8-2019

Text sophistication and sophisticated investors

Juha JOENVAARA

Jari KARPPINEN

Song Wee Melvyn TEO
Singapore Management University, melvynteo@smu.edu.sg

Cristian Ioan TIU

Text Sophistication and Sophisticated Investors

Juha Joenväärä, Jari Karppinen, Melvyn Teo, and Cristian Tiu*

Abstract

We show that two novel measures of text sophistication, applied to hedge fund strategy descriptions, encapsulate incremental information about funds. Consistent with the linguistics literature, hedge funds with lexically diverse strategy descriptions outperform, eschew tail risk, and encounter fewer regulatory problems. In line with the literature, hedge funds with syntactically complex strategy descriptions report more regulatory violations and trigger more severe infractions. Fund investors recognize the dichotomy and direct flows accordingly, but not enough to erode away the alphas of lexically diverse funds. Our findings suggest that text sophistication measures provide texture on the cognitive ability and trustworthiness of sophisticated investors.

*Joenväärä is at Department of Finance, School of Business, Aalto University and Center for Financial Policy, Smith School of Business, University of Maryland. E-mail: juha.joenvaara@aalto.fi. Karppinen is at Department of Economics, Accounting, and Finance, University of Oulu. E-mail: jari.karppinen@oulo.fi. Teo (corresponding author) is at the Lee Kong Chian School of Business, Singapore Management University. Address: 50 Stamford Road, Singapore 178899. E-mail: melvynteo@smu.edu.sg. Tel: +65-6828-0735. Fax: +65-6828-0427. Tiu is at Department of Finance, School of Management, University of Buffalo. E-mail: ctiu@buffalo.edu. We have benefitted from conversations with George Aragon, Serge Darolles, Veljko Fotak, Marcus Ibert, Petri Jylhä, Matti Keloharju, Olga Kolokolova, Pete Kyle, John Lewis, Bing Liang, Kevin McKelvey, Aleksandra Rzeźnik (2019 EFA discussant), Daniel Schmidt, Valeri Sokolovski, Petra Vokata, Russ Wermers, Chengdong Yin (2019 CICF discussant) and Filip Zikes as well as seminar participants at Aalto University, Maryland University, Singapore Management University, the 11th Paris Hedge Fund Conference, the Mutual Funds, Hedge Funds, and Factor Investing Conference at Lancaster University, the Federal Reserve Board of Governors, the 2019 China International Conference in Finance in Guangzhou, China, and the 2019 European Finance Association Meetings in Carcavelos, Portugal. We are also indebted to Kathleen Sheehan of Educational Testing Service for her help with the various measures of text sophistication and to Mikko Kauppila for excellent research assistance.

1. Introduction

Financial economists have used textual analysis to gauge the tone, information content, and readability of corporate disclosures, press releases, media articles, shareholder reports, and Internet message boards, and study their relevance for *stocks* (Antweiler and Frank, 2004; Tetlock, 2007; Li, 2008; Tetlock, Saar-Tsechansky, and Macskassy, 2008; Lehavy, Li, and Merkley, 2011; Jegadeesh and Wu, 2013; Loughran and MacDonald, 2011, 2013, 2014; Bodnaruk, Loughran, and MacDonald, 2015; Hoberg and Maksimovic, 2015; Hwang and Kim, 2017; Buehlmaier and Whited, 2018; Ke, Kelly, and Xiu, 2019). With the exception of Hwang and Kim (2017), who analyze the impact of shareholder report readability on closed-end investment firm discounts, none of these studies relate to delegated portfolio management. Given the assets managed by investment managers globally, the reams of qualitative data that they provide, and the value that investors place on fund selection, the application of textual analysis to delegated portfolio management is an important, albeit relatively unexplored, area of work. In this paper, we employ two novel measures of text sophistication, applied to hedge fund strategy descriptions, to explore their implications for investment management.

We ask the following question: Are sophisticated investors also sophisticated writers? A priori, it is not clear that they should be. On one hand, sophisticated managers with superior investment skills, and who are therefore cognitively gifted, should write in a sophisticated manner. On the other hand, managers who do not possess investment skills may write in a sophisticated fashion to deceptively signal investment prowess. The two measures of text sophistication that we enlist, namely lexical diversity and syntactic complexity, allow us to distinguish between the two cases. Lexical diversity is the propensity of the writer to use multiple synonyms rather than repeated words and has been associated with cognitive ability (Bucks et al., 2000; Thordardottir and Namazi, 2007; Fergadiotis and Wright, 2011) and honesty (Humphreys, 2010; Horne and Adali, 2017). Syntactic complexity is the inclination

by the writer to favor complicated sentences characterized by heavy use of subordination and has been linked to deceptive behavior (Moffitt and Burns, 2009; Levitan, 2019). The differential loadings that the two measures have on deception suggest that only lexical diversity, and not syntactic complexity, provides a honest cue to managerial talent.

We study hedge fund managers, some of the world's most sophisticated investors. Hedge funds often engage in dynamic strategies that involve short sales, complex derivatives, and substantial leverage. They typically charge sizeable incentive fees that are pegged to performance, thereby attracting the best and the brightest in investment management (Perold and Spitz, 1996). At the same time, due to the low levels of transparency and light regulation, the industry also features a disproportionate number of investment frauds (Dimmock and Gerken, 2012). Textual analysis may offer investors a novel methodology for differentiating between skilled fund managers and the potential frauds. We apply textual analysis to hedge fund strategy descriptions, as unlike fund prospectuses and investor newsletters, strategy descriptions are readily available from commercial hedge fund databases. We argue that skilled and, therefore, cognitively gifted managers are more likely to employ rich vocabulary when crafting their strategy descriptions, which in turn engenders lexical diversity. Fraudulent managers are more likely to obfuscate and confuse when describing their strategies, which translates into syntactic complexity.

The analysis reveals substantial differences in expected returns, on decile portfolios of hedge funds sorted by lexical diversity, that are unexplained by the Fung and Hsieh (2004) seven factors. Hedge funds with high lexical diversity outperform those with low lexical diversity by an economically and statistically significant 3.63% per year (t-statistic = 7.45) after adjusting for co-variation with the Fung and Hsieh (2004) factors. The results also manifest in multivariate regressions and cannot be explained by differences in share restrictions and illiquidity (Aragon, 2007; Aragon and Strahan, 2012), incentives (Agarwal, Daniel, and Naik, 2009), fund age (Aggarwal and Jorion, 2010), fund size (Berk and Green, 2004), return smoothing behavior (Getmansky, Lo, and Makarov, 2004), and backfill and incubation bias

(Liang, 2000; Fung and Hsieh, 2009; Bhardwaj, Gorton, and Rouwenhorst, 2014).

Hedge funds with lexically diverse strategy descriptions display several additional attributes that are attractive to fund investors. First, they deliver superior Sharpe ratios, information ratios, and Goetzmann et al. (2007) manipulation-proof performance measures. For example, high-lexical diversity funds exhibit annualized information ratios that are on average 1.07 units higher (t-statistic = 8.37) than those of low-lexical diversity funds. Therefore, the higher alphas of lexically diverse funds cannot be attributed to greater leverage or manager manipulation. Second, lexically diverse funds manage risk more judiciously. They eschew idiosyncratic risk and tail risk. In particular, lexically diverse and lexically homogenous funds exhibit annualized residual volatilities of 3.63% and 4.19%, respectively. The difference in annualized residual volatility is an economically and statistically relevant -0.56% (t-statistic = 2.54). Third, they face fewer regulatory actions, encounter fewer civil or criminal problems, and trigger fewer investment violations. Lexical diversity, therefore, predicts fund quality. These findings support the view that lexical diversity is associated with cognitive ability and trustworthiness in the hedge fund arena.

Our tests further reveal that the second measure of text sophistication we consider, namely syntactic complexity, is associated with deception at hedge funds. Funds whose strategy descriptions are syntactically complex experience more regulatory actions, violate more investment rules, and report more severe infractions, than do funds whose strategy descriptions are syntactically simple. After controlling for other factors including lexical diversity, an increase in syntactic complexity from the bottom 10th to the top 10th percentile is associated with a 4.18% increase in the probability of regulatory actions, a 3.48% increase in the probability of investment related infractions, and a 2.09% increase in the probability of severe violations. While syntactic complexity is positively related to fund returns and alphas when we do not control for lexical diversity, those relations are statistically insignificant and turn negative after controlling for lexical diversity.

Do investors appreciate the discrepant implications of lexical diversity and syntactic

complexity on fund performance and quality? We find that, after controlling for other factors that explain fund flow, including past fund performance, funds with high lexical diversity attract greater investor flow than do funds with low lexical diversity. The coefficient estimates from a multivariate regression on fund flow indicate that increases in lexical diversity and syntactic complexity from the bottom 10th to the top 10th percentile are associated with a 14.52% increase and a 4.39% decrease in annual fund flow, respectively. While investors allocate more capital to lexically diverse funds, given that lexical diversity is still positively related to fund alpha in the univariate setting, the additional capital is not enough to zero out the alphas of lexically diverse funds via capacity constraints (Berk and Green, 2004).

One concern is that our measures of text sophistication may capture strategy sophistication instead. Sun, Wang, and Zheng (2012) show that hedge funds that ply more distinctive strategies outperform hedge funds that operate less distinctive strategies. Descriptions of distinctive strategies could be lexically diverse and syntactically complex. Another concern is that syntactic complexity could be related to readability, which Hwang and Kim (2017) show is inversely related to closed-end fund discounts. The Hwang and Kim (2017) readability measure is based on the number of passive verbs, hidden verbs, legal words, overwriting, wordy phrases, and abstract words. Other proxies for readability include the Fog index and the Flesch Kincaid index employed by Li (2008) and Lehavy, Li, and Merkley (2011). These measures are correlated with sentence length and the number of syllables of the words used. Texts with low readability may, therefore, feature complicated sentences and be syntactically complex. Yet another concern is that lexical diversity may be related to fund manager education, which Li, Zhang, and Zhao (2011) and Chaudhuri et al. (2019) show is helpful for fund performance. Managers who attended higher quality schools or received multiple years of schooling could have built up richer vocabulary and be better equipped to compose lexically diverse text. To address these concerns, we rerun our baseline analysis after adjusting for strategy distinctiveness, readability, and manager education, and find that our inferences remain qualitatively unchanged.

The empirical results indicate that text sophistication measures provide incremental information about the cognitive ability and trustworthiness of sophisticated investors. By doing so, we contribute to the growing literature on the application of textual analysis to finance. Unlike the vast majority of papers in this area that extract textual data from word lists (Tetlock, 2007; Tetlock, Saar-Tsechansky, and Macskassy, 2008; Loughran and MacDonald, 2011, 2013; Jegadeesh and Wu, 2013; Bodnaruk, Loughran, and MacDonald, 2015) or machine learning (Antweiler and Frank, 2004; Jegadeesh and Wu, 2013; Hoberg and Maksimovic, 2015; Buehlmaier and Whited, 2018; Ke, Kelly, and Xiu, 2019), we rely on text sophistication measures from the linguistics literature. Our work is related to papers that analyze text readability via text length (Li, 2008; Loughran and McDonald, 2014), the Fog index (Li, 2008; Lehavy, Li, and Merkley, 2011), and writing errors (Hwang and Kim, 2017). The advantage of the more nuanced text sophistication measures is that they provide insights that go beyond readability and into the richness of the vocabulary as well as the intricacies of the sentence structures used.

Several papers investigate the drivers of hedge fund risk-adjusted performance with quantitative data such as manager option deltas (Agarwal, Daniel, and Naik, 2009), past fund performance (Jagannathan, Malakhov, and Novikov, 2010), fund age (Aggarwal and Jorion, 2010), fund R^2 (Titman and Tiu, 2011), and strategy distinctiveness (Sun, Wang, and Zheng, 2012). We contribute to this literature by relating qualitative textual data embedded in hedge fund strategy descriptions to fund performance. We do not claim that qualitative textual data subsumes or dominates traditional quantitative measures of hedge fund quality that are based on past performance or fund characteristics. Rather, our results suggest that textual analysis, in conjunction with traditional methods involving quantitative data, can be helpful for fund selection and for understanding the factors underlying alpha generation.

The remainder of this paper is organized as follows. Section 2 describes the data and methodology, and illustrates the text sophistication measures that we employ. Section 3 reports the empirical results. Section 4 presents robustness tests while Section 5 concludes.

2. Data and methodology

2.1. Hedge fund data

We evaluate the relation between text sophistication and hedge funds using monthly net-of-fee returns and assets under management (henceforth AUM) data of live and dead hedge funds reported in the BarclayHedge, EurekaHedge, eVestment, and Hedge Fund Research (henceforth HFR) databases from January 1990 to December 2016. Because commercial hedge fund databases started distributing their data in 1994, the data sets do not contain information on funds that died before January 1994. This gives rise to survivorship bias. We mitigate this bias by focusing on data from January 1994 onward.

To merge the BarclayHedge, EurekaHedge, eVestment, and HFR databases, we apply the aggregation procedure of Joenväärä, Kauppila, Kosowski, and Tolonen (2019). This yields a total of 21,379 funds with 1,449,207 monthly time series observations of returns and AUM. For each fund, we collect strategy descriptions as well as fund variables related to compensation structure, share restrictions, leverage, domicile, and investment style. We exclude funds with strategy descriptions that do not contain at least one meaningful sentence in English. While we have access to other hedge fund commercial databases, we do not employ them since they do not provide strategy descriptions for both live and dead hedge funds.[1]

Following Agarwal, Daniel, and Naik (2009), we classify funds into four broad investment styles: Security Selection, Multi-process, Directional Trader, and Relative Value. Security Selection funds take long and short positions in undervalued and overvalued securities, respectively. Usually, they take positions in equity markets. Multi-process funds employ multiple strategies that take advantage of significant events, such as spin-offs, mergers and acquisitions, bankruptcy reorganizations, recapitalizations, and share buybacks. Directional

[1]For example, our version of Lipper TASS provides strategy descriptions for live funds only. We do not include these funds in our sample due to survivorship bias concerns.

Trader funds bet on the direction of market prices of currencies, commodities, equities, and bonds in the futures and cash markets. Relative Value funds take positions on spread relations between prices of financial assets and aim to minimize market exposure.

Hedge fund data are susceptible to many biases (Fung and Hsieh, 2009). These biases stem from the fact that inclusion in hedge fund databases is voluntary. For instance, when a fund is listed on a database, it often includes data prior to the listing date. Because successful funds have a strong incentive to list and attract capital, backfilled returns tend to be higher than non-backfilled returns. Moreover, funds often undergo an incubation period during which they build up a track record using manager's money before listing on commercial databases and seeking capital from outside investors. Since funds with poor track records often do not end up listing on hedge fund databases, this induces an incubation bias. To ameliorate backfill and incubation bias, we drop the first 12 months of return data for each fund. To further alleviate concerns raised by Bhardwaj, Gorton, and Rouwenhorst (2014) and others, we will rerun the tests after removing all return observations that have been backfilled prior to the fund listing date.[2]

To mitigate the impact of strategic delays in reporting by hedge funds (Aragon and Nanda, 2017), while we download hedge fund data in mid-2017, we do not use the last few months of returns, and instead focus on the period ending in December 2016. Another concern is that hedge funds may have updated their strategy descriptions post inception, which imply that strategy descriptions may have been overwritten in subsequent database snapshots. To control for this potential look-ahead bias, as a robustness test, we will utilize multiple database snaphots gathered starting in 2007 and redo our tests.[3]

Throughout this paper, we model the risk of hedge funds using the Fung and Hsieh (2004) seven-factor model. The Fung and Hsieh factors are the excess return on the Standard and

[2]For funds from databases that do not report listing date, we will use the Jorion and Schwarz (2019) algorithm to back out fund listing date.

[3]We are able to collect multiple snapshots for BarclayHedge, EurekaHedge, and HFR. For BarclayHedge, we have snapshots of strategy descriptions for 2010, 2011, 2012, 2013, 2015, and 2017. For Eurekahedge, we have snapshots for 2007, 2008, 2009, 2010, 2011, 2013, 2015, and 2017. For HFR, we have snapshots for 2007, 2009, 2011, 2012, 2013, 2015, and 2017.

Poor's (S&P) 500 index (SNPMRF); a small minus big factor (SCMLC) constructed as the difference between the Russell 2000 and S&P 500 stock indexes; the yield spread of the U.S. ten-year Treasury bond over the three-month Treasury bill, adjusted for duration of the ten-year bond (BD10RET); the change in the credit spread of Moody's BAA bond over the ten-year Treasury bond, also appropriately adjusted for duration (BAAMTSY); and the excess returns on portfolios of lookback straddle options on bonds (PTFSBD), currencies (PTFSFX), and commodities (PTFSCOM), which are constructed to replicate the maximum possible return from trend-following strategies on their respective underlying assets.[4] Fung and Hsieh (2004) show that these seven factors have considerable explanatory power on aggregate hedge fund returns.

2.2. Measures of writing sophistication

The first measure of text sophistication that we employ is lexical diversity, which is related to the number of distinct words used in the text. Texts with high lexical diversity feature rich vocabulary, frequent use of synonyms, and few repeated words. Lexical diversity can be measured in multiple ways (Tweedie and Baayen, 1998; Jarvis, 2013). We focus on a particularly simple measure that has been proposed first by Shannon (1948) in his seminal paper on information theory. To calculate lexical diversity, we first apply standard textual analysis preprocessing to the text.[5] During the preprocessing stage, all non-alphanumeric characters are removed, all letters are converted to lowercase, all stop words are removed, and all remaining words are stemmed. The preprocessing yields V distinct tokens, each token i appearing in text with a frequency p_i, where $i = 1, ..., V$. Lexical diversity is then calculated as in Shannon (1948):

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

[4]David Hsieh kindly supplied these risk factors. The trend-following factors can be downloaded from http://faculty.fuqua.duke.edu/ dah7/DataLibrary/TF-Fac.xls.

[5]See Buehlmaier and Whited (2018) for an example of standard textual analysis preprocessing on text.

The Shannon lexical diversity measure is positively correlated with text length (Fergadiotis, Wright, and Green, 2015). Therefore, it will be important that we control for text length in our empirical analyses. We note that our results remain qualitatively unchanged when we employ alternative measures of lexical diversity such as the Carroll's corrected type-token ratio (Carroll, 1964).[6]

The second measure of text sophistication that we consider is syntactic complexity, which captures complexity in sentence structure. There are several dimensions to sentence complexity. For example, long sentences and long paragraphs are both indicative of complex writing. Complexity in writing can also be captured by the concept of word depth proposed by Yngve (1960). Depth refers to the quantity of left branching contained in the path linking a terminal node to the root node of a grammatical sentence structure.[7] According to Yngve (1960), the memory load imposed by left-branching limits them to around seven in a sentence. Consequently, any sentence with more than this limit of left branches is difficult to understand. We calculate syntactic complexity using the Educational Testing Service *TextEvaluator*® tool (Sheehan, 2015), which combines these three dimensions of sentence complexity. According to Napolitano, Sheehan, and Mundkowsky (2015) and Sheehan (2016), their measure of syntactic complexity encapsulates all information regarding how complex the sentences are within the text. As one of the four major components of the *TextEvaluator*® score proposed by Educational Testing Service to assess text sophistication, syntactic complexity has been widely used to ensure that students are reading appropriately challenging texts, and can be computed via the *TextEvaluator*® website.[8]

In this paper, we apply these two measures of text sophistication to hedge fund strategy

[6]Our baseline results are qualitatively similar when we use the Carroll's corrected type-token ratio, defined as $V/\sqrt{2N}$, where V is the number of distinct tokens in a text and N the total number of tokens. Inferences also do not change when we employ the simple uncorrected type-token ratio V/N as our measure of lexical diversity.

[7]For example, based on Sampson's (1997) preferred interpretation of Yngve (1960) depth, the words "New York" have a depth of five in the following sentence "Two errors by New York Yankee shortstop Tony Kubek in the eleventh inning donated four unearned runs and a 5-to-2 victory to the Chicago White Sox today." See page 142 of Sampson (1997).

[8]See https://www.ets.org/c/23491/

descriptions. The advantage of analyzing strategy descriptions is that, unlike fund prospectuses and investor newsletters, strategy descriptions are widely available from commercial hedge fund databases. One concern is that hedge funds may simply provide boilerplate strategy descriptions that convey little incremental information. We argue that the increasing need for transparency by hedge fund investors prevents hedge funds from doing so. Consistent with this view, Fig. 1 indicates that strategy descriptions have been increasing slightly in length over time. Investors' demand for longer and presumably more detailed strategy descriptions imply that strategy descriptions contain even more incremental information for newly launched hedge funds.

[Insert Fig. 1 and Table 1 here]

Another concern is that lexical diversity and syntactic complexity may vary systematically across different investment strategies. Panels A and B of Table 1 report summary statistics for the two text sophistication measures for funds grouped by investment strategy. They indicate that there are no systematic differences in lexical diversity or syntactic complexity across strategies. Nonetheless, to address any lingering concerns, we will standardize both measures within each investment strategy prior to running our empirical tests.

Yet another concern is that strategy descriptions written by native English speakers may be more lexically diverse and syntactically complex than those composed by non-native English speakers (Lu and Ai, 2015). To investigate, we group funds based on where their management firms are based and report summary statistics of the text sophistication measures for each group. Panels C and D of Table 1 indicate that the text sophistication measures do not differ systematically across geographical regions. To further ameliorate this concern, as a robustness test, we will redo the analysis for funds that are based in English-speaking countries or are advised by managers with English names.

Next, we provide two examples of strategy descriptions found in our combined hedge fund database to illustrate lexical diversity and syntactic complexity. Our first example is that

from the CNH Diversified Opportunities Master Account LP, managed by CNH Partners LLC, whose principals may be familiar to readers who are 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.

This strategy description exhibits a low syntactic complexity score of 54, which places it below the median fund in our sample. It also has a high lexical diversity score of 4.48, which places it in the top 10th percentile of funds in our sample. This is not surprising given that the strategy description features simple syntax structures and contains many distinct words. Simply put, the authors write clearly and showcase a wide vocabulary.

Our second example is that from the Fairfield Sentry Ltd fund, managed by the Fairfield Greenwich Group. Post-2008, this fund gained notoriety as one of the feeders for Bernard

Madoff's fund (Gregoriou and Lhabitant, 2009).

> 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.

Relative to the first example, this strategy description is convoluted and features long sentences with flow interrupted by insertions of terms that are used first but defined later. This is the quintessential example of a fund whose strategy description obfuscates rather than clarifies. It is therefore unsurprising that, while the text features a high lexical diversity score of 4.39, it also exhibits a high syntactic complexity score of 84, which is significantly above that of the median fund.

2.3. Empirical hypotheses

Does writing sophistication matter for investment management? Do we expect sophisticated investors, such as skilled hedge fund managers, to also write in a sophisticated fashion? Skilled hedge fund managers who are presumably cognitively superior should produce more sophisticated text. At the same time, unskilled hedge fund managers may deceptively choose to craft sophisticated text so as to appear sophisticated themselves.

To distinguish between these two sets of managers, we appeal to results on cognitive ability and deception from the linguistics literature. Both lexical diversity and syntactic complexity measure text sophistication. In particular, lexical diversity is synonymous with cognitive ability. For example, preschool children with specific language impairment (Thordardottir and Namazi, 2007), adults with aphasia (Fergadiotis and Wright, 2011), and patients with Alzheimer's disease (Bucks et al., 2000) display lower lexical diversity in their spontaneous speech and discourse relative to healthy control groups. At the same time, lexical diversity is related to trustworthiness while syntactic complexity is associated with deceptive behavior. In a meta analysis of deception detection, Humpherys (2010) finds that deceivers exhibit lower lexical diversity than do truth tellers. Likewise, Horne and Adali (2017) demonstrate that fake news articles feature lower lexical diversity than do real news articles. Moreover, Mofitt and Burns (2009) find that fraudulent 10-Ks feature more qualifying conjunctions and are therefore more syntactically complex, and Levitan (2019, page 48–52) shows that syntactic complexity measures are positively related to deception for a large corpus of recorded dialogue between pairs of subjects who play a lying game.

What drives the link between the two text sophistication measures and trustworthiness/deception? Vrij, Granhag, and Porter (2010) argue that deceptive behavior increases cognitive load, which hampers a person's ability to produce complex language. This theory of deceptive behavior predicts that deception negatively relates to lexical diversity and syntactic complexity, since both lexical diversity and syntactic complexity impose additional

mental load on the individual. However, as shown in Levitan (2019), when deception is premeditated, e.g., when the subjects had time to prepare their responses, the cognitive load theory may not hold. Moreover, in the context of this study, since fund managers typically do not face constraints on time and effort when crafting strategy descriptions, cognitive load may be less relevant. Instead, we argue, as do Moffitt and Burns (2009), that a fund manager who has something to hide will tend to favor complicated and hard-to-decipher sentences in an attempt to obfuscate and confuse. Therefore, syntactic complexity will be synonymous with deception. Further, we contend that it is easier to express with a wide array of words the intricacies of an investment strategy if the fund manager can draw from real-life experiences associated with deploying the said strategy. In addition, fraudulent fund managers are less likely to employ rich vocabulary when describing their strategies so as not to divulge too much information and inadvertently reveal the inconsistencies in their fund operations (Freud, 1901).[9] Consequently, lexical diversity could be associated with trustworthiness.

Therefore, we investigate the following hypotheses:

Hypothesis 1 *Funds with lexically diverse strategy descriptions display higher quality; lexically diverse funds outperform and are more judicious when managing risk.*

Hypothesis 2 *Funds with lexically diverse strategy descriptions are more trustworthy; they encounter fewer legal problems and report fewer regulatory and financial violations.*

Hypothesis 3 *Funds with syntactically complex strategy descriptions are more deceptive; they grapple with more legal issues and trigger more regulatory and financial infractions.*

Hypothesis 4 *Fund investors understand the discrepant implications of lexical diversity and syntactic complexity on fund performance and quality; they direct greater fund flows to lexically diverse funds than to syntactically complex funds.*

[9] According to Freud (1901), although liars have some control over the content of their stories, their underlying state of mind may "leak out" through the way that they tell them. See Newman et al. (2003) for a discussion of this view.

3. Empirical results

3.1. Fund performance

To explore the implications of the text sophistication measures for fund performance, we first study the risk-adjusted performance of funds sorted by lexical diversity. Every year, starting in January 1994, ten hedge fund portfolios are formed by sorting funds on the lexical diversity of their strategy descriptions. The post-formation returns on these ten portfolios over the next 12 months are linked across years to form a single return series for each portfolio. We then evaluate the performance of the portfolios relative to the Fung and Hsieh (2004) model. The lexical diversity measure does not change over time for each fund, except possibly when we analyze the fund sample derived from multiple database snapshots. Still, the sorting procedure allows us to accommodate variation in the composition of the fund sample as funds enter and drop out of the combined database.

The results, reported in Panel A of Table 2, reveal substantial differences in expected returns, on the portfolios sorted by fund lexical diversity, that are unexplained by the Fung and Hsieh (2004) seven factors. Hedge funds with lexically diverse strategy descriptions (Portfolio 1) outperform those with lexically homogenous strategy descriptions (Portfolio 10) by an economically and statistically significant 3.03% per year (t-statistic = 6.11). After adjusting for co-variation with the Fung and Hsieh (2004) factors, the outperformance increases to 3.63% per year (t-statistic = 7.45).[10] As in the rest of the paper, we base statistical inferences on Newey and West (1987) heteroskedasticity and autocorrelation consistent standard errors. We note that the average lexical diversity of Portfolio 1 is 1.41 while that for Portfolio 10 is −2.09. The high-minus-low lexical diversity spread is therefore associated

[10]Small funds may be less relevant to institutional investors who allocate significant capital. It is therefore comforting to note that our findings prevail after dropping funds with AUM less than US$50 million from the sample. The risk-adjusted spread from the equal-weighted sort is 2.40% per year (t-statistic = 3.52) for funds managing at least US$50 million. The portfolio sort results are also robust to value-weighting the funds within each portfolio. The risk-adjusted spread for the value-weighted sort is 2.74% per annum (t-statistic = 2.64).

with a 3.50 unit change in lexical diversity. The results reported in Panel B of Table 2 indicate that our findings are not driven by the greater text length of high lexical diversity strategy descriptions. When we scale lexical diversity by text length and redo the sort, we find that the risk-adjusted performance of the high-minus-low scaled lexical diversity spread is economically and statistically significant at 3.74% per annum (t-statistic = 7.04).

[Insert Table 2 here]

For the portfolio sort, we also report other hedge fund performance measures that are relevant for fund investors, namely, the Sharpe ratio, the information ratio, and the Goetzmann et al. (2007) manipulation-proof performance measure (henceforth MPPM). Sharpe ratio is average fund excess return divided by the standard deviation of fund return. Information ratio is Fung and Hsieh (2004) alpha divided by the standard deviation of the residuals from the Fung and Hsieh (2004) model. The advantage of Sharpe and information ratios over fund alpha as measures of fund performance is that they are invariant to leverage. The Goetzmann et al. (2007) MPPM helps ameliorate the concern is that funds, especially those with complicated strategy descriptions, may use various techniques to manipulate performance measures. These techniques may include writing deep out-of-the-money put options, which could inflate the Sharpe and information ratios. We compute fund portfolio MPPM using a risk aversion parameter $\rho = 3$ as per Goetzmann et al. (2007).[11] The fund performance measures reported in Table 2 indicate that lexically diverse funds deliver higher Sharpe ratios, information ratios, and MPPMs than do lexically homogenous funds. The high-minus-low lexical diversity spread portfolio generates an annualized Sharpe ratio of 0.59, an annualized information ratio of 1.07, and a MPPM of 3.23. These measures are all economically meaningful and statistically distinguishable from zero at the 1% level.

Next, we perform the analogous portfolio sort on the syntactic complexity of fund strategy descriptions. As shown in Panel A of Table 3, while hedge funds with syntactically complex strategy descriptions outperform those with syntactically simple descriptions, the

[11] The results are qualitatively similar when we set the risk aversion parameter ρ to 2 or 4.

risk-adjusted outperformance is economically modest at only 1.20% per year (t-statistic = 1.99). The other performance measures of the spread portfolio from the syntactic complexity sort are also significantly more modest than those from the lexical diversity sort. The high-minus-low syntactic complexity spread exhibits a low annualized Sharpe ratio of 0.12, annualized information ratio of 0.25, and MPPM of 1.50. We obtain even weaker results when we scale syntactic complexity by text length and redo the portfolio analysis. These results, reported in Panel B of Table 3, suggest that the outperformance of syntactically complex funds is not robust to controlling for text length.

[Insert Table 3 and Fig. 2 here]

Fig. 2 complements the results from Panel A of Table 2 and Panel A of Table 3. It illustrates the monthly cumulative abnormal returns (henceforth CARs) from the extreme high- and low-lexical diversity and syntactic complexity decile portfolios. CAR is the cumulative difference between a portfolio's excess return and its factor loadings multiplied by the Fung and Hsieh (2004) risk factors, where loadings are estimated over the entire sample period. The CARs in Fig 2. indicate that the high-lexical diversity portfolio outperforms the low-lexical diversity portfolio over the entire sample period. They also reveal that the high-syntactic complexity portfolio only outperforms the low-syntactic complexity portfolio during the first half of the sample period. During the second half of the sample period, the relation between syntactic complexity and performance reverses and the high-syntactic complexity portfolio underperforms the low-syntactic complexity portfolio. Therefore, unlike lexical diversity, syntactic complexity is not a consistent predictor of fund outperformance.

One concern is that the explanatory power of lexical diversity on fund alpha may be subsumed by other factors, e.g., size (Berk and Green, 2004), fund age (Aggarwal and Jorion, 2010), and incentives (Agarwal, Daniel, and Naik, 2009), that drive fund performance.

Therefore, we estimate the following multivariate regression on monthly fund alpha:

$$\begin{aligned}\text{Alpha}_{im} = \gamma_0 + \gamma_1 \text{Text measures}_{im-1} + \gamma_2 \text{Time-varying controls}_{im-1} \\ + \gamma_3 \text{Time-invariant controls}_i + \epsilon_{im}\end{aligned} \tag{1}$$

where Alpha_{im} is fund alpha for fund i in month m. Monthly fund alpha is the difference between fund excess return and its Fung and Hsieh (2004) factor loadings multiplied by the factor realizations, where factor loadings are estimated over the last 24 months.[12] $\text{Text measures}_{im-1}$ are combinations of the text sophistication measures, namely, lexical diversity, syntactic complexity, and their scaled variants. The text measures are winsorized at the 1st and 99th percentiles and standardized to mean zero and unit standard deviation within investment styles.[13] We do so as strategy descriptions may vary systematically across investment styles. Time-varying controls_{im-1} include the logarithm of lagged fund AUM (Berk and Green, 2004) and lagged fund age in years (Aggarwal and Jorion, 2010). Time-invariant controls_i include (i) share restrictions such as the sum of the redemption and notice periods (Aragon, 2007) and the lockup dummy, (ii) compensation structure variables, such as management fee, performance fee, and the high-water mark dummy, to capture fund incentives (Agarwal, Daniel, and Naik, 2009), and (iii) the leverage dummy. All regression specifications feature year and strategy fixed effects. Statistical inferences are based on robust standard errors that are clustered by fund and month. We also estimate analogous regressions on fund monthly excess return to ensure that our findings are not driven by the risk-adjustment methodology.

The regression estimates reported in Table 4 corroborate the findings from the portfolio sorts. They indicate that after controlling for other factors that can drive fund performance, lexical diversity is positively associated with both fund excess return and fund alpha. The

[12]Inferences remain qualitatively unchanged when we construct monthly alpha using factor loadings estimated over the last 36 months.

[13]We obtain similar results when we (i) do not standardize within styles or winsorize, or (ii) do not standardize within styles and only winsorize over the entire fund sample.

relationship between lexical diversity is economically meaningful and statistically significant at the 1% level. The coefficient estimates from column 6 of Table 4 suggest that an increase in lexical diversity from the bottom 10th percentile to the top 10th percentile, i.e., a 3.50 unit increase, engenders a 2.85% increase in annualized fund alpha (t-statistic = 5.86). In contrast, the relationship between syntactic complexity and fund performance while positive, is only statistically significant at the 10% level, and changes sign once we control for lexical diversity. Inferences do not change when we scale lexical diversity and syntactic complexity by text length. The coefficient estimates on the other fund variables echo the prior literature. They suggest that larger (Berk and Green, 2004) and older (Aggarwal and Jorion, 2010) funds underperform while funds with lock-ups and longer redemption and notice periods outperform (Aragon, 2007).

[Insert Tables 4 and 5 here]

To ensure that our fund performance results are not driven by fund leverage or fund manager manipulation, we estimate analogous regressions on fund Sharpe ratio, information ratio, and MPPM (with $\rho = 3$). These performance measures are computed over each non-overlapping 24-month period.[14] The results showcased in Table 5 indicate that the relationship between lexical diversity and fund performance extends to these performance measures as well. Fund lexical diversity is positively and statistically related to fund Sharpe ratio, information ratio, and MPPM. Specifically, after controlling for syntactic complexity, an increase in lexical diversity from the bottom 10th to the top 10th percentile, i.e., by 3.50 units, is associated with a 0.41, 0.79, and 2.12 increase in annualized Sharpe ratio, annualized information ratio, and MPPM, respectively. These results further bolster the view that lexical diversity is indicative of managerial skill.

[14]For example, if a fund first reports returns in June 1994, the first observation for Sharpe ratio is derived from the June 1994 to May 1996 period, the second observation is derived from the June 1996 to May 1998 period, etc. The results are robust to using non-overlapping two-calendar year periods instead.

3.2. Fund investment risk

Are the fund managers who craft lexically diverse text also more judicious when managing risk? Since bearers of idiosyncratic risk forgo systematic risk premia and bearers of tail risks could face significant drawdowns and sudden fund termination (Duarte, Longstaff, and Yu, 2007), we postulate that the high quality managers who compose lexically diverse text will tend to eschew idiosyncratic risk and tail risk. To test whether lexically diversity is related to fund risk taking, we estimate the following multivariate regression on fund risk:

$$\begin{aligned}\text{Risk measure}_{im,m+23} = \gamma_0 + \gamma_1 \text{Text measures}_{im-1} + \gamma_2 \text{Time-varying controls}_{im-1} \\ + \gamma_3 \text{Time-invariant controls}_i + \epsilon_{im} \end{aligned} \tag{2}$$

where Risk measure$_{im,m+23}$ is fund idiosyncratic or tail risk for fund i estimated from month m to month $m+23$, and the rest of the variables as per defined in Eq. (1). Fund idiosyncratic risk is fund residual volatility or the standard deviation of fund monthly residuals from the Fung and Hsieh (2004) model estimated over each non-overlapping 24-month period. We employ three proxies for tail risk: (i) the maximum monthly loss, (ii) the maximum drawdown, and (iii) the Agarwal, Ruenzi, and Weigert (2017) tail risk measure, all estimated over non-overlapping 24-month periods. The Agarwal, Ruenzi, and Weigert (2017) tail risk measure is the conditional probability that a hedge fund has its two worst individual return realizations exactly when the equity market also has its two worst return realizations in a 24-month period, scaled by the absolute value of their expected shortfalls. All regression specifications feature year and strategy fixed effects. Statistical inferences are based on robust standard errors that are clustered by fund and year.

[Insert Table 6 here]

The coefficient estimates reported in Table 6 suggest that fund managers who compose lexically diverse text tend to avoid idiosyncratic risk and tail risk. The coefficient estimates

on lexical diversity in the regressions on fund residual volatility, maximum monthly loss, maximum drawdown, and the Agarwal, Ruenzi, and Weigert (2017) tail risk measure are all negative and statistically significant at the 1% level. For example, column 9 of Table 6 indicates that, after controlling for syntactic complexity, an increase in lexical diversity from the bottom 10th to the top 10th percentile translates into a 2.33% decrease in maximum drawdown.

3.3. Fund disciplinary disclosures

In this section, we examine whether our two measures of text sophistication, based on fund strategy descriptions, are related to the regulatory, civil, or criminal violations reported by hedge funds. Lexical diversity is associated with honesty (Humphreys, 2010; Horne and Adali, 2017) while syntactic complexity is linked to deception (Moffitt and Burns, 2009; Levitan, 2019). Therefore, we hypothesize that lexically diverse funds report fewer violations while syntactically complex funds trigger more violations.

We study hedge fund disciplinary disclosures that we gather from mandatory regulatory filings. The Securities and Exchange Commission (henceforth SEC) requires all hedge funds exceeding US$100m of AUM to register and file a Form ADV. The Form ADV has to be updated at least annually. However, certain updates such as regulatory violations must be reported promptly as soon as material changes occur. As per Dimmock and Gerken (2012), we use historical filings available at Historical Archive of Investment Advisor Reports filed from August 2001 through June 2017. We are able to obtain Form ADV filings information for 9,789 of the hedge funds in our sample.

Data on fund regulatory, civil, and criminal violations are culled from Item 11 of the Form ADV. There are three categories of violations on the Form ADV: Criminal Action (Items 11.A–11.B), Regulatory Action (Items 11.C–11.G), and Civil Judicial Action (Item 11.H). We define the indicator variable Violation as that which takes a value of one when a fund reports a regulatory, civil, or criminal violation, and is zero otherwise. Following

Dimmock and Gerken (2012), we also create two broad indicator variables: RegViolation and CivilCriminalViolation. RegViolation takes a value of one when a fund reports a regulatory violation, and is zero otherwise. CivilCriminalViolation takes a value of one when a fund reports a civil or criminal violation, and is zero otherwise. To check that our results are not driven by non investment related violations such as drunk driving or drug usage, we form an indicator variable, InvestmentViolation, that takes a value of one if the violation is investment related, and is zero otherwise. To verify that our results also apply to severe violations such as fraud and felonies, we form an indicator variable, SevereViolation, that takes a value of one if the violation is severe, and is zero otherwise.[15]

Item 11 on the Form ADV asks whether an advisor had committed a violation within the past ten years. Therefore, it is important to control for look-ahead bias when analyzing Form ADV violations. To do so, we leverage on Form ADV Disclosure Reporting Pages (henceforth DRPs). Any affirmative response to Item 11 on the Form ADV must be accompanied by a DRP, which details the first date and last date for each violation. The first date corresponds to the initiation date for regulatory violations, the first filing date for civil violations, and the first charging date for criminal violations, while the last date corresponds to the resolution date for the case. For each fund-year observation, the violation variables take a value of one if and only if the year overlaps with the date range for the specific violation.

To investigate the relationship between Form ADV violations and the text sophistication measures, we estimate the following probit regression:

$$\begin{aligned}\text{Violate}_{it} = \gamma_0 + \gamma_1 &\text{Text measures}_{it-1} + \gamma_2 \text{Time-varying controls}_{it-1} \\ &+ \gamma_3 \text{Time-invariant controls}_i + \epsilon_{it} \end{aligned} \tag{3}$$

where Violate_{it} is a placeholder for one of the five violation indicator variables discussed

[15]Specifically, InvestmentViolation equals one if the management company answers "Yes" to any of the questions in Items 11.B.1., 11.C.3, 11.C.4, 11.C.5, 11.D.2, 11.D.3, 11.D.4, 11.D.5, 11.E.3, 11.H.1a, and 11.H.1b, and is zero otherwise. SevereViolation equals one if the management company answers "Yes" to any of the questions in Items 11.A.1, 11.A.2, 11.C.4, 11.C.5, 11.D.4, or 11.D.5, and is zero otherwise.

above, and the rest of the variables as per defined in Eq. (1). All regression specifications feature year and strategy fixed effects. Statistical inferences are based on robust standard errors that are clustered by fund and year.

[Insert Table 7 here]

The results from Table 7 support the view that lexical diversity is related to trustworthiness while syntactic complexity is associated with deception. The coefficient estimates reported in Panel A of column 3 of Table 7 indicate that the likelihood that a hedge fund reports a violation is negatively linked to lexical diversity and positively linked to syntactic complexity. The marginal effects imply that an increase in lexical diversity from the bottom 10th to the top 10th percentile engenders a 3.85% decrease in the probability of triggering a fresh violation. Conversely, a similar percentile increase in syntactic complexity translates to a 4.87% increase in the probability of reporting a new violation. These numbers are economically meaningful as the unconditional probability that a fund triggers a violation in any given year is 7.60%. Similar results apply to regulatory violations, civil or criminal violations, investment violations, and severe violations.[16] Moreover, Panel B of Table 7 indicates that we obtain similar inferences when we analyze variants of the text sophistication measures that are scaled by text length.

3.4. Fund investor response

Do fund investors understand the discrepant implications of lexical diversity and syntactic complexity on fund performance and quality? In this section, we analyze hedge fund flow and test whether fund investors respond to information embedded in fund strategy descriptions.

[16]For example, an increase in syntactic complexity from the bottom 10th to the top 10th percentile is associated with a 4.18% increase in the probability of regulatory actions, a 3.48% increase in the probability of investment related infractions, and a 2.09% increase in the probability of severe violations.

Specifically, we estimate the following multivariate regression on fund flow:

$$\begin{aligned}\text{Flow}_{it} = \gamma_0 + \gamma_1 \text{Text measures}_{it-1} + \gamma_2 \text{Rank}_{it-1} \\ + \gamma_3 \text{Time-varying controls}_{it-1} + \gamma_4 \text{Time-invariant controls}_i + \epsilon_{it}\end{aligned} \tag{4}$$

where Flow_{it} is fund flow for fund i in year t. Rank_{it-1} is past-year fund performance rank derived from fund return in the spirit of Siri and Tufano (1998). The other variables are as per Eq. (1). We also report flow regressions with fund performance rank derived from CAPM alpha and Fung and Hsieh (2004) alpha since fund investors may respond more to fund alpha than to fund return (Agarwal, Green, and Ren, 2018)

[Insert Table 8 here]

The results reported in Table 8 indicate that investors gravitate toward hedge funds with lexically diverse strategy descriptions. The coefficient estimates on lexical diversity in the flow regressions are positive and statistically significant at the 1% level regardless of whether we control for syntactic complexity. Specifically, the coefficient estimate on lexical diversity reported in column 3 of Panel A, Table 8 indicate that after controlling for syntactic complexity, past fund return rank, and other fund variables, an increase in lexical diversity from the bottom 10th to the top 10th percentile elicits a 14.52 percent increase in fund flow.

The explanatory power of syntactic complexity on fund flow is more mixed. The coefficient estimate on syntactic complexity is positive and statistically significant at the 10% level, when we do not control for lexical diversity. However, once we adjust for lexical diversity, the coefficient estimate on syntactic complexity turns negative. This suggests that investors park incremental capital with syntactically complex hedge funds only to the extent that syntactic complexity covaries with lexical diversity. Once we orthogonalize syntactic complexity to lexical diversity, we find that investors eschew syntactically complex funds that are not also lexically diverse.

Columns 4 to 9 of Table 8 reveal that the findings are robust when we control for fund rank estimated using fund CAPM alpha or Fung and Hsieh (2004) alpha. The coefficient estimates on lexical diversity are positive and statistically reliable regardless of which performance rank variable we use as a control. As shown in Panel B of Table 8, the findings are also qualitatively similar when we scale the text measures by text length. Collectively, the results suggest that investors react correctly, but not fully, to the information on fund quality and trustworthiness that is embedded in hedge fund strategy descriptions.

4. Robustness tests

In this section, we conduct a battery of robustness tests to ascertain the strength of our empirical results.

4.1. Strategy distinctiveness

One concern is that funds with complicated strategy descriptions may engage in more distinctive strategies, which Sun, Wang, and Zheng (2012) show outperform. Therefore, strategy distinctiveness may explain the outperformance of funds with lexically diverse strategy descriptions. To allay such concerns, we redo our baseline results after controlling for the Sun, Wang, and Zheng (2012) strategy distinctiveness index (henceforth SDI), which measures the extent to which a fund's return differs from those of its peers. The results reported in Panel A of Table 9 indicate that inferences remain unchanged when we adjust for the explanatory power of SDI.

[Insert Table 9 here]

4.2. Unconventional strategies

A related concern is that funds with complicated strategy descriptions may pursue more unconventional strategies that cannot be explained by standard risk factor models. Titman

and Tiu (2011) argue that such funds tend to outperform. To adjust for unconventional strategies, we control for fund R^2 estimated over the prior 36 months in our baseline performance and disciplinary disclosure regressions. The findings reported in Panel B of Table 9 suggest that our results are not driven by unconventional strategies.

4.3. Manager education

Fund managers who have received multiple years of education or attended higher quality schools may be better placed to craft lexically diverse and syntactically complex text. To control for education level, we leverage on information from fund manager biographies provided by BarclayHedge, Eurekahedge, and eVestment, and construct two indicator variables for whether the manager has obtained a master's or doctoral degree.[17] To control for education quality, we determine the median SAT score of the undergraduate college attended by the fund manager for those who attended U.S. undergraduate institutions as per Chevalier and Ellison (1999).[18] Next, we redo our baseline regressions after controlling separately for manager education level and undergraduate college quality. The results reported in Panels C and D of Table 9 reveal that education level and quality do not explain our findings.

4.4. Text readability

There may be concerns that our text sophistication measures may be subsumed by the readability measures used in the finance literature. To proxy for readability, we compute the Fog index and the Flesch Kincaid index for each strategy description. The Fog index equals to 0.4×(the average number of words in each sentence + the percentage of words with three or more syllables) and has been used by researchers to relate annual report

[17]HFR does not provide manager biographies. Therefore, we are able to conduct the education analysis for only a subset of managers. Note that the vast majority of managers have at least a bachelor's degree.

[18]We identify manager undergraduate institutions from manager biographies reported in hedge fund databases and from manager LinkedIn profiles. For each manager, once we have identified the undergraduate institution, we compute the median SAT score for that college from https://www.compassprep.com/college-profiles-new-sat/. We thank Yan Lu for kindly providing LinkedIn education data, which were collected manually from manager LinkedIn profiles.

readability to earnings persistence (Li, 2008) and analyst earnings forecasts (Lehavy, Li, and Merkley, 2011). The Flesch Kincaid index is $0.39\times$(Total number of words/total number of sentences)$+11.8\times$(total number of syllables/total number of words)-15.59 and has also been employed by Li (2008) and Lehavy, Li, and Merkley (2011). We also compute the Hwang and Kim (2017) readability measure, which is based on the number of passive verbs, hidden verbs, legal words, overwriting, wordy phrases, and abstract words. Like Hwang and Kim (2017), we use StyleWriter, a manuscript editing software, to determine the pervasiveness of these errors for each strategy description.[19] Next, we redo our baseline regressions after controlling for the Fog index, the Flesch Kincaid index, and the Hwang and Kim (2017) readability measure. The results showcased in Panels E, F, and G of Table 9 support the view that our findings are not driven by readability.

4.5. Native English speakers

Native English speakers may be better equipped to craft sophisticated text than are non-native English speakers. We proxy for native English speakers in two ways. First, we focus on funds operated by investment management companies that are based in the U.S. or the U.K. Second, we zero in on funds run by managers with English names (based on a Python algorithm). Next, we redo the baseline regressions for these two sets of funds. The results reported in Panels H and I of Table 9 indicate that our findings are not driven by differences in text sophistication between native and non-native English speakers.

4.6. Legal words

Yet another concern is that syntactic complexity may proxy for the number of legal words embedded in hedge fund strategy descriptions. To measure the number of legal terms, we

[19] As noted by Hwang and Kim (2017), the manual application of StyleWriter to a large number of documents can be very labor-intensive. For comparison, Hwang and Kim (2017) analyze 92 equity closed-end investment companies. To circumvent this issue, we write a Python program to automate the process of applying StyleWriter to our strategy descriptions.

rely on the dictionary of legal terms created by Loughran and McDonald (2011). Next, we reestimate our baseline regressions after controlling for legalese usage. The results showcased in Panel J of Table 9 reveal that our findings are not driven by legal words.

4.7. Look-ahead bias

Fund strategy descriptions may change over time. This induces a look-ahead bias when databases overwrite existing strategy descriptions and report the latest description. To ameliorate this bias, we focus on the 2007 to 2016 period where we have multiple snapshots of BarclayHedge, EurekaHedge, and HFR. Next, we redo the baseline regressions for that period and for funds from those databases. The coefficient estimates reported in Panel K of Table 9 suggest that inferences do not change when we adjust for look-ahead bias.

4.8. Backfill bias

In our baseline empirical analysis, we adjust for backfill and incubation bias by removing the first 12 months of returns for each fund. However, that may not completely remove backfill bias. Therefore, we remove all returns reported prior to the listing date for funds that report to HFR, which is the only database in our sample that provides data on listing dates. For funds that report to the other databases, we employ the Jorion and Schwarz (2019) algorithm to back out listing dates and remove the backfilled returns. The results reported in Panel L of Table 9 with the non-backfilled returns indicate that our findings are not driven by backfill bias.

4.9. Serial correlation

Serial correlation in fund returns could arise from linear interpolation of prices for illiquid and infrequently traded securities or the use of smoothed broker dealer quotes. This could inflate the test statistics used to derive inferences from our empirical tests. To adjust for serial

correlation, we unsmooth returns using the methodology of Getmansky, Lo, and Makarov (2004) and redo the baseline analysis on the unsmoothed returns. The results showcased in Panel M of Table 9 indicate that serial correlation does not drive our findings.

4.10. Omitted risk factors

To allay concerns that our findings may be driven by omitted risk factors, we redo the baseline regressions after augmenting the Fung and Hsieh (2004) model with an emerging markets factor derived from the MSCI Emerging Markets Index, the Pástor and Stambaugh (2003) liquidity risk factor, and the out-of-the-money equity call and put option factors from the Agarwal and Naik (2004) model. The results displayed in Panels N, O, and P of Table 9 indicate that the omitted risk factors do not explain our results.

5. Conclusion

This paper finds that two novel measures of text sophistication from the linguistics literature, namely lexical diversity and syntactic complexity, encapsulate useful information for hedge fund selection and for understanding the factors underlying alpha generation. By doing so, we make several contributions to the finance literature. First, we show that lexical diversity has implications for hedge fund performance. Funds with lexically diverse strategy descriptions deliver greater alphas, Sharpe ratios, information ratios, and manipulation proof performance measures than do funds with lexically homogeneous strategy descriptions. Second, lexically diverse funds display other attributes that are consonant with fund quality. They are more conservative when taking idiosyncratic risk and tail risk. Consequently, they feature lower residual volatilities, maximum monthly losses, and maximum drawdowns. Moreover, their worst return months are less likely to coincide with those of the equity market. Third, lexically diverse funds are more trustworthy. They face fewer regulatory actions, encounter fewer civil or criminal problems, and trigger fewer investment violations. Fourth, unlike

lexical diversity, syntactic complexity is associated with deception. Syntactically complex funds experience more regulatory actions, report more severe infractions, and are more likely to violate investment rules. Also unlike lexical diversity, syntactic complexity is not a reliable predictor of fund outperformance. Fifth, investors react correctly, but not fully, to the textual information on fund quality and trustworthiness that is embedded in hedge fund strategy descriptions. After controlling for past fund performance, we find that investors direct greater flows to lexically diverse funds than to syntactically complex funds. However, the capital that investors allocate to lexically diverse hedge funds is not sufficient to erode away their positive alphas.

These results illuminate the link between text sophistication and sophisticated investors. They indicate that the richness of the vocabulary employed by delegated portfolio managers provides an honest cue to investor sophistication. However, the complexity of the sentence structures used is less reliable as a signal of fund quality. We believe that this is because fraudulent managers are more likely to favor convoluted and complex sentences in an effort to obfuscate and confuse. At the same time, they are less likely to employ rich vocabulary when communicating to their investors so as not to divulge too much information and inadvertently reveal the inconsistencies in their fund operations. Therefore, the relation between text sophistication and investor sophistication is contingent on the specific aspect of writer sophistication captured by the textual sophistication measure. These findings are relevant for investment fiduciaries who allocate capital to hedge funds. They also underscore the importance of assessing manager writing sophistication when conducting operational due diligence in a delegated portfolio management setting.

References

Agarwal, V., Daniel, N., Naik, N. Y., 2009. Role of managerial incentives and discretion in hedge fund performance. Journal of Finance 64, 2221–2256.

Agarwal, V., Green, T.C., Ren, H., 2018. Alpha or beta in the eye of the beholder: what drives hedge fund flows? Journal of Financial Economics 127, 417–434.

Agarwal, V., Naik, N. Y., 2004. Risk and portfolio decisions involving hedge funds. Review of Financial Studies 17, 63–98.

Agarwal, V., Ruenzi, S., Weigert, F., 2017. Tail risk in hedge funds: a unique view from portfolio holdings. Journal of Financial Economics 125, 610–636.

Aggarwal, R. K., Jorion, P., 2010. The performance of emerging hedge funds and managers. Journal of Financial Economics 96, 238–256.

Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? The information content of Internet stock message boards. Journal of Finance 59, 1259–1293.

Aragon, G. 2007. Share restrictions and asset pricing: evidence from the hedge fund industry. Journal of Financial Economics 83, 33–58.

Aragon, G., Nanda, V., 2017. Strategic delays and clustering in hedge fund reported returns. Journal of Financial and Quantitative Analysis 52, 1–35.

Aragon, G., Strahan, P., 2012. Hedge funds as liquidity providers: evidence from the Lehman bankruptcy. Journal of Financial Economics 103, 570–587.

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

Bhardwaj, G., Gorton, G., Rouwenhorst K.G., 2014. Fooling some of the people all of the time: the inefficient performance and persistence of commodity trading advisors. Review of Financial Studies 27, 3099–3132.

Bodnaruk, A., Loughran, T., McDonald, B., 2015. Using 10-K text to gauge financial constraints. Journal of Financial and Quantitative Analysis 50, 623–646.

Bucks, R.S., Singh, S., Cuerden, J.M., Wilcock, G.K., 2000. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: evaluation of an objective technique for analysing lexical performance. Aphasiology 14, 71–91.

Buehlmaier, M.M., Whited, T.M., 2018. Are financial constraints priced? Evidence from textual analysis. Review of Financial Studies 31, 2693–2728.

Carroll, J.B., 1964. Language and Thought. Prentice Hall, New Jersey.

Chaudhuri, R., Ivković, Z., Pollet, J., Trzcinka, C., 2019. A tangled tale of training and talent: PhDs in institutional asset management. Unpublished working paper, Michigan State University.

Chevalier, J., Ellison, G., 1999. Are some mutual fund managers better than others? Cross-sectional patters in behavior and performance. Journal of Finance 54, 875–899.

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

Duarte, J., Longstaff, F., Yu, F., 2007. Risk and return in fixed income arbitrage: nickels in front of a steamroller? Review of Financial Studies 20, 769–811.

Fergadiotis, G., Wright, H.H., 2011. Lexical diversity for adults with and without aphasia across discourse elicitation tasks. Aphasiology 25, 1414–1430.

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, 840–852.

Freud, S., 1901. Psychopathology of Everyday Life. Basic Books, New York.

Fung, W., Hsieh, D., 2004. Hedge fund benchmarks: a risk based approach. Financial Analysts Journal 60, 65–80.

Fung, W., Hsieh, D., 2009. Measurement biases in hedge fund performance data: an update. Financial Analysts Journal 60, 36–38.

Gennaioli, N., Shleifer, A., Vishny, R., 2015. Money doctors. Journal of Finance 70, 91–114.

Getmansky, M., Lo, A., Makarov, I., 2004. An econometric model of serial correlation and illiquidity of hedge fund returns. Journal of Financial Economics 74, 529–610.

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

Gregoriou, G.N., Lhabitant, F.S., 2009. Madoff: a flock of red flags. Journal of Wealth Management 12, 89–97.

Hoberg, G., Maksimovic, V., 2015. Redefining financial constraints: a text-based analysis. Review of Financial Studies 28, 1312–1352.

Horne, B.D., Adali, S., 2017. This just in: fake new packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. Unpublished working paper, Rensselaer Polytechnic Institute

Humpherys, S.L., 2010. A system of deception and fraud detection using reliable linguistic cues including hedging, disfluencies, and repeated phrases. Ph.D. dissertation, University of Arizona.

Hwang, B.H., Kim, H.H., 2017. It pays to write well. Journal of Financial Economics 124, 373–394

Jagannathan, R., Malakhov, A., Novikov, D., 2010. Do hot hands exist among hedge fund managers? An empirical evaluation. Journal of Finance 65, 217–255.

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

Jegadeesh, N., Wu, D., 2013. Word power: a new approach for content analysis. Journal of Financial Economics 110, 712–729.

Joenväärä, J., Kauppila, M., Kosowski, R., Tolonen, P., 2019. Hedge fund performance: are stylized facts sensitive to which database one uses? Critical Finance Review, forthcoming.

Jorion, P., Schwarz, C., 2019. The fix is in: properly backing out backfill bias. Review of Financial Studies, forthcoming.

Ke, Z., Kelly, B., Xiu, D., 2019. Predicting returns with text data. Unpublished working paper, Yale University.

Lehavy, R., Li, F., Merkley, K., 2011. The effect of annual report readability on analyst following and the properties of their earnings forecasts. Accounting Review 86, 1087–1115.

Levitan, S.I., 2019. Deception in spoken dialogue: classification and individual differences. Ph.D. dissertation, Columbia University.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics 45, 221–247.

Li, H., Zhang, X., Zhao, R., 2011. Investing in talents: manager characteristics and hedge fund performances. Journal of Financial and Quantitative Analysis 46, 59–32.

Liang, B., 2000. Hedge funds: the living and the dead. Journal of Financial and Quantitative Analysis 35, 309–326.

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

Loughran, T., McDonald, B., 2013. IPO first-day returns, offer price revisions, volatility, and form S-1 language. Journal of Financial Economics 109, 307–326.

Loughran, T., McDonald, B., 2014. Measuring readability in financial disclosures. Journal of Finance 69, 1643–1671.

Loughran, T., McDonald, B., 2016. Textual analysis in accounting and finance: a survey.

Journal of Accounting Research 54, 1187–1230.

Lu, X., Ai, H., 2015. Syntactic complexity in college-level English writing: differences among writers with diverse L1 backgrounds. Journal of Second Language Writing 29, 16–27.

Moffitt, K., Burns, M.B., 2009. What does that mean? Investigating obfuscation and readability cues as indicators of deception in fraudulent financial reports. American Conference on Information Systems (AMCIS) 2009 Proceedings.

Napolitano, D., Sheehan, K.M., Mundkowsky, R., 2015. Online readability and text complexity analysis with *TextEvaluator*. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 96–100.

Newey, W., West, K., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Newman, M.L., Pennebaker, J.W., Berry, D.S., Richards, J.M., 2003. Lying words: predicting deception from linguistic styles. Personality and Social Psychology Bulletin 29, 665–675.

Pástor, L., Stambaugh, R., 2003. Liquidity risk and expected stock returns. Journal of Political Economy 111, 642–685.

Perold, A., Spitz, W.T., 1996. The Commonfund hedge fund portfolio. Harvard Business School Case 297-014.

Sampson, G., 1997. Depth in English grammar. Journal of Linguistics 33, 131–151.

Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal 27, 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, 1–20

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, 1–15.

Siri, E.R., Tufano, P., 1998. Costly search and mutual fund flows. Journal of Finance 53, 1589–1622.

Sun, Z., Wang, A., Zheng, L., 2012. The road less traveled: strategy distinctiveness and hedge fund performance. Review of Financial Studies 25, 96–143.

Tetlock, P.C., 2007. Giving content to investor sentiment: the role of media in the stock

market. Journal of Finance 62, 1139–1168.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: quantifying language to measure firms' fundamentals. Journal of Finance 63, 1437–1467.

Thordardottir, E.T., Namazi, M., 2007. Specific language impairment in French-speaking children: beyond grammatical morphology. Journal of Speech Language Research 50, 698–715.

Titman, S., Tiu, C., 2011. Do the best hedge funds hedge? Review of Financial Studies 24, 123–168.

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

Vrij, A., Granhag, P.A., Porter, S., 2010. Pitfalls and opportunities in non-verbal and verbal lie detection. Psychological Science in the Public Interest 11, 89–121.

Yngve, V.H., 1960. A model and an hypothesis for language structure. Proceedings of the American Philosophical Society 104, 444–466.

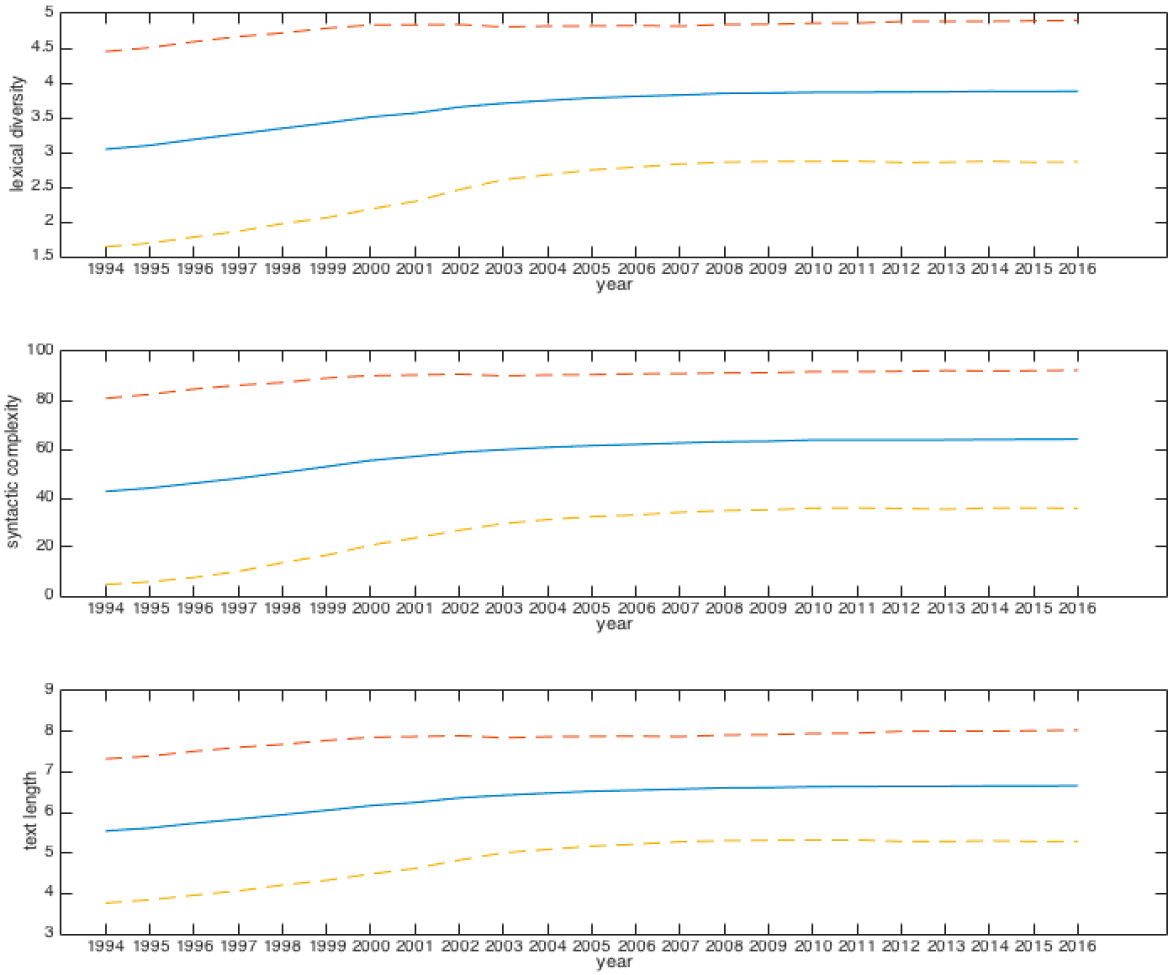

Fig 1. Text sophistication measures and text length over time. The solid line denotes the average lexical diversity, syntactic complexity, and text length of the strategy descriptions in our sample. The dashed lines denote the two-standard deviation upper and lower bounds. Lexical diversity measures the richness of the vocabulary used in the text. Lexical diversity is calculated as per Shannon (1948) and is based on the number of distinct tokens in a text as well as the frequencies at which those tokens appear. Syntactic complexity measures the complexity of the sentence structures used in the text. It encapsulates three dimensions of sentence complexity: paragraph length, sentence length, and word depth (Yngve, 1960). Text length is the natural logarithm of the number of characters in the text. The sample period is from January 1994 to December 2016.

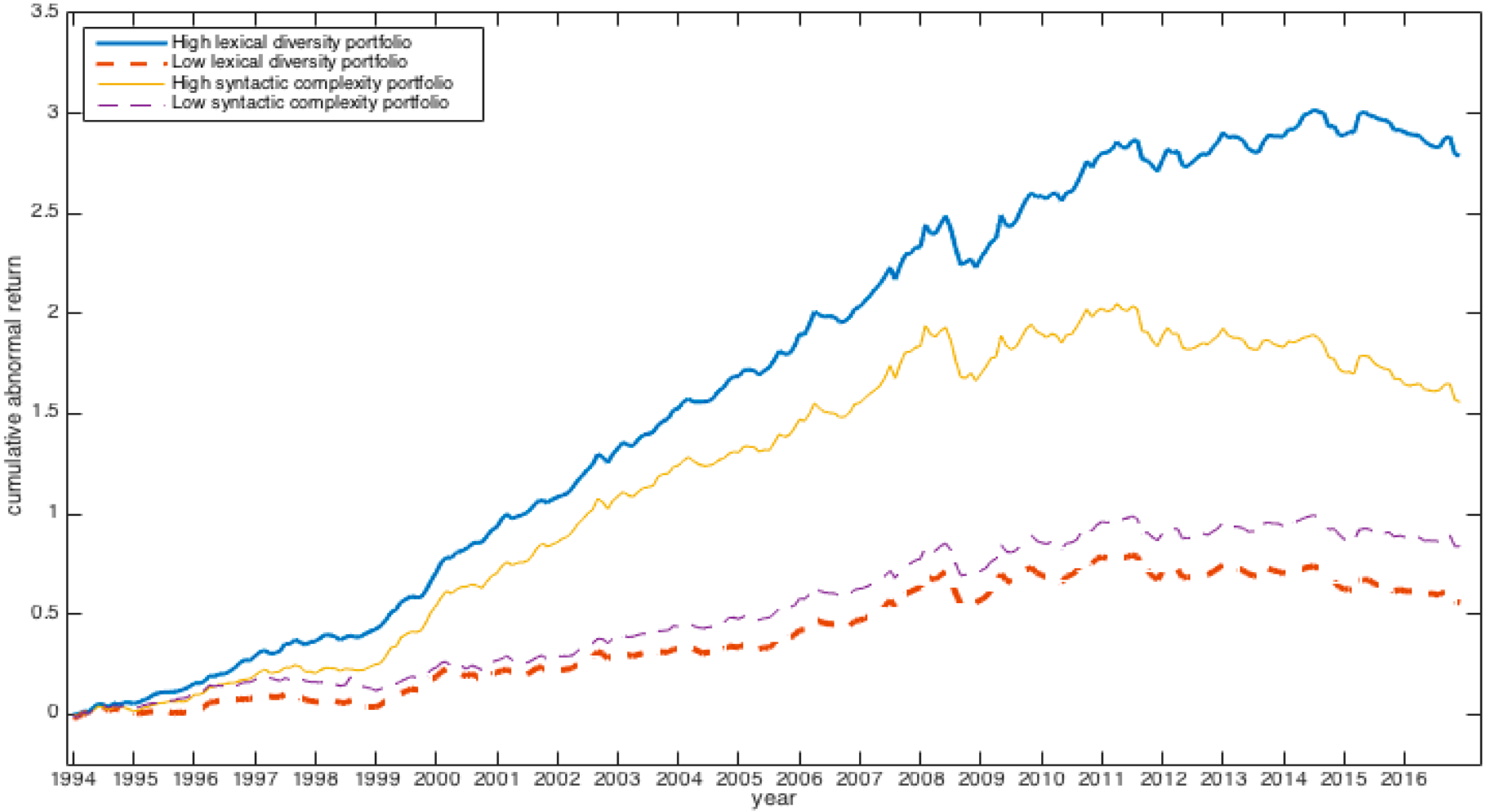

Fig. 2. Cumulative abnormal returns of hedge funds sorted by the lexical diversity and syntactic complexity of their strategy descriptions. Every January 1st, hedge funds are sorted into decile portfolios based on lexical diversity and syntactic complexity. The thick solid line denotes the extreme high lexical diversity portfolio. The thick dashed line denotes the extreme low lexical diversity portfolio. The thin solid line denotes the extreme high syntactic complexity portfolio. The thin dashed line denotes the extreme low syntactic complexity portfolio. Cumulative abnormal return is the difference between a portfolio's excess return and its factor loadings multiplied by the Fung and Hsieh (2004) risk factors, where factor loadings are estimated over the entire sample period. The sample period is from January 1994 to December 2016.

Table 1

Distribution of lexical diversity and syntactic complexity by hedge fund investment strategy and manager domicile

Lexical diversity measures the richness of the vocabulary used in the text. Lexical diversity is calculated as per Shannon (1948) and is based on the number of distinct tokens in a text as well as the frequencies at which those tokens appear. Syntactic complexity measures the complexity of the sentence structures used in the text. It encapsulates three dimensions of sentence complexity: paragraph length, sentence length, and word depth (Yngve, 1960). Security Selection funds take long and short positions in undervalued and overvalued securities, respectively. Usually, they take positions in equity markets. Multi-process funds employ multiple strategies that take advantage of significant events, such as spin-offs, mergers and acquisitions, bankruptcy reorganizations, recapitalizations, and share buybacks. Directional Trader funds bet on the direction of market prices of currencies, commodities, equities, and bonds in the futures and cash markets. Relative Value funds take positions on spread relations between prices of financial assets and aim to minimize market exposure. The sample period is from January 1994 to December 2016.

Investment strategy/manager domicile	Number of funds (1)	Mean (2)	Median (3)	Standard deviation (4)	Minimum (5)	25th Percentile (6)	75th Percentile (7)	Maximum (8)
Panel A: Lexical diversity								
Security Selection	5,561	3.75	3.84	0.62	1.10	3.44	4.14	5.85
Multi-process	3,962	3.71	3.82	0.63	1.10	3.41	4.13	5.39
Directional Trader	6,930	3.75	3.80	0.57	0.69	3.43	4.13	5.37
Relative Value	3,602	3.80	3.86	0.56	0.69	3.52	4.16	5.54
All Funds	20,055	3.75	3.83	0.59	0.69	3.45	4.14	5.85
Panel B: Syntactic complexity								
Security Selection	5,454	60.73	63.00	16.44	1.00	52.00	72.00	100.00
Multi-process	3,893	59.73	62.00	17.22	1.00	50.00	71.00	100.00
Directional Trader	6,827	60.55	63.00	15.83	1.00	51.00	71.00	100.00
Relative Value	3,525	62.67	65.00	15.47	1.00	55.00	73.00	100.00
All Funds	19,699	60.82	63.00	16.25	1.00	52.00	72.00	100.00
Panel C: Lexical diversity								
Caribbean	494	3.67	3.75	0.61	1.39	3.38	4.11	4.93
Europe	5,542	3.76	3.82	0.54	1.10	3.46	4.12	5.30
North America	11,548	3.74	3.83	0.63	0.69	3.43	4.15	5.85
Others	2,439	3.80	3.84	0.53	1.10	3.51	4.14	5.18
All Funds	20,023	3.75	3.83	0.59	0.69	3.45	4.14	5.85
Panel D: Syntactic complexity								
Caribbean	487	60.48	62.00	16.98	1.00	50.00	72.00	100.00
Europe	5,468	61.89	64.00	15.05	1.00	53.00	72.00	100.00
North America	11,366	59.99	63.00	16.95	1.00	51.00	71.00	100.00
Others	2,347	62.54	64.00	14.89	1.00	53.00	72.00	100.00
All Funds	19,668	60.83	63.00	16.23	1.00	52.00	72.00	100.00

Table 2

Portfolio sorts on lexical diversity

Hedge funds are sorted into ten portfolios based on the lexical diversity of fund strategy descriptions. Ret-rf is fund return in excess of the risk free rate. Alpha is Fung and Hsieh (2004) alpha. SR is fund Sharpe ratio. IR is fund information ratio. MPPM is fund manipulation proof performance measure with risk aversion parameter $\rho = 3$ (Goetzmann et al., 2007). RV is fund residual volatility or standard deviation of fund residuals from the Fung and Hsieh (2004) regression. MaxLoss is maximum monthly loss over the entire sample period. The Fung and Hsieh (2004) factors are S&P 500 return minus risk free rate (SNPMRF), Russell 2000 return minus S&P 500 return (SCMLC), change in the constant maturity yield of the U.S. 10-year Treasury bond appropriately adjusted for the duration (BD10RET), change in the spread of Moody's BAA bond over 10-year Treasury bond appropriately adjusted for duration (BAAMTSY), bond PTFS (PTFSBD), currency PTFS (PTFSFX), and commodity PTFS (PTFSCOM), where PTFS is primitive trend following strategy. The t-statistics derived from Newey and West (1987) heteroskedasticity and autocorrelation consistent standard errors are in parentheses. The sample period is from January 1994 to December 2016.

Portfolio	Ret-rf	Alpha	SR	IR	MPPM	RV	MaxLoss	SNPMRF	SCMLC	BD10RET	BAAMTSY	PTFSBD	PTFSFX	PTFSCOM	Adj. R^2
Panel A: Sort on lexical diversity															
1 (High lexical diversity)	8.624	5.567	1.328	1.535	7.944	3.628	0.081	0.281	0.158	0.051	0.236	-0.007	0.009	0.005	0.680
2	8.109	5.079	1.227	1.335	7.412	3.806	0.076	0.293	0.152	0.050	0.218	-0.004	0.010	0.004	0.660
3	7.794	4.783	1.159	1.163	7.077	4.111	0.078	0.287	0.145	0.062	0.243	-0.002	0.015	0.008	0.616
4	7.392	4.247	1.120	1.062	6.703	3.998	0.084	0.278	0.142	0.097	0.259	-0.005	0.013	0.010	0.624
5	7.103	4.038	1.091	1.052	6.428	3.837	0.093	0.271	0.133	0.066	0.283	-0.002	0.010	0.007	0.643
6	6.577	3.382	0.983	0.874	5.873	3.872	0.091	0.282	0.150	0.070	0.269	-0.006	0.011	0.005	0.656
7	5.909	2.515	0.819	0.607	5.102	4.142	0.096	0.305	0.156	0.061	0.295	-0.005	0.011	0.004	0.661
8	5.097	1.720	0.708	0.411	4.300	4.188	0.089	0.306	0.159	0.068	0.288	-0.004	0.011	0.007	0.653
9	5.928	2.386	0.795	0.552	5.065	4.323	0.107	0.318	0.147	0.077	0.321	-0.002	0.012	0.007	0.655
10 (Low lexical diversity)	5.594	1.937	0.742	0.463	4.717	4.187	0.095	0.326	0.170	0.076	0.305	-0.007	0.011	0.009	0.684
Spread (1-10)	3.030	3.630	0.587	1.074	3.227	-0.558	-0.014	-0.045	-0.011	-0.024	-0.069	0.000	-0.002	-0.003	0.158
t-statistic	6.112	7.453	6.540	8.370	18.931	-2.544	-2.030	-3.201	-0.878	-1.114	-3.595	0.135	-0.788	-1.263	
Panel B: Sort on lexical diversity scaled by text length															
1 (High scaled lexical diversity)	8.566	5.537	1.300	1.467	7.872	3.774	0.070	0.293	0.152	0.052	0.224	-0.004	0.013	0.006	0.663
2	8.205	5.253	1.245	1.351	7.511	3.888	0.080	0.280	0.150	0.046	0.252	-0.002	0.014	0.006	0.643
3	7.320	4.453	1.195	1.203	6.722	3.701	0.070	0.267	0.122	0.074	0.224	-0.003	0.011	0.009	0.626
4	6.910	4.030	1.096	1.060	6.280	3.803	0.084	0.259	0.149	0.067	0.249	-0.001	0.008	0.005	0.627
5	6.789	3.759	1.049	0.985	6.129	3.815	0.083	0.264	0.151	0.067	0.271	-0.007	0.014	0.007	0.644
6	6.579	3.357	0.978	0.845	5.868	3.972	0.091	0.285	0.143	0.082	0.267	-0.006	0.011	0.008	0.642
7	6.075	2.769	0.876	0.699	5.326	3.959	0.093	0.294	0.155	0.069	0.286	-0.004	0.009	0.006	0.666
8	6.067	2.791	0.877	0.715	5.323	3.900	0.086	0.309	0.156	0.072	0.252	-0.002	0.012	0.005	0.674
9	5.917	2.678	0.851	0.647	5.168	4.137	0.083	0.291	0.141	0.064	0.289	-0.006	0.015	0.007	0.637
10 (Low scaled lexical diversity)	5.598	1.789	0.713	0.402	4.642	4.455	0.114	0.328	0.172	0.083	0.340	-0.008	0.011	0.008	0.670
Spread (1-10)	2.968	3.748	0.589	1.067	3.230	-0.679	-0.018	-0.035	-0.020	-0.031	-0.116	0.003	0.002	-0.002	0.238
t-statistic	4.606	7.043	8.673	8.925	14.872	-3.149	-1.991	-2.552	-1.967	-1.348	-3.949	1.014	0.799	-0.803	

Table 3

Portfolio sorts on syntactic complexity

Hedge funds are sorted into ten portfolios based on the syntactic complexity of fund strategy descriptions. Ret-rf is fund return in excess of the risk free rate. Alpha is Fung and Hsieh (2004) alpha. SR is fund Sharpe ratio. IR is fund information ratio. MPPM is fund manipulation proof performance measure with risk aversion parameter ρ = 3 (Goetzmann et al., 2007). RV is fund residual volatility or standard deviation of fund residuals from the Fung and Hsieh (2004) regression. MaxLoss is maximum monthly loss over the entire sample period. The Fung and Hsieh (2004) factors are S&P 500 return minus risk free rate (SNPMRF), Russell 2000 return minus S&P 500 return (SCMLC), change in the constant maturity yield of the U.S. 10-year Treasury bond appropriately adjusted for the duration (BD10RET), change in the spread of Moody's BAA bond over 10-year Treasury bond appropriately adjusted for duration (BAAMTSY), bond PTFS (PTFSBD), currency PTFS (PTFSFX), and commodity PTFS (PTFSCOM), where PTFS is primitive trend following strategy. The t-statistics derived from Newey and West (1987) heteroskedasticity and autocorrelation consistent standard errors are in parentheses. The sample period is from January 1994 to December 2016.

Portfolio	Ret-rf	Alpha	SR	IR	MPPM	RV	MaxLoss	SNPMRF	SCMLC	BD10RET	BAAMTSY	PTFSBD	PTFSFX	PTFSCOM	Adj. R^2
Panel A: Sort on syntactic complexity															
1 (High syntactic complexity)	7.311	3.707	0.955	0.869	6.389	4.264	0.102	0.323	0.176	0.048	0.294	-0.009	0.010	-0.001	0.682
2	7.841	4.455	1.085	1.095	7.013	4.069	0.101	0.312	0.163	0.057	0.288	-0.003	0.008	0.008	0.675
3	7.766	4.383	1.131	1.164	7.012	3.764	0.095	0.296	0.158	0.077	0.282	-0.007	0.010	0.005	0.692
4	7.165	3.970	1.055	1.025	6.437	3.874	0.084	0.295	0.135	0.051	0.273	-0.005	0.012	0.006	0.666
5	6.911	3.793	1.039	0.971	6.215	3.905	0.090	0.285	0.149	0.076	0.262	-0.002	0.012	0.005	0.646
6	6.589	3.467	1.000	0.899	5.907	3.857	0.089	0.279	0.131	0.065	0.276	-0.004	0.011	0.004	0.648
7	6.930	3.801	1.037	0.954	6.229	3.985	0.080	0.291	0.137	0.072	0.242	-0.006	0.014	0.008	0.635
8	6.300	3.308	0.959	0.823	5.631	4.017	0.075	0.276	0.170	0.088	0.217	-0.004	0.013	0.007	0.616
9	5.589	2.186	0.783	0.527	4.801	4.150	0.093	0.307	0.135	0.069	0.298	-0.005	0.012	0.011	0.653
10 (Low syntactic complexity)	5.591	2.505	0.831	0.622	4.894	4.027	0.081	0.274	0.155	0.077	0.294	0.000	0.012	0.011	0.632
Spread (1-10)	1.720	1.202	0.124	0.248	1.496	0.237	0.000	0.048	0.021	-0.029	0.000	-0.009	-0.003	-0.012	0.268
t-statistic	2.421	1.989	1.371	1.747	7.015	1.010	0.022	2.714	1.411	-1.454	0.005	-2.420	-1.267	-5.097	
Panel B: Sort on syntactic complexity scaled by text length															
1 (High scaled syntactic complexity)	6.580	2.864	0.843	0.668	5.627	4.285	0.109	0.331	0.180	0.053	0.301	-0.010	0.010	-0.001	0.690
2	7.236	3.576	0.959	0.873	6.340	4.095	0.100	0.334	0.162	0.068	0.303	-0.005	0.009	0.007	0.698
3	7.416	3.979	1.050	1.030	6.623	3.862	0.101	0.309	0.147	0.066	0.301	-0.004	0.009	0.006	0.693
4	7.027	3.773	1.016	0.941	6.273	4.009	0.092	0.291	0.148	0.062	0.287	-0.005	0.012	0.005	0.655
5	7.205	4.010	1.066	1.044	6.486	3.842	0.088	0.296	0.154	0.069	0.254	-0.003	0.010	0.005	0.668
6	7.026	3.919	1.053	1.011	6.324	3.878	0.083	0.286	0.147	0.060	0.254	-0.006	0.016	0.003	0.653
7	7.242	4.275	1.120	1.039	6.582	4.116	0.081	0.260	0.138	0.085	0.252	-0.004	0.011	0.010	0.584
8	6.642	3.775	1.067	1.035	6.034	3.647	0.068	0.272	0.146	0.063	0.219	-0.004	0.014	0.009	0.648
9	5.819	2.685	0.862	0.662	5.111	4.059	0.087	0.283	0.134	0.067	0.280	-0.003	0.011	0.010	0.629
10 (Low scaled syntactic complexity)	5.847	2.767	0.879	0.690	5.161	4.013	0.080	0.275	0.153	0.087	0.278	-0.001	0.012	0.011	0.627
Spread (1-10)	0.734	0.097	-0.036	-0.021	0.466	0.272	0.004	0.056	0.027	-0.034	0.023	-0.010	-0.002	-0.012	0.339
t-statistic	1.060	0.164	-0.435	-0.173	2.350	1.566	0.577	3.263	1.677	-1.825	1.025	-2.369	-1.004	-5.192	

Table 4

Multivariate regressions on fund performance

This table reports results from multivariate regressions on hedge fund performance. The dependent variables include hedge fund excess return and alpha. Excess return is the monthly hedge fund net-of-fee return minus the risk free rate. Alpha is the Fung and Hsieh (2004) seven-factor monthly alpha where factor loadings are estimated over the last 24 months. The primary independent variables of interest are lexical diversity, syntactic complexity, and their scaled equivalents. Scaled lexical diversity is lexical diversity divided by text length. Scaled syntactic complexity is defined analogously. The other independent variables include fund characteristics such as the logarithm of last month fund AUM in US$m, fund age in years, share restriction period in years which equals to the sum of the redemption period and notice period, lockup dummy, management fee as a proportion of AUM, incentive fee as a proportion of AUM, high-water mark dummy, leverage dummy, as well as dummy variables for year and fund investment strategy. The t-statistics, derived from robust standard errors that are clustered by fund and month, are in parentheses. The sample period is from January 1994 to December 2016. * Significant at the 5% level; ** Significant at the 1% level.

	Dependent variables											
	Fund Excess Return			Fund Alpha			Fund Excess Return			Fund Alpha		
Independent variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Lexical Diversity	0.041**		0.043**	0.058**		0.068**						
	(3.54)		(2.86)	(5.91)		(5.86)						
Syntactic Complexity		0.018	-0.004		0.016	-0.018						
		(1.95)	(-0.31)		(1.95)	(-1.90)						
Scaled Lexical Diversity							0.035**		0.035**	0.050**		0.051**
							(3.33)		(3.23)	(5.69)		(5.68)
Scaled Syntactic Complexity								0.004	0.001		-0.005	-0.009
								(0.38)	(0.10)		(-0.65)	(-1.13)
Log(AUM)	-0.054**	-0.053**	-0.054**	-0.044**	-0.043**	-0.043**	-0.053**	-0.053**	-0.053**	-0.043**	-0.042**	-0.043**
	(-5.76)	(-5.71)	(-5.73)	(-5.97)	(-5.84)	(-5.90)	(-5.72)	(-5.67)	(-5.70)	(-5.91)	(-5.79)	(-5.85)
Age	-0.009**	-0.009**	-0.009**	-0.011**	-0.012**	-0.011**	-0.009**	-0.010**	-0.009**	-0.012**	-0.012**	-0.012**
	(-3.16)	(-3.30)	(-3.14)	(-3.75)	(-4.12)	(-3.86)	(-3.34)	(-3.49)	(-3.29)	(-3.97)	(-4.34)	(-4.04)
Restriction period	0.251*	0.241*	0.247*	0.268**	0.256**	0.267**	0.254*	0.243*	0.250*	0.273**	0.260**	0.270**
	(2.40)	(2.32)	(2.39)	(3.22)	(3.12)	(3.24)	(2.43)	(2.35)	(2.42)	(3.27)	(3.16)	(3.27)
Lockup Dummy	0.080*	0.083*	0.081*	0.067**	0.071**	0.068**	0.080*	0.084*	0.081*	0.067**	0.072**	0.068**
	(2.38)	(2.49)	(2.41)	(2.59)	(2.74)	(2.62)	(2.39)	(2.51)	(2.42)	(2.61)	(2.77)	(2.63)
Management Fee	1.663	1.714	1.631	3.691	3.832	3.702	1.613	1.731	1.586	3.617	3.854	3.641
	(0.59)	(0.61)	(0.59)	(1.52)	(1.60)	(1.55)	(0.57)	(0.62)	(0.57)	(1.49)	(1.61)	(1.52)
Incentive Fee	-0.420	-0.419	-0.433	0.941**	0.942**	0.919**	-0.434	-0.428	-0.444	0.920**	0.928**	0.905**
	(-0.79)	(-0.79)	(-0.82)	(4.06)	(4.05)	(3.98)	(-0.82)	(-0.81)	(-0.84)	(3.98)	(4.01)	(3.93)
High-Water Mark Dummy	0.167**	0.178**	0.168**	0.100**	0.115**	0.100**	0.169**	0.180**	0.170**	0.102**	0.117**	0.102**
	(3.88)	(4.20)	(3.85)	(3.23)	(3.69)	(3.19)	(3.89)	(4.22)	(3.88)	(3.28)	(3.73)	(3.26)
Leverage Dummy	-0.002	-0.002	-0.001	0.039*	0.039*	0.039*	-0.002	-0.002	-0.002	0.039*	0.038*	0.039*
	(-0.08)	(-0.07)	(-0.06)	(2.09)	(2.06)	(2.09)	(-0.09)	(-0.08)	(-0.07)	(2.07)	(2.05)	(2.07)
Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Strategy Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Adj. R^2	0.024	0.024	0.024	0.011	0.011	0.011	0.024	0.024	0.024	0.011	0.011	0.011
Number of Observations	921,155	913,523	913,523	912,621	905,190	905,190	921,155	913,523	913,523	912,621	905,190	905,190

Table 5

Multivariate regressions on fund performance measures

This table reports results from multivariate regressions on hedge fund performance measures. The dependent variables include hedge fund Sharpe ratio, information ratio, and MPPM. Sharpe ratio is mean fund excess return divided by standard deviation of fund returns. Information ratio is mean fund abnormal return divided by standard deviation of fund residuals from the Fung and Hsieh (2004) regression. MPPM is fund manipulation proof performance measure with risk aversion parameter $\rho = 3$ (Goetzmann et al., 2007). All performance measures are measured over non-overlapping 24-month periods. The primary independent variables of interest are lexical diversity, syntactic complexity, and their scaled equivalents. Scaled lexical diversity is lexical diversity divided by text length. Scaled syntactic complexity is defined analogously. The other independent variables include fund characteristics such as the logarithm of last year's fund AUM, fund age in years, share restriction period in years which equals to the sum of the redemption period and notice period, lockup dummy, management fee as a proportion of AUM, incentive fee as a proportion of AUM, high-water mark dummy, leverage dummy, as well as dummy variables for year and fund investment strategy. The t-statistics, derived from robust standard errors that are clustered by fund and year, are in parentheses. The sample period is from January 1994 to December 2016. * Significant at the 5% level; ** Significant at the 1% level.

	Dependent variables											
	Sharpe ratio				Information ratio				MPPM			
Independent variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Lexical Diversity	0.053**	0.034*			0.082**	0.065**			0.580**	0.607**		
	(3.63)	(2.15)			(3.94)	(3.12)			(3.65)	(3.77)		
Syntactic Complexity		0.029**				0.023				-0.076		
		(2.08)				(1.32)				(-0.41)		
Scaled Lexical Diversity			0.040**	0.035**			0.056**	0.052**			0.509**	0.496**
			(3.08)	(2.80)			(3.36)	(3.27)			(4.03)	(3.85)
Scaled Syntactic Complexity				0.028*				0.024				0.018
				(2.49)				(1.67)				(0.11)
Log(AUM)	-0.006	-0.006	-0.006	-0.006	-0.025	-0.025	-0.024	-0.024	-0.140	-0.140	-0.135	-0.137
	(-0.63)	(-0.64)	(-0.57)	(-0.60)	(-1.60)	(-1.60)	(-1.56)	(-1.57)	(-1.18)	(-1.18)	(-1.13)	(-1.15)
Age	-0.015**	-0.014**	-0.016**	-0.015**	-0.015*	-0.014*	-0.016*	-0.015*	-0.089**	-0.090**	-0.094**	-0.094**
	(-3.97)	(-3.89)	(-4.19)	(-4.07)	(-2.18)	(-2.10)	(-2.38)	(-2.27)	(-2.99)	(-3.01)	(-3.17)	(-3.15)
Restriction period	0.941**	0.902**	0.944**	0.906**	1.284**	1.239**	1.288**	1.243**	3.590**	3.570**	3.644**	3.605**
	(4.31)	(4.42)	(4.31)	(4.42)	(4.41)	(4.55)	(4.40)	(4.54)	(3.67)	(3.68)	(3.72)	(3.70)
Lockup Dummy	0.146*	0.155**	0.147*	0.155*	0.202*	0.213*	0.204*	0.214*	0.078	0.098	0.080	0.099
	(2.28)	(2.34)	(2.29)	(2.35)	(2.15)	(2.23)	(2.16)	(2.24)	(0.17)	(0.21)	(0.17)	(0.21)
Management Fee	-5.499	-5.530	-5.536	-5.584	1.044	1.096	1.007	1.048	-87.500**	-86.225**	-88.237**	-86.922**
	(-1.84)	(-1.81)	(-1.86)	(-1.83)	(0.23)	(0.25)	(0.23)	(0.24)	(-2.69)	(-2.64)	(-2.72)	(-2.67)
Incentive Fee	-1.197**	-1.204**	-1.213**	-1.218**	0.744	0.724	0.722	0.705	-7.098	-7.201	-7.297	-7.337
	(-2.80)	(-2.80)	(-2.83)	(-2.82)	(1.30)	(1.29)	(1.27)	(1.26)	(-1.02)	(-1.04)	(-1.05)	(-1.06)
High-Water Mark Dummy	0.174**	0.171**	0.177**	0.173**	0.213**	0.208**	0.220**	0.213**	2.345**	2.339**	2.360**	2.356**
	(3.29)	(3.15)	(3.31)	(3.14)	(2.80)	(2.68)	(2.82)	(2.68)	(3.73)	(3.58)	(3.68)	(3.56)
Leverage Dummy	-0.015	-0.005	-0.015	-0.005	0.004	0.015	0.004	0.014	0.225	0.232	0.221	0.228
	(-0.35)	(-0.12)	(-0.36)	(-0.12)	(0.08)	(0.28)	(0.07)	(0.27)	(0.60)	(0.61)	(0.59)	(0.60)
Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Strategy Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Adj. R^2	0.064	0.066	0.064	0.066	0.072	0.074	0.072	0.074	0.135	0.135	0.135	0.135
Number of Observations	32,128	31,906	32,128	31,906	32,128	31,906	32,128	31,906	32,128	31,906	32,128	31,906

Table 6

Multivariate regressions on fund risk

This table reports results from multivariate regressions on hedge fund risk measures. The dependent variables include hedge fund idiosyncratic risk, maximum loss, maximum drawdown, and Agarwal, Ruenzi, and Weigert (2017) tail risk. Idiosyncratic risk is standard deviation of monthly residuals from the Fung and Hsieh (2004) model. Maximum loss is maximum monthly loss. Maximum drawdown is maximum cumulative loss. The Agarwal, Ruenzi, and Weigert (2017) tail risk measure is is the conditional probability that a hedge fund has its two worst individual return realizations exactly when the equity market also has its two worst return realizations in a specific period, scaled by the absolute value of their expected shortfalls. All performance measures are measured over non-overlapping 24-month periods. The primary independent variables of interest are lexical diversity, syntactic complexity, and their scaled equivalents. Scaled lexical diversity is lexical diversity divided by text length. Scaled syntactic complexity is defined analogously. The other independent variables include fund characteristics such as the logarithm of last year's fund AUM, fund age in years, share restriction period in years which equals to the sum of the redemption period and notice period, lockup dummy, management fee as a proportion of AUM, incentive fee as a proportion of AUM, high-water mark dummy, leverage dummy, as well as dummy variables for year and fund investment strategy. The coefficient estimates on the fund control variables are omitted for brevity. The t-statistics, derived from robust standard errors that are clustered by fund and year, are in parentheses. Panel A reports regressions with lexical diversity and syntactic complexity as the primary independent variables of interest. Panel B reports regressions with scaled lexical diversity and scaled syntactic complexity as the primary independent variables of interest. The sample period is from January 1994 to December 2016. * Significant at the 5% level; ** Significant at the 1% level.

	Dependent variables											
	Idiosyncratic risk			Maximum loss			Maximum drawdown			ARW (2017) Tail risk		
Independent variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: Regressions with raw lexical diversity and syntactic complexity												
Lexical Diversity	-0.254**		-0.285**	-0.276**		-0.336**	-0.578**		-0.665**	-1.675**		-2.139**
	(-3.15)		(-3.04)	(-3.77)		(-4.29)	(-3.90)		(-4.36)	(-2.98)		(-3.16)
Syntactic Complexity		-0.073	0.074		-0.045	0.128		-0.155	0.188		-0.139	0.964*
		(-0.79)	(0.70)		(-0.48)	(1.27)		(-0.91)	(1.11)		(-0.37)	(2.15)
Fund Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Strategy Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Adj. R^2	0.134	0.133	0.134	0.145	0.144	0.145	0.177	0.176	0.177	0.105	0.104	0.104
Number of Observations	32,128	31,906	31,906	32,128	31,906	31,906	32,128	31,906	31,906	32,128	31,906	31,906
Panel B: Regressions with scaled lexical diversity and syntactic complexity												
Scaled Lexical Diversity	-0.227**		-0.224**	-0.272**		-0.273**	-0.494**		-0.491**	-1.517**		-1.538**
	(-2.86)		(-2.84)	(-4.14)		(-4.17)	(-4.02)		(-4.03)	(-2.63)		(-2.62)
Scaled Syntactic Complexity		0.005	0.023		0.038	0.060		0.031	0.071		0.457	0.581
		(0.06)	(0.26)		(0.43)	(0.70)		(0.21)	(0.49)		(1.31)	(1.66)
Fund Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Strategy Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Adj. R^2	0.134	0.133	0.134	0.145	0.144	0.145	0.176	0.176	0.177	0.105	0.104	0.104
Number of Observations	32,128	31,906	31,906	32,128	31,906	31,906	32,128	31,906	31,906	32,128	31,906	31,906

Table 7

Multivariate probit regressions on fund disciplinary disclosure

This table reports results from multivariate probit regressions on the probability that hedge funds report violations on their Form ADVs. The dependent variables include Violation, RegViolation, CivilCriminalViolation, InvestmentViolation, and SevereViolation. Violation is an indicator variable that takes a value of one if a fund reports any regulatory, civil, or criminal violation, and is zero otherwise. RegViolation is an indicator variable that takes a value of one if a fund reports a regulatory violation, and is zero otherwise. CivilCriminalViolation is an indicator variable that takes a value of one if a fund reports a civil or criminal violation, and is zero otherwise. InvestmentViolation is an indicator variable that takes a value of one when a fund reports an investment related violation, and is zero otherwise. SevereViolation is an indicator variable that takes a value of one when a fund reports a severe violation, and is zero otherwise. Scaled lexical diversity is lexical diversity divided by text length. Scaled syntactic complexity is defined analogously. The other independent variables include fund characteristics such as the logarithm of last year's fund AUM, fund age in years, share restriction period in years which equals to the sum of the redemption period and notice period, lockup dummy, management fee as a proportion of AUM, incentive fee as a proportion of AUM, high-water mark dummy, leverage dummy, as well as dummy variables for year and fund investment strategy. The coefficient estimates on the fund control variables are omitted for brevity. The t-statistics, derived from robust standard errors that are clustered by fund, are in parentheses. The marginal effects are in brackets. Panel A reports regressions with lexical diversity and syntactic complexity as the primary independent variables of interest. Panel B reports regressions with scaled lexical diversity and scaled syntactic complexity as the primary independent variables of interest. The sample period is from January 1994 to December 2016. * Significant at the 5% level; ** Significant at the 1% level.

	Dependent variables														
	Violation			RegViolation			CivilCriminalViolation			InvestmentViolation			SevereViolation		
Independent variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Panel A: Regressions with raw lexical diversity and syntactic complexity															
Lexical Diversity	-0.030		-0.078**	-0.019		-0.066*	-0.042		-0.070	-0.013		-0.057*	-0.002		-0.044
	(-1.44)		(-3.03)	(-0.85)		(-2.45)	(-1.22)		(-1.86)	(-0.53)		(-2.07)	(-0.07)		(-1.28)
	[-0.004]		[-0.011]	[-0.002]		[-0.008]	[-0.002]		[-0.004]	[-0.001]		[-0.006]	[-0.000]		[-0.003]
Syntactic Complexity		0.063**	0.099**		0.069**	0.100**		0.022	0.055		0.070**	0.096**		0.064**	0.085**
		(3.27)	(4.53)		(3.29)	(4.38)		(0.71)	(1.67)		(3.29)	(4.28)		(2.59)	(3.42)
		[0.009]	[0.014]		[0.008]	[0.012]		[0.001]	[0.003]		[0.008]	[0.010]		[0.005]	[0.006]
Fund Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Strategy Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Pseudo R^2	0.023	0.023	0.025	0.020	0.021	0.022	0.007	0.007	0.007	0.022	0.023	0.024	0.014	0.014	0.014
Number of Observations	45,306	45,061	45,061	45,306	45,061	45,061	45,306	45,061	45,061	45,306	45,061	45,061	45,306	45,061	45,061
Panel B: Regressions with scaled lexical diversity and syntactic complexity															
Scaled Lexical Diversity	-0.065**		-0.066**	-0.055**		-0.056**	-0.046		-0.047	-0.048*		-0.048*	-0.038		-0.041
	(-3.50)		(-3.47)	(-2.93)		(-2.81)	(-1.74)		(-1.78)	(-2.48)		(-2.34)	(-1.60)		(-1.65)
	[-0.009]		[-0.009]	[-0.006]		[-0.007]	[-0.003]		[-0.003]	[-0.005]		[-0.005]	[-0.003]		[-0.003]
Scaled Syntactic Complexity		0.075**	0.074**		0.077**	0.076**		0.036	0.035		0.074**	0.073**		0.061**	0.060**
		(3.94)	(4.04)		(3.84)	(3.91)		(1.26)	(1.28)		(3.85)	(3.91)		(2.83)	(2.87)
		[0.010]	[0.010]		[0.009]	[0.009]		[0.002]	[0.002]		[0.008]	[0.008]		[0.004]	[0.004]
Fund Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Strategy Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Pseudo R^2	0.024	0.024	0.025	0.021	0.021	0.022	0.007	0.007	0.007	0.023	0.023	0.024	0.014	0.014	0.014
Number of Observations	45,306	45,061	45,061	45,306	45,061	45,061	45,306	45,061	45,061	45,306	45,061	45,061	45,306	45,061	45,061

Table 8

Multivariate regressions on fund flow

This table reports results from multivariate regressions on hedge fund annual flow. The dependent variable is fund annual flow in percentage. The primary independent variables of interest are lexical diversity, syntactic complexity, and their scaled equivalents. Scaled lexical diversity is lexical diversity divided by text length. Scaled syntactic complexity is defined analogously. The regressions control for fund performance rank, CAPM rank, or FH rank. Rank is derived from last year fund return. CAPM rank is derived from last year fund CAPM alpha. FH rank is derived from last year fund Fung and Hsieh (2004) alpha. The other independent variables include fund characteristics such as the logarithm of last year's fund AUM, fund age in years, share restriction period in years which equals to the sum of the redemption period and notice period, lockup dummy, management fee as a proportion of AUM, incentive fee as a proportion of AUM, high-water mark dummy, leverage dummy, as well as dummy variables for year and fund investment strategy. The coefficient estimates on the non-rank fund control variables are omitted for brevity. The t-statistics, derived from robust standard errors that are clustered by fund and year, are in parentheses. Panel A reports regressions with lexical diversity and syntactic complexity as the primary independent variables of interest. Panel B reports regressions with scaled lexical diversity and scaled syntactic complexity as the primary independent variables of interest. The sample period is from January 1994 to December 2016. * Significant at the 5% level; ** Significant at the 1% level.

	Dependent variable = Fund flow								
Independent variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Regressions with raw lexical diversity and syntactic complexity									
Lexical Diversity	3.403**		4.148**	3.326**		4.064**	3.406**		4.159**
	(4.81)		(5.63)	(4.69)		(5.42)	(4.79)		(5.55)
Syntactic Complexity		0.889	-1.262**		0.862	-1.245*		0.884	-1.272**
		(1.70)	(-2.73)		(1.65)	(-2.57)		(1.69)	(-2.63)
Rank	52.304**	52.766**	52.478**						
	(11.40)	(11.34)	(11.43)						
CAPM Rank				53.195**	53.734**	53.402**			
				(13.18)	(13.16)	(13.24)			
FH Rank							44.933**	45.461**	45.132**
							(12.17)	(12.06)	(12.13)
Other Fund Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Strategy Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Adj. R^2	0.074	0.073	0.074	0.074	0.073	0.074	0.068	0.067	0.068
Number of Observations	53,688	53,304	53,304	53,688	53,304	53,304	53,688	53,304	53,304
Panel B: Regressions with scaled lexical diversity and syntactic complexity									
Scaled Lexical Diversity	2.567**		2.712**	2.551**		2.692**	2.589**		2.736**
	(4.13)		(4.34)	(4.19)		(4.40)	(4.19)		(4.42)
Scaled Syntactic Complexity		-0.489	-0.698		-0.476	-0.683		-0.493	-0.704
		(-1.25)	(-1.82)		(-1.18)	(-1.71)		(-1.23)	(-1.77)
Rank	52.414**	52.813**	52.592**						
	(11.37)	(11.32)	(11.40)						
CAPM Rank				53.343**	53.783**	53.555**			
				(13.14)	(13.10)	(13.21)			
FH Rank							45.071**	45.516**	45.275**
							(12.09)	(11.96)	(12.07)
Other Fund Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Strategy Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Adj. R^2	0.073	0.073	0.073	0.074	0.073	0.074	0.068	0.067	0.068
Number of Observations	53,688	53,304	53,304	53,688	53,304	53,304	53,688	53,304	53,304

Table 9

Robustness tests

This table reports results from (i) multivariate regressions on hedge fund monthly alpha and (ii) multivariate probit regressions on the probability that hedge funds report violations on their Form ADVs. Alpha is the Fung and Hsieh (2004) seven-factor monthly alpha where factor loadings are estimated over the last 24 months. Violation is an indicator variable that takes a value of one if a fund reports any regulatory, civil, or criminal violation, and is zero othewise. The primary independent variables of interest are lexical diversity, syntactic complexity, and their scaled equivalents. Scaled lexical diversity is lexical diversity divided by text length. Scaled syntactic complexity is defined analogously. The other independent variables include fund characteristics such as the logarithm of last month or last year fund AUM, fund age in years, share restriction period in years which equals to the sum of the redemption period and notice period, lockup dummy, management fee as a proportion of AUM, incentive fee as a proportion of AUM, high-water mark dummy, leverage dummy, as well as dummy variables for year and fund investment strategy.The coefficient estimates on the fund control variables are omitted for brevity. The t-statistics, derived from robust standard errors that are clustered by fund and month (or year), are in parentheses. The sample period is from January 1994 to December 2016. * Significant at the 5% level; ** Significant at the 1% level.

	Dependent variables			Dependent variables	
Independent variables	Alpha	Violation	Independent variables	Alpha	Violation
Panel A: Controlling for the Sun, Wang, and Zheng (2012) Strategy Distinctiveness Index					
Lexical diversity	0.057**	-0.088**	Scaled lexical diversity	0.039**	-0.077**
	(4.89)	(-2.96)		(4.32)	(-3.69)
Syntactic Complexity	-0.016	0.106**	Scaled syntactic complexity	-0.009	0.077**
	(-1.61)	(3.98)		(-1.13)	(3.49)
Panel B: Controlling for fund R^2					
Lexical diversity	0.062**	-0.087**	Scaled lexical diversity	0.042**	-0.076**
	(5.03)	(-2.93)		(4.46)	(-3.65)
Syntactic Complexity	-0.017	0.105**	Scaled syntactic complexity	-0.010	0.077**
	(-1.68)	(3.96)		(-1.19)	(3.47)
Panel C: Controlling for manager education level					
Lexical diversity	0.065**	-0.080**	Scaled lexical diversity	0.049**	-0.068**
	(5.74)	(-2.98)		(5.53)	(-3.55)
Syntactic Complexity	-0.016	0.101**	Scaled syntactic complexity	-0.007	0.075**
	(-1.65)	(4.45)		(-0.88)	(3.97)
Panel D: Controlling for median SAT score of manager's undergraduate institution					
Lexical diversity	0.034**	-0.048	Scaled lexical diversity	0.025**	-0.038
	(2.88)	(-1.36)		(2.67)	(-1.72)
Syntactic Complexity	-0.010	0.061*	Scaled syntactic complexity	-0.007	0.043
	(-0.88)	(2.08)		(-0.75)	(1.79)
Panel E: Controlling for the Fog index					
Lexical diversity	0.065**	-0.067**	Scaled lexical diversity	0.051**	-0.068**
	(5.61)	(-2.68)		(5.70)	(-3.54)
Syntactic Complexity	-0.008	0.062*	Scaled syntactic complexity	-0.002	0.038
	(-0.76)	(2.49)		(-0.20)	(1.81)
Panel F: Controlling for the Flesch Kincaid index					
Lexical diversity	0.068**	-0.069**	Scaled lexical diversity	0.051**	-0.070**
	(5.69)	(-2.76)		(5.69)	(-3.62)
Syntactic Complexity	-0.018	0.065**	Scaled syntactic complexity	-0.010	0.039
	(-1.54)	(2.62)		(-0.99)	(1.86)
Panel G: Controlling for Hwang and Kim (2017) readability					
Lexical diversity	0.066**	-0.079**	Scaled lexical diversity	0.049**	-0.068**
	(5.73)	(-3.05)		(5.51)	(-3.58)
Syntactic Complexity	-0.017	0.099**	Scaled syntactic complexity	-0.010	0.073**
	(-1.70)	(4.45)		(-1.23)	(3.89)
Panel H: Funds based in the US or the UK					
Lexical diversity	0.062**	-0.110**	Scaled lexical diversity	0.049**	-0.090**
	(5.16)	(-3.54)		(5.14)	(-4.28)
Syntactic Complexity	-0.016	0.117**	Scaled syntactic complexity	-0.006	0.089**
	(-1.62)	(4.61)		(-0.76)	(4.18)
Panel I: Funds managed by native English speakers					
Lexical diversity	0.065**	-0.071*	Scaled lexical diversity	0.049**	-0.055*
	(5.46)	(-2.48)		(5.18)	(-2.54)
Syntactic Complexity	-0.012	0.096**	Scaled syntactic complexity	-0.004	0.076**
	(-1.14)	(3.92)		(-0.42)	(3.76)
Panel J: Controlling for legal words					
Lexical diversity	0.068**	-0.078**	Scaled lexical diversity	0.051**	-0.066**
	(5.86)	(-3.03)		(5.66)	(-3.46)
Syntactic Complexity	-0.018	0.100**	Scaled syntactic complexity	-0.009	0.074**
	(-1.90)	(4.54)		(-1.14)	(4.05)
Panel K: Controlling for look ahead bias in strategy descriptions					
Lexical diversity	0.050**	-0.059	Scaled lexical diversity	0.032**	-0.067**
	(3.31)	(-1.86)		(2.66)	(-2.89)
Syntactic Complexity	-0.040**	0.098**	Scaled syntactic complexity	-0.030**	0.069**
	(-3.34)	(3.52)		(-2.95)	(2.92)
Panel L: Controlling for backfill bias					
Lexical diversity	0.044**	-0.075**	Scaled lexical diversity	0.032**	-0.070**
	(3.67)	(-2.67)		(3.49)	(-3.38)
Syntactic Complexity	-0.023*	0.091**	Scaled syntactic complexity	-0.016	0.064**
	(-2.31)	(3.89)		(-1.87)	(3.35)
Panel M: Controlling for serial correlation in returns					
Lexical diversity	0.068**	-0.078**	Scaled lexical diversity	0.050**	-0.066**
	(5.59)	(-3.03)		(5.37)	(-3.47)
Syntactic Complexity	-0.020*	0.099**	Scaled syntactic complexity	-0.011	0.074**
	(-1.98)	(4.53)		(-1.27)	(4.04)
Panel N: Fung and Hsieh (2004) model augmented with an emerging markets factor					
Lexical diversity	0.051**	-0.078**	Scaled lexical diversity	0.042**	-0.066**
	(5.20)	(-3.03)		(5.16)	(-3.47)
Syntactic Complexity	-0.008	0.099**	Scaled syntactic complexity	-0.002	0.074**
	(-0.93)	(4.53)		(-0.22)	(4.04)
Panel O: Fung and Hsieh (2004) model augmented with the Pástor and Stambaugh (2003) liquidity risk factor					
Lexical diversity	0.065**	-0.078**	Scaled lexical diversity	0.048**	-0.066**
	(5.64)	(-3.03)		(5.37)	(-3.47)
Syntactic Complexity	-0.021*	0.099**	Scaled syntactic complexity	-0.012	0.074**
	(-2.21)	(4.53)		(-1.48)	(4.04)
Panel P: Fung and Hsieh (2004) model augmented with OTM equity call and put option factors from the Agarwal and Naik (2004) model					
Lexical diversity	0.062**	-0.078**	Scaled lexical diversity	0.046**	-0.066**
	(5.18)	(-3.03)		(4.97)	(-3.47)
Syntactic Complexity	-0.021*	0.099**	Scaled syntactic complexity	-0.012	0.074**
	(-2.08)	(4.53)		(-1.44)	(4.04)