Measuring investment opportunities using financial statement text

Sudipta Basu
Xinjie Ma
Temple University
Hoa Tran
The Ohio State University

December 18, 2018

ABSTRACT: We textually analyze 10-K texts from EDGAR during 1995-2009 to score firms' investment opportunity sets on multiple dimensions. We identify 646 unique key words that predict future investments and group them into 62 factors. Industry-specific factors include *Bio-Pharmaceutical, Banking, Information Technology, Oil & Gas* and *Semi-conductor*, while more general factors include *Impairment, Debt Intensity, Executive Changes, Preferred Stock Buyback and Capital Seeking*. Our multi-dimensional measures of firms' investment opportunities outperform Tobin's Q and/or industry-fixed effects, in predicting out-of-sample future (2010-15) investments and related corporate policies, and even inform incrementally over lagged dependent variables. Our IOS factors outperform Tobin's Q more in subsamples with less efficient market prices, i.e., when Tobin's Q is a noisier signal of investment opportunities.

Key words: Investment opportunities, IOS, textual analysis, Tobin's Q, corporate policies, Lasso

JEL classification: G31, G32, G35, M41, M21

We thank Dmitri Byzalov, Joshua Lee, Oleg Rytchkov, C. Jack Liebersohn, Debby Su (discussant), Wan-Ting Wu (discussant), Fang Zhao (discussant), workshop participants at Temple University, and conference participants at SEC Doctoral Student Symposium 2018, AAA Mid-Atlantic Region Meeting 2018, Eastern Finance Association Meeting 2018, AAA Financial Accounting and Reporting (FARS) Mid-year Meeting 2018 and Academy of Business Research International Conference 2017 for their helpful suggestions.

I. INTRODUCTION

Firms create value by identifying and exploiting profitable investment opportunities, also called positive net-present-value (NPV) projects. Managers select operating policies to exploit a chosen investment opportunity set (IOS) thereby creating value (Smith and Watts 1992). Even when managers are not optimizing firms' value, their IOSs constrain the set of successful operating policies that they can pursue (Alchian 1950, Becker 1962). The better managers match operating policies to their IOS, the closer firms are to their optimal values (Hayashi 1982), and—in aggregate—the closer the economy is to an optimal allocation of capital.

Researchers need a good proxy for IOS to better understand managers' decisions. Prior research shows that key financial ratios, such as Tobin's Q proxied by a market-to-book ratio, are associated with firms' investment opportunities. However, firms can have similar unidimensional Tobin's Qs but different investment opportunities because the latter are inherently multi-dimensional. A drug store firm like Rite Aid would rely on tangible investment opportunities more than a mass-media firm like Walt Disney, even though Walt Disney and Rite Aid have had similar market-to-book ratios of assets (1.34 vs. 1.32, in 2001). Therefore, we propose a multi-dimensional approach to measuring investment opportunities.

We create a multi-dimensional financial-text-based measure of IOS. We first apply the least absolute shrinkage and selection operator or Lasso (Tibshirani, 1996) to identify a set of key words that best predict future investments in panel data, and then apply factor analysis on these word frequencies to identify latent factors that capture IOS. This approach, unlike commonly used machine learning techniques, allows us to meaningfully interpret the factors that predict future investments. By examining the word lists and the firms that rank highly on each factor, we identify

many different dimensions of IOSs. Specifically, we find that variations in investment opportunities are associated with 62 textual factors, 26 of which are industry-specific. These industry-specific factors indicate systematic IOS differences between major industries, such as *Bio-Pharmaceutical, Banking, Information Technology, Oil & Gas,* and *Semi-conductor*. Some factors, such as *Bio-Pharmaceutical*, also identify IOS differences within a single industry. More general IOS factors that are not specific to any single industry or group of related industries include *Impairment, Debt Intensity, Preferred Stock Buyback, Capital Seeking,* and *Executive Changes.*

Our analysis helps us better understand how IOSs influence the set of corporate policies. For example, our results indicate that firms with a high *Bio-pharma* factor score, i.e. firms that use more words such as *drug, approval, milestone* …, have higher future R&D but lower SG&A expenses, issue more equity and lower their leverage. This suggests that bio-pharmaceutical firms granted more investment opportunities, such as receiving a major FDA drug approval, exploit their higher growth options by not only shifting investments from SG&A to R&D, but also issuing more equity to finance further R&D expansion.

Our textual IOS factors outperform market-to-book ratio (*MTB*, a proxy for Tobin's Q) and industry-fixed effects in predicting future investments. Further, adding the IOS factors to *MTB*, industry-fixed effects, and control variables increases out-of-sample pseudo R^2s from 19% to 33% in predicting firms' one-year-ahead total investment. The three components of total investment—*R&D* (research and development expenditure), *SG&A* (selling, general and administrative expenses, excluding R&D), and *CAPEX* (capital expenditure)—are predicted much better, as pseudo R^2s improve from 30% to 42%, 58% to 63%, and 30% to 41%, respectively. The IOS factors also predict future investment variables incremental to *MTB*, industry-fixed effects, control variables (cash flows, competition…), and lagged investments.

Our textual IOS factors predict other corporate policies that are likely matched to investment opportunities. Pseudo R^2s in predictive regressions of market leverage, debt issuance and free cash flow increase from 23% to 38%, from 5% to 8%, and from 44% to 46%, respectively, when we add IOS factor scores to *MTB*, control variables, and industry-fixed effects as independent variables. Similarly, our IOS factors improve pseudo R^2s from 8% to 16%, and from 7% to 12% in regressions predicting dividend payouts and total payouts, respectively. Again, the IOS factors predict both financing and payout variables incremental to *MTB*, industry-fixed effects, control variables, and lagged financing and payout variables.

We validate our IOS factors by studying their behavior around shocks to IOS. First, we show that after the Bush Steel Tariffs started in March 2002, steel producers' 10-K Forms included more words loading highly on *Metal Manufacturing* Factor 22, a factor specific to U.S. steel producers' IOS, relative to other manufacturing firms. Steel tariffs likely reduced international competition, thus improving IOS for U.S. steel producers. We show that the effect was reversed when the tariffs were lifted in December 2003. Second, we show that our industry-specific textual IOS factors are negatively associated with changes in McLaughlin and Sherouse's (2017) measure of industry regulation, indicating that more regulatory restrictions decrease the IOS of firms in an industry.

We also study subsamples where *MTB* is likely to be a noisy signal of IOS. When market prices are less efficient, as in the case of loss firms or firms with volatile returns, *MTB*'s ability to predict future investments worsens in at least one component of investments (SG&A, R&D or CAPEX). Our textual IOS factors generally retain their predictive power for future investments.

Our results suggest that Form 10-Ks are very informative about firms' IOS. More importantly, we show that IOS is a multi-dimensional construct that the unidimensional Tobin's Q captures poorly.

Firms and industries with similar Tobin's Qs can have very different IOSs. For example, Walt Disney and Rite Aid had similar *MTB*s in 2001 but differed on *Impairment* (-0.24 vs. 0.46), *Debt Intensity* (-0.65 vs. -0.05) and *Preferred Stock Buyback* (-0.16 vs. 0.75). By analyzing textual factors, we understand better why corporate policies vary across firms and over time.

This study makes several contributions. First, while researchers have relied on a unidimensional Tobin's Q measure, we propose a multi-dimensional approach based on textual analysis. This approach can help investors sort faster through investment opportunities across firms, thus improving information processing and capital allocation efficiency. Second, we use a data-driven approach to form and interpret word lists associated with different dimensions of IOS, which can be refined to target different concepts related to investment opportunities. This research advances the literature documenting the information content of Forms 10-K for hard-to-measure constructs, such as competition (Feng, Lundholm, and Minnis, 2013), accruals (Frankel, Jennings, and Lee, 2016) and financial constraints (Buehlmaier and Whited, 2018).

Third, we illustrate a new application of well-established statistical procedures to textual analysis: Lasso followed by factor analysis. Our approach facilitates rich interpretation of word lists produced by standard statistical learning methods and thus allows researchers to better understand what their output captures. This helps counter the criticism that machine learning methods borrowed from computational linguistics are a black box hindering interpretability.

We review the literatures on IOS and textual analysis in Section II, describe data and methodology in Section III, examine the properties of word lists and IOS factor scores in section IV, test the predictive power of the textual factors in Section V, and conclude in Section VI.

II. THEORETICAL DEVELOPMENT

Define Investment Opportunity Set

Smith and Watts (1992) define a firm's IOS as its "prospective investment opportunities and associated payoff distributions," which is in line with Myers' (1977) notion that a firm's value consists of the value of assets in place and the present value of future investment opportunities. Theoretically, all firms have identical IOSs in a frictionless world. Without regulatory, technological, and financial constraints, a firm can invest in any location, any industry and at any scale. Therefore, there is no variation in this theoretical IOS, and thus, no point in measuring it.

Practically, a firm considers various constraints and potential synergies, which restrict its likely next investments to a subset of the theoretical IOS, which we call the **realistic investment opportunity set (realistic IOS)**. We aim to measure the realistic IOS as defined below:

Definition (Realistic IOS): a firm's realistic investment opportunity set at a given time is the set of positive NPV projects that the firm can exploit soon, given the firm's competitive advantages (shaped by its existing constraints and potential synergies in implementing new projects).

We stress a firm's existing competitive advantage, such as regulatory protection and technological know-how, in defining its realistic IOS. Regulators reduce competition by constraining the types of investments of the firm and its peers. For example, financial regulators' minimum capital requirements restrict the risk and scale of banking projects, while Food and Drug Administration (FDA) approvals shapes investment opportunities for pharmaceutical firms. Existing production technology determines current technological feasibility and opportunities to enhance technological competitive advantage. A firm's future investment opportunities are likely close to its existing operations, as suggested by the theory of the "adjacent possible" (Kauffman 1995, Johnson 2010).

Our definition of the realistic IOS implies that it is endogenously shaped by financial constraints. This differs from the extant finance literature's view that IOS is shaped by only technological and regulatory constraints, and thus, exogenous with respect to financial constraints (e.g. Farre-Mensa and Ljungqvist, 2016; Smith and Watts, 1992). Although such a view lets one more closely study the effect of financial constraints, it is inconsistent with measuring IOS using the endogenous market-to-book ratio. Instead, we expect our IOS measures to reflect all three types of constraints.

We assume that financial statement texts contain meaningful data regarding firms' existing competitive advantages, and thus, could help measure realistic IOS. This assumption is reasonable given the regulatory constraints on disclosure and the associated high costs of non-compliance, especially in the U.S., and thereby assures that managers' annual disclosures in financial statement texts (Form 10-Ks) reflect their truthful perception of the firms' realistic IOS.

Our IOS definition suggests that researchers cannot directly observe IOS. Only some choices within the realistic IOS are easily observable ex-post (i.e. those that the managers invested in). Therefore, attempts to measure the realistic IOS assume that a sufficiently large sample can uncover ex-ante unobservable IOS from ex-post observable investments. Without this assumption, any IOS measure's ability to predict investments and other corporate policies should be interpreted as pure association, not causation.

How firms choose investment opportunities

The Q-theory of investment (Kaldor 1966; Tobin 1969) asserts that the rate of investment is an increasing function of the ratio of the market value of additional capital to its replacement cost, called Tobin's Q. Assuming zero adjustment costs for investments, Q-theory predicts that a firm invests until Tobin's Q equals 1, or until the marginal benefit of investment equals the marginal

cost of capital. This rule is equivalent to the Net Present Value rule that a firm should invest in a project only if the project's NPV is positive (Fisher 1930). With positive adjustment costs, the neoclassical theory of investment (Jorgenson 1963) derives an optimal level of investment assuming that firms maximize profits. Hayashi (1982) shows that this neoclassical theory and the Q-theory make similar predictions. For example, if one assumes quadratic adjustment costs, the optimal rate of investment is a linear function of Tobin's Q (Strebulaev and Whited 2011).

Consistent with these theories, Hayashi (1982) shows that a proxy for Tobin's Q can explain 48% of aggregate investments from 1953 to 1976. Peters and Taylor (2017) report that Tobin's Q predicts future investments in both intangible capital (R&D expenses, and selling, general and administrative expenses) and physical capital (capital expenditures) from 1975 to 2011.

From a management viewpoint, choosing a subset of investment opportunities is analogous to committing to a generic business strategy. Porter (1979) indicates that such commitment includes choosing competitive advantage (differentiation or cost leadership) and market scope (industry-wide or focus). Wernerfelt (1984) and Barney (1991), on the other hand, assert that firms utilize their available resources and turn them into sustained competitive advantages. In other words, this view suggests firms should choose the subset of investment opportunities that they have the best capacity to exploit. Newbert (2007), however, shows that the resource-based view receives modest empirical support (53% supported) among 549 related tests in 55 papers.

How investment opportunity sets affect firms' operations

After choosing a subset of investment opportunities, each firm then organizes its operations, internal structures, financing, and other attributes so that these factors complement each other, to maximize efficiency (Milgrom and Roberts, 1995). As a result, a firm's IOS characteristics can

predict systematic variations in corporate policies and market perceptions of those policies (e.g. Myers, 1977; Smith and Watts, 1992).

Investment opportunities are associated with characteristics of mergers and acquisitions (M&As) such as takeover likelihood and leveraged buyout likelihood opportunities (Hasbrouck 1985; Opler and Titman 1993), methods of payment (Martin 1996) and the gain or loss by bidder firms and target firms (Lang, Stulz, and Walkling 1989). Announcement returns of capital expenditures and R&D investments are higher for firms with higher investment opportunities (Szewczyk, Tsetsekos, and Zantout 1996; Chung, Wright, and Charoenwong 1998; and Brailsford and Yeoh 2004). Chen et al. (2000) document a similar effect for international joint venture announcements.

Intangible investment opportunities are associated with less leverage (Myers 1977; Smith and Watts 1992; Gaver and Gaver 1993), fewer debt covenants (Skinner 1993; Billett, King, and Mauer 2007; Nash, Netter, and Poulsen 2003), higher bankruptcy risk (Lyandres and Zhdanov 2013), lower payouts to shareholders (Gaver and Gaver 1993; Smith and Watts 1992; and Dittmar 2000), and higher executive compensation, greater use of market-based performance schemes (Smith and Watts 1992; Gaver and Gaver 1993; Skinner 1993) and higher pay-performance sensitivity (Baber, Janakiraman, and Kang 1996). Firms with higher investment opportunities are more likely to have Big 5 auditors (Lai, 2009) and industry-specialized auditors (Cahan et al., 2008).

Imperfect proxies of investment opportunity sets

Ex-ante measures: Tobin's Q review

Theoretically, firms' IOS are summarized in Tobin's Q, "the ratio of market value of new additional investment goods to their replacement cost" (Hayashi 1982, p. 214). This marginal Q ratio is the marginal benefit of a dollar invested in capital stock, or the shadow value of capital in

a firm's value maximization problem subject to a capital stock constraint. Marginal Q is not observable, and thus researchers use average Q, the ratio of market value to book value of equity or assets to proxy for marginal Q (Szewczyk, Tsetsekos, and Zantout 1996; Lang, Stulz, and Walkling 1989; Hasbrouck 1985; McConnell and Servaes 1990; Brown, James, and Mooradian 1994). A popular alternative measure is cash flow, which predicts future investments well (Carpenter and Guariglia 2008, Gilchrist and Himmelberg 1995, Blundell et al. 1992). Another alternative is to factor analyze various proxies for investment opportunities to reduce error (Baber, Janakiraman, and Kang 1996). Gaver and Gaver (1993) transform common investment measures into an indicator variable that is equal to 1 if and only if the firms are considered as growth firms.

Investment opportunities can also be estimated using structural estimation (Blundell et al. 1992, Gala 2015), but this approach is quite complex and requires many simplifying assumptions. Peters and Taylor (2017) improve measurement by incorporating estimated intangible capital in calculating average Q. Although this method helps to better predict firms' investments in today's knowledge-based economy, it still contains the measurement errors of average Q constructs.

Ex-post measures of investment opportunities

Ex-post measures of investment opportunities use inputs or outputs of investments to proxy for investment opportunities. Investment outputs include realized growth of equity value and return variance (Smith and Watts 1992) and realized revenue growth (Kallapur and Trombley 1999).

Investment inputs include tangible and intangible realized investments. Proxies for physical or tangible investments include property, plant, and equipment (PPE), depreciation (DEP) and capital expenditures (CAPEX), all typically scaled by firm value (Denis 1994; Smith and Watts 1992). Intangible investment variables include research and development expenses (R&D), and selling,

general and administrative expenses (SG&A), both scaled by either sales or total assets, as in Peters and Taylor (2017) and Eisfeldt and Papanikolaou (2013). These studies usually combine SG&A and R&D linearly to represent intangible investments, allowing them to estimate intangible capital by cumulating (weighted) past intangible investments.

III. DATA AND METHODS

Textual Analysis

Economics-based researchers have recently started to use computerized textual analysis to measure constructs that are otherwise hard to capture (Loughran and McDonald 2016). Due to "a strong methodology flavor" from linguistics, early research focused on linguistics topics such as tone and readability (Li 2010, p. 158). Recent works study accounting and finance topics such as accruals (Frankel et al. 2016) and financial constraints (Hoberg and Maksimovic, 2015; Buehlmaier and Whited, 2018). Our paper adds to this growing literature since we analyze texts to measure IOS, arguably the most fundamental construct in corporate finance.

Traditional textual analysis needs researchers' judgement to create word lists targeting a construct. For example, Feng et al. (2013) exert judgement in removing phrases such as "less competitive" before counting the use of "competition" in Forms 10-K to proxy for firm-level competition. Recent work explores what words capture a construct of interest without researcher intervention. For instance, Frankel et al. (2016) use support vector regressions (SVR), a supervised machine learning technique, to measure accruals using MD&A text without creating their own word list.

Machine learning is often criticized for its black-box nature: users do not know in detail how algorithms translate text into a specific measure. We use a statistical (machine) learning technique that facilitates rich interpretability: Lasso (Tibshirani, 1996) combined with factor analysis. Lasso

helps us select words that predict future investments without any ex-ante restrictions on what combination of words managers *might* use to describe IOS. We then use factor analysis to interpret the words selected by Lasso.

Data

We collect Form 10-Ks, which are available from the third quarter of 1994 through the present, from EDGAR. Using EDGAR index files, we download all 163,729 Form 10-K filings released for fiscal years 1995 to 2015 using Python. We merge 10-K data with Compustat financial data using the central index key (CIK). After dropping observations without sufficient Compustat data to construct all dependent variables (including investment, financing, and payout variables), we are left with 66,344 observations. We exclude penny stocks (firms with stock price smaller than $1). Additionally, since we scale most variables by total assets, we drop observations with assets less than $5 million (Fama and French 2015). After requiring sufficient data to control for industry competition, the sample reduces to 53,324 firm-year observations for panel data analysis.

For each Form 10-K, we remove HTML tags, the heading items, tables, and all numbers (including Roman numerals), non-alphabetic symbols, and single letters that are often used to index lists.[1] We then parse the remaining texts into individual words, reduce these words to their initial roots (called word stems), and count each document's frequency of each unique word stem. Following textual analysis conventions, we exclude common stop words such as *a, an, the, of, is* and *are*. We adjust for word stems' relative importance by multiplying each word stem's frequency by the logarithm of inverse document frequency (idf: number of documents in the whole selection scaled

[1] We find that the Beautiful Soup 4 package in Python, which is the standard natural language processing package, does not fully eliminate all html tags. We use regular expressions upon output from the Beautiful Soup 4 package to eliminate html tags and formatting words, such as *bold, div, left, right* and *justified.*

by the number of documents containing that word stem), to obtain the tf-idf (term frequency – inverse document frequency). We call this number the adjusted term frequency for simplicity.

We use a training sample from 1995 to 2009 to form a base word list and factor scores. We then form factors and test out-of-sample predictive ability using the test sample from 2010 to 2015.

Form a base word list

We apply the least absolute shrinkage and selection operator or Lasso (Tibshirani 1996), to select the subset of unique words whose frequencies best predict firms' future actual investments. The Lasso minimizes the sum of squared residuals subject to a constraint on the maximum value of summed absolute coefficients:

$$(\hat{\alpha}, \hat{\beta}) = argmin \sum_{i,t} \left(INV_{it+1} - \alpha - \sum_{j} \beta_j freq_{it,j} \right)^2$$

$$s.t \sum_{j} |\beta_j| \leq C$$

where INV_{it+1} denotes the future actual investment made by the firm i viewed from year t, which we operationalize as the sum of one-year-ahead investments in *CAPEX*, *R&D* and 30% of *SG&A* (after R&D expense is excluded from *SG&A*), scaled by the current end-of-year total assets (Eisfeldt and Papanikolaou 2014; Peters and Taylor 2017); $freq_{it,j}$ denotes the adjusted frequency of the j^{th} word in the Form 10-K of firm i in year t.

The constraint on estimated coefficients shrinks many coefficients and sets others to zero. Intuitively, the Lasso performs variable selection while stabilizing estimation (Tibshirani 1996). Also, standard Lasso packages automatically set the constraint C such that cross-validation

estimation errors are minimized. Given the standard Lasso procedure, we impose no ex-ante judgement on the types of words and the number of words that predict firms' future investments.

Lasso finds words that managers use to describe future opportunities and their past actions that have persistent future implications. These words reflect firms' realistic IOS if three assumptions hold. First, managers on average choose investments optimally form firms' realistic IOS. Second, one-year-ahead investments on average reflect managers' expectation of all future investment opportunities. Third, managers disclose all relevant data about realistic IOS in their Form 10-Ks.

Word classification

To meaningfully categorize words that capture the dimensions of firms' realistic IOS, we factor analyze the word stems chosen by Lasso. Factor analysis assumes that each observed variable (word-stem frequency) is a linear combination of some underlying unobservable factors plus a normally distributed error term. It estimates these factors by exploiting the correlations among the observed variables.[2] We apply promax rotation to the standardized, un-rotated factors to allow correlation between them (Pett, Lackey, and Sullivan 2003). We retain factors with eigenvalues greater than one (Kaiser 1960).

Combining Lasso and factor analysis is essentially a supervised machine learning technique. While commonly used machine learning methods focus solely on predicting some target variable, we produce and interpret factors that predict the target variable – corporate investment. This two-step procedure increases our understanding of the economic process driving corporate investments.

[2] We use the principal component method to identify factors. As commonly used in the literature, we set the priors required by the factor analysis procedure to squared multiple correlations (SMC), which could be interpreted as how much each observed variable is explained by the other observed variables.

IV. INTERPRETING THE FACTORS

Applying Lasso on 3,179 unique word stems, we select 646 words whose frequencies best predict future investments. Factor analysis on these 646 variables results in 62 factors whose eigenvalues exceed one. Throughout the paper, we separately report illustrative results for the 10 factors with the highest eigenvalues, although we include all 62 factors in our regressions.

Word lists

Table 3 presents lists of the word stems that load highly onto each of the 10 factors and lists of the firm-year observations that have the highest scores for each factor. We tabulate the top five word stems for each factor based on their loadings. Analyzing the word stems and firm-years jointly helps us name the factors meaningfully. We divide the factors into two groups: industry-related factors and general factors.

General factors

Table 3 Panel A presents lists of the top five word stems and firm-years that correspond to 10-Ks' descriptions of investment opportunities that are not specific to any single industry or group of related industries. Here we show only the top 5 general factors that explain the most variation in the word frequencies across all the Form 10-Ks in our sample. These factors include Factor 1 (*Impairment*), Factor 4 (*Debt Intensity*), Factor 6 (*Executive Changes*), Factor 8 (*Preferred Stock Buyback*) and Factor 10 (*Seeking Capital*).[3]

[3] The next 10 general factors are: executive compensation, option exercising, auditing, pension compensation, real estate leases, bond issuance, partnership, debt seniority, M&A, and marketing intensity. For more details, please refer to the online appendix (under construction).

Factor 1 is associated with words that describe impairments, such as *fair, result, valu, finance* and *affect*. Firms with Form 10-Ks that score high on this factor belong to many different industries, such as Intersil Corp. (a semiconductor company, among top 5) and Steelcase Inc. (a furniture manufacturer, among top 10). More than 20 of the top 30 10-K documents that score high on this *Impairment* Factor 1 were reported during and after the financial crisis: fiscal years 2008-2010. Poor equity market conditions forced many companies to impair goodwill. Intersil Corp. for example, "recorded an impairment loss of $1,154.7 million" (Form 10-K fiscal year-end January 2009), which was 150% of the company's revenue in the same fiscal year.

Factor 4, *Debt Intensity*, is associated with firm-years that use debt financing heavily, indicating the dominance of tangible investment opportunities or a lack of intangible investment opportunities. These firms' 10-Ks tend to use many debt-financing-related words, such as *borrow, lender, agent, document,* and *lien*. Firm-years that score high on this factor have high debt ratios. For example, ranked third on this factor is Wellman Inc., a multinational Fortune 500 company involved in the recycling and manufacture of fibers and plastic resins that increased its total book value of debt from 55%, its previous 10-year maximum, to 70% of total assets in 2003.

Factor 6, *Executive Changes*, is associated with firm-years that have important executive changes. Form 10-Ks associated with these firm-year observations contain more word stems such as *employ, execut, agreement, caus,* and *confidenti*. Ranked 1st in Factor 6 is Med-Design Corporation in 2002, during which the firm hired David R. Dowsett as its new Chief Operating Officer, who became the firm's Chief Executive Officer two years later. Another example is MetaSolv Inc., a software company whose Executive Vice President in charge of America's Sales changed from Joseph W. Pollard to Philip C. Thrasher in 2002. Intuitively, major executive changes signal firms' anticipation of seizing or losing large investment opportunities.

Factor 8 is associated with *Preferred Stock Buyback*. The top firm-years on this factor have Form 10-Ks containing word stems like *prefer, seri, convers, redeem* and *convert*, suggesting that the firms use cash *redemption* or ordinary stock *conversion* to reduce *series* of *preferred* stocks outstanding. Trinsic Inc., which in 2004 has the highest Factor 8 score in our sample, for example, "consummated a tender offer whereby the firm exchanged common shares for all of its outstanding preferred stock" (Trinsic Inc., Form 10-K, December 2004). This preferred stock event was a major event since it involved exchanging 46,657,636 of the firm's common shares, or 84% of the firm's number of shares outstanding. Buying back preferred stocks, a debt-like financial instrument, is indicative of future investment opportunities.

Factor 10 captures young firms seeking capital. Firm-year observations that rank high on this *Capital Seeking* factor use word stems like *go, raise, concern* and *deficit* to describe their capital deficits. Many of these firms have auditors issuing a Going Concern opinion. The top observations are mostly IPO firms. For example, number one in Factor 10 is Uniontown Energy Inc., which went public in 2009 but ceased operations only two years later because of an inability to raise capital to cover losses and continuing operations. Another example is Kedem Pharmaceuticals, which went IPO in 2007 but soon experienced a capital deficit resulting in a Going Concern audit opinion. Kedem Pharmaceuticals endured its capital deficit for five years to exploit its growth prospects as a public firm.

Industry-specific factors

Panel B of Table 3 reports the word lists that we interpret as indicative of industry-specific investment opportunities together with the corresponding top firm-year observations. This panel

includes Factor 2 (*Banking*), Factor 3 (*Bio-Pharma*), Factor 5 (*IT* or *Information Technology*), Factor 7 (*Oil & Gas*), and Factor 9 (*Semiconductor*).[4]

We name Factor 3 *Bio-Pharma* because the word stems that load highly onto this factor, as shown in Panel A of Table 3, are *preclin, trial, clinic, efficacy,* and *candid*, which are all bio-pharmaceutical terms. The top five firm-year observations on this factor confirm this naming, as these are all bio-pharmaceutical firms. Indeed, all top 30 firm-year observations (not tabulated) are in either Standard Industrial Classification (SIC) code 2834 (Pharmaceutical Preparations) or 2836 (Biological Products, Except Diagnostic Substances).

Ranked first on Factor 3, for example, is Aradigm Corp in 2009, a fast-growing pharmaceutical company. The company got a major U.S. Food and Drug Administration ("FDA") approval in 2009, which explains why the Factor 3 score for this firm, as a measure of its investment opportunities in the bio-pharmaceutical industry, is the highest for this year. The Fortune 500's fastest-growing pharmaceutical companies, such as Merck and Biogen, have Factor 3 scores in the top 10% of all firm-year observations during 1995-2010.[5]

Similarly, we name Factor 2 *Banking*. Its word list in Panel A includes *bank, loan, capit, feder,* and *branch*, which all suggest unique features of the banking industry. The top 30 firm-years on Factor 2 all have SIC code 6020 (commercial banking).

---

[4] The next 15 industry-related factors are: real estate investment trusts, insurance, metal manufacturing, retailers, telecommunication firms, medical device manufacturers, broadcasting firms, construction firms, electric service providers, brokerage firms, short-term savings institutes, licensing businesses, trucking companies, bio-research firms, education firms.

[5]Fastest growing based on 5-year average growth rate of sales: http://fortune.com/2015/06/20/fortune-500-fastest-growing-pharmaceutical-companies/

We name Factor 5 *Information Technology (IT).* The top firm-year observations on Factor 4 are firms whose 10-Ks use many IT-related terms, such as *support, hardware, software, server, solution,* and *computer*. These firms are mostly prepackaged software firms (SIC code 7372) or computer integrated systems design companies (SIC code 7373).

We interpret Factor 7 as an IOS measure for the *Oil & Gas* industry. Words that load highly on this factor include *prove, oil, nature, reserve,* and *drill,* all of which suggest future growth opportunities for oil and gas firms. The 30 Form 10-Ks with the highest Factor 7 scores all belong to firms with SIC code 1311 (Crude Petroleum and Natural Gas).

We interpret Factor 9 as an IOS measure for the *Semiconductor* industry. Word stems that load highly onto this factor are *semiconductor, manufacture, optic, design.*…all of which are terms often used to describe businesses that rely highly on semiconductor materials. The top 30 observations based on their Factor 9 score belong to many SIC 4-digit classifications such as 3674 (Semiconductors and Related Devices), 3827 (Optical Instruments and Lenses) and 3823 (Industrial Instruments for Measurement, Display, and Control of Process). The *Semiconductor* factor thus describes investment opportunities that cross SIC industry boundaries. In other words, if one is to classify firms based on their IOSs, this textual IOS factor could improve upon conventional industry classifications for semiconductor firms.

In short, our factor scores not only correctly pick out industries whose Tobin's Qs tend to be consistently low (e.g. banks) or high (e.g. pharmaceutical firms), but also capture general factors that indicate various aspects of investment opportunities, such as impairment and tangibility.[6]

Factor properties

To validate our interpretation of the factors and to understand their properties, we provide summary statistics for the top firm-year observations for each factor in Table 4.

General factors

Table 4 Panel A reports on the top 5 general factors. Key summary statistics for the top 30 firm-year observations for each factor validate their names. As expected, the top firm-year observations on the *Debt Intensity* factor are associated with lower intangible investment opportunities and thus have relatively low *MTB* (1.41), low *R&D* (almost zero), but high leverage (average market leverage ratio of 48%).[7] Also, firm-years that score high on the *Preferred Stock Buyback* factor tend to have more intangible investment opportunities, and thus, much higher *MTB* (16.7), *SG&A* (110% of book assets) and *R&D* (19% of book assets).

[6] We re-run our analysis using 1995-1999, 2000-2004 and 2005-2009 as shorter training samples and find that the industry-specific factors are stable over time. For example, *Bio-Pharma, IT, Banking, Semi-conductor…* remain in the top 10 factors across analyses using different training samples. For the general factors, top factors, such as *Debt Intensity* and *Executive Changes* remain in the top 10 factors across training samples, while *Impairment* factor only enters the top 10 for the training sample 2005-2009, which spans the financial crisis in 2008. Alternatively, we use first order autoregressive (AR1) models to explore how much our IOS factors vary over time. We find that the mean of the AR1 coefficients is 0.725, indicating that our IOS factors have significant time-series variation. Specifically, we find that the average AR1 coefficient of general factors is 0.587 and the average AR1 coefficient of industry-specific factors is 0.862, consistent with general factors being event-driven and varying more over time than the more stable industry-specific factors.

[7] Using market-to-book value of assets to proxy for Tobin's Q is among the most popular empirical choices in prior research (Adam and Goyal 2008, Kumar and Krishnan 2008, Fama and French 2002).

For firm-years ranked high on the *Capital Seeking* Factor 10, *MTB* is intuitively very high (15,262 on average, albeit reduced to 611 when 5 outliers are removed), as these are likely young firms with abundant growth options but accumulated losses depleting their book value of assets close to zero. The top firm-years on the *Impairment* Factor 1 have lower profitability (ROA).

Table 4 Panel A also reports sales growth (*SG*). Sales growth is the highest for top firm-years in *Capital Seeking* Factor 10, followed by *Preferred Stock Buyback* Factor 8, and *Executive Changes* Factor 6. Sales growths for top observations in *Impairment* Factor 1 and *Debt Intensity* Factor 4 are low. This order of realized sales growths in part validates our textual measures of IOS as Kallapur and Trombley (1999) argue that correlation with future growth is probably the most appropriate benchmark to evaluate alternative IOS proxies.

Industry-specific factors

Table 4 Panel B reports on the top 5 industry-related factors. These factors pick up commonly observed differences across industries for leverage, investments, and *MTB.* For instance, among the top observations on these 5 factors, those of Factor 3 (*Bio-Pharma)* on average have the highest *MTB* (4.61), the highest R&D investment intensity (40% of total assets), and the lowest market leverage ratio (12%). Similarly, the top observations of Factor 7 (*Oil & Gas*) on average have the highest capital expenditure (24% of total assets), while those highest on Factor 2 (*Banking*) have almost zero capital expenditure.

As expected, high tech firms such as those scoring high on Factors 1 (*Bio-Pharma*), 4 (*IT*), and 9 (*Semiconductor*) have low leverage (below 25%). While the latter two industries are quite similar (i.e. *MTB* 2.5 vs. 2.7), the two factors help distinguish them, as *Semiconductor* companies tend to

have more R&D expenses than *IT* firms (27% vs. 11% of total assets). This suggests that the multi-dimensional industry-specific textual factors differentiate IOSs when *MTB* does not.

Within-industry variations

Figure 1 explores whether our factors accurately capture IOS variation within an industry. We plot total investment intensity (INV: the sum of the one-year-ahead *SG&A*, *R&D*, and *CAPEX*, scaled by total assets) against our factor scores and *MTB*. We plot these for pharmaceutical preparations (SIC=2834) and commercial banks (SIC=6020) industries associated with Factor 3 (*Bio-Pharma*) and Factor 2 (*Banking*) in 2002 (an in-sample year) and 2012 (an out-of-sample year).

We find that our Factor 3 (*Bio-Pharma*) outperforms *MTB* in predicting within-industry variations of future investments for pharmaceutical preparations firms, but the Factor 2 (*Banking*) scores do not outperform *MTB* for commercial banking firms. The latter result suggests that the low investment intensity in the banking industry lets Factor 2 (*Banking*) distinguish it from other industries, but investment opportunities vary too little within this industry, possibly due to regulation, for either *MTB* or Factor 2 (*Banking*) to distinguish meaningfully between banks.

We also find that Factor 3 (*Bio-Pharma*) predicts future investments equally well both in sample and out of sample. This suggests that this textual factor's predictive ability is persistent over time and thus can be used for out-of-sample prediction.

Overall, we conclude that our IOS factors predict future investments and revenue growth at least as well as *MTB*, and thus support the use of these factors as measures of investment opportunities. Some of the factors, such as Factor 3 (*Bio-Pharma*), capture both within-industry and between-industry variations of investments, and others, such as Factor 2 (*Banking*), only capture between-industry variations.

V. VALIDATING THE IOS FACTORS

Multiple regression predictive power

To evaluate the power of our factors in predicting subsequent investments, financing policies, and payout policies, we examine six model specifications with different combinations including models with: 1) MTB_t; 2) IOS factors ($Factors_t$); 3) industry fixed effects ($Ind\ FE$); 4) MTB_t and $Ind\ FE$; 5) $Factors_t$ and $Ind\ FE$; and 6) MTB_t, $Factors_t$, and $Ind\ FE$.[8] All models contain year fixed effects and dependent-variable-specific control variables that we discuss later. [9] We test the predictability of IOS factors for the dependent variables one-, three- and five-years ahead, and find qualitatively similar but weaker results for longer horizons. For parsimony, we report the tests of the one-year ahead dependent variables.

First, we run in-sample regression using data from 1994 to 2009 and evaluate the informativeness of the textual factors relative [incremental] to Tobin's Q in predicting firms' future actions.[10] We report the in-sample regression coefficients to facilitate the interpretation of IOS factors and how IOS factors predict future investments, financing policies, and payout policies. Next, we assess the out-of-sample predictive ability of our factors and report out-of-sample pseudo R^2. We use the in-sample regression coefficients, make forecasts for 2010-2015, and estimate mean square error

[8] The industry-fixed effects are based on the Fama-French 48 industry classifications. We use alternative industry classifications, including 2-digit SIC, 3-digit NAICS, and 10-K Text-based fixed industry classification (FIC-100, 200, 300, 400, 500) by Hoberg and Phillips (2016), and find qualitatively similar results. We find that IOS factors outperform industry-fixed effects until the classification separates firms into 300 or more industry groups.

[9] In untabulated tests, we find that the textual factors also have incremental predictive power over the lagged dependent variable, Tobin's Q, and industry-fixed effects.

[10] In untabulated tests, we compare the informativeness of IOS factors and Total Q (Peters and Taylor 2017) and find consistent results. We report the *MTB* results for comparability with prior research.

(MSE). Out-of-sample pseudo R^2 is calculated as one minus the ratio of MSE from a forecasting model to that of the intercept-only model.[11]

Future investments

We start by predicting future investments, so Y_{t+1} denotes future investments, which we capture with three proxies ($SG\&A$: selling, general and administrative expenses excluding R&D expenses; $R\&D$: research and development expenses; $CAPEX$: capital expenditures; all scaled by book value of assets).[12] For future investment variables, we control for operating cash flow (*CF*) and industry competition (*HHI*) (Peters and Taylor, 2017; Hoberg and Phillips, 2016).[13]

Table 5 Panel A presents the in-sample coefficients of the top 10 factors for models (4), (5), and (6). Compared to model (5), most coefficients on the factors remain statistically and economically significant after industry-fixed effects and Tobin's Q are included in model (6). Our factors thus contain information beyond Tobin's Q and industry-fixed effects regarding firm-specific investment opportunities. This incremental explanatory power comes from within-industry variation in our IOS factors. Untabulated analysis confirms that the IOS factors also capture investment variation across industries. For example, average investments are small in Banking and Insurance but big in Oil & Gas and IT, which correspond to our industry-specific factor loadings. The regression coefficients on factors reveal interesting economic insights. For instance, firms with higher *Bio-pharma* score tend to have higher future *R&D* but lower *SG&A*. This suggests that

[11] We choose the intercept-only model as the benchmark model, which captures the predictive power of the historical means of the dependent variables.
[12] In untabulated tests, we also predict the sales growth in year t+1 that results from firms utilizing their investment opportunities. We find that both *MTB* and our IOS factors do not predict future sales growth well, possibly because many factors other than investment opportunities affect sales growth in year t+1.
[13] Herfindahl-Hirschman Index (HHI) concentration metrics based on Text-based Network Industry Classifications (TNIC): available on http://hobergphillips.usc.edu/industryconcen.htm.

bio-pharmaceutical firms endowed with more investment opportunities, indicated by more words like *drug, trial, approval* … in their annual reports, exercise their growth options by cutting on administrative expenses to spend more on drug development. This insight, however, cannot be revealed by examining just *MTB* for the same sample. The coefficients on *MTB* indicate that higher growth options would on average lead to both higher *SG&A* and higher *R&D*. Thus, this emphasizes the power of a multi-dimensional measure of IOS.

Finally, the coefficients on IOS factors are almost unchanged when adding *MTB* to the regressions while the coefficients on *MTB* are halved by adding IOS factors. This suggests that inferences made using textual factors as IOS measures are likely more stable than those made using *MTB*.

Table 5 Panel B presents the out-of-sample pseudo R^2 of models (1) to (6) for investment variables and sales growth. Our 62 textual IOS factors predict firms' subsequent investments much better than Tobin's Q. For example, the out-of-sample pseudo R^2 of future *SG&A* increases from 3.31% in model (1) to 37.55% in model (2). Moreover, our textual factors outperform *MTB* when industry-fixed effects are included in the regressions. The factors capture between-firm variation within each industry, since the factor scores differ between firms, depending on how often each firm uses words that load highly on each factor. As a result, the textual factors have incremental predictive power over industry-fixed effects as indicated by the noticeable differences between out-of-sample pseudo R^2s in models (4) and (5) for all future investment variables. For subsequent *R&D*, the out-of-sample pseudo R^2 improves from 58.12% in model (4) to 62.67% in model (5) when textual factors are included in the regression.[14] We also observe increases of similar

[14] The untabulated *F*-statistics show that the incremental R^2 of unrestricted model (6) relative to restricted model (5) is significant for all dependent variables at 1%.

magnitudes for *CAPEX* and *SG&A* and the overall total investment variable. Finally, our textual factors capture the industry differences in investment opportunities better than *MTB*. For example, the predictive power of model (1) increases significantly (12%) when industry-fixed effects are added as shown in model (4) while the incremental predictive power of industry effects was much less (3%) for textual factors from model (2) to model (5) for total investments.

Financing

To assess the predictive powers of our IOS factors for financing activities and to compare these with those of *MTB*, we examine regression models where Y_{t+1} denotes financing activities, which we capture by four proxies (*MLEV*: market leverage; *DIssue*: funds raised with long-term debt; *EIssue*: funds raised with equity; *FCF*: free cash flow for internal financing). Following Richardson (2006), *FCF* is calculated as the free cash flow from existing assets in place, i.e., the net cash flow from operating activities minus maintenance investment expenditure plus research and development expenditure. Following Fama and French (2002), we control for profitability (*EBIT*), non-debt tax shields including R&D (*R&D*) and depreciation (*DEP*), firm size (*LnAT*), and competition (*HHI*).

Table 6 Panel A presents the in-sample coefficients of the top 10 factors for models (4), (5), and (6). The regression coefficients reveal many economic insights. One example is the *Bio-pharma* factor, which negatively predicts future lower leverage and higher equity financing. Bio-pharmaceutical firms that use more words like *drug, approval, milestone*… in their 10-K Forms are likely firms that have more investment opportunities needing equity financing. These firms, often license their approved drugs to big pharmaceutical firms to earn royalty. By avoiding directly manufacturing the drugs, which often involves debt financing, these firms effectively lower their

leverage. Also, one would expect drug approvals to induce positive market reaction, increasing these firms' equity value, resulting in lower market leverage. Another example is the *Impairment* factor, which is negatively correlated with future leverage and external financing. This finding suggests that firms that are negatively affected by poor market conditions (resulting in high score for *Impairment* factor) tend to lower their future leverage. Intuitively, poor market conditions shrink these firms' investment opportunities, inducing them to write down assets and retire existing long-term debts.

Table 6 Panel B presents the corresponding out-of-sample pseudo R^2 of models (1) to (6) for financing variables. Our factors improve predictive power beyond *MTB* for all four financing variables. The biggest improvement in pseudo R^2 is from 22.98% to 38.19% in leverage regressions from models (4) to (5). For debt issuance and free cash flow, our factors improve pseudo R^2 from 4.86% to 7.99% and 43.48% to 46.02% respectively. IOS factors do not significantly outperform *MTB* for equity issuance potentially because *MTB* captures over-valuation and firms tend to issue equity when their market value is high (Baker and Wurgler, 2002).

Payout

To compare the predictive power of our factors with that of *MTB* for distribution policies, we examine regression models (1) to (10) interpreting Y_{t+1} as denoting corporate distribution activities, which we capture by total dividend (*DIV*), total repurchase (*REPO*), and the sum of both—total payout (*Payout*)—all scaled by book value of assets. Again, we control for profitability (*EBIT*), non-debt tax shields including R&D (*R&D*) and depreciation (*DEP*), firm size (*LnAT*), and competition (*HHI*).

Table 7 Panel A presents the in-sample coefficients of only the top 10 factors for models (4), (5), and (6). Many factors remain economically and statistically even after controlling for Tobin's Q and industry-fixed effects, in predicting payout policies. For example, firms in *Banking*, *IT*, *Oil & Gas*, and *Semiconductor* industries have lower dividend payouts. Firms with high scores on *Preferred stock* tend to repurchase more stocks. These are firms that mention in their Forms 10-K substantial repurchases of preferred stocks. These repurchases reduce the debt-like responsibility of preferred stock dividends.

Table 7 Panel B presents the corresponding out-of-sample pseudo R^2 of models (1) to model (6) for payout variables. Our factors predict payout variables better than *MTB* and industry-fixed effects. Model (4) with *MTB* and industry-fixed effects has pseudo R^2 of 7.95% for dividends, -5.04% for repurchase and 6.83% for total payout. By contrast, model (5) with our factors and industry-fixed effects has pseudo R^2 of 15.70% for dividends, -4.50% for repuchase and 11.73% for total payout.

Validation of IOS factors using the Bush Steel Tariff as a shock

We validate the textual IOS factors as measures of firms' investment opportunities by examining how they respond to shocks that likely affect IOSs. One such shock is President George W. Bush's tariffs placed on imported steel from March 20, 2002 to December 4, 2003. The tariffs targeted imported raw steel and were intended to help U.S. raw steel producers. Thus, the tariffs should improve IOSs of raw steel manufacturers. Conversely, less competition among steel producers would increase raw steel prices and thus decrease the IOSs of raw steel consumers.

To evaluate the effect of Bush Steel Tariffs on steel producers' IOS, we employ a difference-in-differences design for tariff initiation in March 2002 and tariff removal in December 2003:

$$Factor\ 22 = \beta_0 + \beta_1 TREAT + \beta_2 POST_{tariffs\ initiated} + \beta_3 TREAT \times POST_{tariffs\ initiated} + \epsilon \quad (1)$$

$$Factor\ 22 = \alpha_0 + \alpha_1 TREAT + \alpha_2 POST_{tariffs\ lifted} + \alpha_3 TREAT \times POST_{tariffs\ lifted} + \epsilon \quad (2)$$

Factor 22 denotes the textual IOS factor corresponding to the metal manufacturing industry; $TREAT$ equals 1 for firms in Steel Works, Blast Furnaces and Rolling Mills industry (SIC=3312) and equals 0 for other firms in the manufacturing industry (Fama-French 12 industry code FF12=3) excluding raw steel consumers (SIC=3317); $POST_{tariffs\ initiated}$ equals 1 if fiscal year is 2003 and equals 0 if fiscal year is 2001; ϵ, ϵ' are error terms.[15]

We expect β_3 to be positive since tariffs on imported steel would improve IOSs in the domestic steel production industry. We expect the opposite impact when the Bush tariffs were removed and thus expect α_3 to be negative. The regression results reported in the first two columns of Table 8 are as we expect. Estimated β_3 is 2.13 (s.e.=0.61), which is statistically and economically significantly positive. The increase of 2.13 in the Factor 22 score for steel producers relative to other manufacturing firms is large, given that Factor 22 scores range from -3 to 9 in our full sample. Similarly, estimated α_3 is statistically and economically large.

We repeat our difference-in-differences analysis for a raw steel consumer industry (Steel Pipe and Tubes SIC=3317) and report the results in the middle two columns of Table 8. In this sub-analysis, TREAT equals 1 for raw steel consumers (SIC=3317 Steel Pipes and Tubes) and equals 0 for other manufacturing firms (Fama-French 12 industry code FF12=3). The results, though weaker than in

[15] We interpret Factor 22 as an IOS factor for *Metal Manufacturing* firms because words that load highly on Factor 22 are: *raw, steel, ton, product, scrap, …* Firm-years that score high on Factor 22 are mostly metal manufacturers (SIC 3312, 5093, 3317, 3350, 3341…)

the previous analysis, are consistent with our expectation that the Bush tariffs would affect IOS in the steel consumer industry in the opposite manner to that in the steel producing industry.

We also pool together the producer and consumer industries in one analysis by modifying variable $TREAT$ to equal +1 for producers, -1 for consumers and 0 for other manufacturing firms. The last two columns of Table 8 depict the corresponding results, which again confirm our expectations and thus help validate Factor 22 as a measure of IOS in the steel manufacturing industry.

Changes in regulation as shocks

Regulatory restrictions shape the investment opportunities available to different industries. In this section, we examine the IOS factors' response to changes in industries' regulatory environment. We employ McLaughlin and Sherouse (2017)'s text-based industry-year measure of regulation, detailed to 4-digit North American Industry Classification System (NAICS) code. McLaughlin and Sherouse (2017) analyze the Codes of Federal Regulation (1970-2016) to calculate the number of restrictive words, such as *shall* and *must*, in each small part of the Codes and each part's relevance to each specific industry. Their regulation measure is then calculated as the relevance-weighted average number of restrictive words for each industry-year.

We first regress the top 5 industry-specific IOS factors on the regulation measure with firm fixed effects. We expect an increase in regulatory restrictions on an industry to shrink that industry's IOS. Therefore, if the textual IOS factors capture firms' IOSs then the factor scores should fall when the regulation measure increases. Table 9 Panel A displays the predicted results. The coefficients on lagged regulation are consistently significantly negative at the 1% level. The coefficients on concurrent regulation are also all negative, and statistically significant except for *Bio-pharma* Factor 3. This result suggests that most industry-specific factors capture the effects of

regulatory changes on IOS in both the current and subsequent years. For bio-pharmaceutical firms, regulation changes might not have an immediate effect on IOS.

We then repeat the analysis for the top 5 general factors, which we show in Table 9 Panel B. We do not expect the coefficients on regulation to be negative for all general factors. For example, we expect and observe a positive coefficient of regulation in the regression for *Impairment* Factor 1, but a negative coefficient for *Debt Intensity* Factor 4. More regulatory restrictions are more likely to lead to more impairment and lower debt financing. We, however, do not have clear expectations for the effects of regulation on other general factors. Nonetheless, we find that both regulation and lagged regulation have positive effects on *Preferred Stock Buyback* Factor 8, but negative effects on *Executive Changes* Factor 6 and *Seeking Capital* Factor 10.

Where Q performs worse

A vast literature documents measurement errors in Tobin's Q and their consequences (Lewellen and Badrinath, 1997; Whited and Erickson, 2012). In this section, we investigate whether our IOS factors can outperform *MTB* even more when *MTB* is likely to be a poor proxy for IOS, like in the case of loss firms or firms with volatile stock price. For each such condition, we create a training sample and a test sample, and then use the coefficients from the training sample to predict the test sample outcomes. Table 10 compares our IOS factors' incremental power over *MTB* between the full sample (Panel A) and sub-samples where *MTB* is likely to perform poorly (Panels B and C).

We hypothesize that losses complicate valuation and thus render a market-based measure like *MTB* a noisy signal of IOS. Thus, we expect that our text-based measure of IOS to outperform *MTB* even more for loss firms. Indeed, the out-of-sample R^2 for the regressions predicting future SG&A and CAPEX using only *MTB* and control variables (cash flow, *HHI*, and year-fixed effects)

become negative in the sample of loss firms (Panel B). By contrast, we do not observe such a decline in out-of-sample R^2 for regressions using IOS factors to predict future investments. The out-of-sample R^2 for our factors predicting SG&A declines in the loss subsample relative to the full sample too, but the decline is small relative to the factors' high out-of-sample R^2. We see an improvement in out-of-sample R^2 for regressions predicting R&D using either *MTB* (10%) or factors (9%) for loss firms relative to the full sample. This result is consistent with Darrough and Ye's (2007) argument that many loss firms are R&D-intensive. For these firms, as our results indicate, both *MTB* and IOS factors are better at capturing future R&D investment.

Panel C depicts out-of-sample R^2 for the sub-sample of firms with more valuation difficulty proxied by higher-than-median variance of returns. For firms whose stock prices are more volatile, we expect *MTB* to be a noisier signal of IOS, and thus expect a lower ability to predict future investments. Indeed, the out-of-sample R^2 for predicting SG&A using only *MTB* and control variables becomes negative in the sub-sample, whereas IOS factors see an improvement in out-of-sample R^2 (39.7% in the subsample vs. 38.7% in the full sample). For R&D and CAPEX, however, out-of-sample R^2 for both regressions using factors and *MTB* improve in the high-return-variance sub-sample relative to the full sample. Overall, IOS factors predict future investments better than *MTB* to a greater degree in situations when market prices are likely less informative.

VI. CONCLUSION

We argue that firms' IOSs are multi-dimensional and poorly captured by existing unidimensional proxies for investment opportunities, such as Tobin's Q as proxied by *MTB*. We propose that a multi-dimensional measure of investment opportunities can overcome such limitations and suggest one such multi-dimensional measure based on textual analysis of 10-K Forms.

Using textual analysis, we identify key word lists and associated factors that describe and measure firms' investment opportunities on multiple dimensions. General factors include *Impairment, Debt Intensity, Executive Changes, Preferred Stock Buyback and Capital Seeking*, while industry-specific factors include *Bio-Pharmaceutical*, *Banking*, *Information Technology*, *Oil & Gas* and *Semi-conductor*. Our textual factor scores predict firms' future investments, and many related corporate policies, substantially better than *MTB* and industry-fixed effects. The factors also capture exogenous shocks to IOS via regulations, such as the Bush Steel Tariffs in 2002.

We contribute a better measure of IOS variaion, as captured by our 62 textual factors, and a better understanding of the determinants of corporate policies as explained by these factors. Such understanding could help researchers identify settings that pin down reasons for specific changes in corporate policies, since major corporate decisions might be related to different characteristics of firms' IOS. Our multi-dimensional measures of IOS could help investors improve stock screening without reading through 10-K statements in detail. This, in turn, can improve stock and bond market efficiency, and consequently, improve financial resource allocation in the economy. Simultaneously, we show that Forms 10-K have substantial information regarding IOS—arguably the most important input to firms' production function. This insight provides a new perspective for the financial accounting literature, where the main question has been how well accounting provides information about firms' output variables, such as earnings and returns. Finally, our approach of using the Lasso and factor analysis in selecting and classifying key words pertaining to IOS could be used to multi-dimensionally measure different concepts, such as earnings management, bankruptcy probability, and investment efficiency.

We acknowledge several limitations to our findings. First, our textual IOS factors rely on the assumption that managers truthfully disclose all substantive data about realistic IOS in 10-K

Forms. Testing hypotheses about managers' incentives to hide or provide overly optimistic information about IOS could be fruitful for future research. Theory-guided procedures are needed to remove the effect of managers' incentives from the estimated textual IOS.

Second, we rely on the assumption that managers make optimal investments. In other words, we assume there is no over- or under-investment. Although on average, overinvestment and underinvestment could offset each other and thus not affect the textual IOS factors' measurement, we call for more research on situations when sub-optimal investments bias our factor scores.

Third, we assume that one-year-ahead investments on average reflect all investment opportunities at different horizons. While our textual IOS factors predict future investments at different horizons (i.e. one-, three-, five-year ahead) well, we can relax this assumption to improve the predictive power of the factors even further. However, this analysis would require more time-series data, unless future research identifies some clever way to overcome this obstacle.

We also see many directions for future research. For example, one could examine market reaction to changes in the textual IOS factors as a validation test. Since we focus on managers' disclosure of realistic IOS in Forms 10-K, market reactions to the textual IOS factors are beyond our scope In examining market reactions, we must carefully analyze disclosure timing to avoid confounding events, which we leave for future research. Another research direction is to test whether the IOS factors respond to technological and financial constraint shocks. Along this line, one might be able to evaluate various constraints in shaping IOS by decomposing changes in IOS into components associated with financial constraints, technological constraints, and regulatory constraints.

REFERENCES

Adam, Tim, and Vidhan K. Goyal, 2008, The investment opportunity set and its proxy variables, *Journal of Financial Research* 31, 41–63.

Alchian, Armen A., 1950, Uncertainty, Evolution, and Economic Theory, *Journal of Political Economy* (The University of Chicago Press).

Baber, William R., Surya N. Janakiraman, and Sok-Hyon Kang, 1996, Investment opportunities and the structure of executive compensation, *Journal of Accounting and Economics* 21, 297–318.

Baker, Malcolm P, and Jeffrey Wurgler, 2002, Market Timing and Capital Structure, *The Journal of Finance* LVII, 1–32.

Barney, Jay Barry, 1991, Firm resources and sustained competitive advantage, *Journal of Management* 17, 99–120.

Becker, Gary S., 1962, Irrational Behavior and Economic Theory, *Journal of Political Economy* 70, 1–13.

Billett, Matthew T., Tao Hsien Dolly King, and David C. Mauer, 2007, Growth opportunities and the choice of leverage, debt maturity, and covenants, *Journal of Finance* 62, 697–730.

Blundell, Richard, Stephen Bond, Michael Devereux, and Fabio Schiantarelli, 1992, Investment and Tobin ' s Q Evidence from company panel data, *Journal of Econometrics* 51, 233–257.

Brailsford, Timothy J., and Daniel Yeoh, 2004, Agency Problems and Capital Expenditure Announcements, *The Journal of Business* 77, 223–256.

Brown, David T., Christopher M. James, and Robert M. Mooradian, 1994, Asset sales by financially distressed firms, *Journal of Corporate Finance* 1, 233–257.

Buehlmaier, Matthias M. M., and Toni M. Whited, 2018, Are Financial Constraints Priced? Evidence from Textual Analysis, *Review of Financial Studies* 31, 17–20.

Cahan, Steven F., Jayne M. Godfrey, Jane Hamilton, and Debra C. Jeter, 2008, Auditor Specialization, Auditor Dominance, and Audit Fees: The Role of Investment Opportunities, *The Accounting Review* 83, 1393–1423.

Carpenter, Robert E., and Alessandra Guariglia, 2008, Cash flow, investment, and investment opportunities: New tests using UK panel data, *Journal of Banking and Finance* 32, 1894–1906.

Chen, Sheng-Syan, Kim Wai Ho, Cheng-few Lee, and Gillian H.H. Yeo, 2000, Investment opportunities, free cash flow and market reaction to international joint ventures, *Journal of Banking & Finance* 24, 1747–1765.

Chung, Kee H., Peter Wright, and Charlie Charoenwong, 1998, Investment opportunities and market reaction to capital expenditure decisions, *Journal of Banking & Finance* 22, 41–60.

Darrough, Masako, and Jianming Ye, 2007, Valuation of loss firms in a knowledge-based economy, *Review of Accounting Studies* 12, 61–93.

Denis, D., 1994, Investment Opportunities and the Market Reaction to Equity Offerings, *Journal of Financial and Qualitative Analysis* 29, 159–177.

Dittmar, Amy K, 2000, Why Do Firms Repurchase Stock?, *The Journal of Business* 73, 331–355.

Eisfeldt, Andrea L., and Dimitris Papanikolaou, 2013, Organization Capital and the Cross-Section of Expected Returns, *The Journal of Finance* 68, 1365–1406.

Eisfeldt, Andrea L., and Dimitris Papanikolaou, 2014, The Value and Ownership of Intagible Capital, *American Economic Review: Papers and Proceedings* 104, 189–194.

Fama, Eugene F, and Kenneth R French, 2002, Testing Trade-Off and Pecking Order Predictions About Dividends and Debt, *Review of Financial Studies* 15, 1–33.

Farre-Mensa, Joan, and Alexander Ljungqvist, 2016, Do measures of financial constraints measure financial constraints?, *Review of Financial Studies* 29, 271–308.

Feng, Li, Russell Lundholm, and Minnis Michael, 2013, A measure of competition based on 10-k filings, *Journal of Accounting Research* 51, 399–436.

Fisher, Irving, 1930, *The Theory of Interest* (New York: The Macmillan Co.).

Frankel, Richard, Jared Jennings, and Joshua Lee, 2016, Using unstructured and qualitative disclosures to explain accruals, *Journal of Accounting and Economics* 62, 209–227.

Gala, Vito D, 2015, Measuring Marginal q, *Working Paper*.

Gaver, Jennifer J., and Kenneth M Gaver, 1993, Additional evidence on the association between the investment opportunity set and corporate financing, dividend, and compensation policies, *Journal of Accounting and Economics* 16, 125–160.

Gentzkow, Matthew, Stanford Bryan, T Kelly, Chicago Booth, and Matt Taddy, 2017, Text as Data, *Working paper*, 1–53.

Gilchrist, Simon, and Charles P. Himmelberg, 1995, Evidence on the role of cash flow for investment, *Journal of Monetary Economics* 36, 541–572.

Hasbrouck, Joel, 1985, The characteristics of takeover targets q and other measures, *Journal of Banking and Finance* 9, 351–362.

Hayashi, Fumio, 1982, Tobin's Marginal q and Average q: A Neoclassical Interpretation, *Econometrica* 50, 213.

Heston, Steven L., and Nitish Ranjan Sinha, 2013, News versus Sentiment: Comparing Textual Processing Approaches for Predicting Stock Returns, *SSRN Electronic Journal*.

Hoberg, Gerard, and Vojislav Maksimovic, 2015, Redefining financial constraints: A text-based analysis, *Review of Financial Studies*.

Hoberg, Gerard, and Gordon Phillips, 2016, Text-Based Network Industries and Endogenous Product Differentiation, *Journal of Political Economy* 124, 1423–1465.

Jogerson, D. W., 1963, Capital Theory and Investment Behavior, *The American Economic Review* 53, 247–259.

Johnson, Steven, 2010, *Where Good Ideas Come from : The Natural History of Innovation* (Riverhead Books, New York).

Kaiser, Henry F, 1960, The Application of Electronic Computers to Factor Analysis, *Educational and Psychological Measurement* 20, 141–151.

Kaldor, Nicholas, 1966, Marginal Productivity and the Macro-Economic Theories of Distribution, *Review of Economic Studies*.

Kallapur, Sanjay, and Mark A. Trombley, 1999, The association between investment opportunity set proxies and realized growth, *Journal of Business Finance and Accounting* 26, 505–519.

Kauffman, Stuart, 1995, *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity Book*.

Kumar, Krishna R., and Gopal V. Krishnan, 2008, The Value-Relevance of Cash Flows and Accruals: The Role of Investment Opportunities, *The Accounting Review* 83, 997–1040.

Lai, Kam Wah, 2009, Does audit quality matter more for firms with high investment opportunities?, *Journal of Accounting and Public Policy* 28, 33–50.

Lang, Larry H. P., René M. Stulz, and Ralph A. Walkling, 1989, Managerial performance, Tobin's Q, and the gains from successful tender offers, *Journal of Financial Economics* 24, 137–154.

Lewellen, Wilbur G., and S. G. Badrinath, 1997, On the measurement of Tobin's q, *Journal of Financial Economics* 44, 77–122.

Li, Feng, 2010, Textual Analysis of Corporate Disclosures : A Survey of the Literature, *Journal of Accounting Literature*.

Loughran, Tim, and Bill McDonald, 2016, Textual Analysis in Accounting and Finance: A Survey, *Journal of Accounting Research* 54, 1187–1230.

Lyandres, Evgeny, and Alexei Zhdanov, 2013, Investment opportunities and bankruptcy prediction, *Journal of Financial Markets* 16, 439–476.

Martin, Kenneth J., 1996, The Method of Payment in Corporate Acquisitions, Investment Opportunities, and Management Ownership, *The Journal of Finance* 51, 1227–1246.

McConnell, John J., and Henri Servaes, 1990, Additional evidence on equity ownership and corporate value, *Journal of Financial Economics* 27, 595–612.

Mclaughlin, Patrick A, and Oliver Sherouse, 2017, QuantGov — A Policy Analytics Platform, *Working Paper*, 1–7.

Milgrom, Paul, and John Roberts, 1995, Complementarities and fit strategy, structure, and organizational change in manufacturing, *Journal of Accounting and Economics* 19, 179–208.

Myers, Stewart C., 1977, Determinants of corporate borrowing, *Journal of Financial Economics* 5, 147–175.

Nash, Robert C., Jeffry M. Netter, and Annette B. Poulsen, 2003, *Determinants of Contractual Relations between Shareholders and Bondholders: Investment Opportunities and Restrictive Covenants, Journal of Corporate Finance*. Vol. 9.

Newbert, Scott L., 2007, Empirical Research On The Resource-Based View Of The Firm: An Assessment And Suggestions For Future Research, *Strategic Management Journal* 28, 121–146.

Opler, Tim, and Sheridan Titman, 1993, The Determinants of Leveraged Buyout Activity: Free Cash Flow vs. Financial Distress Costs, *The Journal of Finance* 48, 1985–1999.

Peters, Ryan H., and Lucian A. Taylor, 2017, Intangible capital and the investment-q relation, *Journal of Financial Economics* 123, 251–272.

Pett, Marjorie, Nancy Lackey, and John Sullivan, 2003, *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research*.

Porter, Michael E., 1979, How competitive forces shape strategy, *Harvard Business Review* 2, 137–145.

Purda, Lynnette, and David Skillicorn, 2015, Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection, *Contemporary Accounting Research* 32, 1193–1223.

Richardson, Scott A., 2006, Over-investment of free cash flow, *Review of Accounting Studies*.

Skinner, Douglas J., 1993, The investment opportunity set and accounting procedure choice. Preliminary evidence, *Journal of Accounting and Economics* 16, 407–445.

Smith, Clifford W., and Ross L. Watts, 1992, The investment opportunity set and corporate financing, dividend, and compensation policies, *Journal of Financial Economics* 32, 263–292.

Strebulaev, Ilya A, and Toni M. Whited, 2011, Dynamic Models and Structural Estimation in Corporate Finance, *Foundations and Trends{®} in Finance* 6, 1–163.

Szewczyk, Samuel H., George P. Tsetsekos, and Zaher Zantout, 1996, The Valuation of Corporate R & D Expenditures : Evidence from Investment Opportunities and Free Cash Flow, *Financial Management* 25, 105–110.

Tibshirani, Robert, 1996, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 58, 267–288.

Tobin, James, 1969, A general equilibrium approach to monetary theory, *Journal of Money, Credit and Banking* 1, 15–29.

Wernerfelt, Birger, 1984, A resource-based view of the firm, *Strategic Management Journal* 5, 171–180.

Whited, Toni M., and Timothy Erickson, 2012, Treating Measurement Error in Tobin's Q, *The Review of Financial Studies* 25, 1286–1329.

Table 1: Variable description

Variable	Definition
MTB	Market-to-book ratio, defined as the market value of assets scaled by the end-of-year book value of assets
R&D	The ratio of research and development expenditures to the end-of-year book value of assets
SG&A	The ratio of selling, general and administrative (SG&A) expenses (excluding R&D) to the end-of-year book value of assets
CAPEX	The ratio of capital expenditure to the end-of-year book value of assets
INV	The sum of *R&D*, 0.3**SG&A*, and *CAPEX*
SG	The change in sales in year t scaled by the sales in year t-1
MLEV	Market leverage, defined as book-value of debt scaled by the book value of debt plus the market value of equity
DIssue	Funds received from long-term debt issuance divided by the end-of-year book value of assets
EIssue	Funds received from the issuance of common and preferred stock divided by the end-of-year book value of assets
FCF	Free cash flow from existing assets in place, defined as the cash flow from operating activities minus maintenance investment (depreciation and amortization) expenditure plus research and development expenditure divided by the end-of-year book value of assets
DIV	Total dividends declared on all equity capital of the firm divided by the end-of-year book value of assets
REPO	Purchase of common stock divided by the end-of-year book value of assets
Payout	The sum of *DIV* and *REPO*
CF	Operating cash flow, defined as the cash flow from operating activities divided by the end-of-year book value of assets
HHI	Herfindahl-Hirschman index computed using TNIC designations in Hoberg-Phillips database

EBIT	Pre-interest, pretax earnings divided by the end-of-year book value of assets
DEP	Depreciation expense divided by the end-of-year book value of assets
LnAT	Natural logarithm of the end-of-year book value of assets

Variable	Definition
TREAT	For the test on the effects on steel producers, TREAT equals 1 if the firm is a steel producer (SIC=3312); For the test on the effects on steel consumers, TREAT equals 1 if the firm is a steel consumer (SIC=3317); and for the test on the combined effects, TREAT equals +1 if the firm is a steel producer and -1 if the firm is a steel consumer. In all cases, TREAT equals 0 for all other manufacturing firms, defined by FF12=3.
$POST_{tariffs\ initiated}$	Equals 1 if fiscal year equals 2003, and 0 if fiscal year equals 2001
$POST_{tariffs\ lifted}$	Equals 1 if fiscal year equals 2002, and 0 if fiscal year equals 2004
Regulation	A text-based industry-year measure of regulatory restrictions, developed by McLaughlin and Sherouse (2017). We deflate this variable by 100,000.

Table 2: Sample selection

	No. of obs
All available Form 10-K filings on EDGAR database released from fiscal years 1995—2015	163,729
Less:	
Observations without sufficient financial data	(97,385)
Observations with stock price less than $1	(7,798)
Observations with total assets less than $ 5 million	(814)
Observations without Text-based Network Industry Classification (TNIC-3) industry concentration data in Hoberg-Phillips database	(4,408)
Final sample	53,324
Training sample from 1995 to 2009	38,101
Test sample from 2010 to 2015	15,223

Table 3: Word lists and top firm-year observations for top 10 factors

Panel A: Top 5 general factors

Factor 1	Factor 4	Factor 6	Factor 8	Factor 10
Impairment	***Debt Intensity***	***Executive Change***	***Preferred Stock Buyback***	***Seeking Capital***
Top 5 words				
fair	borrow	employ	prefer	go
result	lender	execut	seri	rais
valu	agent	agreement	convers	concern
financi	document	caus	redeem	deficit
affect	lien	confidenti	convert	bulletin
Top 5 firm-year				
Intersil Corp -Cl A - 2009	Modtech Holdings Inc - 2005	Med-Design Corp - 2002	Trinsic Inc - 2004	Uniontown Energy Inc - 2009
Advanced Energy Inds I - 2008	Danka Business Systems - 2006	Metasolv Inc - 2002	Conagra Brands Inc - 1995	Magnegas Corp - 2008
Ddi Corp - 2008	X-Rite Inc - 2005	Market Facts Inc - 1996	On Command Corp - 2001	Lone Star Gold Inc - 2010
Intersil Corp -Cl A - 2008	Wellman Inc - 2003	Sequenom Inc - 2000	Disc Inc - 2001	Kedem Pharmaceuticals - 2007
G&K Services Inc -Cl - 2008	Foamex International I - 2006	Excelon Corp - 1998	Egain Corp - 2000	Kali Inc - 2008

Panel B: Top 5 industry-related factors

Factor 2	Factor 3	Factor 5	Factor 7	Factor 9
Banking	***Bio-pharma***	***IT***	***Oil & Gas***	***Semiconductor***
Top 5 words				
loan	trial	support	oil	semiconductor
bank	drug	user	prove	manufactur
portfolio	collabor	solut	ga	technolog
interest	mileston	function	natur	design
capit	approv	revenu	exploratori	intellectu
Top 5 firm-year				
Chicopee Bancorp Inc - 2009	Aradigm Corp - 2009	Art Technology Group I - 1999	Tengasco Inc - 2002	Semitool Inc - 2004
Lincoln Bancorp/In - 2005	Ariad Pharmaceuticals - 2009	Support.Com Inc - 2002	Petrohawk Energy Corp - 2004	Authentec Inc - 2009
First Bancorp Inc/Me - 2012	Arena Pharmaceuticals - 2008	Microstrategy Inc - 1998	Halcon Resources Corp - 2009	Semitool Inc - 2006
Peoples Bancorp Auburn - 2000	Xoma Corp - 2012	Inktomi Corp - 1998	Carrizo Oil & Gas Inc - 2004	Semitool Inc - 2007
Tib Financial Corp - 2009	Anadys Pharmaceuticals - 2004	Enterprise Informatics - 2002	Evolution Petroleum Co - 2010	Semitool Inc - 2002

This table shows top 10 factors that have the highest associated eigenvalues produced in our factor analysis. We run the factor analysis on tf-idfs (term frequency - inverse document frequency) of 45,061 10-Ks from financial year 1994 to financial year 2009, applying minimum 1 cutoff for eigenvalues, promax rotation, SMC priors and principal component method. This number of observations is different from one reported in table 2 because here we do not exclude 10-Ks with missing control variables needed for subsequent regression tests, such as cash flow, depreciation, HHI, EBIT, lnAT.... The table includes panel A, which shows top 5 industry-related factors, and panel B, which shows top 5 general factors. Each panel shows, for each factor, top 5 tokenized words that have the highest loadings and top 5 firm-year observations that have the highest factor scores, both in descending order.

Table 4: Summary statistics for the top 30 observations on each factor

Panel A: Top 5 general factors

	Factor 1	Factor 4	Factor 6	Factor 8	Factor 10
	Impairment	***Debt Intensity***	***Executive Change***	***Preferred Stock Buyback***	***Seeking Capital***
MTB_t	2.19	1.41	25.17	16.73	15262.81
$SG\&A_{t+1}$	0.33	0.27	0.91	1.10	3.05
$R\&D_{t+1}$	0.06	0.01	0.17	0.19	0.06
$CAPEX_{t+1}$	0.03	0.05	0.03	0.02	0.06
INV_{t+1}	0.17	0.14	0.42	0.49	1.02
$MLEV_{t+1}$	0.22	0.48	0.27	0.53	0.09
ROA_t	(0.04)	0.06	(0.58)	(1.49)	(2.72)
ROA_{t+1}	(0.00)	0.04	(0.80)	(1.39)	(2.49)
SG_{t+1}	0.03	0.11	0.15	0.21	0.77

Panel B: Top 5 industry-related factors

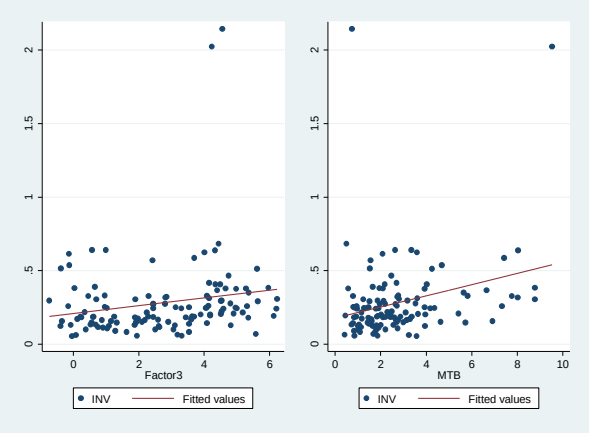
	Factor 2	Factor 3	Factor 5	Factor 7	Factor 9
	Banking	***Bio-pharma***	***IT***	***Oil & Gas***	***Semi-conductor***
MTB_t	0.98	4.61	2.53	1.82	2.78
$SG\&A_{t+1}$	0.02	0.14	0.44	0.06	0.47
$R\&D_{t+1}$	-	0.40	0.11	-	0.27
$CAPEX_{t+1}$	0.00	0.01	0.02	0.24	0.04
INV_{t+1}	0.01	0.34	0.23	0.26	0.37
$MLEV_{t+1}$	0.93	0.12	0.23	0.41	0.19
ROA_t	0.00	(0.31)	0.03	0.04	(0.17)
ROA_{t+1}	0.00	(0.44)	0.02	0.03	(0.17)
SG_{t+1}	(0.02)	(0.00)	0.07	0.29	(0.06)

This table shows summary statistics for top 30 firm-year observations based on factor scores rankings for each factor in the out-of-sample period (2010 onwards). Panel A depicts top 5 industry-related factors while panel B depicts top 5 general factors. We run the factor analysis on tf-idfs (term frequency - inverse document frequency) of 45,061 10-Ks from financial year 1994 to financial year 2009, applying minimum 1 cutoff for eigenvalues, promax rotation, SMC priors and principal component method. This number of observations is different from one reported in table 2 because here we do not exclude 10-Ks with missing control variables needed for subsequent regression tests, such as cash flow, depreciation, HHI, EBIT, lnAT... R&D denotes research and development expenses; SG&A denotes selling, general and administrative expenses; CAPEX denotes capital expenditures; all of these scaled by beginning-of-year total assets. INV is total investment, calculated by summing R&D, SG&A and CAPEX. MLEV denotes market leverage in the next one year. t denotes current year, t+1 denotes the following year.

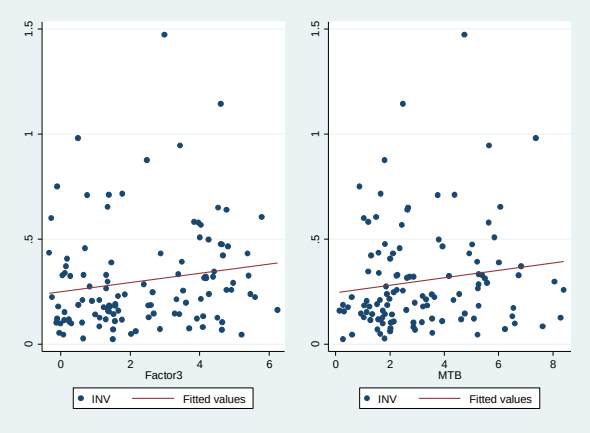
Figure 1: Within-industry variations of factor scores, market-to-book ratios, and future actual investments

A bio-pharmaceutical industry (SIC=2834)

Financial year 2002

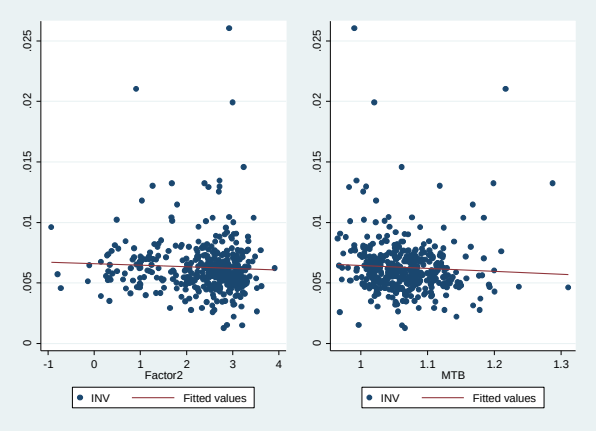


Financial year 2012

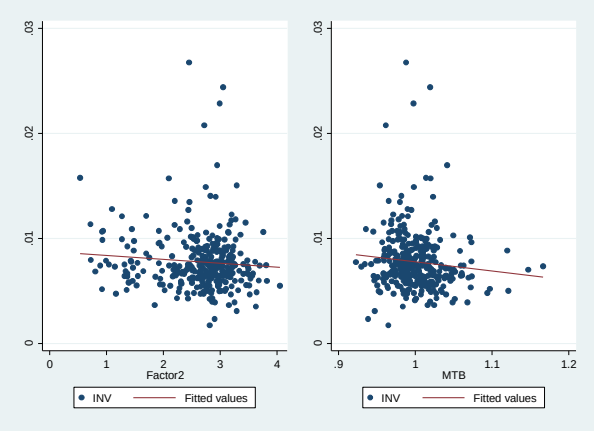


A banking industry (SIC=6020)

Financial year 2002



Financial year 2012



This figure plots the within-industry cross-sectional variations in one bio-pharmaceutical industry (SIC code 2834 Pharmaceutical Preparation) and one banking industry (SIC code 6020 Commercial Banks) in one in-sample year 2002 and one out-of-sample year 2012. For each of the scatter plot, the y-axis denotes total investment (INV: year t+1 investments in SG&A, R&D and CAPEX, scaled by total assets at the end of year t). The x-axis represents either the factor score (*Bio-Pharma* Factor 3 - or *Banking* Factor 2) or end-of-current-year market-to-book (MTB) ratio. We exclude outliers with market-to-book ratios greater than 10.

Table 5: Regressions of investments

Panel A Regression coefficients

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
VARIABLE		$SG\&A_{t+1}$			$R\&D_{t+1}$			$CAPEX_{t+1}$			INV_{t+1}	
$Impairment_t$		0.000	0.000		0.002	0.002*		-0.003***	-0.002***		-0.001	-0.001
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
$Banking_t$		-0.028***	-0.027***		-0.008***	-0.008***		-0.004***	-0.003***		-0.018***	-0.017***
		(0.01)	(0.01)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
$Bio\text{-}pharma_t$		-0.090***	-0.091***		0.039***	0.038***		-0.002**	-0.002***		-0.002	-0.002
		(0.01)	(0.01)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
$Debt\ Intensity_t$		0.018***	0.018***		-0.001*	-0.001		0.002***	0.002***		0.006***	0.007***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
IT_t		0.034***	0.033***		0.027***	0.026***		-0.004***	-0.005***		0.024***	0.022***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
$Executive\ change_t$		0.009***	0.009***		-0.001	-0.001		0.001**	0.001***		0.003***	0.004***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
$Oil\ \&\ Gas_t$		0.001	0.001		-0.003***	-0.003***		0.018***	0.018***		0.020***	0.019***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
$Preferred\ Stock_t$		0.007***	0.007***		-0.001	-0.001		-0.000	-0.000		0.002**	0.001
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
$Semiconductor_t$		-0.043***	-0.043***		0.008***	0.007***		0.006***	0.006***		-0.001	-0.001
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
$Seeking\ Capital_t$		0.058***	0.057***		0.007***	0.006***		0.002**	0.001		0.026***	0.024***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
MTB_t	0.010***		0.006***	0.009***		0.004***	0.004***		0.003***	0.014***		0.007***

	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)
Adj R^2	0.294	0.452	0.453	0.546	0.647	0.649	0.315	0.405	0.411	0.323	0.427	0.436

This panel reports OLS regression estimates on a sample of 38,101 firm-year observations from 1994 to 2009. Specifically, each dependent variable is regressed in ten different models with various explanatory variables. MTB denotes market to book ratio of assets for each firm. R&D denotes research and development expenses, scaled by contemporary assets. $SG\&A_{t+1}$ denotes selling, general and administrative expenses (excluding R&D) in the following year, scaled by end-of-current-year assets. Similarly, $CAPEX_{t+1}$ denotes capital expenditures, and $R\&D_{t+1}$ denotes research and development expenses. INV denotes the sum of 30% of SG&A, R&D, and CAPEX. The numbers in parentheses are clustered standard errors by firm. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively, in two-tailed tests. Cash flow, HHI, year fixed effects, and industry fixed effects are included in all models. The variables from *Impairment* to *Seeking Capital* are factor scores based on factor analysis of tokenized words tf-idf (term frequency - inverse document frequency) for all the available 10-K texts from 1994 to 2009. We also include the remaining IOS factors in the regressions but do not show the corresponding coefficients.

Panel B Out-of-Sample predictive power (Pseudo R^2 %)

		(1)	(2)	(3)	(4)	(5)	(6)
Row	D.V.	MTB_t	$Factors_t$	*Ind FE*	MTB_t + *Ind FE*	$Factors_t$ + *Ind FE*	MTB_t+ $Factors_t$ + *Ind FE*
(I)	$SG\&A_{t+1}$	3.31	37.55	29.59	30.38	41.97	42.17
(II)	$R\&D_{t+1}$	44.82	60.73	55.95	58.12	62.67	63.46
(III)	$CAPEX_{t+1}$	-5.85	30.18	17.59	19.38	32.83	34.06
(IV)	INV_{t+1}	12.64	38.50	24.67	29.55	41.01	43.08

This panel displays the out-of-sample Pseudo R-squared for each OLS regressions in which the future investment dependent variable is specified by row and the explanatory variables are specified by column. Pseudo R-squared is equal to one minus the ratio of MSE from a forecasting model to that of the benchmark model (an intercept-only model). Specifically, each dependent variable is regressed in six different models with various explanatory variables. MTB denotes market to book ratio of assets for each firm. R&D denotes research and development expenses, scaled by contemporary assets. $SG\&A_{t+1}$ denotes selling, general and administrative expenses (excluding R&D) in the following year, scaled by end-of-current-year assets. Similarly, $CAPEX_{t+1}$ denotes capital expenditures, and $R\&D_{t+1}$ denotes research and development expenses. INV denotes the sum of 30% of SG&A, R&D, and CAPEX. Ind FE denotes Fama-French 48 industry fixed effects. Factors denotes the 62 factor scores produced by the factor analysis procedure. Cash flow, HHI, and year fixed effects are included in all models.

Table 6: Regressions of financing policies

Panel A Regression coefficients

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
VARIABLES		$MLEV_{t+1}$			$DIssue_{t+1}$			$EIssue_{t+1}$			FCF_{t+1}	
Impairment$_t$		-0.019***	-0.020***		-0.009**	-0.009**		-0.007***	-0.006***		0.008***	0.009***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Banking$_t$		0.036***	0.035***		0.007	0.007		-0.001	0.001		-0.002	-0.002
		(0.01)	(0.01)		(0.01)	(0.01)		(0.00)	(0.00)		(0.00)	(0.00)
Bio-pharma$_t$		-0.009***	-0.009***		-0.002	-0.002		0.007***	0.008***		-0.005**	-0.004**
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Debt Intensity$_t$		-0.008***	-0.009***		0.006***	0.006***		0.000	0.001		-0.000	-0.000
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
IT$_t$		-0.013***	-0.011***		-0.011***	-0.012***		-0.011***	-0.013***		0.011***	0.011***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Executive change$_t$		0.012***	0.011***		0.002	0.002		-0.001	-0.000		-0.003***	-0.003***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Oil & Gas$_t$		-0.007**	-0.005*		0.006**	0.006**		0.003***	0.002**		0.002*	0.002
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Preferred Stock$_t$		0.007***	0.009***		0.002	0.002		0.007***	0.006***		-0.005***	-0.005***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Semiconductor$_t$		-0.010***	-0.009***		-0.008**	-0.008**		-0.000	-0.001		0.001	0.001
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Seeking Capital$_t$		0.002	0.006*		-0.003	-0.003		0.017***	0.013***		-0.040***	-0.041***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
MTB$_t$	-0.029***		-0.020***	-0.003***		0.002**	0.019***		0.018***	0.002**		0.004***

	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)
Adj R^2	0.307	0.438	0.452	0.049	0.105	0.105	0.383	0.367	0.410	0.372	0.408	0.410

This panel reports OLS regression estimates on a sample of 38,101 firm-year observations from 1994 to 2009. Specifically, each dependent variable is regressed in ten different models with various explanatory variables. MLEV: market leverage ratios averaged in the following year; DIssue: funds raised with long-term debt in the following year; EIssue: funds raised with equity in the following year; FCF: free cash flow for internal financing in the following year. t subscript denotes variables measured at the end of year t. The numbers in parentheses are clustered standard errors by firm. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively, in two-tailed tests. Control variables (profitability, R&D, depreciation, firm size, and HHI), year fixed effects, and industry fixed effects are included in all models. The variables from *Impairment* to *Seeking Capital* are factor scores based on factor analysis of tokenized words tf-idf (term frequency - inverse document frequency) for all the available 10-K texts from 1994 to 2009. We also include the remaining IOS factors in the regressions but do not show the corresponding coefficients.

Panel B Out-of-Sample predictive power (Pseudo R^2 %)

		(1)	(2)	(3)	(4)	(5)	(6)
Row	D.V.	MTB_t	$Factors_t$	*Ind FE*	MTB_t + *Ind FE*	$Factors_t$ + *Ind FE*	MTB_t+ $Factors_t$ + *Ind FE*
(I)	$MLEV_{t+1}$	19.71	37.33	20.23	22.98	38.19	40.41
(II)	$DIssue_{t+1}$	3.26	7.86	5.03	4.86	7.99	8.04
(III)	$EIssue_{t+1}$	45.89	48.03	45.01	47.60	48.07	50.10
(IV)	FCF_{t+1}	41.98	45.88	43.14	43.48	46.02	46.72

This panel displays the out-of-sample Pseudo R-squared for each OLS regressions in which the future investment dependent variable is specified by row and the explanatory variables are specified by column. Pseudo R-squared is equal to one minus the ratio of MSE from a forecasting model to that of the benchmark model (an intercept-only model). Specifically, each dependent variable is regressed in six different models with various explanatory variables. MLEV: market leverage ratios averaged in the following year; DIssue: funds raised with long-term debt in the following year; EIssue: funds raised with equity in the following year; FCF: free cash flow for internal financing in the following year. t subscript denotes variables measured at the end of year t. Ind FE denotes Fama-French 48 industry fixed effects. Factors denotes the 62 factor scores produced by the factor analysis procedure. Control variables (profitability, R&D, depreciation, firm size, and HHI) and year fixed effects are included in all models.

Table 7: Regressions of payout policies

Panel A Regression coefficients

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
VARIABLES		DIV_{t+1}			$REPO_{t+1}$			$Payout_{t+1}$	
Impairment$_t$		0.001**	0.001**		-0.000	-0.000		0.001	0.001*
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Banking$_t$		-0.002***	-0.002***		-0.001**	-0.001**		-0.003***	-0.003***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Bio-pharma$_t$		-0.002***	-0.002***		-0.001**	-0.001**		-0.002***	-0.002***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Debt Intensity$_t$		0.000	0.000		0.000	0.000		0.000	0.000
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
IT$_t$		-0.002***	-0.002***		-0.000	-0.000		-0.002***	-0.003***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Executive change$_t$		-0.001***	-0.001***		-0.000	-0.000		-0.001***	-0.001***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Oil & Gas$_t$		-0.002***	-0.002***		0.000**	0.000**		-0.001*	-0.001**
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Preferred Stock$_t$		0.005***	0.005***		0.004***	0.004***		0.008***	0.008***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Semiconductor$_t$		-0.003***	-0.003***		-0.001*	-0.001*		-0.003***	-0.003***
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Seeking Capital$_t$		0.001	0.000		0.001**	0.001**		0.002**	0.001*
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
MTB$_t$	0.002***		0.002***	0.000***		0.000*	0.003***		0.003***

	(0.00)			(0.00)	(0.00)			(0.00)	(0.00)			(0.00)
Adj R^2	0.086	0.165		0.180	0.007	0.036		0.036	0.058	0.122		0.133

This panel reports OLS regression estimates on a sample of 38,101 firm-year observations from 1994 to 2009. Specifically, each dependent variable is regressed in ten different models with various explanatory variables. DIV denotes total dividends declared on all equity capital of the firm in the following year divided by the end-of-current year book value of assets, REPO denotes share repurchase in the following year scaled by end-of-current-year book value of assets. Payout denotes the sum of DIV and REPO. t subscript denotes variables measured at the end of year t. The numbers in parentheses are clustered standard errors by firm. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively, in two-tailed tests. Control variables (profitability, R&D, depreciation, firm size, and HHI), year fixed effects, and industry fixed effects are included in all models. The variables from *Impairment* to *Seeking Capital* are factor scores based on factor analysis of tokenized words tf-idf (term frequency - inverse document frequency) for all the available 10-K texts from 1994 to 2009. We also include the remaining IOS factors in the regressions but do not show the corresponding coefficients. All other variables are defined as in the description of Panel A above.

Panel B Out-of-Sample predictive power (Pseudo R^2 %)

		(1)	(2)	(3)	(4)	(5)	(6)
Row	D.V.	MTB_t	$Factors_t$	*Ind FE*	MTB_t + *Ind FE*	$Factors_t$ + *Ind FE*	MTB_t+ $Factors_t$ + *Ind FE*
(I)	DIV_{t+1}	4.75	14.40	5.62	7.95	15.70	17.83
(II)	$REPO_{t+1}$	-5.04	-4.19	-5.07	-5.04	-4.50	-4.37
(III)	$Payout_{t+1}$	4.20	10.98	4.63	6.83	11.73	13.86

This panel displays the out-of-sample Pseudo R-squared for each OLS regressions in which the future investment dependent variable is specified by row and the explanatory variables are specified by column. Pseudo R-squared is equal to one minus the ratio of MSE from a forecasting model to that of the benchmark model (an intercept-only model). Specifically, each dependent variable is regressed in six different models with various explanatory variables. DIV denotes total dividends declared on all equity capital of the firm in the following year divided by the end-of-current year book value of assets, REPO denotes share repurchase in the following year scaled by end-of-current-year book value of assets. Payout denotes the sum of DIV and REPO. t subscript denotes variables measured at the end of year t. Ind FE denotes Fama-French 48 industry fixed effects. Factors denotes the 62 factor scores produced by the factor analysis procedure. Control variables (profitability, R&D, depreciation, firm size, and HHI) and year fixed effects are included in all models.

Table 8: Effects of Bush Steel Tariff on IOS in Metal Manufacturing Industry

	Effects on producers		**Effects on consumers**		**Effects on producers - consumers**	
TREAT	3.18***	5.30***	3.44***	2.26**	1.36***	3.35***
	(0.36)	(0.49)	(0.57)	(0.80)	(0.35)	(0.47)
$POST_{tariffs\ initiated}$	0.10		0.10		0.10	
	(0.11)		(0.11)		(0.12)	
$POST_{tariffs\ lifted}$		0.08		0.08		0.07
		(0.11)		(0.10)		(0.12)
$TREAT*POST_{tariffs\ initiated}$	2.13***		-1.18		2.00***	
	(0.61)		(0.99)		(0.59)	
$TREAT*POST_{tariffs\ lifted}$		-1.85**		-0.02		-1.53**
		(0.59)		(0.94)		(0.56)
Number of observations	523	573	512	561	529	580
Adjusted R^2	0.27	0.28	0.07	0.04	0.11	0.13

This table presents regression results of the difference-in-differences design examining the effects of Bush Steel Tariff (March 2002 - December 2003) on Factor 22 *Metal Manufacturing* scores. All regressions have Factor 22 as the dependent variable. In all regressions, the control group (TREAT=0) is the same: all manufacturing firms other than steel producers and consumers (Fama-French 12 industry code FF12=3). However, the treatment groups are different accross regressions: for the first two regressions, TREAT equals 1 if the firm is a steel producer (SIC=3312); for the next two regressions, TREAT equals 1 if the firm is a steel consumer (SIC=3317); and for the last two regressions, TREAT equals +1 if the firm is a steel producer and -1 if the firm is a steel consumer. $POST_{tariffs\ initiated}$ equals 1 if fiscal year is 2003 and equals 0 if fiscal year is 2001. $POST_{tariffs\ lifted}$ equals 1 if fiscal year is 2002 and equals 0 if fiscal year is 2004. Standard errors are in parentheses, significance levels * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Table 9: Factors' response to changes in regulation

Panel A: Top 5 industry-related factors

	Factor 2 *Banking*		Factor 3 *Bio-pharma*		Factor 5 *IT*		Factor 7 *Oil & Gas*		Factor 9 *Semi-conductor*	
$Regulation_{t-1}$	-0.75***		-0.15***		-0.72***		-0.24***		-0.21***	
	(0.047)		(0.040)		(0.042)		(0.044)		(0.032)	
$Regulation_t$		-0.58***		-0.01		-0.66***		-0.16***		-0.23***
		(0.038)		(0.033)		(0.036)		(0.035)		(0.027)
Intercept	0.38***	0.33***	0.11***	0.10***	0.05***	0.06***	0.16***	0.15***	-0.14***	-0.11***
	(0.010)	(0.008)	(0.008)	(0.007)	(0.009)	(0.008)	(0.009)	(0.007)	(0.007)	(0.006)
R^2	0.009	0.006	0.001	0.000	0.010	0.008	0.001	0.001	0.001	0.002

Panel B: Top 5 general factors

	Factor 1 *Impairment*		Factor 4 *Debt Intensity*		Factor 6 *Executive Employment*		Factor 8 *Preferred Stock Buyback*		Factor 10 *Seeking Capital*	
$Regulation_{t-1}$	1.24***		-0.93***		-1.01***		0.27***		-0.40***	
	(0.086)		(0.068)		(0.071)		(0.067)		(0.045)	
$Regulation_t$		1.59***		-0.88***		-0.89***		0.37***		-0.36***
		(0.070)		(0.058)		(0.060)		(0.058)		(0.041)
Intercept	-0.05**	-0.18***	0.00	0.02	0.04**	0.04***	0.04**	0.05***	0.07***	0.12***
	(0.018)	(0.015)	(0.014)	(0.012)	(0.015)	(0.012)	(0.014)	(0.012)	(0.009)	(0.009)
R^2	0.007	0.013	0.006	0.006	0.007	0.006	0.001	0.001	0.003	0.002

This table presents regression results of the IOS factors on regulation index by Mclaughlin and Sherouse (2017). We deflate the regulation index by 100,000. Panel A and panel B represent the results for top 5 industry-specific and top 5 general factors, respectively. All regressions include firm fixed effects. $Regulation_t$ denotes regulation index at year t. Standard errors are in parentheses, significance levels * p<0.05, ** p<0.01, *** p<0.001

Table 10: Out-of-sample predictability (Pseudo R^2 %)

Row	D.V.	(1) MTB_t	(2) $Factors_t$	(3) *Ind FE*	(4) MTB_t + *Ind FE*	(5) $Factors_t$ + *Ind FE*	(6) MTB_t + $Factors_t$ + *Ind FE*
Panel A: Entire sample							
(I)	$SG\&A_{t+1}$	3.01	38.73	29.39	30.22	43.33	43.60
(II)	$R\&D_{t+1}$	40.89	57.95	51.35	54.24	59.98	60.91
(III)	$CAPEX_{t+1}$	-4.55	31.13	16.90	18.62	33.36	34.39
(IV)	INV_{t+1}	10.88	35.79	20.01	25.48	38.30	40.49
Panel B: Loss firms							
(I)	$SG\&A_{t+1}$	-0.52	34.50	24.75	25.47	35.10	35.87
(II)	$R\&D_{t+1}$	51.53	66.16	59.14	61.00	67.44	67.85
(III)	$CAPEX_{t+1}$	-7.82	31.56	19.67	20.48	33.06	33.49
(IV)	INV_{t+1}	19.47	44.62	30.84	34.51	45.50	47.34
Panel C: Firms with high stock return variance							
(I)	$SG\&A_{t+1}$	-1.74	39.70	27.98	28.73	43.58	43.90
(II)	$R\&D_{t+1}$	42.02	58.88	51.94	54.83	60.85	61.79
(III)	$CAPEX_{t+1}$	0.47	34.54	23.79	25.00	37.01	37.87
(IV)	INV_{t+1}	10.90	37.07	22.87	27.77	39.53	41.71

This table displays the out-of-sample Pseudo R-squared for OLS regressions in which the dependent variables are specified by row and the explanatory variables are specified by column. Pseudo R-squared is equal to one minus the ratio of mean squared errors from each OLS regression to that of a benchmark model (an intercept-only regression). Specifically, each dependent variable is regressed in six different models on various explanatory variables. All dependent variables are defined as before. Ind FE denotes Fama-French 48 industry fixed effects. Factors denotes the 62 factor scores produced by the factor analysis procedure. Cash flow, HHI, and year fixed effects are included in all models. Panel A reports the out-of-sample predictability for 14, 153 observations from 2010 to 2015 using the coefficient estimates on a sample of 36,747 firm-year observations from 1994 to 2009. Panel B reports the out-of-sample predictability for 4,176 firm-year observations with negative earnings from 2010 to 2015 using the coefficient estimates on a sample of 11,704 firm-year observations with negative earnings from 1994 to 2009. Panel C focus on a subsample of firms with high stock return variance. It reports the out-of-sample predictability for 4,526 observations from 2010 to 2015 using the coefficient estimates on a sample of 15,166 firm-year observations from 1994 to 2009. We define firms with high stock return variance as observations with prior-year daily stock return variance greater than the sample median.