' strategies are heterogeneous, throughout this paper we standardize the used text characteristic measures within broad strategies. This allows us to address the potential concern that the content of strategy description texts differs systemati-

[6]Our version of Lipper TASS provides strategy descriptions only for active funds. We opt not to use them because it may lead to survivorship bias.

cally between strategies. Furthermore, since the strategy descriptions written by native English speaking managers can be more fluent than the descriptions written by non-native English speaking managers, we group the funds based on their management firm domiciles. In our robustness tests, we show that our conclusions hold even when we conduct our analysis using only the funds whose management firms are domiciled in English-speaking countries.

Our novel database aggregation allows us to address well-known biases in hedge fund data. In particular, since we use both active and defunct funds, our analysis does not suffer from survivorship bias. Furthermore, in our baseline tests, we address the performance backfill bias by dropping out the first 12 return observations. Finally, in order to mitigate the effect of look-ahead bias arising from potential changes in strategy description texts, we have collected the strategy descriptions from old data annual snapshots starting from 2007.[7] In the analysis reported in this study, the strategy descriptions are imputed backwards for years prior 2007, i.e. we assume the strategy descriptions for funds started before 2007 have not changed. As we find that changes in strategy description texts and very rare, this assumption does seem to materially affect our results. Consistent with this assertion, when we re-run our empirical analysis using only the post-2007 data, for which we can capture the changes in strategy description texts, results remain qualitatively the same. This latter testing framework does not use any backfilled information on strategy descriptions or fund returns, and, thereby, we can conclude that our analysis does not suffer from either a look-ahead or backfill

[7]More specifically, for BarclayHedge we have snapshots of strategy descriptions for 2010, 2011, 2012, 2013, 2015 and 2017. For EurekaHedge we have snapshots of strategy descriptions for 2007, 2008, 2009, 2010, 2011, 2013, 2015, 2017. For HFR we have snapshots of strategy descriptions for 2007, 2009, 2011, 2012, 2013, 2015 and 2017. For eVestment the strategy descriptions are collected from a 2017 snapshot.

bias.

3.2 Text characteristics measures

We calculate the three measures of text characteristics described in Section 2.1.

Table 1 Panel C plots the time-series averages and confidence bands of the text characteristics. For all of the three text characteristics, we observe that the most recent strategy descriptions are more complicated than what the strategy text descriptions used to be when hedge fund industry was relatively new.

Descriptive statistics and correlations are shown in Panels D and E of Table 1. One notable detail is that text length and lexical diversity are highly positively correlated. This follows from the fact that for a population of V different words, the Shannon diversity index is always bounded by 0 (in the case that there is only 1 word) and $\ln V$ (in the case that all V words are equally common) and that the strategy texts are comparably short (typically a few paragraphs at maximum). Nevertheless we use it to proxy for richness of vocabulary in written strategy descriptions and further in the analyses show that quality of the content outperforms quantity of the content. The literature on the validity of different measures of lexical diversity for different sample sizes is mixed, see for example Fergadiotis et al. (2015).

4 Fund performance

In this section we present empirical results linking the strategy sophistication scores to hedge fund performance. First, we conduct standard univariate as well as double portfolio sorts in order to examine whether text characteristics are associated

with fund performance. Second, we run a set of panel regressions to confirm that the link between text characteristics and subsequent fund performance is robust to controlling for a variety of fund characteristics. The multivariate regression framework further allows us to study which of the text characteristics is the most important variable in explaining the cross-sectional performance differences in hedge fund returns, and whether any of the text characteristics subsumes the others.

4.1 Portfolio sorts

In order to investigate performance predictability, we sort hedge funds into decile portfolios based on each measure of text sophistication. More specifically, we sort funds at end of each year into decile portfolios and, then evaluate the decile equally-weighted returns for the following year. To construct a single time series spanning the entire sample period, we concatenate the decile portfolio returns across the holding periods. Following the practice in this literature, we use the Fung and Hsieh (2004) seven-factor model to assess the performance of hedge funds. In order to gauge the economic magnitude of performance predictability, we estimate several different performance measures spreads between the top- and the bottom-decile portfolios. For each of the spreads, we conduct statistical significance tests using the heteroskedasticity and autocorrelation consistent (HAC) GMM estimator. In our baseline tests, we correct for backfilling bias by removing the first 12 months of a fund's returns (e.g. Kosowski et al. (2007) and Bollen et al. (2018)).[8]

Panels A, B and C of Table 2 show the results for univariate portfolio sorts by each text characteristic. Our results are supportive of the fact that the funds with high lexical diversity are strongly associated with superior performance, while

[8]In robustness tests, we show that our findings are not sensitive to this assumption.

there is some weaker evidence that higher syntactic complexity is associated higher performance. For lexical diversity, we find that the annualized Fung-Hsieh 7-factor alpha spread between the top and bottom decile portfolios is 3.56% with a t-statistic of 6.10. For text length, the alpha spread is slightly lower, at 3.25% with a t-statistic of 4.84. For syntactic complexity, the alpha spread is considerably lower being equal to 1.54% with a t-statistic of 2.26. For both text length and lexical diversity, we find that the risk measured using residual volatility relative to Fung-Hsieh 7-factor model is lower for funds in the top portfolios. The annualized top-bottom volatility spread are -0.57% and -0.69% with t-statistics of -2.89 and -3.36. For syntactic complexity, the spread in residual volatility is statistically indistinguishable from zero. To measure performance as a unit of risk taken by hedge fund manager, we estimate the appraisal ratio spreads between the top- and bottom decile portfolios. We find that appraisal ratio spread is extremely high for lexical diversity, being 1.00 per annum with a t-statistic of 7.69. The respective spread for text length is 0.89 with t-statistic of 6.84 and for syntactic complexity only 0.36 with t-statistic of 2.63.[9]

In order to gain insight into the type of funds whose strategy descriptions show high lexical diversity, it is useful to analyze fund risk exposures (Panels F, G and H of Table 2). It is interesting to note that equity as well as credit risk exposures are significantly lower for the higher lexical diversity funds compared to the lower lexical diversity funds. For the syntactic complexity, we cannot document any similar consistent patterns in their risk exposures. This is consistent with

[9]We run similar univariate sorts for a sample where the funds are restricted to those managed by firms domiciled in English-speaking countries and for a sample where the data period starts from 2007 (i.e. from the earlier commercial database snapshot). The results are qualitatively unchanged, largest and statistically significant spreads are realized when sorting funds by lexical diversity.

funds writing less lexically diverse strategies using less sophisticated strategies, such as being exposed to more market risk. In contrast to text sophistication being linked positively to the investment strategy sophistication, when we use syntactic complexity as a measure of text sophistication we observe that high syntactic complexity is associated with lower exposure to derivative strategies (the primitive trend following strategies from Fung and Hsieh (2001)). It appears that funds with higher levels of sophistication in their investment strategies can write strategy descriptions that are lexically more diverse, while hedge funds that are less sophisticated and refrain from using derivative strategies write syntactically more complex descriptions.

To study whether a text characteristic contains unique information on fund performance, we report the results for double portfolio sorts in Panels D and E of Table 2. First in Panel D we sort the funds by text length into two groups and next each group is sorted to decile portfolios by lexical diversity. We find that the both alpha and appraisal ratio spreads are positive and statistically significant both below-median and above-median text length groups suggesting that the lexical diversity contain unique information that do not depend on text length. In Panel E the order of the sort variables is reversed, we first sort the funds into two groups by lexical diversity and then into decile portfolios by text length. The spreads in alphas, residual volatilities and appraisal ratios are now smaller in magnitude, but the text length contains some unique information and is not totally redundant. Hence, the double sorts suggest that although the text length and lexical diversity are highly correlated, the effects of the characteristics can be decoupled and that lexical diversity is more significant driver of fund performance.[10]

[10]In the robustness section, we show that these findings remain consistent in double sorts

4.2 Multivariate analysis

Although our univariate and double sorts analysis provides strong evidence that text characteristics are positively associated with fund performance, it is important to show that the other fund characteristics do not subsume them in multivariate regressions. The multivariate regression model also allows us to study which of the text characteristics are the most important variables in explaining the hedge fund cross-sectional performance differences.

To examine the relationship between text characteristics and hedge fund returns, we run a set of annual panel regressions while simultaneously controlling for the role of other fund characteristics. Similarly to Agarwal et al. (2009), we estimate fund-level alphas from time-series regressions of excess net returns on the Fung-Hsieh factors. We measure annual alpha as the sum of monthly alphas in that year, where monthly alpha is given by the sum of the intercept and the monthly residual. Similarly we measure annual residual volatility as the volatility of monthly residuals in that year, and appraisal ratio as the ratio of alpha and residual volatility.

The panel regression model is specified as

$$\begin{aligned}\text{Measure}_{i,t} = \gamma_0 + \gamma_1 &\text{Text characteristics}_{i,t-1} + \gamma_2 \text{Measure}_{i,t-1} \\ &+ \gamma_3 \text{Time-varying controls}_{i,t-1} + \gamma_4 \text{Time-invariant controls}_i + \varepsilon_{i,t},\end{aligned}$$

where $\text{Measure}_{i,t}$ is either hedge fund i Alpha, Residual Volatility or Appraisal Ratio, for year t. $\text{Text characteristics}_{i,t-1}$ are combinations of the lagged text

conducted for the both subsamples (a sample where funds are restricted to those managed by firms domiciled in English-speaking countries, and a sample where the data period starts from 2007).

characteristics. The text characteristics are winsorized at 1% and 99% percentiles and scaled to zero mean and unit standard deviation within investment styles. Time-varying controls include $\text{Measure}_{i,t-1}$ as well as lagged assets under management and age of the fund measured in years from inception. The time-invariant controls include share restrictions, compensation structure variables, and leverage indicator taking value of one when the fund uses leverage and otherwise zero. All of the specifications include both the time fixed effects and style fixed effects, and the standard errors are clustered by fund.

Panels A, B and C of Table 3 report the results from these regressions. From the panels, we observe that lexical diversity is positively (negatively) related to Fung-Hsieh alphas and appraisal ratios (residual volatility), and that this relationship is highly statistically significant. The economic significance is also high, for example, with one standard deviation increase in lexical diversity being associated with 47 to 89 basis points increase in annual alphas. Syntactic complexity and text length also predict outperformance in standalone regressions, but the statistical significance of these relationships wanes when regressions include lexical diversity, suggesting that lexical diversity subsumes other text characteristics. Overall, the multivariate regression results suggest that lexical diversity is the most important text characteristic in explaining cross-sectional differences in hedge fund performance.

4.3 Performance measure manipulation, tail risk and fund failures

Although we document that our text characteristics measures are related to fund performance, the relationship may not be robust for performance measures that are subject to manipulation. As Goetzmann et al. (2007) show, the fund managers can use various techniques to manipulate performance measures. Some of the fund managers that write complicated strategy description texts may also employ complex instruments or strategies in order to manipulate performance measures and report better performance. To address this concern, we evaluate fund performance using the manipulation-proof performance measure of Goetzmann et al. (2007).

For this purpose, we first estimate annual fund-level manipulation-proof performance measures (MPPM) with risk aversion parameter $\rho = 3$ using the monthly returns for each fund in each year. Then, we run the following panel regressions

$$\begin{aligned}\text{MPPM}_{i,t} = \gamma_0 + \gamma_1 &\text{Text Characteristics}_{i,t-1} + \gamma_2 \text{Time-varying controls}_{i,t-1} \\ &+ \gamma_3 \text{Time-invariant controls}_i + \varepsilon_{i,t},\end{aligned}$$

where $\text{MPPM}_{i,t}$ is hedge fund i MPPM at time t. Text characteristics, time-varying and time-invariant controls are defined and specified similarly as in Section 4.2.

Panel A of Table 4 shows that the higher lexical diversity is consistently related to higher utility-based performance. Syntactic complexity and text length appear to have similar relationships in standalone regressions, but those effects disappear or become negative when specification includes lexical diversity. Hence, the lexical diversity appears to subsume two other text characteristics, suggesting that out-

performance by funds with lexically complex strategies is not due to performance manipulation.

Another concern is the presence of nonlinearities in hedge fund returns. Therefore, the conclusions inferred from traditional risk measures that assume normality may be misleading. To address this concern, we measure the risk using maximum loss defined as the most negative return or maximum loss for year t. This measure should better take into account the potential downside risk in hedge fund returns. We run a set of multivariate regressions in which the maximum loss is a dependent variable and text characteristics are main explanatory variables. Panel B of Table 4 shows that the coefficients for lexical diversity are consistently negative and cannot be subsumed be by text length or syntactic complexity. This suggest that non-linearities in hedge fund returns do not drive our conclusion on the relationship between text characteristics and risk.

Finally, we confirm that text characteristics are not associated with a higher likelihood of fund failures. To do so, we estimate annual probit regressions:

$$\begin{aligned}\text{Pr(Attrition)}_{i,t} = \gamma_0 &+ \gamma_1 \text{Text Characteristics}_{i,t-1} \\ &+ \gamma_2 \text{Low rank}_{i,t-1} + \gamma_3 \text{Mid rank}_{i,t-1} + \gamma_4 \text{High rank}_{i,t-1} \\ &+ \gamma_5 \text{Time-varying controls}_{i,t-1} \\ &+ \gamma_6 \text{Time-invariant controls}_i + \varepsilon_{i,t},\end{aligned}$$

where Attrition is a binary variable that takes the value of one, if the fund stops reporting returns in year t, or zero otherwise. The low, mid and high past-year performance ranks of the funds are defined as in Sirri and Tufano (1998). The text characteristics and control variables are as in previous regressions. The results of

these probit analysis are shown in Panel C of Table 4. We observe that higher lexical diversity consistently predicts lower probability of attrition even after controlling for text length and syntactic complexity.

Overall, we conclude that text sophistication, measured in particular as lexical diversity, positively predicts fund performance as well as lower risk-taking even after we control for performance measure manipulation, non-linearities in fund returns and fund failures.

5 Disciplinary disclosures

In this section, we examine whether funds' strategy descriptions' text characteristics predict deceptive behavior. To do so, we first gather hedge funds' disciplinary disclosures from Form ADV reports. Not only do we collect the most recent ADV files, but we also include the historical filings, which allows us to actually conduct predictive tests. In our tests, we analyze whether the text characteristics are associated with regulatory, civil or criminal action disclosures. As we have conjectured, we expect that the syntactic complexity is positively associated with deceptive behavior, while the lexical diversity is associated with true skill, and not with deceptive behavior.

5.1 Form ADV

To measure deceptive behavior, we rely on hedge funds' disciplinary disclosures gathered from mandatory regulatory filings. To be more specific, the Securities and Exchange Commission (SEC) requires hedge funds (exceeding a threshold on assets under management and number of clients) to register as such by filing

an annually updated Form ADV report, which includes rich information about items such as the advisor's assets, clients, employees, investment style, affiliates, and history. Similar to Dimmock and Gerken (2012), we use the historical filings available at Historical Archive of Investment Adviser Reports[11] filed from August 2001 through June 2017. Our analysis is based on the latest Form ADV filing in each year for each advisor.

The Item 11 of the Form ADV asks for past legal or regulatory violations of the advisor or its affiliates, and Brown, Goetzmann, Liang, and Schwarz (2008, 2009) show that these past violations are connected to other information in the Form ADVs, such as conflicts of interest, and also to hedge fund characteristics. From Item 11, we extract two indicator variables (RegAct and CivilOrCriminal) which take a value of one if the advisor (or any of its related persons) has had legal or regulatory violations during the last 10 years causing the managing firm to complete a Regulatory Action Disclosure Reporting Page (DRP) or a Civil Judicial Action DRP or a Criminal Action DRP.

The summary statistics of the Form ADV data are shown in Panel A of Table 5. The total number of funds matched with Form ADV filings is $9,789$.

5.2 Deceptive behavior

In order to examine the relationship between deceptive behavior and text characteristics, we run the following probit estimations of violations reported in Item 11

[11] See https://www.sec.gov/help/foiadocsinvafoiahtm.html.

of From ADV on lagged text characteristics and a set of control variables.

$$\Pr(\text{Disclosures})_{i,t} = \gamma_0 + \gamma_1 \text{Text Characteristics}_{i,t-1} + \gamma_2 \text{Time-varying controls}_{i,t-1} + \gamma_3 \text{Time-invariant controls}_i + \varepsilon_{i,t},$$

where the indicator variable (Disclosures) takes a value of one if the hedge fund advisor (or any of its related persons) has had legal or regulatory violations during the last 10 years, and zero otherwise. Text characteristics$_{i,t-1}$ are combinations of the lagged text characteristics. The text characteristics are winsorized at 1% and 99% percentiles and scaled to zero mean and unit standard deviation within broad styles. Time-varying controls are lagged assets under management and lagged fund age. The time-invariant controls include share restrictions, compensation structure variables and leverage indicator. All of the specifications include both the time fixed effects and style fixed effects, and the standard errors are clustered by fund.

Table 5 shows the results of probit estimations of Form ADV violations on text characteristics. The independent variable is regulatory action disclosures indicator in Panel B, and civil judicial or criminal action disclosures indicator in Panel C. The results show that funds whose strategies are more lexically diverse report fewer legal problems, especially regulatory actions against the fund by a domestic or foreign government agency. Based on these and our previous section results, it appears that lexical diversity predicts both outperformance as well as less deceptive behavior. We also observe that funds whose strategy descriptions are complex in the sense of syntax structure have in fact more deceptive behavior. It appears that text sophistication, as measured by syntactic complexity, is associated with deception, in line with what Moffitt and Burns (2009) or Vrij et al. (2010) suggest.

6 Investors and creditors response

The results that text characteristics associated with either outperformance or with fewer legal legal problems for the fund and its employees are not important as long as investors or creditors in hedge funds are unable to distinguish these funds and allocate their capital accordingly. In order to establish whether investors recognize and reward (or punish) text characteristics, in this section we start by examining the relationship between the strategy description characteristics and fund flows. For that purpose, we run the following panel regression, in which hedge funds' annual flows are explained by lagged text characteristics while simultaneously controlling for the role of other fund variables.

$$\begin{aligned}\text{Flow}_{i,t} = \gamma_0 &+ \gamma_1 \text{Text Characteristics}_{i,t-1} \\ &+ \gamma_2 \text{Low rank}_{i,t-1} + \gamma_3 \text{Mid rank}_{i,t-1} + \gamma_4 \text{High rank}_{i,t-1} \\ &+ \gamma_5 \text{Time-varying controls}_{i,t-1} + \gamma_6 \text{Time-invariant controls}_i + \varepsilon_{i,t}\end{aligned}$$

where low, mid and high rank are the past-year performance ranks of the funds as in Sirri and Tufano (1998). The text characteristics are winsorized at 1% and 99% percentiles and scaled to zero mean and unit standard deviation within broad styles. The time-varying controls include the lagged age, lagged assets under management and lagged flows, while the time-invariant controls include the fund characteristics reported in Table 1. The estimations also include style and time fixed effects, and standard errors are clustered by fund.

Table 6 shows that the coefficients for the lexical diversity and text length are positive and statistically significant in standalone specifications. Coefficient

for syntactic complexity is statistically indistinguishable from zero and becomes negative and statistically significant in models including lexical diversity or text length. Hence, our results suggest that investors respond positively to lexically diverse and longer strategy texts, but investor flows do not respond to syntactic complexity. It is quite surprising to provide evidence that text characteristics are important determinants of fund flows even after controlling for past performance, since prior literature has documented that operational risk factors (Brown et al. (2008) and Bollen and Pool (2012)) do not appear to influence to fund flows, while both the quality and quantity of strategy description text is important for investors when they made their hedge fund investment decision.

Finally, we examine creditors' response to strategy description text. Prior literature has shown that financial institutions such as prime brokers are able to distinguish low operational risk funds from high operational risk funds. Hence, it is important to understand better whether the debt investors respond to strategy descriptions. For that purposes, we estimate a cross-sectional regression of fund leverage on text complexity scores at funds inception, including standard controls. The model specification is as follows.

$$\text{Leverage}_i = \gamma_0 + \gamma_1 \text{Text Characteristics}_i + \gamma_2 \text{Fund Characteristics}_i + \varepsilon_i$$

where the text characteristics are winsorized at 1% and 99% percentiles and scaled to zero mean and unit standard deviation within broad styles. Fund characteristics are reported in Table 1. The results are shown in Table 7. Funds with higher lexical diversity have higher leverage. The effect of syntactic complexity is subsumed in

the specification that includes all complexity measures.

The interpretation of these results is consistent with the idea that hedge fund investors and creditors are sophisticated and are able to understand (and show preference for) the funds whose strategy descriptions show both high quantity and quality of text, but they are indifferent to syntactic complexity of text. In turn, as we showed in the previous sections, that especially funds with high lexical diversity tend to subsequently outperform and also have lower risk as well as fewer legal problems.

7 Robustness checks

To ensure that out results are robust, we conduct our analysis separately using two data subsamples. First, in order to control for possible language effects, we re-run the analysis presented in previous sections using a subset of funds whose management firm is domiciled in English-speaking countries. The rationale is that managers whose native language is not English may write less sophisticated strategy description texts than their English-speaking peers. Although the number of funds reduces from $21,379$ funds to $15,277$ funds, the results summarized in Table 8 provide consistent evidence that text characteristics are associated with performance and legal problems in the same way as we document in our baseline analysis.

Second, we re-run our whole empirical analysis using only the post-2007 data, for which we can capture the changes in strategy description texts. The post-2007 sample is based on snapshots and, therefore, it does not suffer from a backfill bias or look-ahead bias. For each of the snapshots we have the time stamp for both fund

returns and strategy description texts. Hence, this test is a perfect way control for two potential issues that can contaminate our empirical analysis. The post-2007 subsample contains 16,992 funds suggesting that for the majority of funds we have very accurate information even in our base sample. Table 8 shows that the main conclusions are unchanged even for that post-2007 subsample.

8 Conclusions

When a hedge fund manager writes something for investors, besides conveying quantitative information or economic insights that writing sample also reveals other characteristics via writing style and the level of text sophistication. As the financial economics literature points out, a more sophisticated text may signal fund quality, while as the psychology literature points out, more complexity is related to the intention to deceive.

Our study shows that managers unambiguously use a single measure of text sophistication to write strategy descriptions that solely signal quality. Namely, we document that hedge funds whose strategy descriptions are lexically more diverse outperform, are more likely to survive and report having fewer legal problems.

Funds of lower quality may also prefer to signal that they too have investment talent, and write strategy descriptions that are complex. However, the complexity exhibited by this latter type of fund consists more of convoluted, difficult to read sentences that confuse the reader about the nature of the strategies described. These funds in turn have more legal problems - consistent with their managers attempting to deceive investors.

Since written information volunteered by hedge funds, such as that contained

in strategy descriptions, is not audited or regulated, our study's findings caution against investors putting too much weight on written materials from hedge funds, as these materials could either be unadulterated communiques correlated with the manager's skill or construed texts designed to deceive.

References

Agarwal, V., Daniel, N. D., & Naik, N. Y. (2009). Role of managerial incentives and discretion in hedge fund performance. *The Journal of Finance*, *64*(5), 2221–2256.

Agarwal, V., Gay, G. D., & Ling, L. (2014). Window Dressing in Mutual Funds. *The Review of Financial Studies*, *27*(11), 3133–3170.

Bollen, N. P., Joenväärä, J., & Kauppila, M. (2018). Picking winners? selecting hedge funds for a diversified portfolio. Retrieved from https://ssrn.com/abstract=3034283

Bollen, N. P., & Pool, V. K. (2012). Suspicious patterns in hedge fund returns and the risk of fraud. *The Review of Financial Studies*, *25*(9), 2673–2702.

Brown, S., & Goetzmann, W. N. (2003). Hedge funds with style. *The Journal of Portfolio Management*, *29*(2), 101–112.

Brown, S., Goetzmann, W., Liang, B., & Schwarz, C. (2008). Mandatory disclosure and operational risk: Evidence from hedge fund registration. *The Journal of Finance*, *63*(6), 2785–2815.

Brown, S., Goetzmann, W., Liang, B., & Schwarz, C. (2009). Estimating operational risk for hedge funds: The ω-score. *Financial Analysts Journal*, *65*(1), 43–53.

Buehlmaier, M. M., & Whited, T. M. (2018). Are financial constraints priced? evidence from textual analysis. *The Review of Financial Studies*, *31*(7), 2693–2728.

Dimmock, S. G., & Gerken, W. C. (2012). Predicting fraud by investment managers. *Journal of Financial Economics*, *105*(1), 153–173.

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research*, *58*(3), 840–852.

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 157–214.

Fung, W., & Hsieh, D. A. (1997). Empirical characteristics of dynamic trading strategies: The case of hedge funds. *The Review of Financial Studies*, *10*(2), 275–302.

Fung, W., & Hsieh, D. A. (2001). The risk in hedge fund strategies: Theory and evidence from trend followers. *The Review of Financial Studies*, *14*(2), 313–341.

Fung, W., & Hsieh, D. A. (2004). Hedge fund benchmarks: A risk-based approach. *Financial Analysts Journal*, *60*(5), 65–80.

Goetzmann, W., Ingersoll, J., Spiegel, M., & Welch, I. (2007). Portfolio performance manipulation and manipulation-proof performance measures. *Review of Financial Studies*, *20*(5), 1503–1546.

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

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, *63*, 87–106.

Joenväärä, J., Kosowski, R., & Tolonen, P. (2016). Hedge fund performance: What do we know? Retrieved from https://ssrn.com/abstract=1989410

Joenväärä, J., & Tiu, C. I. (2017). Hedge fund flows and name gravitas. Retrieved from https://ssrn.com/abstract=2939028

Kosowski, R., Naik, N. Y., & Teo, M. (2007). Do hedge funds deliver alpha? a bayesian and bootstrap analysis. *Journal of Financial Economics*, *84*(1), 229–264.

Li, H., Zhang, X., & Zhao, R. (2011). Investing in talents: Manager characteristics and hedge fund performances. *Journal of Financial and Quantitative Analysis*, *46*(1), 59–82.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65.

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, *69*(4), 1643–1671.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*(4), 1187–1230.

Meucci, A. (2009). Managing diversification. *Risk*, *22*(5), 74.

Moffitt, K., & Burns, M. B. (2009). What does that mean? investigating obfuscation and readability cues as indicators of deception in fraudulent financial reports. *AMCIS 2009 Proceedings*, 399.

Napolitano, D., Sheehan, K., & Mundkowsky, R. (2015). Online readability and text complexity analysis with textevaluator. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Demonstrations* (pp. 96–100).

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423 and 623–656.

Sheehan, K. M. (2015). Aligning TextEvaluator® scores with the accelerated text complexity guidelines specified in the common core state standards. *ETS Research Report Series*, *2015*(2), 1–20. doi:10.1002/ets2.12068

Sheehan, K. M. (2016). A review of evidence presented in support of three key claims in the validity argument for the textevaluator® text analysis tool. *ETS Research Report Series*, *2016*(1), 1–15.

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

Stein, J. C. (2005). Why are most funds open-end? competition and the limits of arbitrage. *The Quarterly Journal of Economics*, *120*(1), 247–272.

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

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, *32*(5), 323–352.

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological science in the public interest*, *11*(3), 89–121.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, *104*(5), 444–466.

Table 1: Summary statistics

This table presents summary statistics of fund characteristics. Panel A (and respectively, Panel B) presents time-varying (time-invariant) fund characteristics. "N" is number of annual observations (that are time-varying) or number of funds (that are time-invariant). "Mean" ("Std") is the cross-sectional average (standard deviation) of a particular fund characteristic. "Median" is the 50th percentile of that characteristic. "R" is the annual average of a hedge fund's simple return. "Flow" is the fund's annual flow measured in percentage of total assets. "AUM" is the fund's asset under management, measured in millions U.S. dollars. "Age" is the fund's age from inception, measured in years. "AliveDead" refers to the portion of funds that are still reporting at the period's end. "HighWaterMark" indicates the percentage of funds imposing a high-water mark provision. "ManagementFee" gives the management fee charged by the funds, while "IncentiveFee" is the performance-based fee charged by the funds. "LeverageDummy" is an indicator variable that takes on value of 1 if the fund applies leverage or 0 otherwise. "LockupDummy" is an indicator variable that takes on value of 1 if the fund has a lockup period or 0 otherwise. "Restriction" is defined as the sum of redemption period and notice period. "MinimumInvestment" is the minimum amount of money in U.S. dollars that has to be invested in the fund. "Offshore" is an indicator variable that is 1 if the fund is domiciled in offshore location and 0 otherwise. Panel C presents the time-series average of text characteristics by year. "TextLength" is the logarithm of the length of the strategy description text in characters. "LexicalDiversity" is the Shannon index calculated based on the frequencies of words in the fund's strategy description. "SyntacticComplexity" is calculated using TextEvaluator. The shaded area corresponds to one standard deviation. Panel D presents the descriptive statistics for the text characteristics and Panel E shows the correlations. Panel F presents the descriptive statistics for the text characteristics when funds are grouped by style, firm domicile and fund domicile.

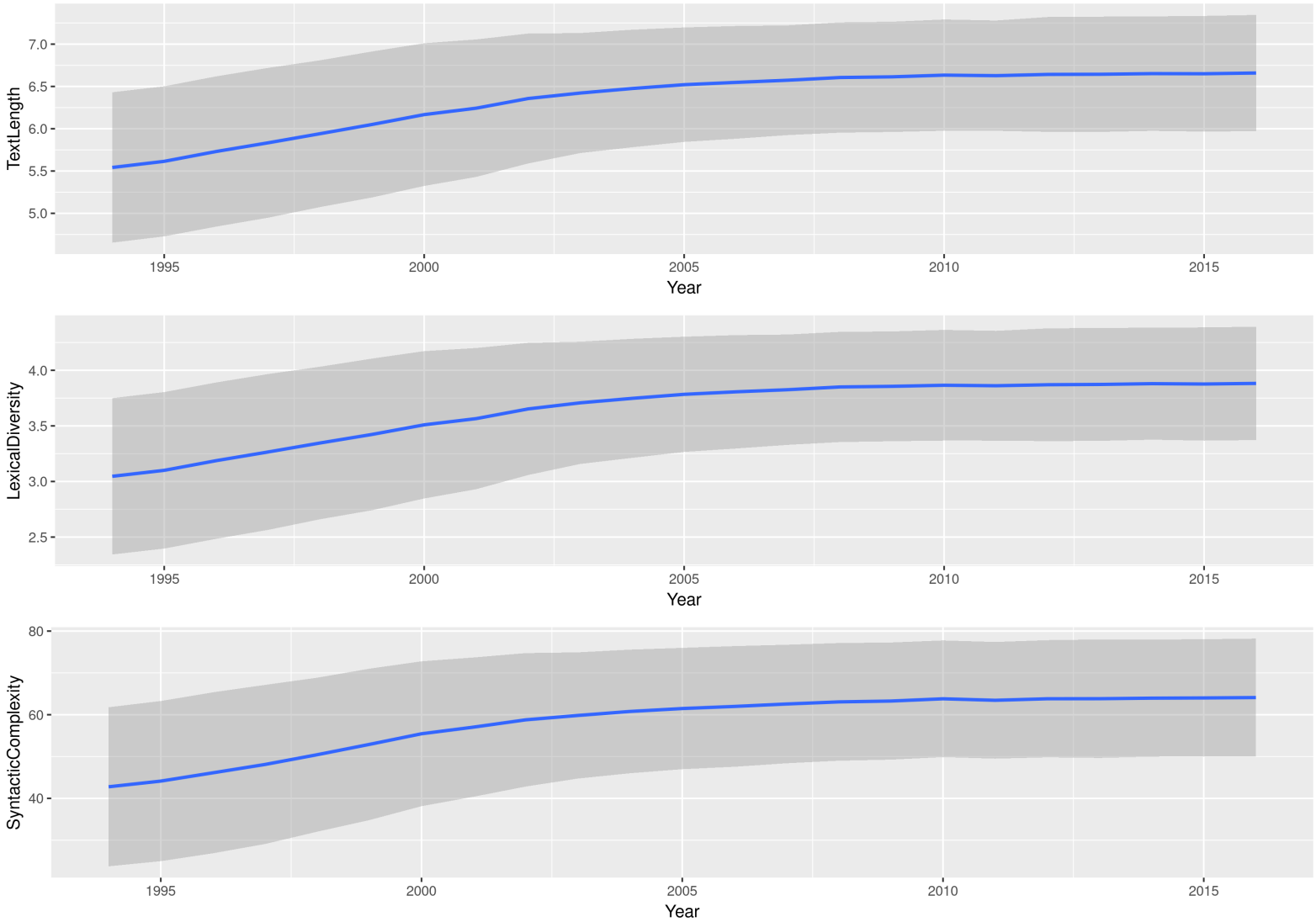
Panel A: Time-varying fund characteristics

Variable	N	Mean	Median	Std
Age	121618	5.27	3.91	4.75
AUM	108833	141.09	23.00	367.89
Flow	79542	0.43	0.01	1.63
R	106054	0.10	0.07	0.33

Panel B: Time-invariant fund characteristics

Variable	N	Mean	Median	Std
AliveDead	21379	0.32	0.00	0.47
HighWaterMark	19807	0.70	1.00	0.46
IncentiveFee	20760	17.31	20.00	7.13
LeverageDummy	18955	0.47	0.00	0.50
LockupDummy	17777	0.24	0.00	0.42
ManagementFee	20928	1.52	1.50	0.79
MinimumInvestment	18859	8.01	0.50	111.00
Offshore	20534	0.32	0.00	0.47
Restriction	16869	83.81	60.42	90.54

Panel C: Average text characteristics by year



Panel D: Descriptive statistics of text characteristics

Measure	N	Mean	Median	Std
LexicalDiversity	120430	3.76	3.82	0.58
SyntacticComplexity	119262	60.95	63.00	15.76
TextLength	120430	6.49	6.57	0.75

Panel E: Correlations of text characteristics

	TextLength	LexicalDiversity	SyntacticComplexity
TextLength	1.00	0.97	0.62
LexicalDiversity		1.00	0.58
SyntacticComplexity			1.00

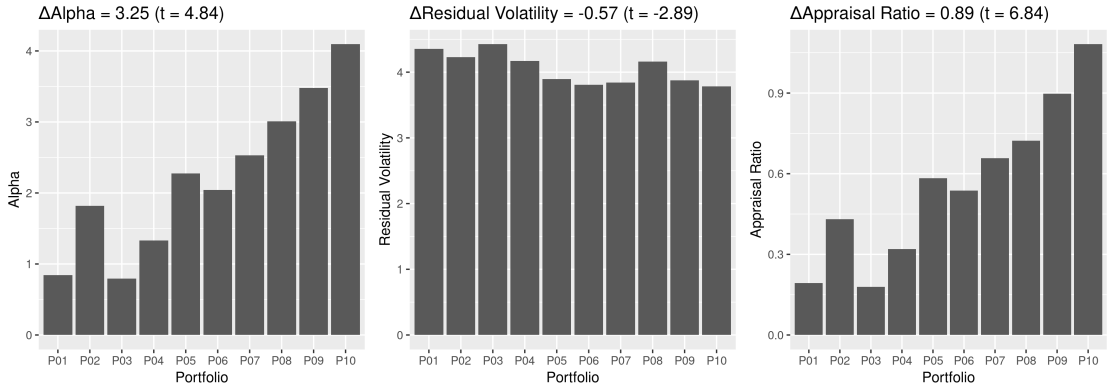
Panel F: Descriptive statistics of text characteristics within groups

Measure	CoarseStyle	N	Mean	Median	Std
LexicalDiversity	DIRECTIONAL TRADERS	39449	3.75	3.80	0.55
LexicalDiversity	MULTI-PROCESS	25643	3.75	3.83	0.58
LexicalDiversity	RELATIVE VALUE	20156	3.80	3.85	0.55
LexicalDiversity	SECURITY SELECTION	35182	3.75	3.83	0.60
SyntacticComplexity	DIRECTIONAL TRADERS	39159	60.78	63.00	15.25
SyntacticComplexity	MULTI-PROCESS	25426	60.49	63.00	16.35
SyntacticComplexity	RELATIVE VALUE	19907	62.89	65.00	15.06
SyntacticComplexity	SECURITY SELECTION	34770	60.35	63.00	16.18
TextLength	DIRECTIONAL TRADERS	39449	6.47	6.52	0.73
TextLength	MULTI-PROCESS	25643	6.48	6.56	0.76
TextLength	RELATIVE VALUE	20156	6.56	6.63	0.73
TextLength	SECURITY SELECTION	35182	6.48	6.59	0.79
Measure	FirmDomicile	N	Mean	Median	Std
LexicalDiversity	Canada	3104	3.80	3.84	0.51
LexicalDiversity	Other	34795	3.76	3.81	0.53
LexicalDiversity	UK	16450	3.73	3.80	0.56
LexicalDiversity	US	66081	3.76	3.84	0.60
SyntacticComplexity	Canada	3091	63.77	65.00	14.58
SyntacticComplexity	Other	34376	61.47	63.00	14.76
SyntacticComplexity	UK	16309	62.27	64.00	15.42
SyntacticComplexity	US	65486	60.21	63.00	16.34
TextLength	Canada	3104	6.57	6.62	0.67
TextLength	Other	34795	6.48	6.54	0.69
TextLength	UK	16450	6.46	6.53	0.74
TextLength	US	66081	6.50	6.58	0.79
Measure	FundDomicile	N	Mean	Median	Std
LexicalDiversity	CARIBBEAN	38246	3.77	3.84	0.57
LexicalDiversity	EUROPE	22361	3.77	3.79	0.50
LexicalDiversity	NORTH AMERICA	50460	3.76	3.84	0.60
LexicalDiversity	OTHERS	4491	3.84	3.88	0.52
SyntacticComplexity	CARIBBEAN	37946	61.37	63.00	15.36
SyntacticComplexity	EUROPE	22090	62.53	64.00	14.51
SyntacticComplexity	NORTH AMERICA	50042	60.28	63.00	16.38
SyntacticComplexity	OTHERS	4425	61.99	63.00	13.81
TextLength	CARIBBEAN	38246	6.50	6.59	0.73
TextLength	EUROPE	22361	6.50	6.54	0.67
TextLength	NORTH AMERICA	50460	6.50	6.57	0.79
TextLength	OTHERS	4491	6.58	6.61	0.69

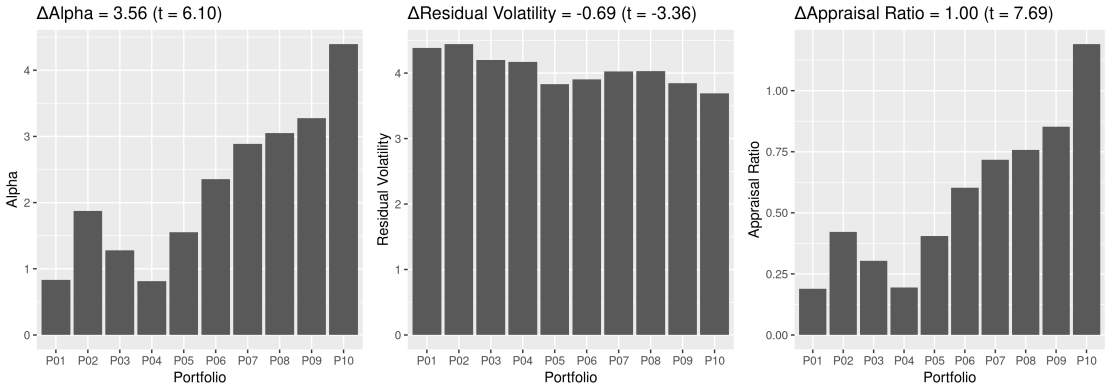
Table 2: Portfolio sorts

This table presents out-of-sample performance results from portfolios sorted by various ways. Funds are sorted at end of each year into decile portfolios and equally-weighted returns are evaluated for the following year. Panel A (and respectively, Panels B and C) show the Fung-Hsieh 7-factor model alphas, residual volatilities and appraisal ratios for portfolios sorted by Text Length (and respectively, Lexical Diversity and Syntactic Complexity). The text characteristics are winsorized at 1% and 99% percentiles and scaled to zero mean and unit standard deviation within styles. We also test the differences in alphas, residual volatilities and appraisal ratios for the extreme portfolios. Panel D (and respectively, Panel E) shows results from double sorts first by Text Length and then by Lexical Diversity (in Panel E funds are first sorted by Lexical Diversity and then by TextLength). Panel F (and respectively, Panels G and H) shows the Fung-Hsieh 7-factor model estimation results for the decile portfolios sorted by Text Length (and respectively, Lexical Diversity and Syntactic Complexity) and the spread portfolio. Values for "Alpha" are defined as the annualized intercept of the regression model. The seven factors are: the S&P 500 return minus the risk-free rate (SP); returns on the Russell 2000 index minus the S&P 500 index return (SCLC); excess return on 10-year US Treasury bonds (CGS10); the yield spread between 10-year T-bonds and Moody's Baa-rated bonds (CREDSPR); and the so-called primitive trend-following strategy for bonds (PTFSBD), currency (PTFSFX), and commodities (PTFSCOM). "Adjusted R^2" shows the adjusted R-squared of the regression. To correct for backfilling bias, the first 12 months of each fund's returns are removed.

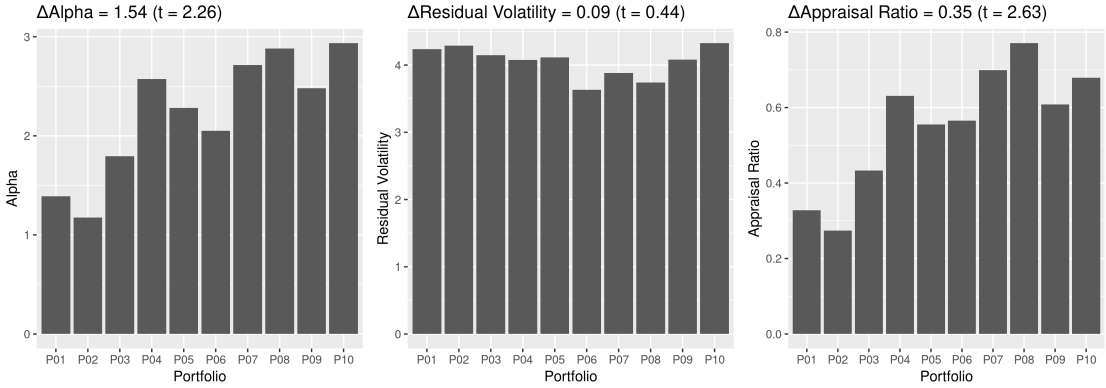
Panel A: Sorts by TextLength



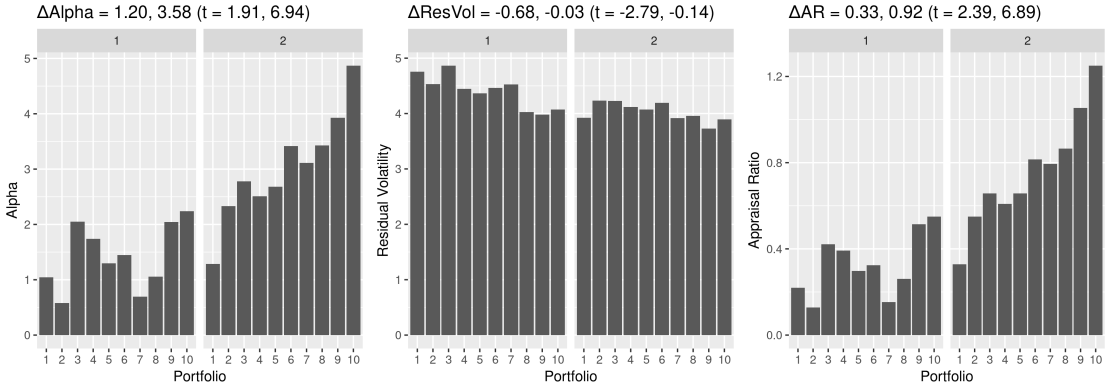
Panel B: Sorts by LexicalDiversity



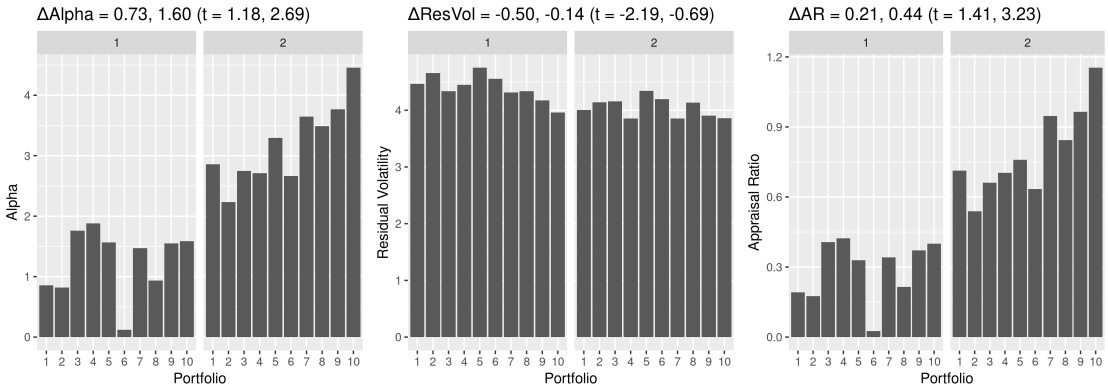
Panel C: Sorts by SyntacticComplexity



Panel D: Double Sort by TextLength and LexicalDiversity



Panel E: Double sort by LexicalDiversity and TextLength



Panel F: Factor model estimations for portfolios sorted by Text Length

Portfolio	Alpha	SP	SCLC	CGS10	Credspr	PTFSBD	PTFSFX	PTFSCOM	Adjusted R^2
Bottom	0.84	0.33	0.18	0.11	0.29	-0.01	0.01	0.01	0.69
2	1.82	0.31	0.15	0.08	0.29	-0.01	0.01	0.00	0.67
3	0.79	0.32	0.15	0.09	0.29	-0.01	0.01	0.01	0.65
4	1.33	0.30	0.18	0.12	0.27	-0.01	0.01	0.01	0.66
5	2.27	0.27	0.15	0.08	0.26	-0.01	0.01	0.01	0.65
6	2.04	0.28	0.13	0.09	0.27	-0.01	0.01	0.00	0.67
7	2.53	0.27	0.12	0.08	0.27	-0.01	0.01	0.01	0.65
8	3.01	0.29	0.15	0.09	0.28	-0.01	0.01	0.01	0.65
9	3.48	0.31	0.18	0.07	0.23	-0.01	0.01	0.00	0.70
Top	4.09	0.30	0.16	0.08	0.24	-0.01	0.01	0.00	0.69
Top - Bottom	3.25	-0.04	-0.02	-0.04	-0.06	0.00	-0.00	-0.00	0.12
tHAC	4.84	-2.28	-1.10	-1.71	-2.23	1.38	-0.30	-1.37	

Panel G: Factor model estimations for portfolios sorted by Lexical Diversity

Portfolio	Alpha	SP	SCLC	CGS10	Credspr	PTFSBD	PTFSFX	PTFSCOM	Adjusted R^2
Bottom	0.83	0.34	0.18	0.11	0.28	-0.02	0.01	0.01	0.69
2	1.88	0.31	0.15	0.09	0.32	-0.00	0.01	0.01	0.65
3	1.28	0.31	0.15	0.09	0.28	-0.01	0.01	0.01	0.66
4	0.81	0.33	0.16	0.10	0.27	-0.01	0.01	0.01	0.68
5	1.55	0.28	0.16	0.07	0.28	-0.01	0.01	0.01	0.68
6	2.35	0.28	0.15	0.10	0.29	-0.01	0.01	0.00	0.66
7	2.89	0.27	0.13	0.09	0.26	-0.01	0.01	0.01	0.63
8	3.05	0.28	0.16	0.08	0.24	-0.00	0.01	0.01	0.64
9	3.27	0.30	0.16	0.06	0.24	-0.01	0.01	0.00	0.69
Top	4.39	0.29	0.16	0.07	0.22	-0.01	0.01	0.01	0.69
Top - Bottom	3.56	-0.05	-0.02	-0.04	-0.06	0.01	-0.00	-0.00	0.17
tHAC	6.10	-3.40	-1.43	-2.18	-2.62	1.98	-0.64	-0.83	

Panel H: Factor model estimations for portfolios sorted by Syntactic Complexity

Portfolio	Alpha	SP	SCLC	CGS10	Credspr	PTFSBD	PTFSFX	PTFSCOM	Adjusted R^2
Bottom	1.39	0.28	0.17	0.11	0.28	-0.00	0.01	0.01	0.63
2	1.17	0.31	0.14	0.09	0.29	-0.01	0.01	0.01	0.65
3	1.79	0.29	0.17	0.09	0.23	-0.01	0.01	0.01	0.63
4	2.57	0.29	0.14	0.08	0.23	-0.01	0.01	0.00	0.63
5	2.28	0.28	0.14	0.10	0.29	-0.01	0.01	0.01	0.63
6	2.05	0.30	0.14	0.08	0.26	-0.01	0.01	0.00	0.71
7	2.71	0.29	0.15	0.08	0.25	-0.01	0.01	0.01	0.67
8	2.88	0.30	0.15	0.08	0.28	-0.01	0.01	0.00	0.71
9	2.48	0.32	0.18	0.09	0.29	-0.01	0.01	0.01	0.70
Top	2.93	0.32	0.17	0.07	0.30	-0.01	0.01	0.00	0.69
Top - Bottom	1.54	0.04	0.00	-0.04	0.02	-0.01	-0.01	-0.01	0.26
tHAC	2.26	2.50	0.14	-2.01	0.86	-1.61	-1.70	-4.61	

Table 3: Fund performance

This table shows multivariate performance panel regressions. The independent variables are Fung-Hsieh 7-factor model Alpha (in Panel A), Residual Volatility (in Panel B), and Appraisal ratio (in Panel C). Main (lagged) explanatory variables are Lexical Diversity, Syntactic Complexity, and Text Length. The text characteristics are winsorized at 1% and 99% percentiles and scaled to zero mean and unit standard deviation within styles. All estimations include the lagged value of independent variable as regressor. Other lagged controls are same as in Table 1. To correct for backfilling bias, we remove the first 12 months of a fund's returns. All regressions include style and time fixed effects, and standard errors clustered by fund are shown in parentheses.

Panel A: Regressions of Fung-Hsieh 7-factor Alpha

	Alpha	Alpha	Alpha	Alpha	Alpha	Alpha	Alpha
LexicalDiversity	0.0047***			0.0055***	0.0089***		0.0087***
	(0.0009)			(0.0011)	(0.0032)		(0.0032)
SyntacticComplexity		0.0015*		−0.0014		−0.0014	−0.0010
		(0.0009)		(0.0010)		(0.0011)	(0.0011)
TextLength			0.0042***		−0.0044	0.0051***	−0.0035
			(0.0009)		(0.0031)	(0.0011)	(0.0033)
Alpha	0.1815***	0.1820***	0.1816***	0.1813***	0.1814***	0.1815***	0.1813***
	(0.0092)	(0.0092)	(0.0092)	(0.0092)	(0.0092)	(0.0092)	(0.0092)
AUM	−0.0000*	−0.0000	−0.0000*	−0.0000*	−0.0000*	−0.0000*	−0.0000*
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Age	−0.0006***	−0.0007***	−0.0006***	−0.0006***	−0.0006***	−0.0006***	−0.0006***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
ManagementFee	0.1657	0.1795	0.1687	0.1813	0.1625	0.1852	0.1785
	(0.1518)	(0.1524)	(0.1516)	(0.1524)	(0.1519)	(0.1522)	(0.1524)
IncentiveFee	0.0944***	0.0974***	0.0958***	0.0928***	0.0935***	0.0943***	0.0922***
	(0.0126)	(0.0126)	(0.0125)	(0.0126)	(0.0126)	(0.0126)	(0.0127)
LockupDummy	0.0074***	0.0078***	0.0075***	0.0074***	0.0074***	0.0075***	0.0074***
	(0.0020)	(0.0020)	(0.0020)	(0.0020)	(0.0020)	(0.0020)	(0.0020)
Restriction	0.0163***	0.0159***	0.0162***	0.0163***	0.0163***	0.0162***	0.0163***
	(0.0033)	(0.0033)	(0.0033)	(0.0033)	(0.0033)	(0.0033)	(0.0033)
LeverageDummy	0.0076***	0.0074***	0.0076***	0.0075***	0.0076***	0.0075***	0.0075***
	(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0016)
Num. obs.	54704	54260	54704	54260	54704	54260	54260

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Panel B: Regressions of Residual Volatility

	ResidualVolatility	ResidualVolatility	ResidualVolatility	ResidualVolatility	ResidualVolatility	ResidualVolatility	ResidualVolatility
LexicalDiversity	−0.0014***			−0.0014***	−0.0026*		−0.0026*
	(0.0004)			(0.0005)	(0.0014)		(0.0014)
SyntacticComplexity		−0.0006		0.0001		0.0001	−0.0000
		(0.0004)		(0.0004)		(0.0005)	(0.0005)
TextLength			−0.0013***		0.0013	−0.0012**	0.0014
			(0.0004)		(0.0013)	(0.0005)	(0.0014)
ResidualVolatility	0.6525***	0.6521***	0.6526***	0.6518***	0.6525***	0.6519***	0.6517***
	(0.0170)	(0.0170)	(0.0170)	(0.0170)	(0.0170)	(0.0170)	(0.0170)
AUM	−0.0000***	−0.0000***	−0.0000***	−0.0000***	−0.0000***	−0.0000***	−0.0000***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Age	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
ManagementFee	0.5321***	0.5390***	0.5311***	0.5388***	0.5330***	0.5378***	0.5400***
	(0.0842)	(0.0844)	(0.0842)	(0.0846)	(0.0841)	(0.0845)	(0.0845)
IncentiveFee	0.0438***	0.0431***	0.0434***	0.0444***	0.0441***	0.0439***	0.0446***
	(0.0056)	(0.0056)	(0.0056)	(0.0056)	(0.0056)	(0.0056)	(0.0057)
LockupDummy	0.0034***	0.0032***	0.0034***	0.0033***	0.0034***	0.0033***	0.0033***
	(0.0010)	(0.0010)	(0.0010)	(0.0010)	(0.0010)	(0.0010)	(0.0010)
Restriction	−0.0065***	−0.0064***	−0.0065***	−0.0065***	−0.0066***	−0.0065***	−0.0065***
	(0.0015)	(0.0015)	(0.0015)	(0.0015)	(0.0015)	(0.0015)	(0.0015)
LeverageDummy	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008
	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)
Num. obs.	54874	54433	54874	54433	54874	54433	54433

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Panel C: Regressions of Appraisal Ratio

	AppraisalRatio	AppraisalRatio	AppraisalRatio	AppraisalRatio	AppraisalRatio	AppraisalRatio	AppraisalRatio
LexicalDiversity	0.0464***			0.0470***	0.0722***		0.0752***
	(0.0069)			(0.0082)	(0.0250)		(0.0252)
SyntacticComplexity		0.0236***		−0.0011		−0.0012	0.0020
		(0.0069)		(0.0080)		(0.0083)	(0.0084)
TextLength			0.0429***		−0.0271	0.0434***	−0.0313
			(0.0070)		(0.0253)	(0.0087)	(0.0266)
AppraisalRatio	0.2500***	0.2504***	0.2502***	0.2495***	0.2500***	0.2497***	0.2494***
	(0.0051)	(0.0052)	(0.0051)	(0.0052)	(0.0051)	(0.0052)	(0.0052)
AUM	−0.0000	−0.0000	−0.0000	−0.0000	−0.0000	−0.0000	−0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Age	−0.0010	−0.0015	−0.0010	−0.0010	−0.0010	−0.0010	−0.0010
	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0013)
ManagementFee	3.7595***	3.8753***	3.7896***	3.8936***	3.7400***	3.9255***	3.8684***
	(1.1611)	(1.1676)	(1.1598)	(1.1656)	(1.1625)	(1.1644)	(1.1674)
IncentiveFee	1.0223***	1.0486***	1.0352***	1.0099***	1.0165***	1.0231***	1.0051***
	(0.0944)	(0.0946)	(0.0943)	(0.0948)	(0.0948)	(0.0945)	(0.0950)
LockupDummy	0.0610***	0.0643***	0.0616***	0.0607***	0.0609***	0.0614***	0.0606***
	(0.0164)	(0.0164)	(0.0164)	(0.0164)	(0.0164)	(0.0164)	(0.0164)
Restriction	0.3097***	0.3068***	0.3088***	0.3103***	0.3100***	0.3096***	0.3104***
	(0.0312)	(0.0314)	(0.0312)	(0.0315)	(0.0313)	(0.0315)	(0.0316)
LeverageDummy	0.0529***	0.0506***	0.0529***	0.0514***	0.0529***	0.0513***	0.0513***
	(0.0130)	(0.0131)	(0.0130)	(0.0131)	(0.0131)	(0.0131)	(0.0131)
Num. obs.	54704	54260	54704	54260	54704	54260	54260

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 4: Performance measure manipulation, tail risk and fund attrition

This table shows multivariate financial risk panel regressions. The independent variables are manipulation-proof performance measure (MPPM, with risk aversion parameter $\rho = 3$) of Goetzmann et al. (2007) (in Panel A), and Maximum Loss (in Panel B). In Panel C the independent variable is attrition (1, if fund stops reporting returns, or 0 otherwise). Main (lagged) explanatory variables are Lexical Diversity, Syntactic Complexity, and Text Length. The text characteristics are winsorized at 1% and 99% percentiles and scaled to zero mean and unit standard deviation within styles. All estimations include the lagged value of independent variable as regressor. Other lagged controls are same as in Table 1. To correct for backfilling bias, we remove the first 12 months of a fund's returns. All regressions include style and time fixed effects. Standard errors clustered by fund are shown in parentheses. Average marginal effects are shown in brackets.

Panel A: Regressions of MPPM ($\rho = 3$)

	MPPM	MPPM	MPPM	MPPM	MPPM	MPPM	MPPM
LexicalDiversity	0.8741***			0.9625***	1.8663***		1.8536***
	(0.1665)			(0.1919)	(0.6089)		(0.6138)
SyntacticComplexity		0.3468**		−0.1575		−0.1399	−0.0588
		(0.1732)		(0.1977)		(0.2021)	(0.2023)
TextLength			0.7685***		−1.0410*	0.8525***	−0.9904
			(0.1652)		(0.6051)	(0.1941)	(0.6229)
MPPM	0.0161	0.0165	0.0163	0.0160	0.0160	0.0161	0.0159
	(0.0160)	(0.0160)	(0.0160)	(0.0160)	(0.0160)	(0.0160)	(0.0160)
AUM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Age	0.0100	−0.0011	0.0083	0.0095	0.0074	0.0083	0.0085
	(0.0284)	(0.0284)	(0.0284)	(0.0287)	(0.0285)	(0.0286)	(0.0287)
ManagementFee	−58.4666	−57.8715	−57.8964	−57.6377	−59.2264	−56.9898	−58.4458
	(41.9526)	(42.1560)	(41.9131)	(42.2127)	(41.7727)	(42.1821)	(42.0350)
IncentiveFee	−8.2684***	−7.6549**	−7.9994**	−8.4794***	−8.4953***	−8.1792**	−8.6336***
	(3.1442)	(3.1630)	(3.1314)	(3.2056)	(3.1948)	(3.1807)	(3.2358)
LockupDummy	0.1086	0.2082	0.1238	0.1329	0.1038	0.1497	0.1312
	(0.3762)	(0.3775)	(0.3766)	(0.3765)	(0.3760)	(0.3772)	(0.3765)
Restriction	3.4826***	3.3823***	3.4653***	3.4497***	3.4946***	3.4342***	3.4518***
	(0.6480)	(0.6547)	(0.6467)	(0.6568)	(0.6505)	(0.6551)	(0.6577)
LeverageDummy	0.5065	0.4938	0.5057	0.5092	0.5054	0.5076	0.5073
	(0.3257)	(0.3282)	(0.3259)	(0.3278)	(0.3260)	(0.3281)	(0.3282)
Num. obs.	54704	54260	54704	54260	54704	54260	54260

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Panel B: Regressions of Maximum Loss

	MaxLoss	MaxLoss	MaxLoss	MaxLoss	MaxLoss	MaxLoss	MaxLoss
LexicalDiversity	−0.0015***			−0.0015***	−0.0028**		−0.0028**
	(0.0003)			(0.0004)	(0.0011)		(0.0011)
SyntacticComplexity		−0.0007**		0.0001		0.0000	−0.0001
		(0.0003)		(0.0004)		(0.0004)	(0.0004)
TextLength			−0.0013***		0.0014	−0.0013***	0.0015
			(0.0003)		(0.0011)	(0.0004)	(0.0012)
MaxLoss	0.4214***	0.4220***	0.4217***	0.4212***	0.4213***	0.4215***	0.4211***
	(0.0082)	(0.0082)	(0.0082)	(0.0082)	(0.0082)	(0.0082)	(0.0082)
AUM	−0.0000*	−0.0000*	−0.0000*	−0.0000*	−0.0000*	−0.0000*	−0.0000*
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Age	−0.0001*	−0.0001*	−0.0001*	−0.0001**	−0.0001*	−0.0001*	−0.0001*
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
ManagementFee	0.3128***	0.3151***	0.3117***	0.3151***	0.3139***	0.3140***	0.3164***
	(0.0650)	(0.0651)	(0.0650)	(0.0653)	(0.0650)	(0.0652)	(0.0653)
IncentiveFee	0.0020	0.0010	0.0015	0.0022	0.0023	0.0018	0.0025
	(0.0047)	(0.0047)	(0.0047)	(0.0047)	(0.0047)	(0.0047)	(0.0047)
LockupDummy	0.0025***	0.0023***	0.0025***	0.0024***	0.0025***	0.0024***	0.0024***
	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
Restriction	−0.0062***	−0.0060***	−0.0062***	−0.0061***	−0.0062***	−0.0061***	−0.0061***
	(0.0013)	(0.0014)	(0.0013)	(0.0014)	(0.0013)	(0.0014)	(0.0014)
LeverageDummy	−0.0010*	−0.0010	−0.0010*	−0.0010*	−0.0010*	−0.0010*	−0.0010*
	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Num. obs.	54704	54260	54704	54260	54704	54260	54260

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Panel C: Attrition

	Attrition	Attrition	Attrition	Attrition	Attrition	Attrition	Attrition
LexicalDiversity	−0.0960***			−0.0868***	−0.0992***		−0.1037***
	(0.0077)			(0.0090)	(0.0268)		(0.0272)
	[−0.0184]			[−0.0167]	[−0.0190]		[−0.0199]
SyntacticComplexity		−0.0575***		−0.0140		−0.0117	−0.0158*
		(0.0077)		(0.0088)		(0.0091)	(0.0092)
		[−0.0111]		[−0.0027]		[−0.0022]	[−0.0030]
TextLength			−0.0920***		0.0034	−0.0833***	0.0187
			(0.0077)		(0.0269)	(0.0093)	(0.0283)
			[−0.0176]		[0.0007]	[−0.0160]	[0.0036]
LowRank	−1.2105***	−1.2392***	−1.2123***	−1.2214***	−1.2105***	−1.2230***	−1.2215***
	(0.1628)	(0.1634)	(0.1628)	(0.1635)	(0.1628)	(0.1634)	(0.1635)
	[−0.2320]	[−0.2386]	[−0.2324]	[−0.2346]	[−0.2320]	[−0.2350]	[−0.2346]
MidRank	−0.7162***	−0.7157***	−0.7166***	−0.7119***	−0.7162***	−0.7123***	−0.7119***
	(0.0438)	(0.0439)	(0.0438)	(0.0439)	(0.0438)	(0.0439)	(0.0439)
	[−0.1373]	[−0.1378]	[−0.1374]	[−0.1367]	[−0.1373]	[−0.1368]	[−0.1367]
HighRank	−0.1150	−0.1040	−0.1150	−0.1308	−0.1148	−0.1303	−0.1302
	(0.2039)	(0.2043)	(0.2038)	(0.2045)	(0.2039)	(0.2045)	(0.2045)
	[−0.0220]	[−0.0200]	[−0.0220]	[−0.0251]	[−0.0220]	[−0.0250]	[−0.0250]
AUM	−0.0005***	−0.0006***	−0.0005***	−0.0005***	−0.0005***	−0.0005***	−0.0006***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
	[−0.0001]	[−0.0001]	[−0.0001]	[−0.0001]	[−0.0001]	[−0.0001]	[−0.0001]
Age	−0.0245***	−0.0236***	−0.0244***	−0.0248***	−0.0245***	−0.0247***	−0.0248***
	(0.0017)	(0.0017)	(0.0017)	(0.0017)	(0.0017)	(0.0017)	(0.0017)
	[−0.0047]	[−0.0045]	[−0.0047]	[−0.0048]	[−0.0047]	[−0.0047]	[−0.0048]
Flow	−0.0548***	−0.0551***	−0.0548***	−0.0545***	−0.0548***	−0.0545***	−0.0545***
	(0.0078)	(0.0078)	(0.0078)	(0.0078)	(0.0078)	(0.0078)	(0.0078)
	[−0.0105]	[−0.0106]	[−0.0105]	[−0.0105]	[−0.0105]	[−0.0105]	[−0.0105]

Continued on next page

	Attrition	Attrition	Attrition	Attrition	Attrition	Attrition	Attrition
HighWaterMark	-0.1414^{***}	-0.1553^{***}	-0.1445^{***}	-0.1397^{***}	-0.1413^{***}	-0.1424^{***}	-0.1396^{***}
	(0.0252)	(0.0252)	(0.0252)	(0.0253)	(0.0252)	(0.0253)	(0.0254)
	[−0.0271]	[−0.0299]	[−0.0277]	[−0.0268]	[−0.0271]	[−0.0274]	[−0.0268]
ManagementFee	5.9021^{***}	5.6899^{***}	5.8014^{***}	5.8236^{***}	5.9054^{***}	5.7297^{***}	5.8405^{***}
	(1.3841)	(1.3911)	(1.3855)	(1.3874)	(1.3838)	(1.3886)	(1.3870)
	[1.1312]	[1.0955]	[1.1123]	[1.1184]	[1.1319]	[1.1008]	[1.1217]
IncentiveFee	1.7117^{***}	1.6121^{***}	1.6921^{***}	1.6735^{***}	1.7122^{***}	1.6554^{***}	1.6758^{***}
	(0.1457)	(0.1459)	(0.1456)	(0.1460)	(0.1459)	(0.1458)	(0.1461)
	[0.3281]	[0.3104]	[0.3244]	[0.3214]	[0.3282]	[0.3180]	[0.3218]
LockupDummy	0.0424^{**}	0.0375^{**}	0.0412^{**}	0.0427^{**}	0.0425^{**}	0.0414^{**}	0.0428^{**}
	(0.0175)	(0.0177)	(0.0176)	(0.0176)	(0.0175)	(0.0176)	(0.0176)
	[0.0081]	[0.0072]	[0.0079]	[0.0082]	[0.0081]	[0.0080]	[0.0082]
Restriction	0.1199^{***}	0.1240^{***}	0.1231^{***}	0.1153^{***}	0.1198^{***}	0.1177^{***}	0.1150^{***}
	(0.0323)	(0.0330)	(0.0323)	(0.0327)	(0.0323)	(0.0327)	(0.0327)
	[0.0230]	[0.0239]	[0.0236]	[0.0221]	[0.0230]	[0.0226]	[0.0221]
LeverageDummy	0.0662^{***}	0.0687^{***}	0.0667^{***}	0.0677^{***}	0.0662^{***}	0.0681^{***}	0.0676^{***}
	(0.0144)	(0.0145)	(0.0144)	(0.0145)	(0.0144)	(0.0145)	(0.0145)
	[0.0127]	[0.0132]	[0.0128]	[0.0130]	[0.0127]	[0.0131]	[0.0130]
(Intercept)	-1.7216^{***}	-1.6294^{***}	-1.7083^{***}	-1.7149^{***}	-1.7216^{***}	-1.7005^{***}	-1.7156^{***}
	(0.1169)	(0.1166)	(0.1168)	(0.1169)	(0.1169)	(0.1169)	(0.1169)
Num. obs.	59283	58816	59283	58816	59283	58816	58816

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 5: Disciplinary disclosures

This table presents results from regressions in which annual Form ADV violations flags are regressed on lagged text characteristics and a set of lagged control variables. Panel A presents the summary statistics for the data set used for Form ADV violations analysis. Total number of funds matched with Form ADV filings is 9,789. The table shows the total number of funds in the sample for each year, the number funds that were required to complete a Regulatory Action Disclosure Reporting Page (DRP) (funds whose management company answered "Yes" to any of the questions in Items 11.C–11.G of Form ADV), a Civil Judicial Action DRP (Item 11.H) or a Criminal Action DRP (Items 11.A–11.B). The independent variable in Panel B is regulatory action disclosures flag, and in Panel C civil judicial action or criminal action disclosures flag. All regressions include style and time fixed effects. Standard errors clustered by fund are shown in parentheses. Average marginal effects are shown in brackets.

Panel A: Summary statistics of Form ADV data

Year	Total	Regulatory	CivilOrCriminal
2000	2810	5	3
2001	3102	139	29
2002	3397	143	39
2003	3893	147	56
2004	5108	218	72
2005	5818	295	76
2006	6573	336	103
2007	7236	415	175
2008	7740	438	195
2009	8117	477	231
2010	7911	486	220
2011	9023	570	272
2012	9262	917	479
2013	9361	954	434
2014	8885	1041	448
2015	8596	1023	412
2016	8056	992	372
2017	7192	833	301

Panel B: Regulatory action disclosures

	RegAct	RegAct	RegAct	RegAct	RegAct	RegAct	RegAct
LexicalDiversity	−0.0364*			−0.0868***	−0.3438***		−0.3098***
	(0.0215)			(0.0250)	(0.0674)		(0.0684)
	[−0.0098]			[−0.0234]	[−0.0925]		[−0.0833]
SyntacticComplexity		0.0700***		0.1108***		0.1000***	0.0843***
		(0.0209)		(0.0234)		(0.0244)	(0.0248)
		[0.0189]		[0.0299]		[0.0270]	[0.0227]
TextLength			−0.0055		0.3144***	−0.0550**	0.2407***
			(0.0216)		(0.0670)	(0.0260)	(0.0710)
			[−0.0015]		[0.0845]	[−0.0148]	[0.0647]
AUM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
Age	0.0027	0.0058	0.0036	0.0046	0.0037	0.0051	0.0049
	(0.0037)	(0.0037)	(0.0037)	(0.0037)	(0.0037)	(0.0037)	(0.0037)
	[0.0007]	[0.0016]	[0.0010]	[0.0013]	[0.0010]	[0.0014]	[0.0013]
ManagementFee	−9.9795***	−9.5306**	−9.9298***	−9.3923**	−9.5611**	−9.4889**	−9.2345**
	(3.7656)	(3.8187)	(3.7706)	(3.8233)	(3.7738)	(3.8225)	(3.8184)
	[−2.6942]	[−2.5746]	[−2.6824]	[−2.5306]	[−2.5715]	[−2.5605]	[−2.4832]
IncentiveFee	−0.7365***	−0.7898***	−0.7677***	−0.7176**	−0.6356**	−0.7638***	−0.6463**
	(0.2838)	(0.2817)	(0.2822)	(0.2862)	(0.2836)	(0.2845)	(0.2855)
	[−0.1988]	[−0.2133]	[−0.2074]	[−0.1934]	[−0.1710]	[−0.2061]	[−0.1738]
LockupDummy	−0.2595***	−0.2599***	−0.2602***	−0.2584***	−0.2581***	−0.2593***	−0.2575***
	(0.0507)	(0.0511)	(0.0508)	(0.0511)	(0.0507)	(0.0511)	(0.0511)
	[−0.0700]	[−0.0702]	[−0.0703]	[−0.0696]	[−0.0694]	[−0.0700]	[−0.0692]
Restriction	−0.1596	−0.1665	−0.1609	−0.1676	−0.1694	−0.1661	−0.1721
	(0.1056)	(0.1078)	(0.1062)	(0.1071)	(0.1063)	(0.1074)	(0.1072)
	[−0.0431]	[−0.0450]	[−0.0435]	[−0.0451]	[−0.0456]	[−0.0448]	[−0.0463]
LeverageDummy	0.0793**	0.0799**	0.0810**	0.0732*	0.0780*	0.0757*	0.0741*
	(0.0397)	(0.0399)	(0.0397)	(0.0400)	(0.0398)	(0.0400)	(0.0401)
	[0.0214]	[0.0216]	[0.0219]	[0.0197]	[0.0210]	[0.0204]	[0.0199]
(Intercept)	0.1194	0.2249	0.1552	0.1598	0.1337	0.1866	0.1608
	(0.3988)	(0.3907)	(0.3974)	(0.3923)	(0.3950)	(0.3919)	(0.3915)
Num. obs.	33102	32828	33102	32828	33102	32828	32828

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Panel C: Civil judicial action and criminal action disclosures

	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal
LexicalDiversity	−0.0504*			−0.1047***	−0.3185***		−0.2844***
	(0.0263)			(0.0309)	(0.0846)		(0.0863)
	[−0.0065]			[−0.0136]	[−0.0413]		[−0.0369]
SyntacticComplexity		0.0699**		0.1162***		0.1096***	0.0946***
		(0.0285)		(0.0311)		(0.0325)	(0.0329)
		[0.0091]		[0.0151]		[0.0143]	[0.0123]
TextLength			−0.0205		0.2733***	−0.0760**	0.1930**
			(0.0270)		(0.0851)	(0.0325)	(0.0904)
			[−0.0027]		[0.0354]	[−0.0099]	[0.0250]
AUM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
Age	−0.0144***	−0.0107**	−0.0135***	−0.0126***	−0.0136***	−0.0120***	−0.0123***
	(0.0042)	(0.0043)	(0.0042)	(0.0042)	(0.0042)	(0.0042)	(0.0042)
	[−0.0019]	[−0.0014]	[−0.0018]	[−0.0016]	[−0.0018]	[−0.0016]	[−0.0016]
ManagementFee	−16.2243***	−15.3806***	−16.1488***	−15.4339***	−15.7797***	−15.5048***	−15.2229***
	(5.6719)	(5.7831)	(5.6961)	(5.7489)	(5.6305)	(5.7744)	(5.7062)
	[−2.1083]	[−2.0046]	[−2.1002]	[−2.0044]	[−2.0444]	[−2.0169]	[−1.9744]
IncentiveFee	−0.4844	−0.5504	−0.5161	−0.4704	−0.3823	−0.5219	−0.4061
	(0.3543)	(0.3545)	(0.3530)	(0.3596)	(0.3531)	(0.3582)	(0.3577)
	[−0.0629]	[−0.0717]	[−0.0671]	[−0.0611]	[−0.0495]	[−0.0679]	[−0.0527]
LockupDummy	−0.1966***	−0.2000***	−0.1974***	−0.1989***	−0.1942***	−0.2000***	−0.1969***
	(0.0626)	(0.0630)	(0.0627)	(0.0628)	(0.0628)	(0.0628)	(0.0628)
	[−0.0256]	[−0.0261]	[−0.0257]	[−0.0258]	[−0.0252]	[−0.0260]	[−0.0255]
Restriction	−0.2737*	−0.2902*	−0.2749*	−0.2906*	−0.2885*	−0.2879*	−0.2976*
	(0.1622)	(0.1677)	(0.1633)	(0.1653)	(0.1632)	(0.1660)	(0.1654)
	[−0.0356]	[−0.0378]	[−0.0358]	[−0.0377]	[−0.0374]	[−0.0374]	[−0.0386]
LeverageDummy	0.0346	0.0349	0.0363	0.0256	0.0336	0.0280	0.0270
	(0.0481)	(0.0484)	(0.0480)	(0.0484)	(0.0483)	(0.0483)	(0.0485)
	[0.0045]	[0.0045]	[0.0047]	[0.0033]	[0.0044]	[0.0036]	[0.0035]
(Intercept)	−0.0641	0.0495	−0.0322	−0.0226	−0.0591	0.0032	−0.0284
	(0.4355)	(0.4343)	(0.4362)	(0.4349)	(0.4315)	(0.4359)	(0.4328)
Num. obs.	33102	32828	33102	32828	33102	32828	32828

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 6: Flows

This table presents results from regressions in which annual fund flows are regressed on lagged text characteristics and a set of lagged control variables. Past performance controls consists of the variables Low rank, Mid rank and High rank that are defined following Sirri and Tufano (1998), by using a fractional rank (FRANK) representing a fund's percentile performance relative to other funds in the same investment strategy during the quarter. The lowest performance tercile (Low rank) is defined as Min (0.2, FRANK); the middle performance tercile (Mid rank) is defined as Min (0.6, FRANK - Low rank); and the highest performance tercile (High rank) is defined as FRANK - Low rank - Mid rank. The rest of the variables are defined in Table 1. All regressions include style and time fixed effects, and standard errors clustered by fund are shown in parentheses.

	Flow	Flow	Flow	Flow	Flow	Flow	Flow
LexicalDiversity	0.0294***			0.0385***	0.0228		0.0155
	(0.0044)			(0.0052)	(0.0152)		(0.0155)
SyntacticComplexity		0.0039		−0.0157***		−0.0190***	−0.0183***
		(0.0041)		(0.0048)		(0.0050)	(0.0051)
TextLength			0.0289***		0.0070	0.0407***	0.0254
			(0.0045)		(0.0154)	(0.0055)	(0.0163)
LowRank	0.3660***	0.3772***	0.3665***	0.3695***	0.3660***	0.3697***	0.3694***
	(0.0922)	(0.0924)	(0.0921)	(0.0926)	(0.0921)	(0.0925)	(0.0925)
MidRank	0.5720***	0.5755***	0.5723***	0.5727***	0.5721***	0.5731***	0.5729***
	(0.0244)	(0.0246)	(0.0244)	(0.0245)	(0.0244)	(0.0245)	(0.0245)
HighRank	0.1936	0.1809	0.1937	0.1910	0.1939	0.1915	0.1916
	(0.1220)	(0.1228)	(0.1220)	(0.1227)	(0.1220)	(0.1227)	(0.1227)
AUM	−0.0001***	−0.0001***	−0.0001***	−0.0001***	−0.0001***	−0.0001***	−0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Age	−0.0200***	−0.0206***	−0.0200***	−0.0202***	−0.0200***	−0.0202***	−0.0202***
	(0.0009)	(0.0009)	(0.0009)	(0.0009)	(0.0009)	(0.0009)	(0.0009)
Flow	0.0629***	0.0634***	0.0629***	0.0630***	0.0629***	0.0630***	0.0630***
	(0.0041)	(0.0041)	(0.0041)	(0.0041)	(0.0041)	(0.0041)	(0.0041)
HighWaterMark	0.0067	0.0128	0.0075	0.0059	0.0068	0.0065	0.0061
	(0.0153)	(0.0153)	(0.0153)	(0.0154)	(0.0153)	(0.0154)	(0.0154)
ManagementFee	−1.2666	−1.2346	−1.2334	−1.2757	−1.2594	−1.2295	−1.2480
	(0.8274)	(0.8310)	(0.8276)	(0.8290)	(0.8276)	(0.8291)	(0.8292)
IncentiveFee	−0.2248***	−0.1986**	−0.2195***	−0.2215***	−0.2237***	−0.2158***	−0.2186***
	(0.0831)	(0.0838)	(0.0833)	(0.0832)	(0.0831)	(0.0835)	(0.0834)
LockupDummy	−0.0248***	−0.0238***	−0.0244***	−0.0258***	−0.0247***	−0.0255***	−0.0257***
	(0.0091)	(0.0091)	(0.0091)	(0.0091)	(0.0091)	(0.0091)	(0.0091)
Restriction	−0.0642***	−0.0641***	−0.0650***	−0.0607***	−0.0644***	−0.0613***	−0.0610***
	(0.0150)	(0.0150)	(0.0150)	(0.0150)	(0.0150)	(0.0150)	(0.0150)
LeverageDummy	0.0087	0.0088	0.0087	0.0097	0.0087	0.0098	0.0098
	(0.0082)	(0.0082)	(0.0082)	(0.0082)	(0.0082)	(0.0082)	(0.0082)
Num. obs.	52609	52179	52609	52179	52609	52179	52179

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 7: Leverage

This table presents results from regressions in which the fund leverage is regressed on text characteristics at funds inception and a set of control variables.

	Leverage	Leverage	Leverage	Leverage	Leverage	Leverage	Leverage
LexicalDiversity	11.2773***			10.2717**	5.9978		6.3991
	(3.8394)			(4.6892)	(13.8355)		(14.0445)
SyntacticComplexity		7.5379*		1.9342		1.3113	1.5242
		(3.8943)		(4.6588)		(4.8426)	(4.8653)
TextLength			11.2310***		5.4824	10.5029**	4.2553
			(3.8304)		(13.8029)	(4.8569)	(14.5467)
HighWaterMark	21.7551**	24.1593**	22.2591**	22.4175**	21.9443**	22.8913**	22.5605**
	(11.0210)	(11.1321)	(11.0074)	(11.1584)	(11.0317)	(11.1456)	(11.1696)
ManagementFee	81.1077	141.7258	97.6959	79.5992	87.5340	93.8115	83.6092
	(592.4660)	(597.0366)	(592.2240)	(597.5996)	(592.7109)	(597.3409)	(597.7834)
IncentiveFee	208.1507***	214.4413***	210.9508***	209.5495***	209.5632***	211.6058***	210.2450***
	(60.4995)	(61.2777)	(60.5194)	(61.3071)	(60.6064)	(61.2808)	(61.3559)
LockupDummy	0.7450	1.1889	0.7192	0.8552	0.7201	0.8396	0.8395
	(9.1926)	(9.2955)	(9.1928)	(9.2950)	(9.1932)	(9.2952)	(9.2955)
Restriction	−41.7583**	−43.8072**	−41.8327**	−41.9421**	−41.7353**	−41.9715**	−41.9015**
	(16.9364)	(17.1696)	(16.9350)	(17.1875)	(16.9372)	(17.1875)	(17.1888)
Num. obs.	10393	10277	10393	10277	10393	10277	10277

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 8: Robustness checks

This table summarises results from empirical analyses conducted using two restricted subsamples. The first subsample is restricted to funds that are managed by firms domiciled in US, UK or Canada (i.e. English-speaking countries) to control for possible geographical effects. The second subsample is restricted to data starting from 2007, i.e. the year that we start collecting snapshots of funds' strategy description texts. This is to control for effects of imputing the strategy descriptions backwards for funds that have been incepted prior 2007.

Panel A: Spread-tests with univariate sorts

Funds managed by firms domiciled in US, UK or Canada			
Estimate	TextLength	LexicalDiversity	SyntacticComplexity
ΔAlpha	3.22	3.28	1.36
t-stat	5.17	5.27	1.90
ΔResidualVolatility	-0.38	-0.50	0.04
t-stat	-1.82	-2.35	0.19
ΔAppraisalRatio	1.00	1.07	0.36
t-stat	6.95	7.52	2.38
Funds with data period starting in 2007			
Estimate	TextLength	LexicalDiversity	SyntacticComplexity
ΔAlpha	1.33	1.91	-1.85
t-stat	2.22	2.99	-5.74
ΔResidualVolatility	-0.97	-1.11	0.04
t-stat	-6.30	-6.02	0.32
ΔAppraisalRatio	0.30	0.45	-0.45
t-stat	2.31	3.55	-5.14

Panel B: Spread-tests with double sorts

Funds managed by firms domiciled in US, UK or Canada				
Estimate	Sort by TextLength × LexicalDiversity		Sort by LexicalDiversity × TextLength	
ΔAlpha	1.06	2.78	1.39	1.35
t-stat	1.67	5.48	1.83	2.06
ΔResidualVolatility	-0.51	-0.07	-0.33	-0.26
t-stat	-1.82	-0.29	-1.22	-1.16
ΔAppraisalRatio	0.35	0.82	0.41	0.47
t-stat	2.23	5.43	1.99	2.74
Funds with data period starting in 2007				
Estimate	Sort by TextLength × LexicalDiversity		Sort by LexicalDiversity × TextLength	
ΔAlpha	0.62	3.07	-1.39	0.72
t-stat	1.12	6.07	-2.83	1.66
ΔResidualVolatility	-0.94	-0.27	-0.45	-0.22
t-stat	-7.04	-1.15	-2.61	-1.40
ΔAppraisalRatio	0.10	0.84	-0.40	0.21
t-stat	0.86	5.09	-3.31	1.46

Panel C: Multivariate regressions for funds managed by firms domiciled in US, UK or Canada

	Alpha	Alpha	Alpha	Alpha	Alpha	Alpha	Alpha
LexicalDiversity	0.0048***			0.0058***	0.0119***		0.0114***
	(0.0010)			(0.0012)	(0.0035)		(0.0036)
SyntacticComplexity		0.0014		−0.0018		−0.0017	−0.0011
		(0.0010)		(0.0012)		(0.0013)	(0.0013)
TextLength			0.0041***		−0.0074**	0.0051***	−0.0062*
			(0.0010)		(0.0035)	(0.0013)	(0.0037)
	ResidualVolatility	ResidualVolatility	ResidualVolatility	ResidualVolatility	ResidualVolatility	ResidualVolatility	ResidualVolatility
LexicalDiversity	−0.0015***			−0.0014***	−0.0009		−0.0009
	(0.0004)			(0.0005)	(0.0015)		(0.0015)
SyntacticComplexity		−0.0009**		−0.0002		−0.0001	−0.0001
		(0.0005)		(0.0005)		(0.0005)	(0.0005)
TextLength			−0.0015***		−0.0007	−0.0015***	−0.0005
			(0.0004)		(0.0015)	(0.0005)	(0.0016)
	AppraisalRatio	AppraisalRatio	AppraisalRatio	AppraisalRatio	AppraisalRatio	AppraisalRatio	AppraisalRatio
LexicalDiversity	0.0517***			0.0542***	0.0916***		0.0890***
	(0.0085)			(0.0101)	(0.0306)		(0.0309)
SyntacticComplexity		0.0222***		−0.0074		−0.0077	−0.0034
		(0.0084)		(0.0098)		(0.0103)	(0.0104)
TextLength			0.0470***		−0.0417	0.0501***	−0.0386
			(0.0086)		(0.0310)	(0.0107)	(0.0327)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Panel D: Multivariate regressions for funds with data period starting in 2007

	Alpha	Alpha	Alpha	Alpha	Alpha	Alpha	Alpha
LexicalDiversity	0.0019*			0.0038***	0.0114***		0.0106***
	(0.0012)			(0.0014)	(0.0039)		(0.0040)
SyntacticComplexity		−0.0017		−0.0035***		−0.0033**	−0.0028**
		(0.0011)		(0.0013)		(0.0013)	(0.0014)
TextLength			0.0010		−0.0099***	0.0028**	−0.0075*
			(0.0011)		(0.0038)	(0.0014)	(0.0040)
	ResidualVolatility	ResidualVolatility	ResidualVolatility	ResidualVolatility	ResidualVolatility	ResidualVolatility	ResidualVolatility
LexicalDiversity	−0.0015***			−0.0017***	−0.0034**		−0.0031*
	(0.0005)			(0.0006)	(0.0016)		(0.0016)
SyntacticComplexity		−0.0002		0.0006		0.0006	0.0004
		(0.0004)		(0.0005)		(0.0005)	(0.0005)
TextLength			−0.0012**		0.0020	−0.0014**	0.0016
			(0.0005)		(0.0015)	(0.0006)	(0.0016)
	AppraisalRatio	AppraisalRatio	AppraisalRatio	AppraisalRatio	AppraisalRatio	AppraisalRatio	AppraisalRatio
LexicalDiversity	0.0350***			0.0447***	0.0918***		0.0890***
	(0.0091)			(0.0105)	(0.0305)		(0.0309)
SyntacticComplexity		0.0028		−0.0183*		−0.0174*	−0.0132
		(0.0089)		(0.0101)		(0.0106)	(0.0107)
TextLength			0.0280***		−0.0596*	0.0380***	−0.0491
			(0.0092)		(0.0306)	(0.0110)	(0.0322)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Panel E: Financial risk regressions for funds managed by firms domiciled in US, UK or Canada

	MPPM	MPPM	MPPM	MPPM	MPPM	MPPM	MPPM
LexicalDiversity	0.9827***			0.9278***	1.8729***		1.9605***
	(0.1799)			(0.2122)	(0.6778)		(0.6754)
SyntacticComplexity		0.5990***		0.0925		0.1178	0.2128
		(0.1956)		(0.2284)		(0.2320)	(0.2299)
TextLength			0.8847***		−0.9293	0.8069***	−1.1469
			(0.1852)		(0.6985)	(0.2217)	(0.7061)
	MaxLoss	MaxLoss	MaxLoss	MaxLoss	MaxLoss	MaxLoss	MaxLoss
LexicalDiversity	−0.0016***			−0.0015***	−0.0022*		−0.0022*
	(0.0004)			(0.0004)	(0.0013)		(0.0013)
SyntacticComplexity		−0.0009**		−0.0001		−0.0001	−0.0002
		(0.0004)		(0.0004)		(0.0005)	(0.0005)
TextLength			−0.0015***		0.0006	−0.0014***	0.0008
			(0.0004)		(0.0013)	(0.0005)	(0.0014)
	Attrition	Attrition	Attrition	Attrition	Attrition	Attrition	Attrition
LexicalDiversity	−0.1023***			−0.0904***	−0.1239***		−0.1315***
	(0.0089)			(0.0105)	(0.0315)		(0.0319)
	[−0.0198]			[−0.0175]	[−0.0240]		[−0.0255]
SyntacticComplexity		−0.0665***		−0.0193*		−0.0184*	−0.0238**
		(0.0091)		(0.0105)		(0.0109)	(0.0110)
		[−0.0129]		[−0.0038]		[−0.0036]	[−0.0046]
TextLength			−0.0967***		0.0226	−0.0842***	0.0454
			(0.0089)		(0.0315)	(0.0109)	(0.0332)
			[−0.0188]		[0.0044]	[−0.0163]	[0.0088]

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Panel F: Financial risk regressions for funds with data period starting in 2007

	MPPM	MPPM	MPPM	MPPM	MPPM	MPPM	MPPM
LexicalDiversity	0.2767			0.7019***	2.6691***		2.4026***
	(0.2351)			(0.2677)	(0.8233)		(0.8283)
SyntacticComplexity		−0.5475**		−0.8801***		−0.7955***	−0.6833**
		(0.2496)		(0.2816)		(0.2866)	(0.2860)
TextLength			0.0386		−2.5082***	0.4664*	−1.8855**
			(0.2315)		(0.8118)	(0.2669)	(0.8300)
	MaxLoss	MaxLoss	MaxLoss	MaxLoss	MaxLoss	MaxLoss	MaxLoss
LexicalDiversity	−0.0012***			−0.0015***	−0.0038***		−0.0035***
	(0.0004)			(0.0005)	(0.0013)		(0.0014)
SyntacticComplexity		0.0001		0.0008		0.0007	0.0005
		(0.0004)		(0.0005)		(0.0005)	(0.0005)
TextLength			−0.0009**		0.0028**	−0.0012**	0.0023
			(0.0004)		(0.0013)	(0.0005)	(0.0014)
	Attrition	Attrition	Attrition	Attrition	Attrition	Attrition	Attrition
LexicalDiversity	−0.0300***			−0.0363***	−0.0501		−0.0432
	(0.0098)			(0.0112)	(0.0316)		(0.0320)
	[−0.0063]			[−0.0077]	[−0.0106]		[−0.0092]
SyntacticComplexity		−0.0000		0.0164		0.0175	0.0156
		(0.0096)		(0.0108)		(0.0112)	(0.0113)
		[−0.0000]		[0.0035]		[0.0037]	[0.0033]
TextLength			−0.0263***		0.0210	−0.0342***	0.0076
			(0.0097)		(0.0314)	(0.0116)	(0.0331)
			[−0.0056]		[0.0044]	[−0.0072]	[0.0016]

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Panel G: Disciplinary disclosures regressions for funds managed by firms domiciled in US, UK or Canada

	RegAct	RegAct	RegAct	RegAct	RegAct	RegAct	RegAct
LexicalDiversity	−0.0421*			−0.0946***	−0.3396***		−0.2970***
	(0.0242)			(0.0281)	(0.0752)		(0.0768)
	[−0.0110]			[−0.0247]	[−0.0886]		[−0.0774]
SyntacticComplexity		0.0763***		0.1206***		0.1121***	0.0967***
		(0.0234)		(0.0261)		(0.0274)	(0.0279)
		[0.0200]		[0.0315]		[0.0293]	[0.0252]
TextLength			−0.0123		0.3047***	−0.0659**	0.2186***
			(0.0243)		(0.0750)	(0.0292)	(0.0801)
			[−0.0032]		[0.0795]	[−0.0172]	[0.0570]
	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal
LexicalDiversity	−0.1008***			−0.1681***	−0.2186**		−0.1634
	(0.0292)			(0.0359)	(0.0976)		(0.1007)
	[−0.0113]			[−0.0187]	[−0.0244]		[−0.0182]
SyntacticComplexity		0.0738**		0.1470***		0.1557***	0.1475***
		(0.0316)		(0.0347)		(0.0364)	(0.0370)
		[0.0083]		[0.0164]		[0.0174]	[0.0164]
TextLength			−0.0834***		0.1209	−0.1610***	−0.0051
			(0.0293)		(0.0963)	(0.0371)	(0.1042)
			[−0.0093]		[0.0135]	[−0.0179]	[−0.0006]

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Panel H: Disciplinary disclosures regressions for funds with data period starting in 2007

	RegAct	RegAct	RegAct	RegAct	RegAct	RegAct	RegAct
LexicalDiversity	−0.0400*			−0.0844***	−0.3415***		−0.3116***
	(0.0241)			(0.0275)	(0.0735)		(0.0744)
	[−0.0111]			[−0.0235]	[−0.0946]		[−0.0864]
SyntacticComplexity		0.0638***		0.1023***		0.0910***	0.0747***
		(0.0236)		(0.0260)		(0.0271)	(0.0274)
		[0.0178]		[0.0284]		[0.0253]	[0.0207]
TextLength			−0.0074		0.3072***	−0.0506*	0.2442***
			(0.0239)		(0.0722)	(0.0282)	(0.0762)
			[−0.0021]		[0.0851]	[−0.0141]	[0.0677]
	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal	CivilOrCriminal
LexicalDiversity	−0.0650**			−0.1073***	−0.3353***		−0.3106***
	(0.0304)			(0.0347)	(0.0957)		(0.0974)
	[−0.0090]			[−0.0148]	[−0.0461]		[−0.0429]
SyntacticComplexity		0.0471		0.0939***		0.0857**	0.0686*
		(0.0318)		(0.0341)		(0.0357)	(0.0362)
		[0.0065]		[0.0130]		[0.0119]	[0.0095]
TextLength			−0.0321		0.2741***	−0.0738**	0.2175**
			(0.0311)		(0.0958)	(0.0365)	(0.1014)
			[−0.0044]		[0.0377]	[−0.0102]	[0.0300]

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Panel I: Flows and leverage

Funds managed by firms domiciled in US, UK or Canada							
	Flow	Flow	Flow	Flow	Flow	Flow	Flow
LexicalDiversity	0.0275***			0.0382***	0.0305*		0.0223
	(0.0051)			(0.0062)	(0.0183)		(0.0185)
SyntacticComplexity		0.0016		−0.0187***		−0.0216***	−0.0205***
		(0.0048)		(0.0058)		(0.0060)	(0.0061)
TextLength			0.0263***		−0.0031	0.0397***	0.0176
			(0.0051)		(0.0184)	(0.0065)	(0.0195)
	Leverage	Leverage	Leverage	Leverage	Leverage	Leverage	Leverage
LexicalDiversity	7.2974***			6.6405**	−7.5369		−8.5824
	(2.4530)			(3.0373)	(9.0744)		(9.1853)
SyntacticComplexity		4.3635*		0.5517		−0.7977	−1.0770
		(2.5290)		(3.0712)		(3.1938)	(3.2078)
TextLength			8.1269***		15.3614*	8.3153***	16.7026*
			(2.4452)		(9.0471)	(3.1446)	(9.5114)
Funds with data period starting in 2007							
	Flow	Flow	Flow	Flow	Flow	Flow	Flow
LexicalDiversity	0.0142***			0.0222***	0.0149		0.0066
	(0.0055)			(0.0063)	(0.0172)		(0.0173)
SyntacticComplexity		−0.0060		−0.0164***		−0.0185***	−0.0182***
		(0.0050)		(0.0057)		(0.0060)	(0.0060)
TextLength			0.0133**		−0.0008	0.0236***	0.0171
			(0.0055)		(0.0172)	(0.0065)	(0.0180)
	Leverage	Leverage	Leverage	Leverage	Leverage	Leverage	Leverage
LexicalDiversity	6.3179			4.4903	15.9373		17.8574
	(5.3582)			(6.1380)	(16.9422)		(17.2923)
SyntacticComplexity		5.8979		3.9089		4.6562	5.3964
		(5.2027)		(5.8705)		(6.0982)	(6.1401)
TextLength			4.8740		−9.9255	2.4349	−14.5278
			(5.2452)		(16.5840)	(6.2368)	(17.5700)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$