Textual Analysis and Valuation of Human Capital in the Stock Market

Hyuna Park*

December 17, 2024

Abstract

Investments in the knowledge and skills of employees are essential for the success of corporations, but most stock market research does not measure human capital, which is not recorded as assets on balance sheets. This paper presents a methodology to estimate human capital (HCap), using eXtensible Business Reporting Language (XBRL) tags that show stock-based compensation expenses, a forward-looking profit model, and natural language processing and machine learning tools applied to financial statement texts. Firm-level and portfolio-level tests show that estimated human capital scaled by market capitalization (HCap/ME) explains cross-sectional variation in future stock returns after controlling for size, value, profitability, investments, and momentum effects.

Keywords: Unrecorded intangibles, human resources, natural language processing, financial constraints
JEL Classification: G11, G12, G14, B41

* Hyuna Park is Professor of Finance and Herb Kurz Chair in Finance and Risk Management at Brooklyn College of the City University of New York; e-mail hyuna.park38@brooklyn.cuny.edu, phone (413) 348-9116. I would like to express my gratitude to Dannis Kim, James Lynch, and SungIn Moon for helpful discussions on leadership, corporate culture, growth, and human capital management. I thank Ha-Chin, Yi and Session 109 participants of the Financial Management Association 2024 Annual Meeting for their comments and suggestions.

Textual Analysis and Valuation of Human Capital in the Stock Market

Abstract

Investments in the knowledge and skills of employees are essential for the success of corporations, but most stock market research does not measure human capital, which is not recorded as assets on balance sheets. This paper presents a methodology to estimate human capital (HCap), using eXtensible Business Reporting Language (XBRL) tags that show stock-based compensation expenses, a forward-looking profit model, and natural language processing and machine learning tools applied to financial statement texts. Firm-level and portfolio-level tests show that estimated human capital scaled by market capitalization (HCap/ME) explains cross-sectional variation in future stock returns after controlling for size, value, profitability, investments, and momentum effects.

Keywords: Unrecorded intangibles, human resources, natural language processing, financial constraints
JEL Classification: G11, G12, G14, B41

1. Introduction

Artificial intelligence and other sophisticated technologies are rapidly transforming how we work and live, and all technological changes and other types of innovations require knowledge embodied in people. The knowledge and skills employees possess have been increasingly important for the success of corporations especially since the digital economy started in the United States in the mid-1990s.

However, most stock valuation tools and asset pricing models that financial analysts and economists currently use do not analyze human capital. The main reason is the lack of available information because human capital and other internally developed intangible assets are not recorded as assets on the balance sheets of companies under the US Generally Accepted Accounting Principles (GAAP).

I propose a new methodology to estimate human capital US corporations with publicly traded stocks have accumulated by utilizing machine-readable numerical and textual information in the digitalized financial statement data files the Securities and Exchange Commission (SEC) makes available to anyone for download for free. A company's balance sheet does not include information on human capital, but it is only one of the six types of financial statements the SEC provides.

Statements of equity, cash flow statements, income statements, and note sections provide stock-based compensation and other relevant information. Digitalized financial statements and rapidly developing computer science tools make it feasible to process the expense data and textual information as inputs to an economic model we need to estimate the depreciation rate for capitalizing the expenses.

In other words, instead of regarding stock-based compensation as short-term operating expenses by applying the depreciation rate of one hundred percent as the accounting standards require, I estimate an alternative discount rate using a forward-looking profit model. An underlying assumption is that investments in human capital proxied by compensation generate not only one-year but also longer-term appropriable returns and its contribution to the firm's future profit decreases when the appropriable return declines.

I apply the estimated depreciation rate from the model to stock-based compensation expenses reported to the SEC through Electronic Data Gathering, Analysis, and Retrieval (EDGAR) to estimate the human capital to market capitalization ratio of each firm every year. I use this data to test if human capital scaled by market capitalization explains future stock returns at individual firm level and portfolio level by applying standard empirical asset pricing research methods as in Fama and MacBeth (1973) and Fama and French (1992, 1993, 2008, 2012, 2015, and 2016).

In the firm-level test, Fama-MacBeth regressions show that the human capital to market capitalization ratio has a strong positive relationship with future stock returns and the results are significant at the one percent level after controlling for operating profitability, size, and past stock returns. Portfolio-level tests show consistent results. When I form stock portfolios based on human capital and rebalance them every year at the end of June following the standard practice in the asset pricing literature, portfolios with a high human capital to market capitalization ratio significantly outperform those with a low ratio after adjusting for the market, size, book-to-market, profitability, investments, and momentum factors.

When using the depreciation rate based on the forward-looking profit model for estimating human capital, an implicit assumption is that companies will have the same amount of human capital if their prior investment amounts are the same regardless of whether their businesses are doing well or not. However, successful firms in good financial conditions are more likely to attract and keep talent and thus have higher human capital than financially constrained firms. Textual information in financial statements includes discussions on the financial condition and other management topics that may affect human capital. Thus, natural language processing (NLP), and machine learning this information may be useful when developing a measure of human capital.

I first use the dictionaries of negative and financially constraining words developed by Loughran and McDonald (2011) and Bodnaruk *et al.*(2015) and find firms with a high proportion of these words in their financial statements. I assume that these firms with more negative or financially constraining words have higher depreciation rates of prior investments in human capital than others.

I use the human capital to market capitalization ratio updated with these dictionaries and repeat the firm-level and portfolio-level tests and find stronger results, especially with the financially constraining words. The coefficient on the human capital variable becomes more significant in Fama-MacBeth regressions and the risk-adjusted performance difference between the high human capital portfolio and the low human capital portfolio in factor models becomes more significant. That is, textual information adds value to the estimation of human capital for stock portfolio management.

I also test a machine learning tool, Word2Vec, to expand the dictionaries in the literature. Bag-of-words is the approach these dictionaries for sentiments and financial constraints use, meaning that they are limited to single words and do not include phrases. Note that companies use not only single words but also phrases when they explain financial constraints and investments in human capital. I feed the existing dictionary as seed words to Word2Vec and expand the dictionary for financial constraints.

Word2Vec first converts the relationships between words and phrases in financial statements to vectors and uses their cosine similarities to expand the seed word dictionary. For example, “unavailable” is a seed word from the financially constraining word list in Bodnaruk *et al.*(2015) and Word2Vec identifies “acceptable terms” as a closely related phrase, and thus I add it to the new dictionary. See Appendix A for more technical details on Word2Vec and Table 6 for a list of newly identified words and phrases and the corresponding seed words.

I use the expanded dictionary for financial constraints from Word2Vec to update human capital estimates, repeat Fama-MacBeth regressions, and find consistent results. The coefficient for the human capital to market capitalization ratio is positive and significant at the one percent level. Based on these results, I argue that prior investments in human capital in the form of stock-based compensation explain cross-sectional variation in stock returns and textual information on financial constraints in financial statements are useful for improving the human capital estimates.

I also apply Word2Vec to a textual analysis of human capital disclosure. Note that a new SEC rule on human capital became effective in November 2020. The rule requires that companies disclose material measures or objectives related to human capital, but it uses a principle-based approach. Instead of defining human capital management clearly, the SEC takes a position that the definition would evolve over time (Engel, 2021). Using Word2Vec applied to financial statement texts, I developed a dictionary of words and phrases companies use to explain human capital as shown in Table 8. The word clouds presented in Figure 5 clearly show the large impact of the new SEC rule on human capital disclosure. As the SEC expected, the words and phrases companies use to explain human capital management have expanded significantly after the new rule became effective in November 2020.

The rest of the paper is organized as follows. Section 2 reviews the literature on human capital, intangibles for asset pricing, textual analysis of financial statements, and measuring financial constraints. Section 3 describes data and explains methodologies for estimating human capital. Section 4 presents firm-level and portfolio-level tests of the relationship between human capital to market capitalization ratio and future stock returns. Section 5 explains textual analysis and machine learning of financial statements to improve human capital estimates and firm-level and portfolio-level tests using the updated estimates, and Section 6 concludes.

2. Literature review

2.1. Human capital and other intangibles for asset pricing

Friedman (1956) points out that wealth includes all sources of income, and the productive capacity of human beings is one form of holding wealth. The Oxford English Dictionary defines human capital as "*the skills, knowledge, and experience possessed by an individual or population, viewed in terms of their value or cost to an organization or country.*" This concept goes back at least to Smith (1776), but economists hesitated to use the term human capital until the 1950s because of the risk of being criticized for comparing free people with marketable assets (Mincer, 1958; Schultz, 1961; Becker, 1964; Goldin, 2016).

Lev and Schwartz (1971) point out the fundamental distinction between human capital and physical capital in accounting and suggest using economic concepts to measure human capital to be recorded on financial statements. They argue that disclosing the human capital values of corporations will be valuable for users of financial statements and reported human capital values and the ratio to physical capital may shed light on changes in the labor force.

However, US GAAP does not allow recording investments in human capital and most other internally generated intangibles such as technologies as assets. For example, SFAS 2 (Accounting for Research and Development Costs, 1974) requires corporations to expense their R&D expenditures immediately instead of capitalizing on them. It is now Accounting Standards Codification (ASC) 730. The high degree of uncertainty about the future benefits of intangible investments is the rationale behind the immediate expensing decision (Kothari *et al.*, 2002).

Lev and Sougiannis (1999) examine whether the off-balance sheet innovative capital proxied by R&D expenditures can be used to predict future abnormal earnings and stock returns. Using a sample of around 1,200 companies from 1972 to 1989, they show that companies with a low book-to-market ratio have large amounts of R&D capital. They also show that the R&D capital-to-market variable subsumes the role of the book-to-market ratio, using a regression of stock returns on lagged fundamentals.

Daniel and Titman (2006) analyze the impact of the changing business environment and accounting information on the book-to-market effect. They show that a stock's future returns are unrelated to its past accounting-based performance but are strongly negatively related to intangible returns, the component of past return that is orthogonal to the firm's past performance. Extending the findings of Daniel and Titman, Jiang (2010) shows that institutions tend to buy shares in response to positive intangible information and the book-to-market effect is significant in stocks with intense past institutional trading but nonexistent in stocks with moderate institutional trading.

Edmans (2011) is a seminal paper that introduces employee satisfaction to asset pricing research. He points out the conflicting views of traditional theories developed in capital-intensive firms and human relations theories that view employees as core assets. Using *Fortune* magazine's *100 Best Companies to Work For* list as a data source, he provides empirical evidence on the positive relationship between employee satisfaction and long-term stock returns.

Eisfeldt and Papanikolaou (2013) use a perpetual inventory method to capitalize on selling, general, and administrative expenses and call it organization capital. They show that firms with more organization capital have higher average stock returns than others. Peters and Taylor (2017) use a similar method to capitalize R&D and call it knowledge capital. They use the knowledge capital and organization capital to adjust Tobin's q and then analyze the impact of intangibles on the investment-q relation. Using the standard asset pricing research methods as in Fama and MacBeth (1973) and Fama and French (1993 and 2015), Park (2022) shows that a book-to-market ratio adjusted with knowledge capital and organization capital explains the cross-sectional variation in future stock returns better than the unadjusted ratio.

Bernstein and Beeferman (2015) show that there is a material relationship between human capital and corporate financial performance by reviewing over ninety empirical studies that examine the relationship between corporate human capital policies and performance measures such as return on equity and profit margin. They argue that the evidence is compelling to justify investor requests for firms with publicly traded stocks to report systematically on their human capital management policies and practices.

Eisfeldt *et al.* (2023) analyze human capital in the manufacturing industry and the implications for the distribution of income and wealth. They supplement the wage data in the National Bureau of Economic Research and the US Census Bureau's Center for Economic Studies (NBER-CES) database by equity-based compensation measured by shares reserved for conversion and future grant of employee stock options and other equity-based compensation. They show that including equity-based compensation eliminates the decline in the high-skilled labor share in manufacturing, providing evidence of complementarity between physical capital and high-skilled labor.

The SEC amended Regulation S-K in 2020 and included a principles-based human capital disclosure mandate under Item 101(c). SEC (2020) explains that the amendment requires a description of the registrant's human capital resources to the extent such disclosures would be material to an understanding of the registrant's business. It includes human capital measures of objectives that the registrant focuses on in managing the business.

Bourveau *et al.* (2023) compare financial statements before and after the principle-based mandate and show that considerable heterogeneity remains in human capital disclosure. Eisfeldt *et al*., (2022) point out that there is significant variability across industries in how they record intangible investments.

High heterogeneity and significant variability in disclosure mean that we may need not only numerical information provided through XBRL tags but also textual information to analyze unrecorded intangible assets including human capital using textual analysis and machine learning tools computer scientists have developed.

2.2. Textual analysis and machine learning of financial statements

Loughran and McDonald (hereafter LM, 2011) is a seminal paper in the textual analysis of financial statements and LM (2016) provides a review of the literature. I thank them for making their dictionaries, financial statement texts, and other data available for download from their website. They show that dictionaries for textual analysis developed in other disciplines do not work well in finance because many English words have multiple meanings. Finance has many unique expressions, as other disciplines do. LM (2011) presents dictionaries that reflect the tone of financial statements.

The financial constraints of a firm may affect human capital but are not directly observable and thus measuring it is a challenge in empirical tests. Prior research develops indexes based on firm characteristics such as age, size, leverage, accounting profitability, and valuation (Kaplan and Zingales, 1997; Lamont et al., 2001; Whited and Wu, 2006; Hadlock and Pierce, 2010). However, there is growing evidence that the performance of these indexes is deteriorating.

Farre-Mensa and Ljungqvist (2016) examine Kaplan-Zingales, Whited-Wu, and Hadlock-Pierce indexes. They show that firms typically classified as constrained by these indexes do not behave as if they are financially constrained. As increasing unrecorded intangibles may be a reason for the poor performance of the indexes, prior research presents a text-based measure as an alternative to the indexes. Kaplan and Zingales (1997) and Hadlock and Pierce (2010) read selected annual financial statements manually to find where managers explain challenges in external financing. Bodnaruk et al. (2015) build on the idea and apply natural language processing tools to all annual financial statements for 1996 – 2011 and develop a list of 184 constraining words. They find that the frequency of constraining words in financial statements predicts future liquidity events better than other financial constraint measures based on size, age, and numerical accounting data.

The dictionaries in LM (2011) and Bodnaruk et al. (2015) are based on a bag-of-words (BOW) approach to textual analysis because it regards a financial statement as a bag of all words in the document, regardless of how words are combined to explain the meanings of sentences. An alternative is Word2Vec, a machine learning tool that learns the meaning of words in a set of documents, called corpus in computer science, by converting the relationships between words and phrases into a series of mathematical vectors.

The method is based on the idea that words with similar meanings tend to appear with similar neighboring words (Harris, 1954). Mikolov *et al.* (2013) developed a method that trains a corpus to learn relationships between words and phrases. Li *et al.* (2021 a and b) apply this method to earnings call transcript data for textual analysis of corporate culture and the impact of Covid-19 on businesses. This paper uses Word2Vec to improve human capital estimates and explains the method in Section 5.

3. Estimating human capital

3.1. Data

I use *Financial Statement Data Sets* and *Financial Statement and Notes Data Sets* downloaded from the SEC website as a primary source of data for estimating human capital. The former is updated quarterly and the latter monthly. The SEC provides the data by extracting the information directly from the exhibits to the financial reports registrants filed with them using XBRL tags, which work like barcodes. I use the annual stock-based compensation expense of a corporation as a measure of its investments in human capital.

For example, Amazon reports $24 billion as its stock-based compensation in its consolidated statements of cash flows for the fiscal year ending December 31, 2023, using the standard XBRL tag, *us-gaap:ShareBasedCompensation*. As they report the previous two years' data along with the current one, the same statement shows that the compensation expense was $19.6 billion in 2022 and $12.8 billion in 2021. The same tags in Amazon's cash flow statement in earlier years show that their stock-based compensation was $ 9.2 billion in 2020, $6.86 billion in 2019, and $5.4 billion in 2018.

Using an economic model described in the next subsection, I estimate that 17.75% is the depreciation rate. By applying this depreciation rate to the compensation data and a perpetual inventory method as in Peters and Taylor (2017) and Park (2022), I estimate that 24 + 0.8225*(19.6 + *0.8225(12.8 + 0.8225*(9.2 + 0.8225*(6.86 + 0.8225*(5.4 + … + (0.000036 + 0.8225*0.00013)))))) = $62.3 billion is the human capital of Amazon as of December 31, 2023.

Note that this method requires an assumption on the initial stock of human capital when the stock-based compensation data starts for each firm. I assume that the initial human capital stock is the first-year data divided by the sum of the depreciation rate and the growth rate and that the growth rate is 10 percent. Amazon was founded in 1994, and its stock-based compensation records start at $36,000 in 1996. Thus, the initial human capital stock is estimated to be 0.036/(0.1775+0.1) = $0.13 million.

Each *Financial Statement Data* folder includes four text files along with a data manual explaining the scope, organization, file formats, and definitions. The EDGAR assigns a unique accession number to each submission and the Central Index Key (CIK) is a ten-digit number the SEC assigns to each registrant that submits filings. Some companies use a statement of equity, income statement, or note sections to report compensation. For example, Alphabet reports $22.6 billion as its stock-based compensation expense in the consolidated statement of stockholders equity for the fiscal year ending December 31, 2023, using the standard XBRL tag of us-gaap:AdjustmentsToAdditionalPaidInCapitalSharebased CompensationRequisiteServicePeriodRecognitionValue. See Appendix A for more details on the XBRL tag data.

As the XBRL reporting became mandatory in 2009, I used the Compustat database and financial statement files downloaded from either SEC/EDGAR or Investor Relations websites of companies as supplemental data sources either to obtain data for earlier years or to double-check different sources for consistency. 1998 is the first year when the share-based compensation expense records appear in Compustat. I downloaded the stock prices and returns from the Center for Research in Security Prices (CRSP) database and the Compustat data through Wharton Research Data Services (WRDS).

As law and culture have large impacts on human capital, this paper limits the analysis to corporations registered in the United States. I exclude American depositary receipts, real estate investment trusts, and units of benefits interest. Only ordinary common equity shares that have a share code of 10 or 11 are included, following the standard practice in the asset pricing literature.

Table 1 presents the summary statistics of the estimated human capital scaled by market capitalization as of December 31 of the year when each fiscal year ends. There are 109,626 firm-years for the sample period of 1998 – 2023, and 72% have *i*HCap estimates, which are based on the depreciation of 17.75% applied to all firms. For example, Amazon's *i*HCap/ME is 62.3/1,570 = 3.97% because its *i*HCap estimate was $62.3 billion and its market cap was $1.57 trillion on December 31, 2023. I will reevaluate the assumption on depreciation rates in Section 5 using textual analysis and machine learning tools.

Note that there is an increasing trend in the human capital to market capitalization ratio as well as the proportion of firm-years with *i*HCap during the past two decades. For example, the average *i*HCap/ME has increased sharply from 1.02% in 2003 to 8.92% in 2023. The proportion of firm-years with *i*HCap was 60% in 2003 and it is over 98% in 2023. *i*HCap/ME also shows a large variability across industries and business services and pharmaceutical have many firm-years with *i*HCap and a high ratio of *i*HCap/ME. For example, the average *i*HCap/ME of the business services industry has increased from 1.97% in 2003 to 12.28% in 2023, and the average *i*HCap/ME of the pharmaceutical firms is over 14% as of December 2023.

3.2. Estimating the Depreciation Rate for Investments in Human Capital

I use a forward-looking profit model to estimate the depreciation rate for capitalizing stock-based compensation expenditures. The book value of prior compensation is zero because it is classified as short-term operating expenses under US GAAP, meaning that the depreciation rate was assumed to be one hundred percent when preparing financial statements. However, I assume that stock-based compensation generates not only short-term but also long-term benefits, and thus the economic value is not zero. Estimating the economic value is an empirical question and a crucial aspect of this valuation process involves determining the appropriate depreciation rate to capitalize the compensation expense. I use a profit model to estimate this depreciation rate.

I assume that the economic value of prior compensation expenses depreciates as its contribution to the company's profit diminishes over time. Compensation expenses generate appropriable returns, but the impact becomes smaller when its appropriable return declines.

Li and Hall (2020) use a similar model to estimate the depreciation of R&D capital of firms in high-tech industries such as computers, software, semiconductors, and pharmaceuticals. Firms decide compensation expenses to maximize the net present value, and the following equation shows the profit maximization model.

$$\begin{aligned}\max_{C_t} E_t[\pi_t] &= -C_t + E_t\left[\sum_{j=0}^{\infty} \frac{q_{t+j+d} I(C_t)(1-\delta)^j}{(1+r)^{j+d}}\right] \\ &= -C_t + C_\Omega[1-\exp\left(-\frac{C_t}{\theta_t}\right)\sum_{j=0}^{\infty} \frac{E_t[q_{t+j+d}](1-\delta)^j}{(1+r)^{j+d}}]\end{aligned} \tag{1}$$

where C_t is the compensation expense in period *t*, π_t is the net present value of the investment, q_t is the sales revenue in period *t,* δ is the depreciation rate, and *r* is the cost of capital. $I(C_t)$ is the profit rate from the compensation expenditure in period *t*. The profit rate, $I(C_t)$, is unobservable, and we use a concave function, $I(C) = I_\Omega[1-\exp\left(-\frac{C_t}{\theta_t}\right)]$, as in Li and Hall (2020). I assume that sales revenue (q_{t+j}) for *j* periods later than *t,* grow at a constant rate of *g*, $q_{t+j} = q_t(1+g)^j$. θ_t is the investment scale in period *t,* which grows at *G* where *G* is the growth rate of $\theta_t, \theta_t \equiv \theta_0(1+G)^t$ The parameter *d* represents the gestation lag, defined as the duration it takes for compensation expense to start contributing to the operating profit.

The growth rate *G* can be estimated by fitting the data for compensation to the equation, $C_t = C_0(1+G)^t$, C_t is the compensation expenditure in period *t* and $C_0$is the initial expenditure. I also assume that a firm's compensation expenditure in period *t* contributes to the profits, but its contribution in later periods declines at a geometrically decreasing rate.

These assumptions lead equation (1) to the following optimization problem to find an optimal choice of C_t.

$$\max_{R_t} \pi_t = -C_t + I_\Omega[1 - \exp\left(-\frac{C_t}{\theta_0(1+G)^t}\right)] \frac{q_t(1+g)^d}{(1+r)^{d-1}(r+\delta-g+g\delta)} \qquad (2)$$

The following equation shows the first order condition.

$$\frac{\partial \pi_t}{\partial C_t} = -1 + \exp\left(-\frac{C}{\theta_0(1+G)^t}\right) \frac{I_\Omega}{\theta_0(1+G)^t} \frac{q_t(1+g)^d}{(1+r)^{d-1}(r+\delta-g+g\delta)} = 0 \qquad (3)$$

The above condition leads to the following equation.

$$\begin{aligned} \varepsilon_t &\equiv \frac{\left(1+\hat{G}\right)^t}{I_\Omega} \theta_0 \exp\left(-\frac{C_t}{\theta_0\left(1+\hat{G}\right)^t}\right)^{-1} - \frac{q_t(1+\hat{g})^d}{(1+r)^{d-1}(r+\delta-\hat{g}+\hat{g}\delta)} \\ &= \frac{\left(1+\hat{G}\right)^t}{I_\Omega} \theta_0 \exp\left(\frac{C_t}{\theta_0(1+G)^t}\right) - \frac{q_t(1+\hat{g})^d}{(1+r)^{d-1}(r+\delta-\hat{g}+\hat{g}\delta)} \end{aligned} \qquad (4)$$

To estimate the depreciation rate δ in the above equation, we need assumptions on profitability inputs and the gestation lag. I assume that I_Ω is equal to the median return on assets during the sample period. I set r equal to I_Ω assuming that the marginal cost and return are the same in equilibrium. I tested the gestation lag of zero and two. I compute the average growth rate of total sales revenue (*G*) and stock-based compensation (*g*). Using these input data, I estimate θ_0 and δ by a nonlinear least squares method applied to equation (4).

I first apply this method to the business services industry because it has the largest number of firm-year observations as shown in Table 1. The nonlinear least squares method shows that 17.75% is the depreciation rate that makes the error term in equation (4) converge to zero. Next, I applied the model to the pharmaceutical industry which has the second largest firm-year observations, and found that 1.71% is the depreciation rate that makes the error term close to zero. That is, the economic value of investments in human capital depreciates more slowly in the pharmaceutical industry and others. I also applied the model to the aggregated data of all firm-years, but the error term did not converge to zero.

4. Firm-level and Portfolio-level Tests

I use the standard methods in the asset pricing literature to test if the estimated human capital to market capitalization ratio explains the cross-sectional variation in future stock returns at individual firm and portfolio levels. The sample period is July 2003 – December 2023 because of a small sample size in earlier years as shown in Table 1.

4.1. Firm-level Tests using Fama-MacBeth regressions

I test the explanatory power of human capital at the firm level by comparing Fama–MacBeth regressions of monthly returns on log(*i*HCap/ME) with those based on book-to-market ratio log(BE/ME), knowledge capital log (Kcap/ME), organization capital log(Ocap/ME), and an intangible-adjusted book-to-market ratio log (iBE/ME) as in Peters and Taylor (2017) and Park (2022). I include control variables such as size, momentum, short-term reversal, and profitability as commonly used in the literature. log(ME) is the natural logarithm of the market value of equity in the previous month. r_{12-1} is the prior year's return skipping the last month to consider the momentum effect, and $r_{1,1}$ is the prior month's return to control the short-term reversal effect. COP is cash-based operating profitability scaled by the book value of total assets as in Ball *et al.* (2016). As in prior research, financial statement data is updated annually in June with a lag of at least six months to make sure the data is publicly available.

Panel A of Table 2 confirms that the human capital to market capitalization ratio has a significant positive relationship with future stock returns. The coefficient of log(*i*HCap/ME) in the regression is positive and significant at the five percent level while log (BE/ME) and log (Kcap/ME) show insignificant results. Control variables in the regressions confirm results consistent with the literature. COP measuring profitability shows a significant positive relationship with future stock returns and the coefficient on $r_{1,1}$ is significantly negative, confirming the short-term reversal effect.

Prior research shows that microcap stocks behave differently in the Fama-MacBeth regressions of future stock returns and thus I divide the sample into two size groups: ABM (All-but-microcaps) and Micro. Following Fama and French (2008), Micro is defined as NYSE, AMEX, and Nasdaq stocks below the 20th percentile of the market capitalization of NYSE stocks and ABM is all else. Panel B and Panel C of Table 2 confirm the differences between ABM and Micro in the relationship between future stock returns and explanatory variables. The positive relationship between log(*i*HCap/ME) and future stock returns is significant at the one percent level in ABM but it is not significant in Micro. Cash operating profitability (COP) has a significant positive relationship with future stock returns only in Micro and the relationship is not significantly different from zero in ABM.

4.2. Portfolio-level Tests

I implement value-weighted portfolio-level tests in addition to Fama-MacBeth regressions because firm-level regressions are sensitive to outliers, impose a potentially misspecified parametric relation between variables, and overly emphasize nano- and micro-cap stocks by weighing each firm equally. In portfolio-level tests, I no longer split the data into ABM and Micro because microcap stocks have only a small effect on value-weighted portfolio returns.

Following Fama and French (1993 and 2015), I constructed six value-weighted portfolios based on size and *i*HCap/ME. The size breakpoint for each year is the median market capitalization of NYSE stocks, and the *i*HCap/ME breakpoints are the 30th and 70th percentiles. This procedure is like how Fama and French construct the small and big and high and low book-to-market portfolios. I apply the same method to construct Big High, Big Low, Small High, and Small Low *i*HCap/ME portfolios.

Panel A of Table 3 shows that the portfolios of stocks with a higher human capital-to-market capitalization ratio have a higher average return than those with a lower ratio regardless of size. Small stock portfolios have a higher average return than large stock portfolios when controlling for the human capital to market capitalization ratio. The average return on Big High is 1.01% per month, 0.81% on Big Low, 1.05% on Small High, and 0.84% on Small Low.

Figure 1 presents the growth of $100 each invested in high *i*HCap/ME and low *i*HCap/ME portfolios on June 30, 2003, in comparison to the S&P 500 index. Returns on high *i*HCap/ME portfolio are the average returns on Big High and Small High and returns on low *i*HCap/ME portfolio are the average returns on Big Low and Small Low. Note that the value of the high *i*HCap/ME portfolio grows much faster than the S&P500 index, and the growth of the low *i*HCap/ME portfolio is the worst among the three ($794, $733, and $597, respectively on December 31, 2023).

I also compare Big High, Big Low, Small High, and Small Low portfolios using the five-factor model of Fama and French (2015 and 2016) augmented with the momentum factor and report the results in Panel B of Table 3. I find that the six-factor model alpha of

Big Low is significantly negative and the alphas of the other three portfolios are not significantly different from zero.

5. Textual Analysis to Improve Human Capital Estimates

An assumption in the estimation of *i*HCap is that the same dollar amount of stock-based compensation results in the same human capital estimate because the deprecation rate is assumed to be the same. Whether those who received the compensation stay with the company or not, *i*HCap is the same. It is based on the idea that individual employees can resign, but the labor force as a group is constantly associated with the company if the firm can hire others with comparable skills and experience, as Lev and Schwartz (1971) point out. However, this assumption may not be valid if the company is subject to significant operating challenges or severe financial constraints, and the textual information in financial statements may be useful to identify such firm years for improving human capital estimates.

5.1. Using Dictionaries to Measure Sentiments and Financial Constraints

When I estimate *i*HCap, I use only numerical information in digitalized financial statements even though there is a lot of textual information that may be related to the valuation of human capital. One way to use the textual information is by applying dictionaries developed in prior finance research.

LM (2011) points out that dictionaries developed for textual analysis in other disciplines do not work well in financial analysis and presents “liability” as an example. The term has a negative meaning in other disciplines, but not in finance. LM (2011) presents unique dictionaries for finance research and one of them is a dictionary to measure negative sentiments in financial statements, which I test for improving the estimation of human capital.

Another dictionary I use is from Bodnaruk et al. (2015). They apply a parsing algorithm to all annual financial statements for the fifteen years from 1996 to 2011 and develop a list of 184 constraining words. They show that the frequency of constraining words in annual financial statements predicts future liquidity events better than other financial constraint measures based on age, size, and accounting ratios.

It is an empirical question whether the negative sentiment and the financial constraining word dictionaries are useful for detecting firms that have difficulty in recruiting or retaining core talents and thus finding a better depreciation rate for improving human capital estimates. I apply the two dictionaries to the estimation of human capital and present the results in Table 4. Panel A shows the average total number of words, the averages and 70th percentiles of the negative sentiment words (P_negative) and financially constraining words (P_constraining) scaled by the total number of words in the financial statements for the fiscal year ends from 1998 to 2022. For example, the 70th percentile of P_negative and P_constratining was 2.27% and 0.97% in 2022, respectively.

Note that financial statements have become longer, providing a larger amount of textual information to users during the past two decades. The average number of total words in cleaned financial statements has increased from less than 30,000 in 1995 to over 58,000 in 2022. This means over a million pages of financial statement texts become available annually making it impossible for a human analyst to read them all, and thus upskilling financial analysts with computer science tools for textual analysis may increase their human capital and productivity. The CFA Institute recently expanded its Chartered Financial Analyst (CFA) program to include this type of computer science skills in its curriculum.

I compute the P_negative and P_constraining of each firm year and compare them with the corresponding 70th percentile as a threshold. If the firm year has a higher proportion of negative words or financially constraining words than the cutoff, I use it as an indicator for the impairment of human capital from prior investments in the form of stock-based compensation. Using this indicator, I redefine human capital and call it *t*HCap. *t*HCap is equal to *i*HCap only if a textual indicator such as P_negative or P_constraining is below the threshold.

I use *t*HCap instead of *i*HCap and repeat Fama-MacBeth regressions and portfolio-level tests and find stronger results, especially when using the financially constraining word dictionary to build the impairment indicator. Panel B of Table 4 shows that the coefficient on log (*t*HCap/ME) is positive at the one percent level for ABM stocks for both P_negative and P_constraining indicators and the strongest result is obtained with the P_constraining indicator. When comparing these results with the *i*HCap regressions in Table 2, the significance of the positive relationship between human capital in ABM stocks improved from the 5 percent to the 1 percent level and the explanatory power of the model measured by the adjusted R-square increased from around 3 percent to over 5 percent. That is, firm-level tests confirm that textual information in the financial statement of a firm is useful for estimating the value of human capital the company has built.

Portfolio-level tests in Table 5 also confirm that *t*HCap is a better measure of human capital than *i*HCap in stock portfolio management. Panel A shows a higher average return of portfolios with high *t*HCap stocks than low *t*HCap stocks in both large and small stocks, and Panel B shows that high *t*HCap portfolios have a significantly positive alpha after adjusting for other factors such as the market, size, value, profitability, investments, and momentum.

Recall that *i*HCap showed much weaker results. The alphas of high *i*HCap portfolios in Table 3 are not significantly different from zero but the alphas of high *t*HCap portfolios in Table 5 are significantly positive.

These results confirm that we need to fully utilize the textual information in financial statements when we analyze the human capital of companies for stock valuation. I have so far shown the usefulness of the dictionaries prior research has built when analyzing human capital, but it is not clear whether those dictionaries that are based on a bag-of-words (BOW) approach are the best tools or if there is room for improvement.

P_negative and P_constraining are based on BOW because the dictionaries were built under the assumption that a financial statement is a bag of all words in the document, not considering how those words are combined to explain the concepts. BOW is a simple and easy-to-use method, but it may have limitations in some applications because we need to know not only the list of single words but also how those words are combined in sentences. An alternative to BOW is a machine learning tool such as Word2Vec and thus we use it for estimating human capital and explain the method in the next subsection.

5.2. Machine Learning of Financial Statements to Improve Human Capital Estimates

Word2Vec converts sentences in documents into mathematical vector structures and uses cosine similarities of the vectors to expand a list of seed words to include similar words and phrases. See Appendix B for more technical details including numerical examples of Word2Vec. This paper uses Word2Vec for two purposes. One is to improve the measure of financial constraints and the other is to analyze human capital disclosure.

First, I feed the financially constraining word dictionary as the seed words to the Word2Vec models built with financial statement texts. Due to computational capacity constraints, it is not feasible to build a model that uses all the financial statements simultaneously. Thus, I randomly selected 1,000 financial statements from the business services industry because it has the largest firm-years with *i*HCap and a higher average *i*HCap/ME than most other industries. Table 6 shows the list of words and phrases the Word2Vec model identified along with the corresponding seed words in the financially constraining word dictionary. For example, the seed words "comply, abide, and strict" led to the identification of "adhere" as a new keyword to describe financial constraints. I add these forty-one newly identified words and phrases to expand the financial constraining word dictionary and calculate the proportion of financial constraining words and phrases for each firm-year and call it P_constraining_W2V.

Panel A of Table 7 shows the average and the 70th percentile of P_constraining_W2V and I use it to redefine human capital and call it *w*HCap. If P_constraining_W2V is below the 70th percentile threshold, *w*HCap is set equal to *i*HCap. If the proportion of the constraining words and phrases in the financial statement is above the threshold, I take it as a signal to impair human capital capitalized from prior compensation expenses and remove the firm-year from the high *w*HCap/ME portfolio.

Panel B of Table 7 presents the Fama-MacBeth regressions of ABM stocks and Microcap stocks using *w*HCap/ME in comparison with corresponding regressions with *t*HCap/ME. The results show that both *w*HCap/ME and *t*HCap/ME have a strong positive relationship with future stock returns and their explanatory power is much stronger than *i*HCap/ME.

The second application of Word2Vec is to identify words and phrases US corporations use to explain human capital. Since the SEC adopted a new rule on human capital disclosure in 2020, financial statements are increasingly including more information on how companies manage their human capital (Ising *et al*, 2023).

However, it is an empirical question to quantify the new information for testing whether it helps explain the cross-sectional variation in stock returns. The first step is to identify seed words by reading research papers on corporate human capital management, disclosure rules, and surveys, and then feed the seed words to Word2Vec to develop a dictionary.

See Table 8 for the list of seed words, words and phrases newly identified by Word2Vec, and which seed word served as the root for each of the newly identified words and phrases. Figure 4 presents how the total counts of these words and phrases and their ratio to the total number of words in each financial statement has changed over time using firms with their fiscal year ending in December. As shown in the figure, the total counts and the proportion of human capital words and phrases increased sharply after the SEC mandated the principle-based disclosure rule in 2020.

Figure 5 compares the word cloud from applying the dictionary in Table 8 to the financial statements filed with the SEC from 2021 to 2023 with the word cloud for the entire sample period of 1994 – 2023. As shown in the figure, compensation, employee, and health are the most frequently used words in all periods, but the terminology explaining human capital has significantly expanded following the SEC's new disclosure rule.

Panel A of Table 9 presents the summary statistics of words and phrases explaining human capital. P_HCulture is defined as the words and phrases related to human capital as in Table 8 as a percentage of the total number of words in each financial statement. The panel presents the time variation in the average and the 70th percentile of P_HCulture. Both the average and the 70th percentile have increased sharply, especially after the new disclosure rule.

Fama-MacBeth regressions in Panel B of Table 9 are to test if the textual information on the corporate culture related to human capital helps explain the cross-sectional variation in future stock returns. dHCulture is a dummy variable that is one if P_HCulure is above the 70th percentile and zero otherwise. As shown in the table, the coefficient on dHCulture is significantly positive in the ABM sample, meaning that companies that mentioned words and phrases related to human capital more often in their financial statements have a higher return in the stock market than other firms.

However, when adding log (*t*HCap/ME), a quantitative human capital measure based on compensation, the text-based measure, dHCulture, loses explanatory power. The coefficient is still positive but not significantly different from zero. The weak explanatory power of the text-based measure compared to the human capital measure based on compensation may be attributable to the short history of the mandated disclosure that was introduced in 2020. These results show that human capital based on share-based compensation scaled by market capitalization explains the cross-sectional variation in future stock returns. Natural language processing of textual information in financial statements is useful especially when measuring financial constraints to improve the human capital estimates.

6. Conclusion

Rapidly developing technologies are transforming the world, and all innovations rely on the knowledge, skills, and experience embodied in human beings. In this environment, how corporations manage and accumulate their human capital is critical information for stock valuation, but most stock valuation tools and asset pricing models do not analyze human capital.

To fill this gap, I developed a new methodology to estimate the human capital US publicly traded corporations have accumulated using both numerical and textual information included in annual financial statements. Using the conventional asset pricing research methods at both individual firm and portfolio levels, I test the relationship between the estimated human capital and future stock returns. I find that estimated human capital from capitalizing prior stock-based compensation expenses is useful information in stock portfolio management. Textual information in financial statements is useful to signal financial constraints each firm faces and using the signal as an indicator for impairing estimated human capital improves the risk-adjusted performance of human-capital-enhanced stock portfolios.

References

Ball, R., Gerakos, J., Linnainmaa, J. T., & Nikolaev, V. V. (2015). Deflating profitability. *Journal of Financial Economics*, 117, 225-248.

Ball, R., Gerakos, J., Linnainmaa, J. T., & Nikolaev, V. V. (2016). Accruals, cash flows, and operating profitability in the cross-section of stock returns. *Journal of Financial Economics*, 121, 28-45.

Ball, R., Gerakos, J., Linnainmaa, J. T., & Nikolaev, V. V. (2018). Earnings, retained earnings, and book-to-market in the cross-section of expected returns. *Journal of Financial Economics*, forthcoming.

Becker, G. (1964) Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. Harvard University Press.

Bernstein, A. & Beeferman, L. (2015). The materiality of human capital to corporate financial performance, IRRC Institute, the Labor and Worklife Program (LWP) Harvard Law School

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-K text to gauge financial constraints. Journal of Financial and Quantitative Analysis, 50(4), 623–646.

Bourveau, T., Chowdhury, M., Le, A., & Rouen, E. (2023). Human capital disclosures. Working paper, Columbia Business School & Harvard Business School.

Daniel, K., & Titman, S. (2006). Market reactions to tangible and intangible information. *Journal of Finance,* 61(4), 1605–1643.

Edmans, A. (2011). Does the stock market fully value intangibles? Employee satisfaction and equity prices. *Journal of Financial Economics*, 101, 621-640.

Eisfeldt, A., Falato, A, & Xiaolan, M. (2023). Human capitalist. *Macroeconomics Annual*, 37, National Bureau of Economic Research.

Eisfeldt, A., Kim, E. T., & Papanikolaou, D. (2022). Intangible value. *Critical Finance Review* 11(2), 299-332.

Eisfeldt, A. L. & Papanikolaou, D. (2013). Organization Capital and the Cross-Section of Expected Returns. *Journal of Finance*, 68(4), 1365 - 1406.

Engel, M. (2021). New Human Capital Disclosure Requirements, Compensation Advisory Partners, *Harvard Law School Forum on Corporate Governance*.

Fama, E., & French, K. (1992). The cross-section of expected stock returns. *Journal of Finance,* 47, 427-465.

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

Fama, E., & French, K. (2008). Average returns, B/M, and share issues. *Journal of Finance,* 63, 2971–2995.

Fama, E., & French, K. (2012). Size, value, and momentum in international stock returns. *Journal of Financial Economics,* 105, 457–472.

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

Fama, E., & French, K. (2016). Dissecting anomalies with a five-factor model. *The Review of Financial Studies,* 29(1), 69-103.

Fama, E., & French, K. (2018). Choosing factors. *Journal of Financial Economics,* 128, 234-252.

Fama, E., & MacBeth, J. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy,* 81, 607–636.

Farre-Mensa, J. & A. Ljungqvist. (2016). Do Measures of Financial Constraints Measure Financial Constraints? The Review of Financial Studies, 29(2), 271-308.

Friedman, M. (1956). The quantity theory of money – A restatement. Studies in the Quantity Theory of Money. The University of Chicago Press.

Goldin, C. (2016) Human Capital.Handbook of Cliometrics, ed. Claude Diebolt and Michael Haupert, 55-86. Heidelberg, Germany: Springer Verlag.

Hadlock, C., and J. Pierce (2010). New Evidence on Measuring Financial Constraints: Moving Beyond the KZ Index. Review of Financial Studies, 23, 1909–1940.

Harris, Z. S. (1954) Distribution of Structure. *Word* 10: 146-162.

Ising, E. A, Titera, M. A., Lapitskaya, J., Klein, Zoe, & Sherley M. (2023) Form 10-K Human Capital Disclosures Continue to Evolve, Gibson Dunn

Kaplan, S., and L. Zingales. (1997). Do Financing Constraints Explain Why Investment Is Correlated with Cash Flow? Quarterly Journal of Economics, 112, 169–216.

Kothari, S. P., Laguerre, T. E., & Leone, A. J. (2002). Capitalization versus expensing: Evidence on the uncertainty of future earnings from capital expenditures versus R&D outlays. *Review of Accounting Studies,* 7, 355–382.

Jiang, H. (2010) Institutional investors, intangible information, and the book-to-market effect. *Journal of Financial Economics*, 96(1), 98-126.

Lamont, O., C. Polk, and J. Saa-Requejo (2001). Financial constraints and stock returns. Review of Financial Studies, 14, 529 – 544.

Lev, B. & Schwartz, A. (1971). On the use of the economic concept of human capital in financial statements. The Accounting Review, 46(1), 103-112.

Lev, B., & Sougiannis, T. (1999). Penetrating the book-to-market black box: The R&D effect. *Journal of Business, Finance and Accounting,* 26, 419–449.

Li, W. C.Y. & Hall, B. H. (2020). Depreciation of Business R&D Capital. *The Review of Income and Wealth* 66, pp. 161-180.

Li, K., X. Liu, F. Mai, & T. Zhang. (2021). The Role of Corporate Culture in Bad Times: Evidence from the Covid-19 Pandemic. *Journal of Financial and Quantitative Analysis* 56, 2545-2583.

Li, K., F. Mai, R. Shen, & X. Yan. (2021). Measuring Corporate Culture Using Machine Learning. *Review of Financial Studies* 34, 3265-3315

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

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

Mikolav, T., I. Sutskever, K. Chen, G. S. Corrado, & J. Dean. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, 3111-3119.

Mincer, J. (1958) Investment in Human Capital and Personal Income Distribution, *Journal of Political Economy* 66, 281-302.

Newey, W. K, & West, K. D. (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 704-708.

Park, H. (2022) An intangible-adjusted book-to-market ratio still predicts stock returns, *Critical Finance Review* 11(2), 265-297.

Peters, R. H. & Taylor, L. A. (2017). Intangible capital and the investment-q relation. *Journal of Financial Economics*, 123, 251-272.

Schultz, T. W. (1961) Investment in Human Capital. *American Economic Review* 51, 1-17.

SEC (2020) Modernization of Regulations S-K Items 101, 103, and 105, 17 CFR 229, 239, and 240. The Securities and Exchange Commission.

Smith, Adam (1776) An Inquiry into the Nature and Causes of the Wealth of Nations.

Whited, T., and G. Wu. (2006). Financial constraints risk. Review of Financial Studies, 19, 531-559.

Appendix A. XBRL tag data

XBRL tags make it convenient to machine-read and process many numbers in financial statements quickly, akin to using barcodes for electric information tracking. The SEC initially launched the XBRL reporting as a voluntary financial reporting program in 2003, eventually making it mandatory in 2009. We can download the *Financial Statement Data Sets* from the SEC website updated quarterly and *Financial Statement and Notes Data Sets* monthly.

It has been over fifteen years since the data became public, but it is rare to see prior research in finance or economics use XBRL tags. This paper contributes to the literature by showing the latent value of information XBRL tags provide in improving the valuation of unrecorded intangible assets in US companies with publicly traded stocks.

Each quarterly financial statement data folder includes four text files, SUB, NUM, TAG, PRE, along with a data manual explaining the scope, organization, file formats, and definitions. The submission data set (SUB) includes one record for each XBRL submission during the quarter and shows information about the submission and the filing entity, such as ADSH and CIK. ADSH is an accession number the SEC assigns to each submission to its EDGAR system. Central Index Key (CIK) is a ten-digit number the SEC assigns to each registrant that submits filings.

The number data set (NUM) includes one row for each amount for all line-item values from each submission in SUB. As ADSH is common to both datasets, it serves as the link between them. The tag data set (TAG) shows each tag's definitions, descriptions, versions, and other information. For example, *us-gaap:ShareBasedCompensation* tag is in the financial reporting taxonomy, its custom variable is zero because this tag is not for a specific company but from US GAAP, the datatype is monetary, the label is Share-Based

Payment Arrangement, Noncash Expense and the definition is the amount of noncash expense for share-based payment arrangement. I use the tag, custom, and datatype in this data set to sort out all monetary tags. The presentation data set (PRE) shows where each tag was presented in the primary financial statements. The monthly financial statement and notes folder has four additional text files, DIM (Dimensions), TXT (Plain Text), REN (Rendering), and CAL (Calculations).

Hoitash *et al.* (2021) provide a literature review in the use of XBRL tags. They point out the pros and cons of Compustat and XBRL data, noting that Compustat distributing less granular and standardized data that may be different from what was originally reported in XBRL tags. Chychyla and Kogan (2015) analyze the discrepancies between Compustat data and XBRL records and find that 17 out of 30 analyzed variables in Compustat are significantly different from corresponding XBRL tag values.

Dong *et al.* (2016) uses the adoption of XBRL as a natural experiment to test the theoretical reasoning of underinvestment in the production of expensive firm-specific information. They find that the effect of XBRL adoption on stock return synchronicity is significant for complex firms that have financial information that is inherently more difficult to process. Park and Baek (2024) use XBRL tags to measure innovative investments in the finance industry. They examine the monetary XBRL tags in the income statements of financial firms that have multiple patents and identified tags related to innovation.

Appendix B. Machine Learning of Financial Statement Texts Using Word2Vec (W2V)

This appendix explains details on how to apply a W2V model to financial statement text files such as 10Ks as a corpus. The first step is to download the raw 10K files using the SEC/EDGAR full index directory. The next step is to clean the raw texts to remove tables, HML and XBRL tags, and other non-text contents that are not relevant to textual analysis. The regex version 2021.8.3 and *beautifulsoup* 4 in Python 3.8 work well for this cleaning process.

The third step is to convert all alphabet in the cleaned files to lower case to facilitate word search, delete numbers and special characters that are not relevant to textual analysis, and to use the *Phraser* and *Phrases* modules in the genism library to form multiword ngrams that we need to learn how words are combined to explain concepts. The modules automatically detect phrases longer than one word using collocation statistics. For example, the Phrases module makes "write_down' out of "write" and "down" and adds the bigram, the combined word using the underscore symbol _, to the corpus. W2V treats each ngram concatenated with an underscore as if it is a single word.

As in most textual analysis projects, I remove stopwords such as "are," for example, that are used in most documents but do not add value in analyzing the meaning of words and phrases. For example, *"training is assigned to newly hired employees upon joining the firm and to current employees periodically thereafter"* is converted to *"training assigned newly hired employees joining firm current employees periodically"* after removing stopwords.

Note that stemming such as converting "competition" and "compete" to "compet", for example, is also used often in many textual analysis projects. However, I find that the cost of lost information outweighs the benefit of reduced dimension when stemming 10Ks by comparing the results with and without stemming. I think it is because currently available stemming tools have been developed mostly outside of finance and thus do not consider the characteristics of the terms used frequently in financial statements. Thus, the results reported in this paper are from textual analyses without stemming.

After removing stopwords, the remaining words and ngrams are trained using the W2V model in the genism library version 4.1.2. The model depends on word embedding that represents the meaning of a word using a numeric vector so that we can use vector arithmetic to measure the relationship between words and phrases. W2V uses the cosine similarity between two word-vectors to measure how close the two words and phrases are.

For example, when we use vector arithmetic with how often [training, skills, inflation, operating_efficiencies, analytics, compensation] appear close to (talent, succession_planning, supply_chain) in thirty 10-Ks to examine the relationships among the three words, we first need the following three vectors.

talent = [31, 28, 3, 10, 18, 25], succession_planning = [14, 16, 0, 5, 17, 16], and supply_chain = [10, 9, 24, 15, 18, 1]. The 31 and 28 in the talent vector mean "training" and "skills" appear close to "talent" 31 and 28 times, respectively. The window size in a W2V model defines what is regarded as appearing close. The window size of 3, for example, means three or fewer between two words after removing stopwords and forming ngrams are regarded as being close and thus count toward the vectors. The definition of the cosine similarity between vectors is as follows.

Cosine similarity between vectors A and B **= $\Sigma A_i B_i$ / ($\sqrt{\Sigma A_i^2}\sqrt{\Sigma B_i^2}$)**

Cosine similarity between talent and succession_planning

= 1638/(sqrt(2803)*sqrt(1022)) = 0.97

Cosine similarity between talent and supply_chain

= 1133/(sqrt(2803)*sqrt(1307)) = 0.59

The higher cosine similarity of 0.97 vs. 0.59 means that talent is more closely related to succession_planning than to supply_chain in the 10K texts.

This numerical example provides an intuitive explanation of how to use vectors to quantify the relationship between any pair of words and phrases and what kind of challenges we face when applying this method to financial statements. We had only six components in the above vectors meaning that we represented the word "talent" using only six words, but over fifty thousand words appear in a financial statement on average and the dimension grows exponentially with phrases that are combinations of words.

This example also provides a clue to reduce the dimension to make this vectorization method practical. When using simple counting of words for forming vectors as in the previous example, the implicit assumption is the index words [training, skills, inflation, operating_efficiencies, analytics, compensation] are orthogonal, meaning no relationship between "training" and "skills", for example, which is not true, leading to unnecessarily many zeros or smaller numbers with higher-dimensional vectors. We can reduce dimension as well as unnecessary zeros significantly by using combinations instead of all words and ngrams in the corpus.

Mikolov et al. (2013) is a seminar paper addressing the issue by word embedding and this model is called W2V. They applied a training algorithm called backpropagation. The algorithm is common in neural networks to make parameters in the network adjusted and become an effective vector representation of a word when the learning is complete after iterations through the corpus. The neural network of word embedding works like concatenated regressions. Hidden layers receive output from the previous layer as an input and feed the output forward to the next layer. The weight matrix randomly selected initially for the vectors continually improves as a backpropagation algorithm in a feed-forward neural network learns from mistakes and adjusts. The learning is complete after the neural network is adept at the task after passing through the entire corpus iteratively and the result is a final vector representation for the trained corpus. Li *et al.* (2021a and b) apply W2V to earnings call transcripts to analyze corporate culture and to examine the impact of Covid-19 on businesses and their responses. When I apply W2V to 10Ks for this paper, I set the window size to 5, the number of iterations to 30, and the minimum word count in the corpus to be considered to be 3.

Table 1: Summary statistics

This table presents the mean, median, standard deviation, and the 25th and 75th percentiles of human capital scaled by market capitalization (*i*HCap/ME). *i*HCap is estimated by capitalizing stock-based compensation reported in annual financial statements and ME is the market capitalization as of December 31 of the year when the fiscal year ends. The presented statistics are after winsorizing outliers at the 95th percentile.

Panel A: Time series of cross-sectional distribution: 1998 - 2023

Fiscal year ends	Number of firm-years	% of firm-years with *i*HCap	*i*HCap/ME (%) for the firm-years with *i*HCap				
			Mean	Standard Deviation	Percentiles		
					25th	Median	75th
1998-2023	109,626	71.69	4.64	6.84	0.62	1.92	5.05
2000	5,842	0.29	3.73	6.69	0.65	1.39	2.91
2003	**4,787**	**60.31**	**1.02**	**2.91**	**0.00**	**0.17**	**0.73**
2010	3,772	97.35	4.31	5.53	1.12	2.35	5.06
2015	3,561	97.84	5.71	7.27	1.28	2.77	6.41
2023	**3,396**	**98.67**	**8.92**	**9.53**	**1.74**	**4.33**	**14.07**

Panel B: Cross-section by 2-digit Standard Industrial Classification code (SIC2): 2003 vs. 2023

Cross-section/ Top 5 industries with 50+ firms (SIC2)	Firm-years with iHCap	*i*HCap/ME (%) for the firm-years with iHCap				
		Mean	Standard Deviation	Percentiles		
				25th	Median	75th
2003						
Communications (48)	101	2.20	5.21	0.00	0.31	0.90
Business services (73)	**411**	**1.97**	**4.72**	**0.04**	**0.34**	**1.50**
Management services (87)	71	1.75	4.66	0.02	0.18	1.15
Electronic & electrical (36)	256	1.03	2.40	0.00	0.14	0.76
Measuring & analyzing (38)	198	1.00	3.46	0.00	0.11	0.58
2023						
Pharmaceutical (28)	443	14.30	10.39	4.50	11.86	27.07
Business services (73)	**467**	**12.28**	**9.77**	**3.60**	**9.00**	**22.16**
Health services (80)	68	12.23	10.36	3.23	7.48	26.30
Credit unions (61)	55	11.85	9.77	3.54	8.54	20.08
Retail (59)	61	11.66	10.65	2.87	10.31	27.07

Table 2: Fama-MacBeth regressions to test the relationship between *i*HCap/ME and future stock returns

This table reports average Fama and MacMeth (1973) regression slopes (multiplied by 100) and the Newey-West *t*-statistics (in parentheses) from cross-sectional regressions that predict monthly stock returns. The sample period for the monthly regressions is from July 2003 to December 2023 (246 months). Panel A is for all stocks and then the sample is divided into two size groups: All-but-microcaps (ABM) in Panel B and microcap stocks (Micro) in Panel C. Micro is for stocks with a market value of equity below the 20th percentile of NYSE market capitalization distribution, and ABM is for all else. Kcap is knowledge capital, Ocap is organization capital, and Gdwl is goodwill. Control variables are cash-based operating profitability scaled by total assets (cop), size defined as a log of market capitalization in the previous month, short-term reversal ($r_{1,1}$), and momentum (r_{12-1}). The numbers in parentheses show *t*-statistics and ***, **, and * denote that the *t*-statistics are significant at the 1 percent, 5 percent, and 10 percent levels, respectively.

Panel A: All stocks

Explanatory variable	(1)	(2)	(3)	(4)	(5)
log (*i*HCap/ME)	0.12 (2.02)**				
log (BE/ME)		0.13 (1.29)			
log (Kcap/ME)			0.02 (1.45)		
log (Ocap/ME)				0.14 (2.84)***	
log (iBE/ME)					0.30 (2.90)***
cop	2.43 (6.56)***	2.27 (6.18)***	2.33 (6.06)***	2.25 (5.90)***	2.30 (6.03)***
log (ME)	-0.07 (-1.30)	-0.07 (-1.34)	-0.09 (-1.52)	-0.05 (-0.85)	-0.02 (-0.30)
$r_{1,1}$	-2.08 (-3.69)***	-2.12 (-3.82)**	-2.05 (-3.69)***	-2.10 (-3.78)***	-2.15 (-3.85)***
r_{12-1}	-0.09 (-0.29)	-0.06 (-0.21)	-0.09 (-0.28)	-0.08 (-0.27)	-0.10 (-0.33)
Adj-R^2	3.02%	3.19%	2.94%	2.99%	3.12%

Panel B: All-but-microcaps (ABM)

Explanatory variable	(1)	(2)	(3)	(4)	(5)
log (*i*HCap/ME)	0.30 (5.55)***				
log (BE/ME)		-0.09 (-0.79)			
log (Kcap/ME)			0.03 (2.82)***		
log (Ocap/ME)				0.12 (2.85)***	
log (iBE/ME)					0.30 (2.44)**
cop	-0.64 (-1.21)	-0.44 (-0.88)	-0.63 (-1.23)	-0.57 (-1.14)	-0.37 (-0.73)
log (ME)	-0.59 (-9.05)***	-0.54 (-8.42)***	-0.58 (-8.75)***	-0.56 (-8.69)***	-0.53 (-8.91)***
$r_{1,1}$	-2.06 (-2.81)***	-1.99 (-2.59)**	-1.95 (-2.53)**	-1.97 (-2.55)**	-2.10 (-2.76)***
r_{12-1}	-0.16 (-0.52)	-0.18 (-0.53)	-0.16 (-0.49)	-0.14 (-0.42)	-0.15 (-0.48)
Adj-R^2	5.61%	4.91%	4.77%	4.77%	5.21%

Panel C: Microcap stocks (Micro)

Explanatory variable	(1)	(2)	(3)	(4)	(5)
log (*i*HCap/ME)	0.05 (0.70)				
log (BE/ME)		0.10 (0.87)			
log (Kcap/ME)			0.03 (1.72)*		
log (Ocap/ME)				0.10 (1.26)	
log (iBE/ME)					0.26 (2.03)**
cop	2.70 (6.97)***	2.80 (7.33)***	2.78 (7.05)***	2.66 (6.72)***	2.67 (6.72)***
log (ME)	-0.96 (-6.70)***	-0.96 (-7.27)***	-0.97 (-6.91)***	-0.92 (-6.15)***	-0.89 (-5.94)***
$r_{1,1}$	-1.83 (-2.76)***	-1.82 (-2.73)***	-1.78 (-2.70)***	-1.84 (-2.78)***	-1.87 (-2.81)***
r_{12-1}	0.02 (0.07)	-0.04 (-0.10)	-0.03 (-0.09)	-0.02 (-0.06)	-0.06 (-0.17)
Adj-R^2	3.27%	3.26%	3.15%	3.25%	3.30%

Table 3: Risk and Return of Portfolios Formed on *i*HCap/ME

This table presents the risk-adjusted performance of value-weighted portfolios formed on human capital (*i*HCap) and market capitalization (ME). Portfolios are formed at the end of June each year *t* using NYSE median market capitalization and 30th and 70th percentiles of *i*HCap/ME. Big High *i*HCap/ME is the portfolio of stocks with the above median ME and the iHCap/ME ratio above the 70th percentile. Big Low *i*HCap/ME is the portfolio of stocks with above median ME and the *i*HCap/ME ratio below the 30th percentile. Small High *i*HCap/ME is the portfolio of stocks with below median ME and the *i*HCap/ME ratio above the 70th percentile. Small Low *i*HCap/ME is the portfolio of stocks with below median ME and the *i*HCap/ME ratio below the 30th percentile. Panel A compares the average return, standard deviation, and *t*-statistic of the four portfolios. Panel B presents the factor models applied to the excess return on the four portfolios by regressing the excess return on the market (MFA), size (SMB), value (HML), profitability (RMW), investments (CMA), and momentum (UMD) factors as in Fama and French (1993, 2015, and 2018). The sample period is July 2003 – December 2023 and the Kenneth French Data Library is the data source for factor returns and the risk-free rate. The numbers in parentheses show *t*-statistics and ***, **, and * denote that the *t*-statistics are significant at the 1 percent, 5 percent, and 10 percent levels, respectively.

Panel A: Risk and return of portfolios formed on *i*HCap/ME

Portfolios	Big High	Big Low	Small High	Small Low
Average return	1.01	0.81	1.05	0.84
Standard deviation	5.83	3.93	6.53	5.15
t-statistic	2.70***	3.23***	2.53**	2.55**

Panel B: Regressions of *i*HCap/ME portfolio excess returns on seven factors

	Big High	Big Low	Small High	Small Low
Intercept	0.14 (1.56)	-0.11 (-2.07)***	0.04 (0.59)	-0.04 (-0.59)
MFA	1.17 (48.27)***	0.93 (71.83)***	1.09 (70.45)***	0.85 (48.51)***
SMB	-0.05 (-1.10)	-0.12 (-5.41)***	0.98 (35.56)***	0.81 (25.85)***
HML	0.49 (12.45)***	-0.04 (-1.92)*	0.02 (0.94)	0.21 (7.13)***
RMW	-0.37 (-7.30)***	0.18 (6.29)***	-0.24 (-7.03)***	0.05 (1.21)
CMA	-0.35 (-5.82)***	0.06 (1.88)*	0.18 (4.58)***	-0.14 (-3.03)***
UMD	-0.08 (-3.45)***	0.03 (1.98)**	-0.03 (-1.64)	0.04 (2.26)**
Adj-R^2	94.39%	96.16%	98.01%	95.93%

Table 4: Textual analysis of financial constraints applied to Fama-MacBeth regressions

This table presents firm-level tests of the relationship between prior investments in human capital on future stock returns after adjusting the depreciation rate of the investments using additional information extracted from financial statements such as the proportion of negative words (P_negative) and financially constraining words (P_constraining). P_negative is the number of negative words as a percentage of the total words in each financial statement and P_constraining is the number of financially constraining words as a percentage of the total words. Panel A shows the summary statistics of the variables and Panel B reports average Fama and MacMeth (1973) regression slopes (multiplied by 100) and the Newey-West *t*-statistics (in parentheses) from cross-sectional regressions that predict monthly stock returns. *t*HCap is defined as the human capital estimate with a depreciation rate from a forward-looking profit model if P_negative or P_constraining are below the 70th percentile. ME is the market capitalization of the stock as of the last trading day of the year when the firm's fiscal year ends. The sample period for the monthly regressions is from July 2003 to December 2022. The sample is divided into two size groups: All-but-microcaps (ABM) and microcap stocks (Micro). Micro is for stocks with a market value of equity below the 20th percentile of the NYSE market capitalization distribution. ABM includes all other stocks. Control variables are cash-based operating profitability scaled by total assets (cop), size defined as a log of market capitalization in the previous month, short-term reversal ($r_{1,1}$), and momentum ($r_{12\text{-}1}$). The numbers in parentheses show *t*-statistics and ***, **, and * denote that the *t*-statistics are significant at the 1 percent, 5 percent, and 10 percent levels, respectively.

Panel A: Summary statistics of negative and financially constraining words from a textual analysis of financial statements

Fiscal year ends	Average number of words	Average P_negative	70th percentile of P_negative	Average P_constraining	70th percentile of P_constraining
1998-2022	58,602	1.76	1.96	0.82	0.91
2003	51,487	1.59	1.76	0.75	0.83
2010	59,185	1.78	1.97	0.83	0.93
2015	63,588	1.83	2.01	0.86	0.94
2022	64,201	2.07	2.27	0.90	0.97

Panel B: Fama MacBeth regressions using *t*HCap/ME

Textual Analysis	P_negative and P_constraining below the 70th percentile		P_negative below the 70th percentile		P_constraining below the 70th percentile	
Size	ABM	Micro	ABM	Micro	ABM	Micro
log (*t*HCap/ME)	0.20 (3.44)***	-0.10 (-1.41)	0.22 (3.77)***	-0.09 (-1.11)	0.21 (4.07)***	-0.07 (-0.93)
cop	-1.19 (-2.66)***	1.15 (3.65)***	-0.97 (-2.56)**	1.10 (3.78)***	-1.45 (-3.50)***	1.10 (3.96)***
log (ME)	-0.46 (-8.55)***	-0.76 (-6.15)***	-0.46 (-8.64)***	-0.75 (-6.40)***	-0.53 (-9.06)***	-0.80 (-6.55)***
$r_{1,1}$	-1.57 (-2.09)**	-2.15 (-2.97)***	-1.19 (-1.63)	-1.64 (-2.37)**	-0.94 (-1.39)	-1.77 (-2.90)***
$r_{12\text{-}1}$	-0.25 (-0.56)	0.61 (1.89)*	-0.21 (-0.48)	0.76 (2.36)**	-0.27 (-0.70)	0.37 (1.32)
Adj-R^2	5.91%	3.97%	5.74%	4.08%	5.35%	3.26%

Table 5: Portfolios Formed on *t*HCap/ME using textual analysis of financial constraints

This table presents the risk-adjusted performance of value-weighted portfolios formed on human capital adjusted with textual information scaled by market capitalization (*t*HCap/ME). Portfolios are formed at the end of June each year t using NYSE median market capitalization and 30th and 70th percentiles of *t*HCap/ME. Big High *t*HCap/ME is the portfolio of stocks with the above median ME and the *t*HCap/ME ratio above the 70th percentile. Big Low *t*HCap/ME is the portfolio of stocks with above median ME and the *t*HCap/ME ratio below the 30th percentile. Small High iHCap/ME is the portfolio of stocks with below median ME and the *t*HCap/ME ratio above the 70th percentile. Small Low *t*HCap/ME is the portfolio of stocks with below median ME and the *t*HCap/ME ratio below the 30th percentile. Panel A compares the average return, standard deviation, and *t*-statistic of the four portfolios. Panel B presents the factor models applied to the excess return on the four portfolios by regressing the excess return on the market (MFA), size (SMB), value (HML), profitability (RMW), investments (CMA), and momentum (UMD) factors as in Fama and French (1993, 2015, and 2018). The numbers in parentheses show *t*-statistics and ***, **, and * denote that the *t*-statistics are significant at the 1 percent, 5 percent, and 10 percent levels, respectively.

Panel A: Risk and return of portfolios formed on *t*HCap/ME

Portfolios	Big High	Big Low	Small High	Small Low
Average return	0.98	0.64	1.03	0.76
Standard deviation	5.87	4.74	6.48	5.71
t-statistic	2.48**	2.00**	2.38**	1.99**

Panel B: Regressions of *t*HCap/ME portfolio excess returns on seven factors

	Big High	Big Low	Small High	Small Low
Intercept	0.20 (1.89)*	-0.18 (-2.36)***	0.12 (2.17)**	-0.08 (-1.34)
MFA	1.10 (43.93)***	0.99 (52.21)***	1.07 (78.60)***	0.93 (67.79)***
SMB	-0.13 (-2.92)***	-0.20 (-5.84)***	0.92 (36.73)***	0.29 (11.54)***
HML	0.47 (10.66)***	0.20 (6.04)***	0.07 (3.05)***	0.31 (12.94)***
RMW	-0.39 (-6.80)***	-0.09 (-1.97)*	-0.18 (-5.60)***	-0.09 (-2.94)***
CMA	-0.29 (-4.23)***	-0.10 (-1.92)*	-0.01 (-0.23)	-0.10 (-2.78)***
UMD	-0.09 (-3.35)***	-0.06 (-3.16)***	-0.01 (-0.89)	-0.01 (-0.73)
Adj-R^2	93.86%	94.69%	98.48%	97.48%

Table 6: Machine Learning of Financially Constraining Words and Phrases

This table presents the words and phrases identified through Word2Vec, a machine learning tool, applied to financial statement texts using financially constraining words in Bodnaruk et al. (2015) as seed words.

Words and phrases identified by Word2Ved	Seed words
abuse	forbid
acceptable_terms	unavailable
adhere	comply | abide | strict
adverse	impair
alienate	encumber
avoid	imposition
banning	forbid
blacklist	stipulations
boycott	restraint
cancel	noncancelable
criticizing	precluding
damage	impair
delay	prevent
discontinue	unavailable
dismiss	stipulations
fail	require
forced	require | compel
goodwill	impairment
goodwill_impairment	impair
harm	impair
heavily_dependent	dependent
hinder	limit | restrict | impair | inhibit
illegal	prohibited
inconsistent	stricter
indemnify	precondition
interfere	forbidden
interruption	unavailability
jeopardize	inhibiting
outage	unavailability
penalties	impositions
pressures	constraints
recover	impair
refuse	compel | obligating
shortfalls	constraints
shortages	constraints
unforeseen	entail
unplanned	prohibitively
unsuccessful	prevent
vary	depending
violate	abide | prohibit
write_downs	impairments

Table 7: Machine Learning of financial constraints and Fama-MacBeth regressions

This table presents firm-level tests of the relationship between prior investments in human capital on future stock returns after adjusting the depreciation rate of the investments using additional information extracted from financial statements using a machine learning tool, Word2Vec. P_constraining_W2V is the number of financially constraining words and phrases identified using Word2Vec as a percentage of the total words. Panel A shows the summary statistics and Panel B reports average Fama and MacMeth (1973) regression slopes (multiplied by 100) and the Newey-West *t*-statistics (in parentheses). *w*HCap is defined as the human capital estimate with a depreciation rate from a forward-looking profit model if P_constraining_W2V is below the 70th percentile. ME is the market capitalization of the stock as of the last trading day of the year when the firm's fiscal year ends. The sample period for the monthly regressions is from July 2003 to December 2022. The sample is divided into two size groups: All-but-microcaps (ABM) and microcap stocks (Micro). Control variables are cash-based operating profitability scaled by total assets (cop), size defined as a log of market capitalization in the previous month, short-term reversal ($r_{1,1}$), and momentum (r_{12-1}). The numbers in parentheses show *t*-statistics and ***, **, and * denote that the *t*-statistics are significant at the 1 percent, 5 percent, and 10 percent levels, respectively.

Panel A: Summary statistics of negative and financially constraining words from a textual analysis of financial statements

Fiscal year ends	Average number of words	Average P_constraining	70th percentile of P_constraining	Average P_constraining_W2V	70th percentile of P_constraining_W2V
1998-2022	58,602	0.82	0.91	1.02	1.12
2003	51,487	0.75	0.83	0.92	1.01
2010	59,185	0.83	0.93	1.02	1.12
2015	63,588	0.86	0.94	1.06	1.17
2022	64,201	0.90	0.97	1.14	1.24

Panel B: Fama MacBeth regressions using *w*HCap/ME

Textual Analysis Word2Vec	P_constraining below the 70th percentile		P_constraining_W2V below the 70th percentile	
Size	ABM	Micro	ABM	Micro
log (*w*HCap/ME)	0.21 (4.07)***	-0.07 (-0.93)	0.20 (3.62)***	-0.07 (-0.96)
cop	-1.45 (-3.50)***	1.10 (3.96)***	-1.28 (-2.99)***	1.13 (4.33)***
log (ME)	-0.53 (-9.06)***	-0.80 (-6.55)***	-0.48 (-8.68)***	-0.75 (-6.17)***
$r_{1,1}$	-0.94 (-1.39)	-1.77 (-2.90)***	-0.70 (-0.96)	-1.97 (-3.08)***
r_{12-1}	-0.27 (-0.70)	0.37 (1.32)	-0.16 (-0.40)	0.44 (1.36)
Adj-R^2	5.35%	3.26%	5.51%	3.63%

Table 8: Machine Learning of Words and Phrases Describing Human Capital

This table presents the words and phrases identified through Word2Vec, a machine learning tool, applied to financial statement texts using seed words and phrases related to human capital management.

Seed words	Words and phrases identified by Word2Vec	Root
accountability attract_best best_talent building_skill career_growth changing_way chief_people colleague_experience continuous_development culture employee_experience empowerment fair_equal feedback future_fit growth_development growth_our health heart highly_engage internal_candidates key_positions labor_relations leadership_development our_people people_strategy people_success reskilling right_skills succession_planning sustainable talent_attraction team_members total_rewards upskilling well_being retention	career	talent_attraction
	career_advancement	reskilling
	career_mobility	talent_attraction
	career_path	upskilling
	career_progression	reskilling
	coaching	succession_planning
	collaborate	team_members
	colleague	talent_attraction
	colleague_engagement	talent_attraction
	compensation	succession_planning
	compensation_philosophy	succession_planning
	continuous_learning	upskilling
	conversations	feedback
	core_values	culture
	creativity	culture
	employees	team_members
	empower	team_members
	engagement	feedback
	ethical	accountability
	excellent	empowerment
	foster	empowerment
	goal_setting	succession_planning
	healthcare	health
	honesty	accountability
	honor	labor_relations
	innovation	culture
	insight	feedback
	integrity	accountability
	guidelines	accountability
	knowledgeable	empowerment
	leaders	team_members
	leadership	succession_planning
	learning	upskilling
	lifelong_learning	upskilling
	listen	feedback
	loyalty	retention
	mentorship	reskilling
	morale	retention
	people	labor_relations
	privacy_protection	accountability
	recommendations	feedback
	recruitment	talent_attraction
	resilience	accountability
	responsible	succession_planning
	retraining	labor_relations
	safety	accountability
	strategic_planning	succession_planning
	stewardship	accountability
	succession_plans	succession_planning
	suggestions	feedback
	surveys	feedback
	talent	succession_planning
	talent_pipeline	reskilling
	team	team_members

	teammates	team_members
	team	team_members
	teamwork	culture
	thrive	culture
	top_performers	talent_attraction
	transparency	accountability
	turnover	retention
	wellness	health
	work_closely	team_members

Table 9: Machine learning of human capital applied to Fama-MacBeth regressions

This table presents firm-level tests of the relationship between prior investments in human capital on future stock returns using additional information extracted from financial statements such as the proportion of words and phrases describing corporate culture related to human capital management (P_HCulture) and financially constraining words (P_constraining). P_HCulture is the number of words and phrases related to human capital as a percentage of the total words. P_constraining is the number of financially constraining words as a percentage of the total words. Panel A shows the average and 70th percentile of P_HCulture and Panel B reports average Fama and MacMeth (1973) regression slopes (multiplied by 100) and the Newey-West *t*-statistics (in parentheses) from cross-sectional regressions that predict monthly stock returns. dHCulture is a dummy variable that is one if P_HCulure is above the 70th percentile and zero otherwise. *t*HCap is defined as the human capital estimate with a depreciation rate from a forward-looking profit model if P_constraining is below the 70th percentile. ME is the market capitalization of the stock as of the last trading day of the year when the firm's fiscal year ends. The sample period for the monthly regressions is from July 2003 to December 2022. The sample is divided into two size groups: All-but-microcaps (ABM) and microcap stocks (Micro). Micro is for stocks with a market value of equity below the 20th percentile of the NYSE market capitalization distribution. ABM includes all other stocks. Control variables are cash-based operating profitability scaled by total assets (cop), size defined as a log of market capitalization in the previous month, short-term reversal ($r_{1,1}$), and momentum (r_{12-1}). The numbers in parentheses show *t*-statistics and ***, **, and * denote that the *t*-statistics are significant at the 1 percent, 5 percent, and 10 percent levels, respectively.

Panel A: Summary statistics of words and phrases explaining human capital

Fiscal year ends	Average P_HCulture	70th percentile of P_HCulture	Fiscal year ends	Average P_HCulture	70th percentile of P_HCulture
2003	0.38	0.42	2020	0.42	0.47
2010	0.37	0.41	2021	0.45	0.51
2015	0.37	0.41	2022	0.46	0.52

Panel B: Fama MacBeth regressions to test the impact of corporate culture on the valuation of investments in human capital

Explanatory variable	ABM			Micro		
dHCulture	0.25 (2.70)***		0.28 (1.07)	0.01 (0.06)		0.23 (0.81)
log (*t*HCap/ME)		0.21 (3.89)***	0.20 (3.41)***		-0.08 (-1.03)	-0.08 (-1.03)
dHCulture *log (*t*HCap/ME)			0.03 (0.50)			0.01 (0.15)
cop	-1.31 (-3.47)***	-1.45 (-3.43)***	-1.48 (-3.54)***	1.18 (4.47)***	1.12 (4.01)***	1.12 (4.03)***
log (ME)	-0.56 (-8.95)***	-0.52 (-8.87)***	-0.52 (-8.87)***	-0.78 (-5.79)***	-0.79 (-6.32)***	-0.78 (-6.31)***
$r_{1,1}$	-0.70 (-1.05)	-0.86 (-1.25)	-0.88 (-1.28)	-1.52 (-2.55)**	-1.90 (-3.07)***	-1.92 (-3.09)***
r_{12-1}	-0.17 (-0.45)	-0.28 (-0.72)	-0.29 (-0.75)	0.55 (1.86)*	0.37 (1.28)	0.36 (1.23)
Adj-R^2	4.98%	5.38%	5.67%	3.12%	3.29%	3.40%

Figure 1. Cumulative Returns on Portfolios Formed on Human Capital

This figure shows the growth of portfolios formed on human capital to market capitalization ratio (*i*HCap/ME) along with S&P 500 total return as a benchmark from June 30, 2003, to December 31, 2023. The *i*HCap/ME portfolios are constructed using the NYSE median size and the 30th and 70th percentiles of *i*HCap/ME and have a starting value of 100 on June 30, 2003.

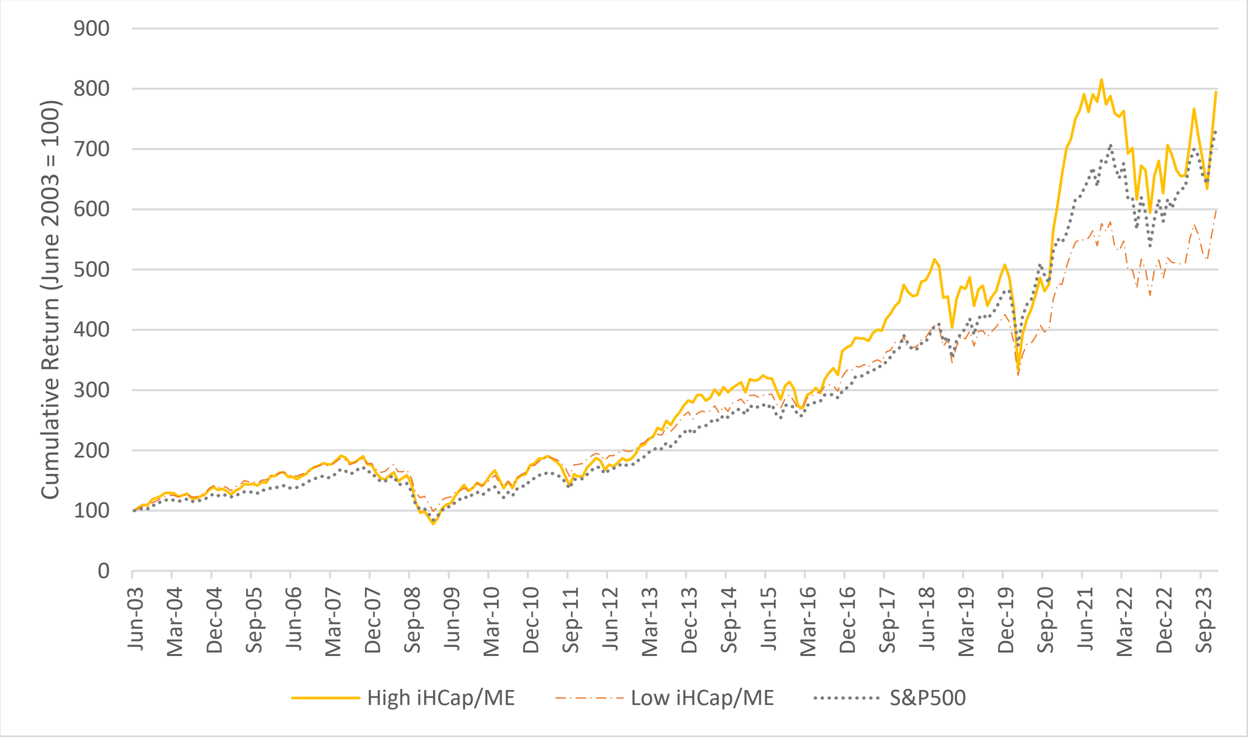


Figure 2. The average number of words in 10-Ks submitted to the SEC EDGAR

This figure presents the cross-sectional average number of words in 10-K annual financial statements submitted to the SEC by the calendar year when each fiscal year ends from 1995 to 2022 and there are 214,576 firm-years.

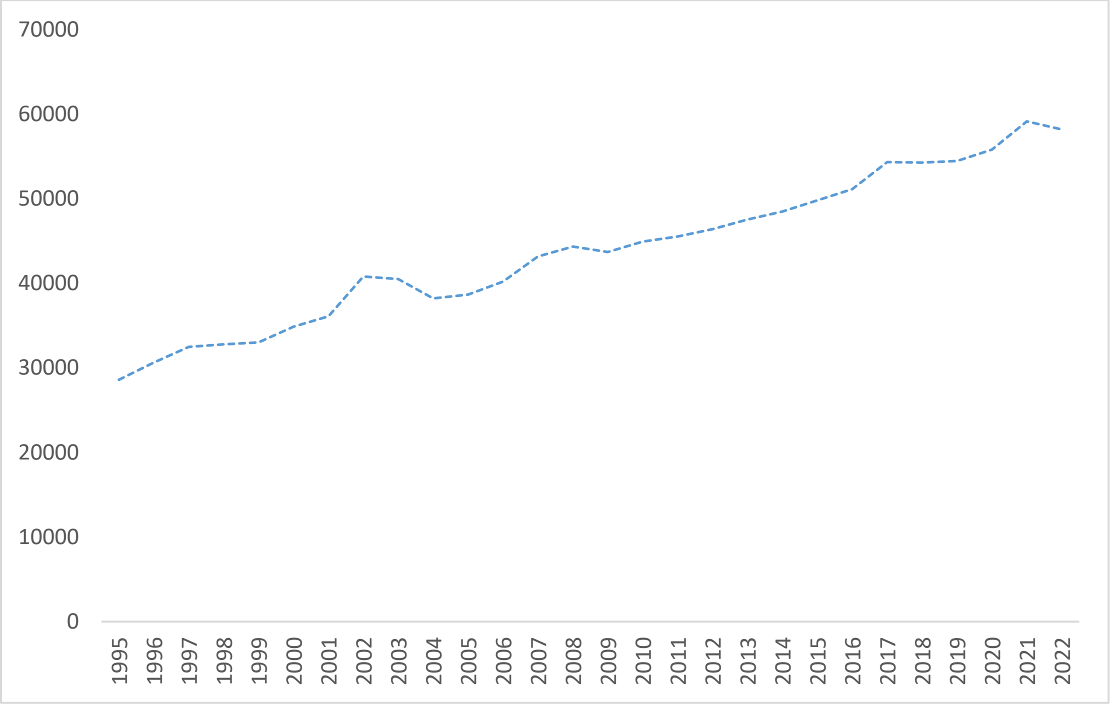


Figure 3. Performance of Portfolios Formed on Human Capital Adjusted with Financial Constraints

This figure shows the growth of portfolios formed on human capital to market capitalization ratio adjusted with textual information on financial constraints (*t*HCap/ME) along with S&P 500 total return as a benchmark from June 30, 2003, to December 31, 2023. The *t*HCap/ME portfolios are constructed using the NYSE median size and the 30th and 70th percentiles of *t*HCap/ME and have a starting value of 100 on June 30, 2003.

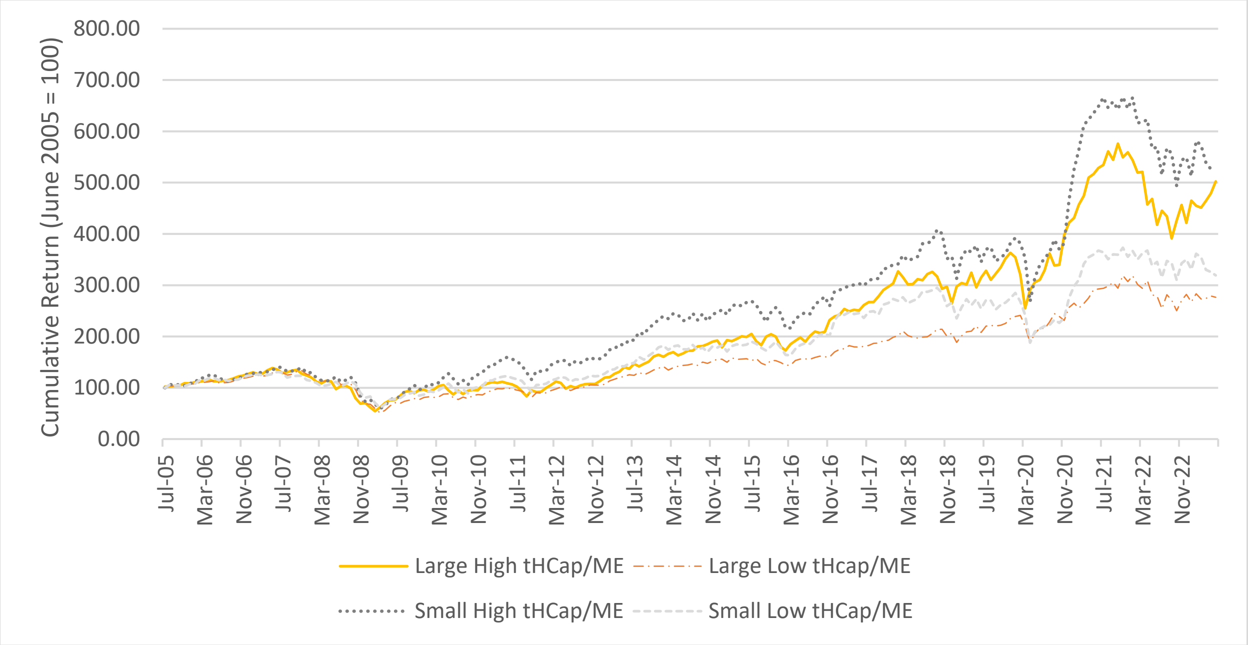


Figure 4. Words and Phrases Related to Human Capital in Financial Statement Texts

This figure presents the time series of the total and the proportion of human capital-related words and phrases as in Table 8 using 10K financial statements of firms that have a fiscal year ending in December. It shows that the total counts and the proportion of human capital words and phrases increased sharply after the new SEC rule on human capital disclosure became effective in November 2020.

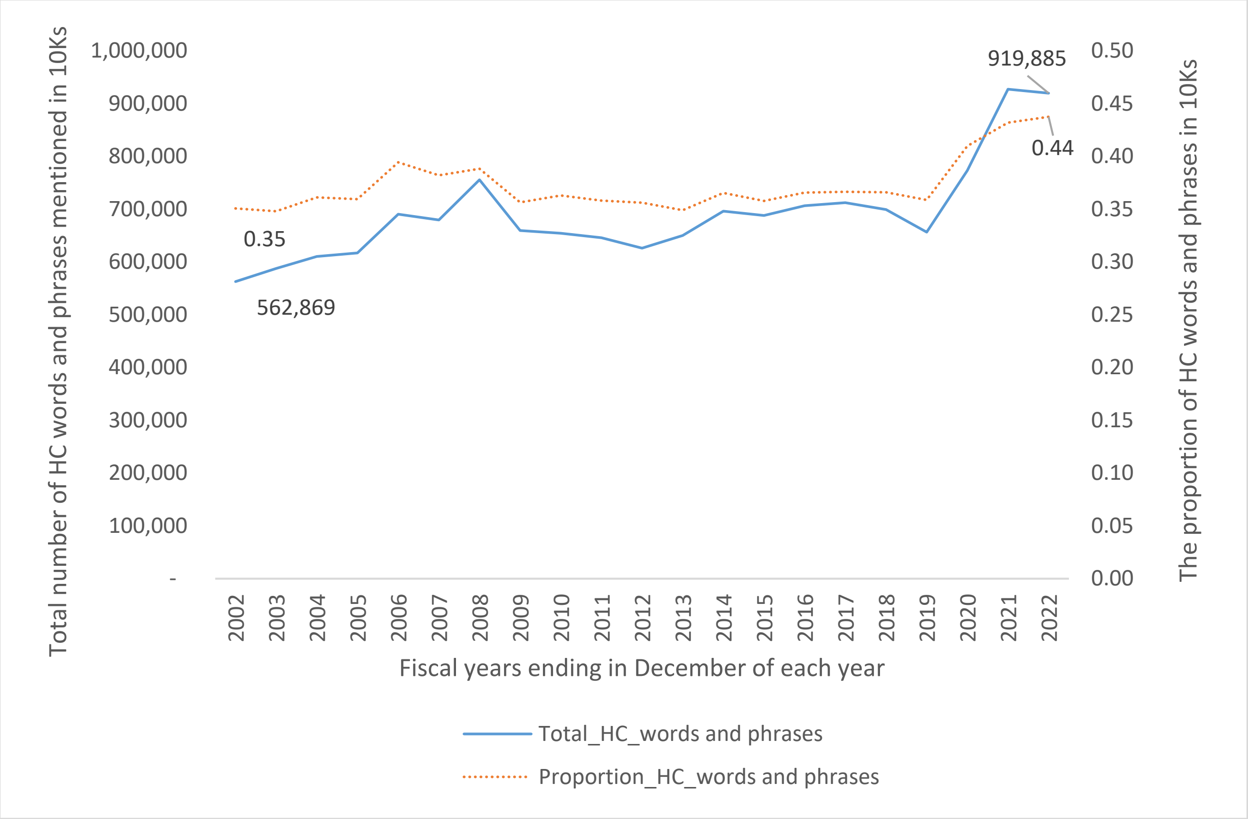
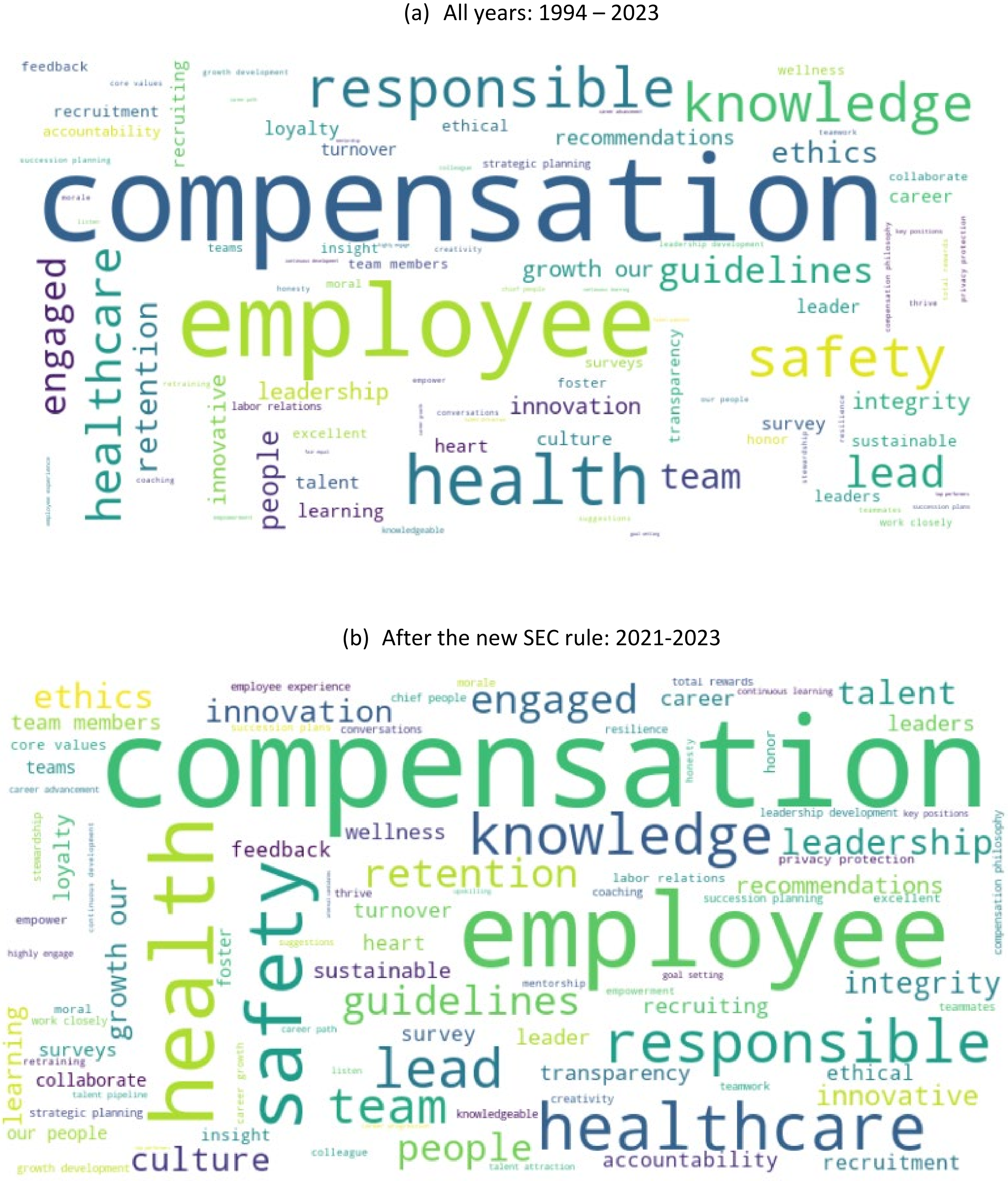


Figure 5. The Impact of the New Disclosure Rule on the Word Clouds Describing Human Capital

This figure compares the word cloud from applying the dictionary in Table 8 to the financial statements filed with the SEC from 2021 to 2023 with the word cloud for the entire sample period of 1994 – 2023.



(a) All years: 1994 – 2023

(b) After the new SEC rule: 2021-2023