# Unveiling the Risk Premium of Knowledge-Intensive Firms: Introducing the HKR Factor

Pedro Vallocci*
University of California, Santa Cruz

December 5, 2024

**Abstract**

This paper introduces a novel measure of knowledge capital risk derived from textual analysis of firms' 10-K filings. Using cosine similarity to R&D-related terms, I construct a High-Knowledge Risk (HKR) factor that captures firms' exposure to uncertainty from R&D outcomes. I demonstrate that the HKR factor offers significant explanatory power for cross-sectional equity returns, even after controlling for established equity pricing factors. This explanatory power persists in out-of-sample tests. Portfolios formed on HKR exhibit a substantial return premium, with high-HKR equities outperforming low-HKR equities over a 14-year period. A trading strategy based on the HKR factor also displays the highest Sharpe ratio among compared factors, indicating an attractive risk-adjusted return. These findings suggest that knowledge capital risk represents a distinct source of priced risk not fully captured by traditional asset pricing models and highlight the importance of knowledge capital risk in asset pricing.

**Keywords:** Innovation, firm risk factors, productivity, intangible capital, natural language processing, empirical asset pricing

**JEL classification:** C43, C55, D2, E22, G11, O3

## 1 Introduction

Recent literature highlights the growing importance of knowledge capital – a firm's cumulative investment in innovation – for firm valuation (Belo et al., 2022). Previous measures of knowledge capital, while informative, have not fully addressed how to quantify

*I am very grateful to my advisor, Galina Hale, for her continuous support and invaluable mentorship throughout my Ph.D. I also thank my committee members Grace Gu and Alonso Villacorta for all their guidance. Finally, I appreciate the helpful comments from John Fernald, Chenyue Hu, Michael Leung, Roberto Mauad, Brenda Samaniego de la Parra, Juliana Salomo, Bhavyaa Sharma, and all the UCSC Macro Workshop participants.

either the risks specific to knowledge-intensive firms or the risk premium that investors demand for exposure to companies with higher knowledge capital risk (KKR). This paper investigates whether a risk premium is associated with firms exhibiting higher KKR. I leverage firms' annual reports (10-Ks) to create a novel textual-based measure of KKR. I find that high-KKR firms consistently display higher returns and tail risks than low-KKR firms. Additionally, a derived high-knowledge risk (HKR) asset pricing factor has significant pricing power, even after controlling for established factors. Notably, the HKR factor can be interpreted as a risk rather than as a mispricing anomaly within the Intertemporal Capital Asset Pricing Model (ICAPM), and an HKR-based trading strategy exhibits the highest Sharpe ratio among the factors studied.

The focus on knowledge capital in this study is motivated by the broader transition towards a service- and knowledge-based economy. Existing literature has established the importance of intangibles in firm valuation and economic growth (Corrado et al., 2009b,a; Eisfeldt and Papanikolaou, 2013; Bloom and Van Reenen, 2007). However, the measurement and pricing of KKR remain challenging due to its distinct characteristics. Investments in R&D are associated with uncertainty in research outcomes, obsolescence risk, vulnerability to loss of key talent, and financial constraints. Not all innovation investments succeed, leading to potential cash flow jumps and increased Tobin's q dispersion (Andrei et al., 2019). Firms may need to write off knowledge capital due to competitive forces, resulting in variable R&D capital depreciation rates across time, industries, and firms (Peters and Taylor, 2017; Li and Hall, 2020). Figure 1 illustrates a patent-driven market value jump, showing Amazon's stock price spiking compared to the NASDAQ index on November 13, 2001, when the company was granted a key e-commerce patent.

In this paper, I use natural language processing techniques, including OpenAI's pretrained text embeddings and topic modeling techniques, to extract a numerical measure of KKR from the Risk Factors section (Item 1A) of 10-K reports. My measure offers three advantages over existing approaches. First, it captures firm-specific factors in KKR that are not fully reflected in R&D expenditures or patent counts alone. Second, it addresses inconsistencies in R&D reporting across industries and firms: since all public companies are required to disclose risk factors, this approach provides a more uniform basis for comparison. Third, this technique accounts for the varying efficacy with which firms convert R&D into innovations and increased cash flows. The language used in risk disclosures can provide insights into a firm's confidence in its R&D processes and potential outcomes.

This paper demonstrates how knowledge capital significantly influences equity valuations and associated risk patterns. I show that the HKR factor offers significant explanatory power for the cross-section of equity returns, even after controlling for well-

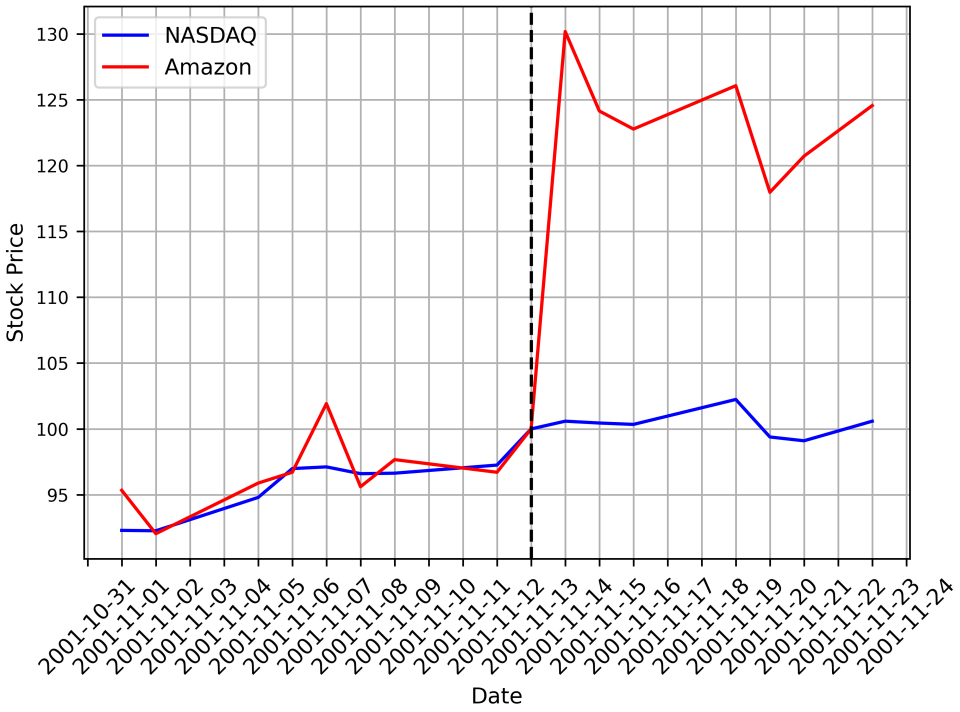



**Figure 1: NASDAQ vs. Amazon Stock Prices in November 2001 (11/13/01 = 100).**
On November 13th, 2001, Amazon was granted a patent for the "use of electronic shopping carts to generate personal recommendations," leading to a significant spike in Amazon's stock price.

established Fama-French factors, and that it improves asset pricing even when the HKR factor is estimated out of sample.

Portfolios formed on high HKR consistently outperform those with low HKR over a 14-year period, suggesting a substantial risk premium associated with knowledge capital intensity. Importantly, firms with higher HKR exhibit higher kurtosis ("fat tails") in their equity returns. This indicates that these firms experience more frequent extreme returns, both positive and negative, compared to firms with lower HKR. This finding aligns with the notion that R&D-intensive firms face more uncertain outcomes, potentially leading to breakthrough innovations or significant setbacks, which are reflected in their equity price movements.

I conduct extensive tests to validate the HKR factor within the framework of the Intertemporal Capital Asset Pricing Model (ICAPM). These tests reveal that the HKR factor satisfies key ICAPM constraints, including its ability to forecast market returns and volatility, as well as its consistent pricing in cross-sectional analyses. The HKR factor also exhibits the highest Sharpe ratio among compared factors.

By introducing this text-based measure of KKR, this study contributes to the growing literature on intangible capital valuation. It highlights the critical role of innovation-related risk in modern financial markets and offers a new tool for researchers and practitioners to understand better and price the risks associated with R&D-intensive firms.

Existing literature has primarily classified firms by KKR using R&D or patent intensity (Andrei et al., 2019). Peters and Taylor (2017) estimate a firm's knowledge capital stock using the perpetual inventory method applied to past R&D spending. Kogan et al. (2017) group firms based on accumulated patent value, estimated from stock market reactions around patent grant dates. While this captures market perceptions of patent value, it does not directly measure KKR. My measure aims to complement these approaches by addressing additional dimensions of KKR. It accounts for varying R&D conversion efficacy across firms and industries, as well as knowledge development that may not result in patent filings. By introducing a text-based measure, I provide a tool that works alongside established metrics to offer a more comprehensive view of KKR, incorporating factors not fully captured by traditional measures (Mezzanotti and Simcoe, 2023; Li and Hall, 2020).

Recent research has also explored the use of textual data and machine learning techniques to improve financial and economic forecasting. Studies have shown that incorporating text data from sources like newspapers and financial filings can enhance predictions of equity returns, volatility, and macroeconomic indicators (Lima and Godeiro, 2023; Kalamara et al., 2022; Rahimikia et al., 2021; Du and Tanaka-Ishii, 2020; Zheng et al., 2024). Word embeddings learned from financial texts have proven useful for various fi-

nancial NLP tasks and portfolio optimization (Sehrawat, 2019; Du and Tanaka-Ishii, 2020). Similarly, stock embeddings derived from news articles and price history can capture meaningful relationships between assets (Dolphin et al., 2022). The integration of textual data with traditional economic indicators has demonstrated improved forecasting performance, particularly during periods of economic stress (Kalamara et al., 2022). These advancements in text-based analysis offer promising avenues for enhancing economic research and financial decision-making (Gentzkow et al., 2019). This paper also contributes to this literature.

The remainder of this paper is organized as follows: Section 2 examines the challenges in quantifying knowledge capital risk, highlighting issues with R&D reporting and the distinction between measuring knowledge capital levels versus associated risks. Section 3 describes the data and methodology used to construct the text-based measure of knowledge capital risk from 10-K filings. Section 4 demonstrates the significant impact of knowledge capital risk on equity valuations and risk patterns, introducing the HKR factor. Section 5 validates the HKR factor within the Intertemporal Capital Asset Pricing Model framework. Section 6 concludes.

# 2 Knowledge capital

The measurement of KKR poses significant challenges due to the unique characteristics of intangible assets and inconsistencies in R&D reporting. Traditional measures based solely on R&D expenditures or patent counts fail to fully capture firm-specific factors in KKR or account for varying efficacy in converting R&D into innovations. Additionally, the relationship between KKR and R&D intensity is likely non-linear, further complicating measurement efforts. These challenges necessitate a more nuanced approach to quantifying KKR that can incorporate firms' own assessments of their innovation-related risks and potential vulnerabilities. The text-based measure developed in this paper aims to address these challenges by leveraging firms' self-reported risk factors in 10-K filings, providing a more comprehensive view of KKR that complements existing metrics.

## 2.1 Characteristics of intangible capital

Intangible assets possess risk profiles that differ systematically from those of tangible assets (Hansen et al., 2005). Key characteristics of intangible assets include partial ownership by key talent, limited collateralizability, and volatility in valuation. Shareholders cannot fully appropriate the cash flows generated by a firm's key talent, as the firm

must compensate this talent at its outside option value (Eisfeldt and Papanikolaou, 2013). Furthermore, high-skilled labor often receives a significant portion of compensation as equity-based pay (Eisfeldt et al., 2023). Some categories of intangibles cannot be used as collateral, potentially providing insurance against aggregate economic shocks and earning lower expected returns (Ai et al., 2020). Companies with substantial knowledge capital experience more volatile valuations due to the unpredictable nature of innovation (Andrei et al., 2019).

Corrado et al. (2009b) estimated total intangible capital in 2003 at $3.6 trillion, with knowledge capital accounting for half of this figure. Crouzet and Eberly (2023) found that overlooking knowledge capital and its associated rents can explain up to two-thirds of the investment gap in R&D-intensive sectors. Tobin's Q has greater explanatory power for investment decisions in R&D-intensive firms (Andrei et al., 2019).

Intangible capital, particularly accumulated R&D expenses, plays a crucial role in both supply-side economic models (Belo et al., 2013) and economic growth models featuring expanding varieties (Romer, 1990) or quality ladders (Grossman and Helpman, 1991; Atkeson and Burstein, 2019).

Companies with high knowledge capital are also more likely to list "loss of key talent" as a risk factor in their 10-K reports (Eisfeldt and Papanikolaou, 2013). This vulnerability extends to immigration-related risks, such as H-1B visa quota shortages (Peri et al., 2015). R&D-intensive firms, especially younger ones, often face challenges in securing debt finance due to unpredictable returns and potential adverse selection and moral hazard in the R&D financing market (Li, 2011; Brown et al., 2009).

## 2.2 Quantifying Knowledge Capital Risk: Challenges

While R&D-based measures of knowledge capital, such as the one described in Peters and Taylor (2017), provide valuable insights, capturing KKR presents additional challenges that warrant further exploration. This section examines two key issues: the complexities of R&D reporting in Compustat and the distinction between measuring knowledge capital levels versus associated risks.

Aboody and Lev (2000) highlight the challenges in valuing innovative firms due to patterns in R&D data reporting. Despite U.S. Financial Accounting Standards mandating the disclosure of relevant R&D expenditures, Koh and Reeb (2015) find that approximately half of NYSE-listed companies do not report any R&D expenditures. Non-random patterns of missing R&D data in annual reports, coupled with substantial patent activity among some non-reporting firms, suggest that R&D alone may not fully capture KKR.

Koh and Reeb (2015) identify several potential factors affecting R&D reporting: difficulties in separating R&D from general expenses, variations in R&D reporting requirements enforcement, absence of materially relevant R&D expenditures, differences in innovation activity structuring, or strategic decisions regarding innovation disclosure (Koh et al., 2018). Adding another layer of complexity to the measurement landscape, Economic policy uncertainty and other macroeconomic variables can also influence R&D disclosures (Amore, 2020). These factors contribute to the observed patterns in R&D data.

Specifically, the pattern of missing R&D data in financial databases is unlikely to be random or to mean negligible R&D expenses. Table 1 reveals substantial differences between firms reporting zero R&D and those with non-reported R&D. These firms differ markedly in industry skill levels, total assets, property, plant, and equipment (PPE), and patent activity. Consequently, equating missing R&D values to zero may mischaracterize a significant subset of innovative firms.

**Table 1: Descriptive Statistics by R&D Reporting Type**
PYVA stands for patent-related yearly value added (to the firm's market value) and is defined as in Kogan and Papanikolaou (2019). Skill is a SIC-industry variable and is retrieved from Belo et al. (2017). PPE stands for property, plant, and equipment. R&D, Assets, PPE, and PYVA in millions of dollars. Skill is defined as a ratio.

| | | R&D | Skill | Assets | PPE | PYVA |
|---|---|---|---|---|---|---|
| R&D Type | Statistic | | | | | |
| Missing | median | – | 0.292 | 2001 | 531 | 0 |
| | mean | – | 0.311 | 17941 | 4622 | 68 |
| | std | – | 0.160 | 110033 | 15000 | 896 |
| Zero | median | 0.000 | 0.082 | 1334 | 618 | 0 |
| | mean | 0.000 | 0.144 | 5565 | 3440 | 36 |
| | std | 0.000 | 0.150 | 16692 | 12067 | 554 |
| Positive | median | 38.000 | 0.372 | 761 | 197 | 4 |
| | mean | 261.000 | 0.385 | 6916 | 3008 | 668 |
| | std | 1203.000 | 0.187 | 31621 | 20000 | 4391 |

Figure 2 illustrates the relationship between knowledge capital intensity and patent intensity using a binscatter plot. Knowledge capital intensity is defined as the ratio of knowledge capital to total assets, following Peters and Taylor (2017). They estimate a firm's knowledge capital $G_{it}$ at the end of period $t$ using the perpetual inventory method:

$$G_{it} = (1 - \delta_{R\&D})G_{i,t-1} + R\&D_{it} \tag{1}$$

where $\delta_{R\&D}$ is the depreciation rate of R&D capital, $G_{i,t-1}$ is the previous period's knowledge capital, and $R\&D_{it}$ is the current period's R&D spending.

Patent intensity, a proxy for reliance on innovation for firm value, is measured as the ratio of a firm's accumulated patent-related private economic value to its total assets. I obtain patent values from the dataset provided by Kogan et al. (2017), which estimates the economic value of each patent based on stock market reactions around patent issuance dates. These values are aggregated annually to construct a metric representing the total added patent value for each firm-year observation. Missing patent values are replaced with zero, and all values are deflated to 1982 dollars using the Consumer Price Index (CPI).

The binscatter plot reveals a non-linear and non-monotonic relationship between knowledge capital intensity and patent intensity, suggesting that the connection between R&D investments and patenting outcomes is neither linear nor monotonic.

Table 2 compares firms that signed a letter supporting high-skilled immigration with non-signatories, providing insight into characteristics of firms self-identifying as vulnerable to KKR. The letter, addressed to President Trump and cabinet secretaries in May 2020, was signed by 324 employers and organizations across various sectors of the American economy, including 45 publicly traded firms. Signatories emphasized the critical role of employment visas (L-1, H-1B, F-1, and H-4) in maintaining a highly-skilled, college-educated science and engineering workforce essential for innovation, production, research, development, and leadership. They argued that constraints on this human capital could result in substantial economic uncertainty and unintended consequences, particularly highlighting their reliance on foreign talent in STEM fields.

A comparison with Table 1 reveals that signatory firms predominantly occupy the uppermost quantiles of R&D expenditure and patent revenue. This finding suggests that the relationship between KKR and R&D expenditure is likely non-linear, challenging the assumption of a simple proportional correlation.

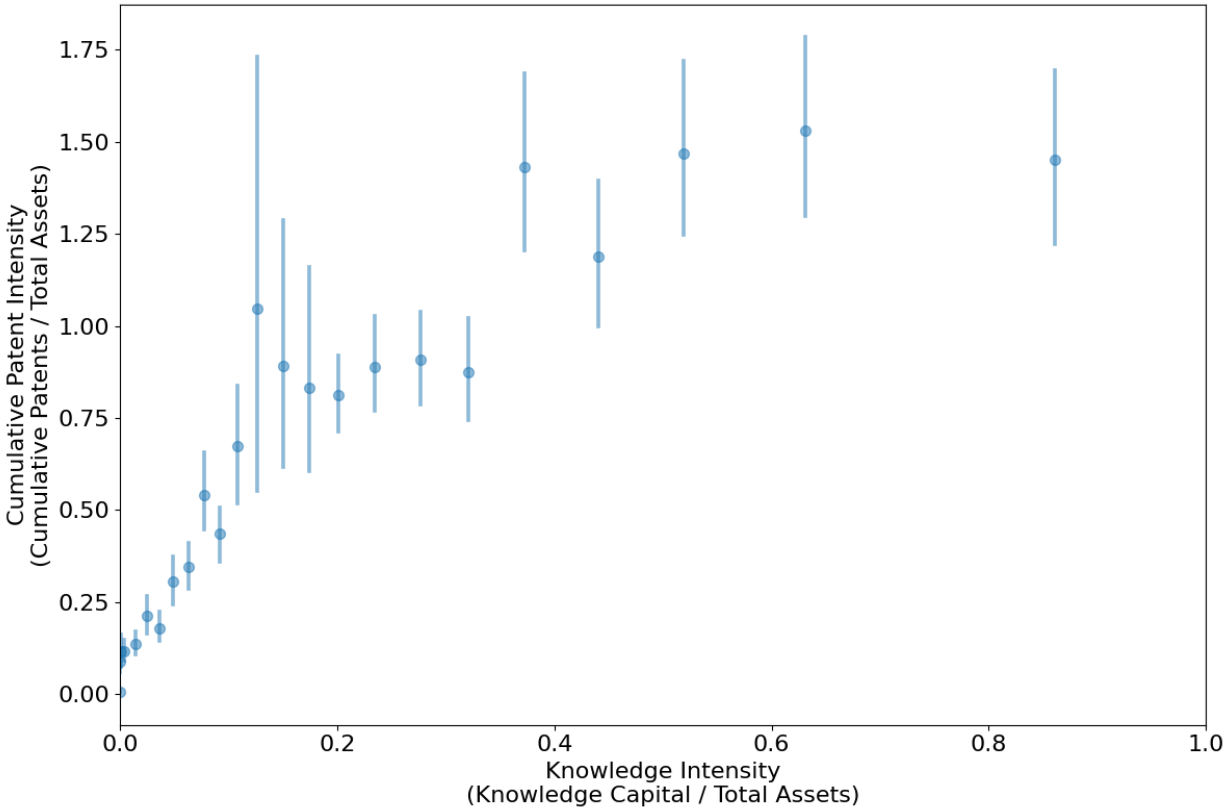



**Figure 2: Bin scatter plot of knowledge capital intensity vs. patent intensity**.
This binscatter plot illustrates the relationship between knowledge capital intensity and patent intensity. Knowledge capital intensity is defined as the ratio of knowledge capital, defined as in Peters and Taylor (2017), to total assets. Patent intensity is measured as the ratio of a firm's accumulated patent-related private economic value to its total assets. Patent values are obtained from Kogan et al. (2017), aggregated annually, and deflated to 1982 dollars using the Consumer Price Index (CPI). Missing patent values are replaced with zero. The non-linear and non-monotonic relationship revealed in this plot suggests that the connection between R&D investments and patenting outcomes is complex and not straightforward.

**Table 2: Descriptive Statistics by R&D Reporting Status**
This table compares firms that signed a letter opposing constraints to high-skilled immigration with non-signatories. The letter, addressed to President Trump and cabinet secretaries in May 2020, was signed by 324 employers and organizations across various sectors of the American economy, including 45 publicly traded firms. R&D, Assets, PPE (Property, Plant, and Equipment), and PYVA (Patent-related Yearly Value Added) are reported in millions of dollars. KKR (Knowledge Capital Risk) and Skill are reported as ratios. PYVA is defined following Kogan and Papanikolaou (2019), while Skill is an industry-level variable retrieved from Belo et al. (2017). The stark differences between signatories and non-signatories, particularly in R&D expenditure and patent revenue, suggest a non-linear relationship between KKR and R&D expenditure.

| Signed | Statistic | R&D | Skill | Assets | PPE | PYVA |
|---|---|---|---|---|---|---|
| Unsigned (0) | median | 23 | 0.292 | 1313 | 344 | 0 |
| | mean | 162 | 0.324 | 11912 | 3617 | 219 |
| | std | 737 | 0.179 | 79893 | 17332 | 2084 |
| Signed (1) | median | 881 | 0.550 | 9049 | 1875 | 426 |
| | mean | 2686 | 0.561 | 30435 | 9990 | 5709 |
| | std | 5058 | 0.155 | 52368 | 22473 | 15730 |

Measuring KKR presents challenges that go beyond quantifying the level of knowledge capital itself. The core of KKR lies in assessing the probability of a firm achieving groundbreaking innovations or, conversely, falling behind its competitors in the race for technological advancement. Therefore, accurately measuring KKR requires insight into the current state and potential of a firm's ongoing research projects, as well as those of its competitorsinformation that is typically highly confidential and closely guarded.

Given these challenges in accessing and evaluating proprietary information, an empirical approach to measuring KKR becomes necessary. This paper adopts such an approach by relying on firms' self-reported risk factors disclosed in their 10-K filings. These disclosures, while not perfect, provide insights into how firms perceive and communicate their innovation-related risks to investors. This method allows for a more nuanced understanding of KKR that incorporates firms' assessments of their innovative capabilities, competitive landscape, and potential vulnerabilities in their research and development efforts.

# 3 Measuring Knowledge Capital Risk: Data and Methodology

I utilize 10-K filings from 2006 to 2022 to construct a text-based measure of KKR. The methodology employs OpenAI's pre-trained text embeddings to analyze the Risk Factors section (Item 1A) of 10-K reports. KKR is quantified using cosine similarity between document embeddings and key phrases related to research and development. This approach captures firm-specific factors in innovation-related risk, addressing inconsistencies in R&D reporting and accounting for varying efficacy in converting R&D into innovations.

## 3.1 Data

Equity return data is obtained from the Center for Research in Security Prices (CRSP), from January 2006 to December 2020. Equities priced below $5 (based on June average prices) are excluded from the analysis. The Consumer Price Index for All Urban Consumers (CPI-U), sourced from the Federal Reserve Economic Data (FRED) database, is used for inflation adjustment. Equity returns are collected at daily, weekly, and monthly frequencies. Only firms filtered according to criteria outlined in Section 3.1 are kept. Fama-French factors (SMB, HML), risk-free rates (one-month Treasury bill rate from Ibbotson Associates), and excess market returns are sourced from Kenneth French's website. Unless stated otherwise, equity return frequency is weekly.

Firm-level accounting data is retrieved from CRSP-Compustat's Fundamentals Annual database. The CRSP-Compustat Merged Database link table is used to merge databases with different identifiers (e.g., CRSP PERMNO, GVKEY). Market capitalization is computed as the product of Compustat's *csho* (Common Shares Outstanding) and *prcc_f* (Price Close - Annual - Fiscal). Market-to-book ratios are computed as the ratio of market capitalization and *ceq* (Book Equity).

I utilize data from 10-K filings to construct my HKR measure. I scrape 121,839 annual reports (10-Ks) from U.S. publicly listed firms from 2006 to 2022, obtained from the Securities and Exchange Commission's EDGAR (Electronic Data Gathering, Analysis, and Retrieval) system. This electronic database is used for the submission and public disclosure of financial statements and other reports by American publicly listed companies.

A 10-K provides a comprehensive overview of a company's financial performance and operations over a year. U.S. laws and regulations strictly prohibit false or misleading statements in 10-Ks to ensure transparency and accuracy. The Sarbanes-Oxley Act further

enhances corporate accountability by requiring Chief Financial Officers (CFOs) and Chief Executive Officers (CEOs) to certify the accuracy of 10-Ks.

Within the 10-K, Item 1A ("Risk Factors") mandates registrants to disclose, in plain English, risk factors that make investments in the firm speculative or risky. This section must be logically organized, with adequately described risks and explanations of how they affect the registrant. Generic risks applicable to any company are discouraged (17 CFR 229.105). Additionally, norms governing text structure, writing style, and readability ensure that risk factors are communicated effectively to all stakeholders, including those without specialized financial or legal knowledge (17 CFR 230.421).

Following Golubov and Konstantinidi (2019); Stambaugh and Yuan (2016), I filter firms by considering only ordinary common shares (CRSP share codes 10 and 11) traded on NYSE, AMEX, and NASDAQ exchanges (CRSP exchange codes 1, 2, and 3). Additionally, I exclude firms whose prices are less than $5 in 2016 dollars to avoid micro-structure effects.

While reporting risk factors is mandatory for most firms, the SEC allows asset-backed issuers and smaller reporting companies to omit them. Asset-backed issuers are entities whose reporting obligation stems from registering an offering of asset-backed securities under the Securities Act or registering a class of asset-backed securities. A company qualifies as a "smaller reporting company" if it has a public float of less than 250 million dollars, or if it has less than 150 million dollars in annual revenues, and either no public float or a public float of less than 700 million dollars.

Therefore, non-reporting firms need to be excluded from the corpus. These firms typically either leave Item 1A empty or include a brief placeholder text stating their exemption from disclosure requirements. This practice results in an unusual distribution of word counts in 10-K filings, as illustrated in Figure 18. Panel (a) shows the overall cumulative distribution of word counts, while panel (b) provides a zoomed-in view of the lower word count range.

The sharp initial rise in the cumulative distribution seen in panel (b) indicates a high frequency of filings with very low word counts, likely corresponding to these non-reporting firms. To address this, I implement a word count threshold in subsequent analysis stages. Specifically, Item 1A texts containing fewer than 200 words are excluded from further consideration. This cutoff, visible in panel (b), effectively removes the majority of placeholder texts and empty filings, ensuring that my analysis focuses on substantive risk factor disclosures from reporting firms.

Table 12 shows the total count of filtered firms by year. A total of 40,174 firm-year combinations are represented in the final topic map.

## 3.2 Methodology

I transform each text in the 1A corpus into high-dimensional vector representations, known as text embeddings. This process utilizes OpenAI's pre-trained `text-embedding-3-small` model, a state-of-the-art transformer-based architecture accessed via OpenAI's API.

Text embeddings are dense vector representations that encapsulate the semantic and contextual information of words and phrases within a continuous vector space. These embeddings result from training large language models on extensive text corpora, enabling the capture of rich, nuanced linguistic representations. In this implementation, each embedding is represented as a sequence of 1536 floating-point numbers, providing a compact yet informative representation of the input text. The distance between two vectors measures text relatedness (Kusupati et al., 2022)).

Each text in the corpus is broken down into individual tokens by the `cl100k_base` tokenizer. Each token typically corresponds to a word or a part of a word, depending on the complexity and structure of the language used. OpenAI's API imposes a limit of 8192 tokens (approximately ten pages of text) per input. Consequently, any text exceeding this limit is truncated.

To quantify KKR, I compute the cosine similarity between each document's embedding and the embeddings of key phrases associated with research and development. This approach yields a ranking equivalent to Euclidean distance but offers superior computational efficiency. The key phrases I focus on are "intellectual property," "patents," "knowledge capital," and "research and development." This method allows for a nuanced assessment of how closely each document aligns with concepts relevant to a company's innovative capacity and potential future value.

For my primary analysis, I define KKR as the cosine similarity between each document's embedding and the embedding of the term "intellectual property." To ensure the robustness of the results, I later conduct sensitivity analyses using the alternative terms mentioned above. Results are shown in the Appendix.

Alternative measures for measuring KKR using topic models, such as LDA (Latent Dirichlet Allocation) and HDP (Hierarchical Dirichlet Allocation), are described in Appendixes B and C.

## 3.3 Validation

The text-based KKR measure exhibits moderate to strong positive correlations with existing innovation proxies, validating its relevance while suggesting it captures additional dimensions of knowledge-related risk. KKR displays substantial cross-sectional varia-

tion across firms and industries, with sectors like Health and Chemicals tending to have higher KKR values.

### 3.3.1 Comparison with Existing Innovation Measures

Table 3 presents the correlation matrix between KKR and two other proxies for knowledge capital risk: Knowledge Capital Intensity (KK Int.) and Patent Intensity (Pat. Int.)

**Table 3: Spearman (rank) correlation between different metrics of innovation intensity: textual KKR, knowledge capital intensity, patent intensity, and industry's skill.**
Knowledge capital intensity is defined as the ratio of knowledge capital to assets. Knowledge capital is defined as in Peters and Taylor (2017). Patent intensity is defined similarly. Firm patent value is retrieved from Kogan et al. (2017). Skill is defined as in Belo et al. (2017).

| | KK Int. | Pat.Int. | Skill |
|---|---|---|---|
| Pat.Int. | 0.661 | | |
| Skill | 0.331 | 0.193 | |
| KKR | 0.598 | 0.440 | 0.236 |

Industry-specific labor skill data is sourced from Frederico Belo's website, following the definition provided by Belo et al. (2017). In their study, the skill level within an industry is quantified as the proportion of employees engaged in occupations necessitating extensive training and preparation, defined as occupations with a Specific Vocational Preparation (SVP) level of 7 or higher. Industries are categorized using 4-digit NAICS codes. For the years post-2013, the established labor skill level as of 2013 is applied.

KKR shows a strong, yet imperfect, positive correlation with Knowledge Capital Intensity (0.598) and a moderate positive correlation with Patent Intensity (0.440), which validates the measure and also suggests that it captures additional dimensions of knowledge-related risk.

### 3.3.2 Cross-sectional properties of knowledge capital risk (KKR)

Table 4 presents summary statistics for KKR alongside other key firm characteristics and innovation measures. KKR exhibits a mean of 0.23 and a standard deviation of 0.052, indicating substantial cross-sectional variation. Notably, the distribution of KKR is relatively symmetric, in contrast to the measures of patent intensity and knowledge capital intensity, which display strong left-skew.

**Table 4: Descriptive Statistics of Firm Characteristics and Innovation Measures**
This table reports summary statistics for key firm characteristics and innovation measures. Total Assets and Gross PPE are in millions of dollars. Intangible Capital (PT) and Knowledge Capital (PT) are calculated following Peters and Taylor (2017) and reported in millions of dollars. Knowledge Capital (PT) Intensity is the ratio of Knowledge Capital (PT) to Total Assets. KKR is the knowledge capital risk measure derived from textual analysis of 10-K filings. Skill is the industry-level measure of labor skill from Belo et al. (2017). Patent Intensity is the ratio of accumulated patent value to total assets. The sample period is 2006–2020.

| | Count | Mean | SD | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| Total Assets | 34763 | 12133 | 79647.6 | 0.4 | 393.1 | 1336.1 | 4773.3 | 3.38607e+06 |
| Gross PPE | 29899 | 3705.04 | 17429.2 | 0 | 71.984 | 353.198 | 1589.91 | 511400 |
| Intangible Cap. (PT) | 27618 | 2649.02 | 11227.8 | 0 | 89.7768 | 329.601 | 1322.55 | 278772 |
| Knowledge Cap. (PT) | 27618 | 425.674 | 3003.95 | 0 | 0 | 0 | 92.5669 | 78910.2 |
| Knowledge Cap. (PT) Intensity | 27616 | 0.156247 | 0.496888 | 0 | 0 | 0 | 0.131775 | 35.9782 |
| KKR | 40528 | 0.23 | 0.052 | 0.04 | 0.192 | 0.229 | 0.264 | 0.51 |
| Skill | 39287 | 0.327 | 0.18 | 0.008 | 0.199 | 0.294 | 0.439 | 0.765 |
| Patent Intensity | 34763 | 0.36 | 1.689 | 0 | 0 | 0 | 0.082 | 116.528 |

Figure 3 displays the distribution of KKR across the entire set of firms, showing a roughly normal unconditional distribution with a slight right skew.

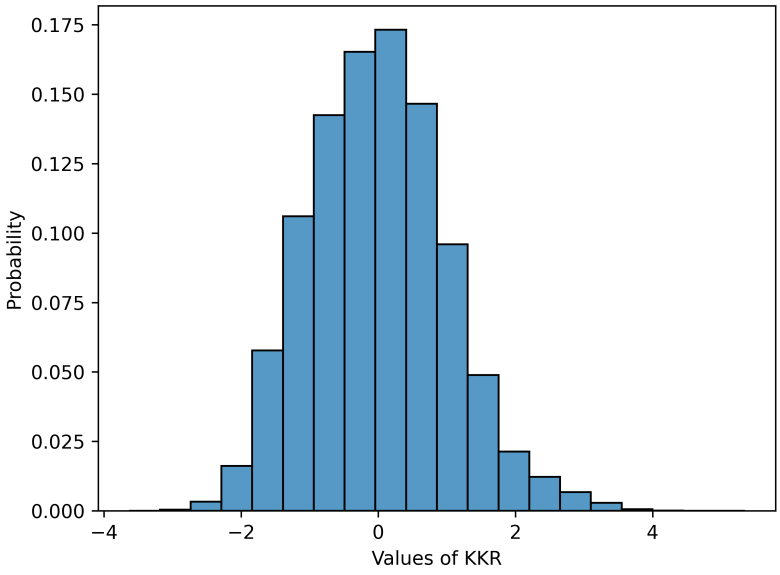



**Figure 3: Distribution of knowledge capital risk (KKR) across the entire set of firms**
This histogram shows the distribution of the KKR measure across all firms in the sample. The KKR measure is derived from textual analysis of firms' 10-K filings using cosine similarity to R&D-related terms. The distribution is approximately normal with a slight right skew.

When examining the distribution of KKR across different industries, as illustrated in

Figure 4, most industries exhibit unimodal distributions but with distinct centers and spreads. Notably, sectors such as Health (Hlth) and Chemicals (Chems) tend to have higher KKR values, with their distributions shifted slightly to the right. This suggests these industries may face greater risks associated with their knowledge capital. In contrast, the Utilities (Utils) and Energy (Enrgy) sectors display narrower distributions, indicating more homogeneity in KKR within these industries.

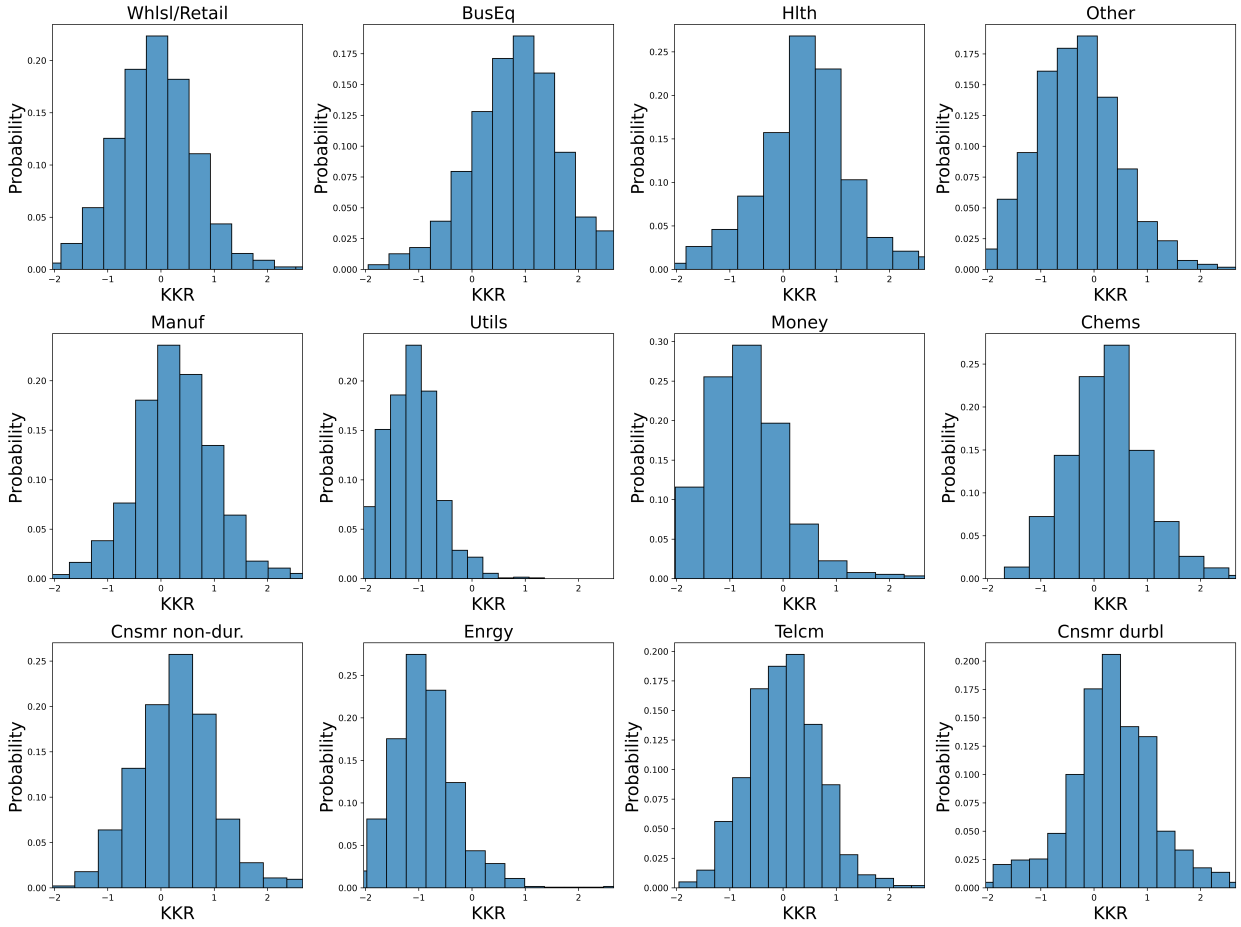
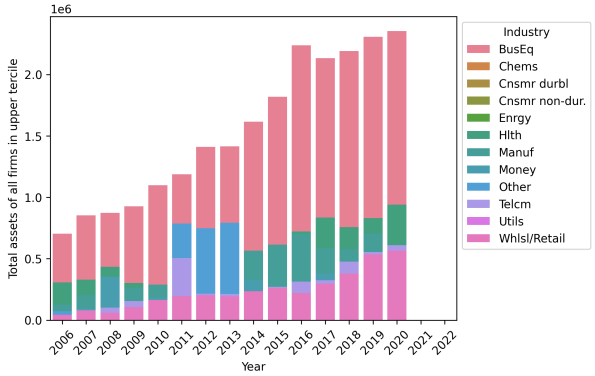
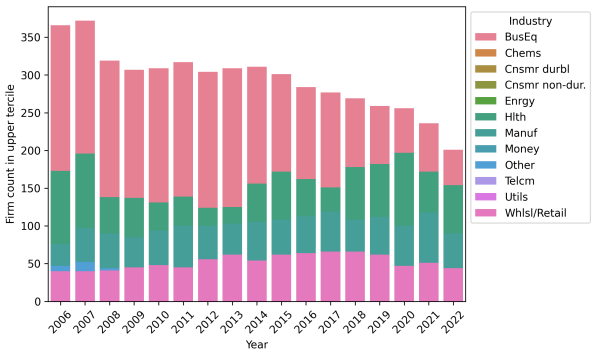



**Figure 4: Distribution of KKR by industry**
This figure presents the distribution of Knowledge Capital Risk (KKR) across industries, as defined by Fama and French's 12-industry classification. The histograms reveal distinct patterns of KKR across different sectors, with industries such as Health and Chemicals exhibiting higher KKR values on average. The variation in distributions highlights the industry-specific nature of knowledge capital risk and its potential importance in cross-sectional asset pricing.

Figure 6 presents the evolution of value-weighted KKR over time across firms. The data reveals a steady increase in knowledge capital risk over the measurement period. The upward momentum intensified during 2018-2020, suggesting firms' growing exposure to knowledge-related risks in recent years.




**(a)** Total assets of firms in the upper tercile of KKR.




**(b)** Number of firms in the upper tercile of KKR.

**Figure 5: Sector composition of high-KKR firms**
This figure depicts the sector distribution of firms in the upper tercile of KKR. The contrast between asset concentration and firm count reveals distinct patterns across industries. Technology and healthcare sectors dominate in both metrics, reflecting their R&D-intensive nature. Manufacturing shows a larger share in assets than firm count, suggesting larger firms in this sector tend to have higher KKR. Conversely, consumer-related industries exhibit a smaller presence, indicating lower knowledge capital intensity.

# 4 Pricing knowledge capital risk

This section demonstrates the significant impact of KKR on equity valuations and risk patterns. Portfolios formed on high KKR consistently outperform those with low KKR over a 14-year period, indicating a substantial risk premium. The HKR factor exhibits significant explanatory power for cross-sectional equity returns, even after controlling for established Fama-French factors. Importantly, the HKR factor satisfies key constraints of the Intertemporal Capital Asset Pricing Model (ICAPM) and displays the highest Sharpe ratio among compared factors.

## 4.1 Risk and return characteristics of KKR-based portfolios

To examine the risk and return characteristics of portfolios formed on KKR, I construct decile portfolios based on firms' KKR scores. Figure 7 illustrates the cumulative asset-weighted weekly returns for the highest and lowest KKR decile portfolios from 2006 to 2020. Both portfolios exhibit an overall upward trend over the 14-year period, indicating positive returns across the KKR spectrum. However, the highest KKR decile consistently outperforms the lowest decile, with the performance gap widening significantly over time. This suggests a substantial risk premium associated with higher levels of KKR.

The portfolios' behavior during market stress periods is particularly informative. Dur-

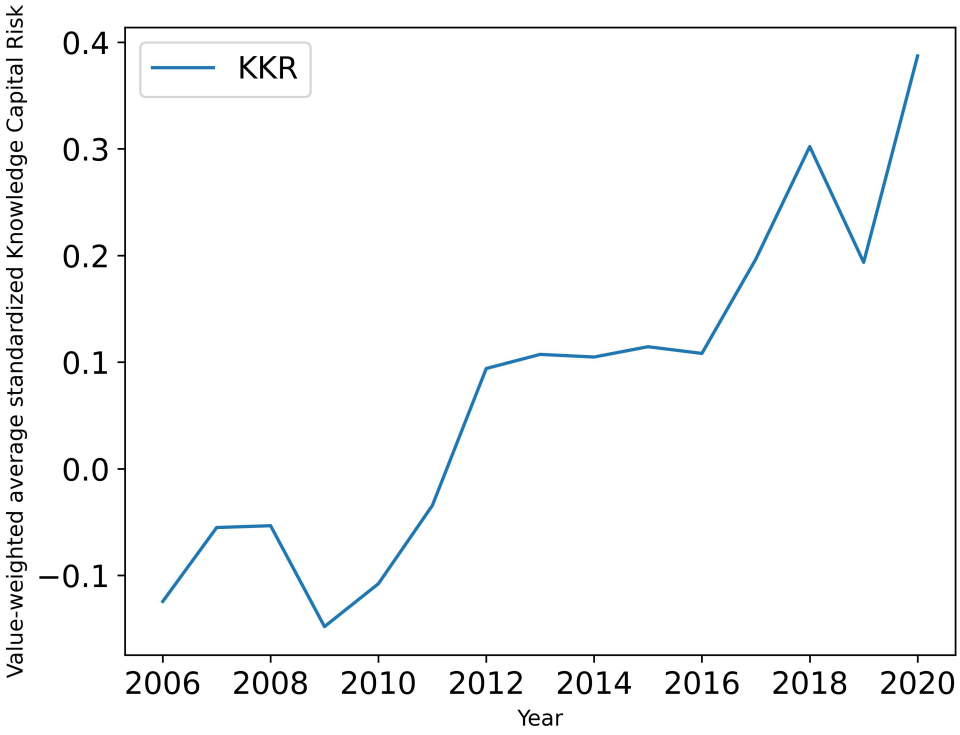



**Figure 6: Mean annual knowledge capital risk**
This figure displays the value-weighted average standardized KKR across all firms in the sample from 2006 to 2020. The figure shows a consistent upward trend over the sample period, starting from negative values in 2006 and rising to positive values after 2012, with particularly sharp increases in 2018-2020, indicating a steady growth in aggregate knowledge capital risk exposure over time.

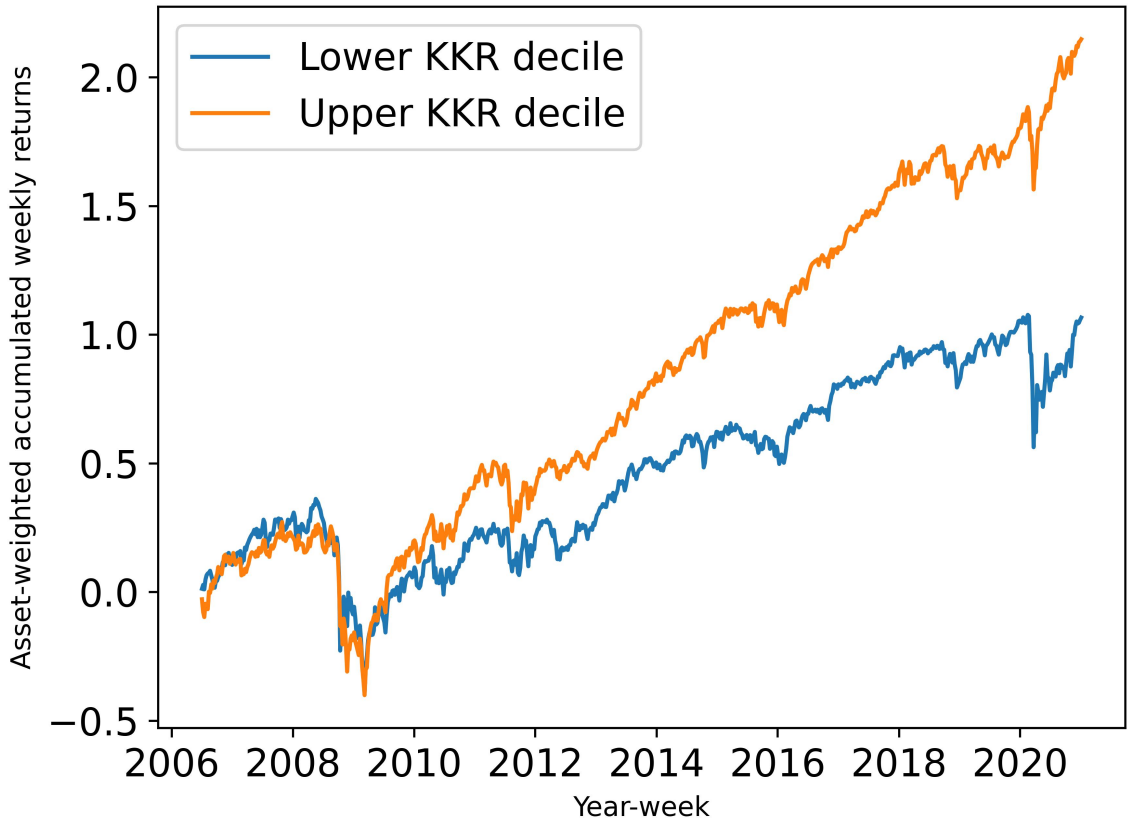



**Figure 7: Cumulative returns of extreme KKR portfolios**
This figure displays the cumulative asset-weighted weekly returns for portfolios formed from the lowest (1st decile) and highest (10th decile) KKR equities from 2006 to 2020. The high-KKR portfolio consistently outperforms the low-KKR portfolio, with the performance gap widening over time. This pattern suggests a substantial risk premium associated with knowledge capital intensity. The divergence is particularly pronounced following market stress periods, such as the 2008 financial crisis, indicating that high-KKR firms may offer greater potential for recovery and growth in subsequent periods despite higher volatility during downturns.

ing the 2008 financial crisis, both portfolios experienced sharp declines. However, the high-KKR portfolio recovered more quickly and rebounded more dramatically in subsequent years. This pattern suggests that while high-KKR equities may be more volatile during downturns, they also offer greater potential for recovery and growth in subsequent periods.

Figure 8 depicts the weekly return volatility for the highest and lowest KKR decile portfolios. The high-KKR portfolio exhibits consistently higher volatility throughout most of the sample period, aligning with the expectation that firms with higher KKR tend to have more volatile returns. Both portfolios show increased volatility during known market stress events, with the most notable spike occurring during the 2008-2009 global financial crisis. The volatility increase is more pronounced for the high-KKR portfolio.

While the high-KKR portfolio generally maintains higher volatility, the gap between the two portfolios varies over time. Towards the end of the sample period in 2020, coinciding with the COVID-19 pandemic, both portfolios experience a sharp increase in volatility. Interestingly, during this period, the low-KKR portfolio briefly shows higher volatility than the high-KKR portfolio. This may be linked to a higher perceived value of innovation and R&D during the pandemic, particularly among pharmaceutical and technology firms, providing a stabilizing effect on high knowledge capital equities.

Figure 9 illustrates the evolution of kurtosis in equity returns over time for the highest and lowest KKR decile portfolios. Throughout the observed period, the high-KKR portfolio consistently exhibits higher kurtosis compared to the low-KKR portfolio, with few exceptions. This pattern suggests that equities with higher KKR tend to have more extreme return distributions, characterized by heavier tails and more frequent outliers. These characteristics are consistent with the nature of R&D-intensive firms, which often engage in research activities with highly uncertain outcomes, potentially leading to more frequent ”breakthrough” or ”failure” events.

The low-KKR portfolio maintains a relatively stable kurtosis range between 4 and 10, while the high-KKR portfolio exhibits much wider fluctuations, ranging from about 6 to 18. This stability in low-KKR equities suggests that traditional, less R&D-intensive sectors of the economy provide a more predictable investment environment. The convergence of kurtosis levels towards the end of 2020 likely reflects the broad economic impact of the COVID-19 pandemic, which temporarily reduced the distinctions between high and low-KKR firms as all sectors faced unprecedented challenges and uncertainties.

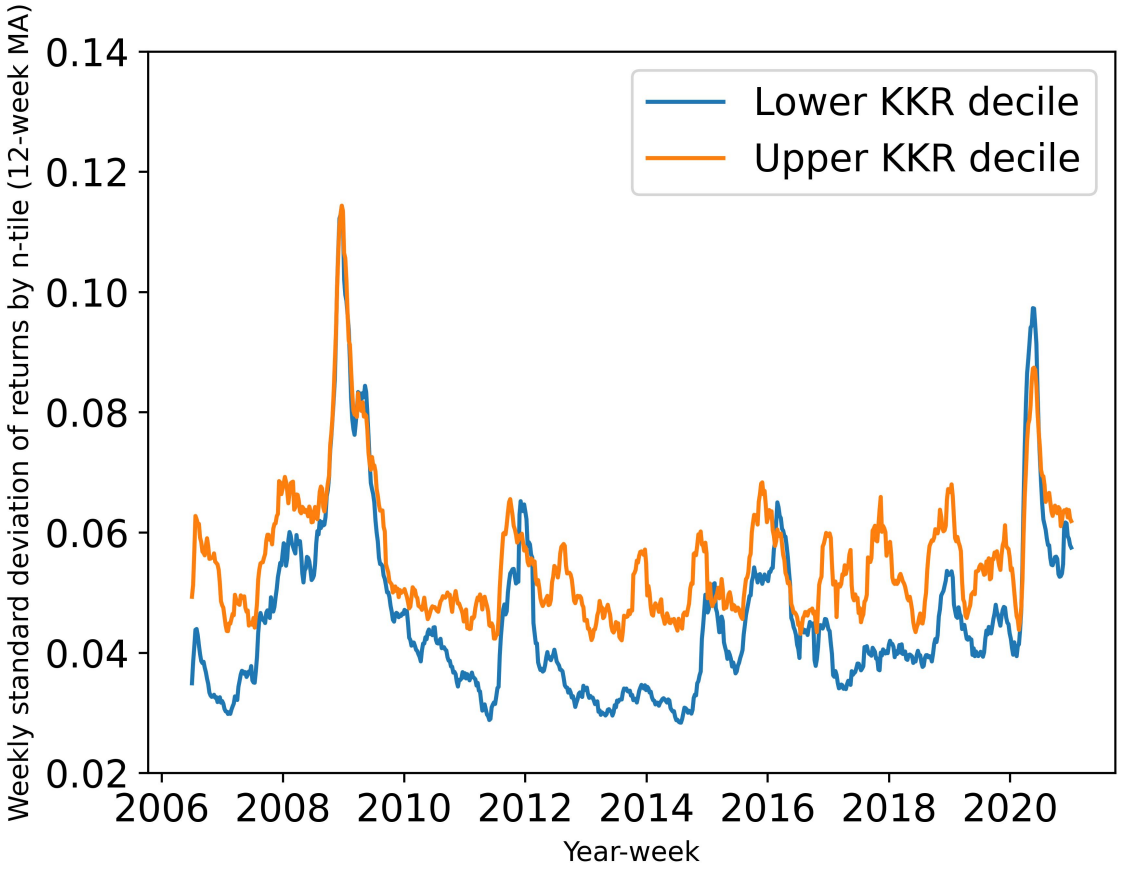
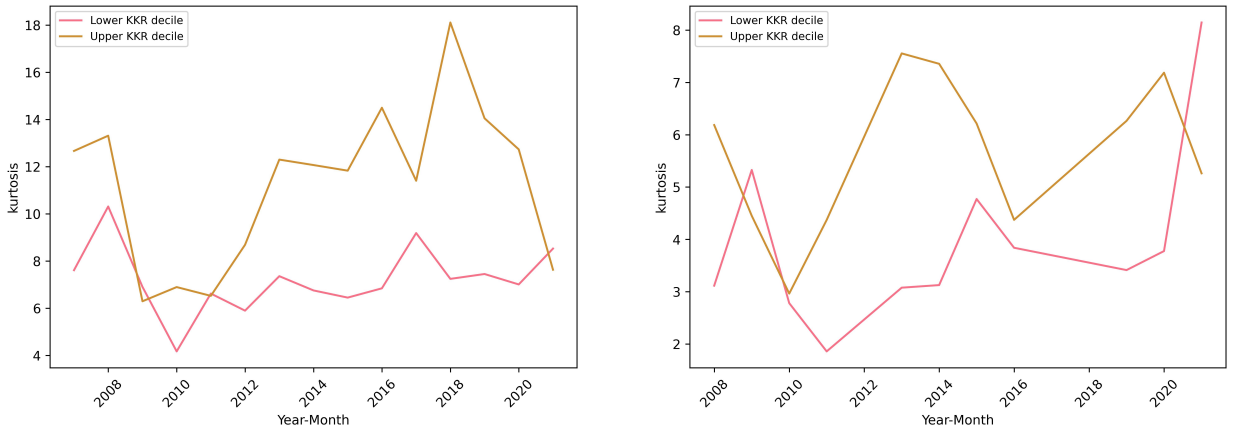



**Figure 8: Weekly return volatility of extreme KKR portfolios**
This figure illustrates the 12-week moving average of weekly return volatility for portfolios formed from the lowest (1st decile) and highest (10th decile) KKR equities from 2006 to 2020. The high-KKR portfolio exhibits consistently higher volatility throughout most of the sample period, aligning with the expectation that firms with higher KKR tend to have more volatile returns. Both portfolios show increased volatility during known market stress events, with the most notable spike occurring during the 2008-2009 global financial crisis. The volatility gap between high and low-KKR portfolios varies over time, narrowing significantly during the 2020 COVID-19 pandemic, potentially reflecting a shift in the perceived value of innovation during this period.




**(a)** Non-Asset-Weighted Kurtosis

**(b)** Asset-Weighted Kurtosis

**Figure 9: Kurtosis of Returns for Extreme Deciles of KKR**
Annual kurtosis is calculated as the mean of firm-level daily return kurtoses within each calendar year. The figures illustrate the consistently higher and more volatile kurtosis of high-KKR equities, indicating more frequent extreme returns for innovation-intensive firms. The left panel shows non-asset-weighted kurtosis; the right panel shows asset-weighted kurtosis.

These findings have important implications for asset pricing and portfolio management. The persistent differences in return distributions between high and low-KKR equities suggest that traditional risk measures based on normal distributions may be inadequate for accurately pricing assets in knowledge-intensive sectors. Moreover, the higher returns and volatility associated with high-KKR firms indicate a risk-return tradeoff that investors must carefully consider when constructing portfolios.

## 4.2 Defining the high-knowledge factor (HKR)

The high-knowledge factor (HKR) is constructed as the difference between the returns of high and low-knowledge capital portfolios, controlling for size effects. Following the methodology of Fama and French (2015) and Leung et al. (2020), I use the NYSE median market capitalization as the annual size breakpoint. For each week $t$, the factor is formally defined as:

$$HKR_t = \frac{(SH_{KK,t} + BH_{KK,t})}{2} - \frac{(SL_{KK,t} + BL_{KK,t})}{2} \tag{2}$$

where $SH_{KK}$ and $BH_{KK}$ represent the market-weighted returns of small and big stock portfolios with high knowledge capital (top tercile), respectively. Similarly, $SL_{KK}$ and $BL_{KK}$ denote the market-weighted returns of small and big stock portfolios with low knowledge capital (bottom tercile)

Notably, HKR boasts the highest Sharpe ratio of all factors presented, as shown in Table 5. HKR exhibits the second-highest mean monthly return among all factors, surpassed only by the market factor. This indicates that HKR offers the best risk-adjusted return, potentially providing efficient exposure to a unique risk premium. This suggests that firms with higher knowledge capital tend to outperform those with lower knowledge capital on average, which is also evident from Figure 7. HKR is one of the few factors showing a positive central tendency, indicating consistent positive performance.

The HKR factor also shows considerable variability, with a standard deviation of 8.45%, reflecting the dynamic nature of returns to knowledge capital. This volatility is lower than that of the market factor (19.02%) but higher than most other factors, suggesting that while HKR captures significant return potential, it also carries substantial risk. The positive median (50th percentile) return of 6.2% for HKR does not differ significantly from its mean, indicating a relatively symmetric distribution of returns. [1]

**Table 5: Summary Statistics of Factor Returns.**
This table presents key statistics for various asset pricing factors, including the market factor (Mkt.RF), size (SMB), value (HML), investment (CMA), profitability (RMW), and the high knowledge factor (HKR). Values are converted to yearly frequency.

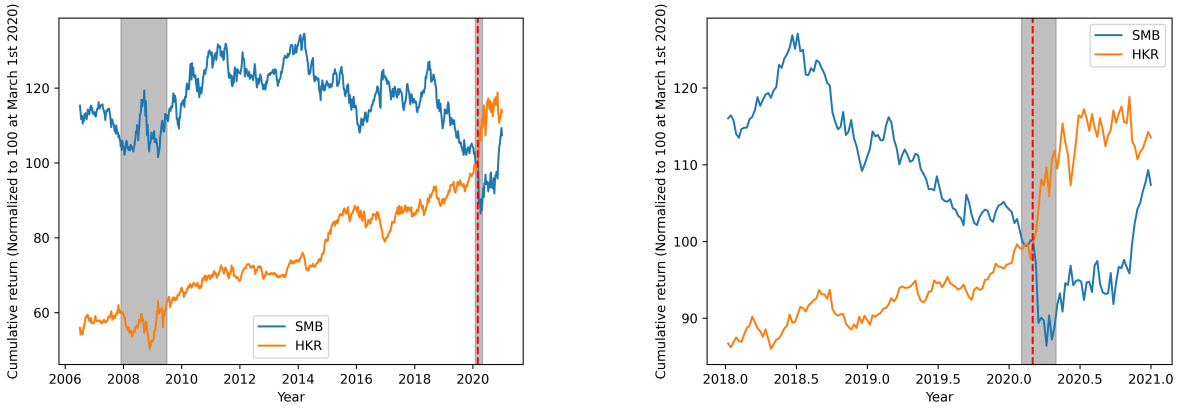
| Factor | Count | Mean | SD | Sharpe | 25% | 50% | 75% |
|---|---|---|---|---|---|---|---|
| HKR | 757 | 5.19% | 8.45% | 0.61 | -22.3% | 6.2% | 45.6% |
| Mkt.RF | 757 | 9.39% | 19.02% | 0.49 | -36.5% | 17.8% | 125.5% |
| SMB | 757 | -0.15% | 9.37% | -0.02 | -33.4% | 0.8% | 47.4% |
| HML | 757 | -5.00% | 12.42% | -0.40 | -34.5% | -7.4% | 36.5% |
| RMW | 757 | 2.41% | 6.04% | 0.40 | -20.1% | 0.6% | 30.3% |
| CMA | 757 | -0.51% | 4.90% | -0.10 | -21.3% | -1.6% | 22.4% |

Figure 10 shows the performance of the HKR factor over time, with the SMB factor added for comparison. The graph illustrates the divergent performance of these factors, particularly in the post-COVID period, with HKR showing a notable outperformance.

## 4.3 Portfolio Return Predictability Tests

To assess the predictive power of the HKR factor, I construct 27 portfolios ($3 \times 3 \times 3$) by independently sorting equities into Size (market capitalization) terciles and Value (book-

[1] This is confirmed by its skewness plot in Figure 19.




**(a)** Full range: 2006-2021

**(b)** Zoomed in: 2018-2021

**Figure 10: Comparative performance of SMB and HKR factors**
This figure presents the cumulative returns of the Fama-French Size factor (SMB) and the High-Knowledge Risk (HKR) factor. Panel (a) shows the full sample period from 2006 to 2021, while panel (b) focuses on the period from 2018 to 2021. The red dashed line marks the onset of the COVID-19 pandemic in March 2020. U.S. recessions are shaded. The graph illustrates the divergent performance of these factors, particularly in the post-COVID period. The HKR factor demonstrates notable outperformance relative to SMB, especially after March 2020, suggesting a potential shift in the market's valuation of knowledge-intensive firms during the pandemic and its aftermath.

to-market ratio) terciles, then sorting into KKR terciles within each of the resulting 9 Size-Value bins. This approach helps mitigate potential correlations between KKR and Size or Value, which could lead to poorly diversified portfolios or portfolios with extreme characteristic tilts. Portfolio assignments are updated annually at the end of June, remaining fixed for the subsequent 12 months until the next rebalancing.

I employ five-factor Fama-MacBeth regressions to analyze the predictive power of the HKR factor, alongside the market factor (MktRF) and established Fama-French factors, on portfolio weekly excess returns. The analysis uses a two-stage approach with three model specifications of increasing complexity.

In the first stage, I estimate time-varying betas through rolling time series regressions. A 2-year (104-week) rolling window balances the incorporation of recent market dynamics with statistical robustness. Within each rolling window $w$, weekly excess returns $R_t^{ei}$ for each portfolio $i$ are regressed against the set of Fama-French factors $F_t$, augmented by the $HKR_t$ factor. Heteroskedasticity-robust White covariance estimation is used.

$$R_t^{ei} = \alpha_{it} + B_{it}' F_t + \epsilon_{iwt} \tag{3}$$

The second stage involves a cross-sectional regression each period of returns on esti-

mated betas:

$$R_t^{ei} = \alpha_t + \widehat{B}_{it}' \lambda_t \tag{4}$$

The factor risk premium and pricing error estimates are then computed as simple time averages:

$$\widehat{\lambda} = \frac{1}{T} \sum_{i=1}^{T} \widehat{\lambda_t}, \widehat{\alpha} = \frac{1}{T} \sum_{i=1}^{T} \widehat{\alpha_t} \tag{5}$$

The Fama-MacBeth regression results, presented in Table 7, provide evidence for the significance of the HKR factor in predicting portfolio returns across the cross-section. Remarkably, this predictive power persists even after controlling for the established Fama-French 3 and 5 factors, suggesting that HKR captures a unique dimension of the R&D anomaly not fully accounted for by traditional asset pricing models.

A closer examination of the portfolio characteristics, detailed in Table 6, reveals that time-averaged $\beta_{HKR}$ exhibits a consistent transition from negative values in low KKR terciles to positive values in high-KKR terciles, a relationship that holds across all Size and Value categories. The impact of KKR on portfolio returns is further illustrated in Figure 11. As the KKR factor increases from 0 to 2, a general trend of increasing returns is observed across most categories.

**Table 6: Alphas, $\beta_{HKR}$, and weekly excess returns for equities sorted into 27 portfolios.**
Portfolios are based on three dimensions: Size (Small, Medium, Big), Value (Low, Medium, High), and Knowledge Capital Tercile (0, 1, 2) within each Size-Value bin. Returns are weekly. Alphas and $\beta_{HKR}$ are calculated for the extended three-factor Fama-French model, which includes excess market return, SMB, HML, and HKR.

| Portfolio | $\alpha$ | t($\alpha$) | $\beta_{\text{HKR}}$ | t($\beta_{\text{HKR}}$) | RET |
|---|---|---|---|---|---|
| S/L/0 | 0.0% | 0.598 | -0.306 | -32.921 | 0.08% |
| S/M/0 | 0.04% | 9.037 | -0.319 | -42.910 | 0.15% |
| S/H/0 | 0.09% | 16.508 | -0.283 | -27.917 | 0.25% |
| M/L/0 | -0.04% | -10.367 | -0.728 | -64.600 | 0.04% |
| M/M/0 | 0.06% | 23.401 | -0.412 | -126.264 | 0.18% |
| M/H/0 | 0.18% | 49.716 | -0.304 | -55.644 | 0.42% |
| B/L/0 | -0.06% | -13.628 | -0.418 | -32.354 | 0.04% |
| B/M/0 | 0.03% | 8.410 | -0.781 | -136.613 | 0.11% |
| B/H/0 | 0.09% | 39.928 | -0.327 | -65.024 | 0.3% |
| S/L/1 | -0.01% | -3.786 | -0.289 | -26.284 | 0.04% |
| S/M/1 | 0.0% | 1.208 | -0.084 | -8.383 | 0.17% |
| S/H/1 | 0.08% | 12.173 | -0.026 | -2.258 | 0.27% |
| M/L/1 | -0.08% | -23.847 | -0.252 | -35.379 | 0.06% |
| M/M/1 | 0.04% | 14.505 | -0.020 | -5.347 | 0.2% |
| M/H/1 | 0.15% | 75.477 | 0.228 | 53.022 | 0.42% |
| B/L/1 | 0.01% | 2.186 | -0.096 | -8.399 | 0.13% |
| B/M/1 | 0.01% | 5.471 | -0.319 | -54.150 | 0.15% |
| B/H/1 | 0.1% | 46.778 | 0.027 | 7.206 | 0.3% |
| S/L/2 | -0.11% | -22.142 | 0.130 | 8.518 | 0.09% |
| S/M/2 | -0.05% | -18.713 | 0.288 | 30.695 | 0.15% |
| S/H/2 | 0.1% | 12.269 | 0.229 | 17.918 | 0.38% |
| M/L/2 | -0.14% | -35.577 | 0.466 | 68.711 | 0.06% |
| M/M/2 | 0.0% | 1.046 | 0.562 | 54.026 | 0.2% |
| M/H/2 | 0.14% | 46.440 | 0.396 | 54.324 | 0.4% |
| B/L/2 | -0.01% | -3.493 | 0.132 | 12.107 | 0.14% |
| B/M/2 | -0.03% | -19.186 | 0.295 | 42.869 | 0.18% |
| B/H/2 | 0.1% | 48.162 | 0.574 | 101.912 | 0.35% |

To further validate the robustness of HKR's cross-sectional explanatory power, I conduct out-of-sample tests using a split-sample approach. Data is sampled weekly, with 50% of data computing HKR and 50% for portfolio construction. The incremental effect of incorporating HKR as a factor is evaluated across three model specifications: the market

| | L | M | H |
|---|---|---|---|
| S | 0.08% | 0.15% | 0.25% |
| M | 0.04% | 0.18% | 0.42% |
| B | 0.04% | 0.11% | 0.30% |

**(a)** KKR = 0 (Low)

| | L | M | H |
|---|---|---|---|
| S | 0.04% | 0.17% | 0.27% |
| M | 0.06% | 0.20% | 0.42% |
| B | 0.13% | 0.15% | 0.30% |

**(b)** KKR = 1 (Medium)

| | L | M | H |
|---|---|---|---|
| S | 0.09% | 0.15% | 0.38% |
| M | 0.06% | 0.20% | 0.40% |
| B | 0.14% | 0.18% | 0.35% |

**(c)** KKR = 2 (High)

**Figure 11: Returns by Size, Value, and KKR**
This table presents average returns for portfolios formed on Size (S: Small, M: Medium, B: Big), Value (L: Low, M: Medium, H: High), and KKR (0: Low, 1: Medium, 2: High) terciles.

**Table 7: Fama-MacBeth Regressions of Portfolio Weekly Excess Returns**
This table presents the results of Fama-MacBeth regressions estimating the pricing of the HKR factor alongside traditional risk factors. The dependent variable is the weekly excess return of 27 portfolios sorted on size, value, and KKR. The independent variables include the market factor (MktRF), size factor (SMB), value factor (HML), profitability factor (RMW), and investment factor (CMA). Columns (1), (2), and (3) show results for the market model, Fama-French three-factor model, and Fama-French five-factor model, respectively. Newey-West robust standard errors are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

| | Mkt+ | FF3+ | FF5+ |
|---|---|---|---|
| | (1) | (2) | (3) |
| HKR | 0.00164*** | 0.00141*** | 0.00127** |
| | (0.00057) | (0.00049) | (0.00049) |
| MktRF | -0.00093 | 0.00083 | 0.00015 |
| | (0.00173) | (0.00135) | (0.00131) |
| SMB | | 0.00024 | -0.00017 |
| | | (0.00051) | (0.00049) |
| HML | | -0.00280*** | -0.00269*** |
| | | (0.00071) | (0.00073) |
| RMW | | | -0.00008 |
| | | | (0.00048) |
| CMA | | | -0.00147*** |
| | | | (0.00044) |
| Intercept | 0.00297*** | 0.00139* | 0.00215** |
| | (0.00095) | (0.00077) | (0.00092) |
| Observations | 17658 | 17658 | 17658 |
| N. of groups | 27 | 27 | 27 |
| $R^2$ | 0.61391 | 0.76029 | 0.76052 |

model, the Fama-French three-factor model (FF3), and the Fama-French five-factor model (FF5). For each specification, predicted returns are computed and regressed against actual returns. Table 8 shows that the inclusion of HKR consistently enhances the predictive power of returns across all specifications, indicating improved model adherence. As expected, the coefficient of predicted returns on actual returns closely approximates unity and the intercept approaches zero. To account for potential heteroskedasticity and autocorrelation in the panel data, the analysis employs robust standard errors with four lags. The results demonstrate that HKR maintains its pricing power in out-of-sample tests. Additional robustness checks using different train-test splits are shown in Appendix E.

**Table 8: Out-of-Sample Performance of HKR Factor**
This table presents out-of-sample results for the High-Knowledge Risk (HKR) factor using a weekly-sampled split-sample approach. The analysis compares model performance with and without HKR across market, Fama-French three-factor (FF3), and Fama-French five-factor (FF5) specifications. Predicted returns are regressed against actual returns, with heteroskedasticity and autocorrelation robust standard errors (4 lags).

| | Mkt | Mkt+ | FF3 | FF3+ | FF5 | FF5+ |
|---|---|---|---|---|---|---|
| Intercept | -0.0005 | -0.0004 | -0.0003 | -0.0004 | -0.0002 | -0.0002 |
| | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Predicted | 1.0570 | 1.0477 | 1.0067 | 1.0056 | 1.0101 | 1.0098 |
| | (0.0097) | (0.0085) | (0.0062) | (0.0060) | (0.0056) | (0.0054) |
| R-squared | 0.7439 | 0.7769 | 0.8614 | 0.8667 | 0.8756 | 0.8801 |
| R-squared Adj. | 0.7439 | 0.7769 | 0.8614 | 0.8667 | 0.8756 | 0.8801 |

# 5 Validating HKR as a risk factor within the Intertemporal Capital Asset Pricing Model (ICAPM) Framework

In Section 4, I demonstrated that the HKR factor has significant explanatory power for the cross-section of equity returns, even after controlling for established asset pricing factors. However, as Maio and Santa-Clara (2012) argue, the ability to explain anomalous returns is not sufficient to justify a factor as a measure of priced risk within the framework of the Intertemporal Capital Asset Pricing Model (ICAPM). Following their approach, this section examines whether the HKR factor satisfies the key restrictions imposed by the ICAPM from Merton (1973). This analysis builds on the work of Leung et al. (2020), who highlight the importance of distinguishing between risk and mispricing explanations for

the R&D premium.

The ICAPM extends the consumption-based model, investors seek to maximize the expected utility of lifetime consumption and can trade continuously in time. The ICAPM generates linear models for the investor's discount factor $m_{t+1}$ in which pricing factors $f_{t+1}$ are state variables for the investor's consumption-portfolio decision (Cochrane, 2009), as shown in Equation 6. Even though the ICAPM does not identify a specific set of pricing factors $f_t$ and related state variables, state variables as in Merton (1973) should be related to changes in the investment opportunity set for the average investor.

$$m_{t+1} = a + b' f_{t+1} \tag{6}$$

In the following sections, I evaluate whether different models with HKR satisfy three necessary criteria outlined by Maio and Santa-Clara (2012) for interpretation as an ICAPM variant: (1) the ability to forecast changes in the investment opportunity set, proxied by future market returns or volatility; (2) consistency between the sign of its predictive slope and its risk price in cross-sectional tests; and (3) economically plausible estimates of relative risk aversion implied by the market price of risk. These tests aim to validate HKR as a legitimate risk factor capturing time-varying investment opportunities, distinguishing it from an *ad hoc* empirical factor merely explaining an anomaly.

## 5.1 First constraint: Forecasting power over the first or second moments of market returns

The first key constraint in Maio and Santa-Clara (2012) is that potential ICAPM state variables must demonstrate predictive power for either the mean or variance of stock market returns. To test the first constraint, I start by conducting predictive regressions for market returns. As independent variables, I use quadriweekly (four-week) factors, including HKR, SMB, and HML. The dependent variable is the aggregated market return over the subsequent q four-week periods, where $q$ ranges from 1 to 65 (representing forecasts from 4 weeks to 260 weeks ahead).

I estimate the following predictive regression:

$$R_{m,t+1:t+q} = \alpha_q + B_q F_t + \varepsilon_{t+1:t+q} \tag{7}$$

where $R_{m,t+1:t+q}$ is the cumulative market return from $t+1$ to $t+q$, $F_t$ is the set of factors, and the factors are measured at time $t$. I use robust standard errors as outlined by Newey and West (1987), with $q$ lags.

Table 9 presents the estimation results for the univariate and multivariate regressions of the HKR factor on future market returns. HKR shows significant predicting power in short horizons ($q = 1$ and $q = 3$). Significance remains only for $q = 1$ when controlling for SMB and HML.

**Table 9: ICAPM Tests: Predictive regressions for market returns**
This table reports the results for single and multiple long-horizon regressions for the monthly continuously compounded return on the value-weighted market index, at horizons of 1, 3, 13, 26, 39, 52, and 65 four-week periods ahead. The independent variable is quadriweekly factors (HKR, SMB, HML). The dependent variable is the aggregated market return $q$ four-week periods ahead. $^{*}$, $^{**}$, and $^{***}$ denote statistical significance at the 10%, 5%, and 1% levels, respectively.

**(a)** Panel A: Single predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| HKR | 0.5220* | 0.6447* | 0.6149 | 0.7599 | 0.5512 | 0.6690 | 0.8027 |
| | (0.3165) | (0.3774) | (0.6199) | (0.8020) | (0.7809) | (0.6618) | (0.6968) |
| R-squared | 0.0470 | 0.0301 | 0.0064 | 0.0051 | 0.0034 | 0.0047 | 0.0057 |
| R-squared Adj. | 0.0420 | 0.0248 | 0.0007 | -0.0010 | -0.0033 | -0.0026 | -0.0024 |

**(b)** Panel B: Multivariate predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| SMB | -0.2088 | -0.0942 | -0.2467 | -0.7720** | 0.1300 | 0.0940 | -0.0930 |
| | (0.1903) | (0.2545) | (0.4974) | (0.3798) | (0.4157) | (0.4463) | (0.5800) |
| HML | 0.1313 | -0.3371 | 0.3127 | 0.5129 | -0.1192 | 0.0358 | 0.0141 |
| | (0.2541) | (0.4091) | (0.5741) | (0.7131) | (0.6210) | (0.5877) | (0.4315) |
| HKR | 0.6260* | 0.2984 | 0.8790 | 1.2020 | 0.4526 | 0.6956 | 0.8183 |
| | (0.3585) | (0.5135) | (0.7457) | (0.8606) | (0.9567) | (0.9013) | (0.6818) |
| R-squared | 0.0538 | 0.0423 | 0.0080 | 0.0099 | 0.0036 | 0.0049 | 0.0057 |
| R-squared Adj. | 0.0384 | 0.0266 | -0.0092 | -0.0087 | -0.0168 | -0.0174 | -0.0189 |

As a variant of Maio and Santa-Clara (2012)'s first necessary condition, I test in Table 10 if the HKR factor forecasts changes in the second moment of market returns. As a proxy for market volatility, I calculate the realized stock market variance (SVAR). For each four-week period within the sample, SVAR is computed as the sum of the squared daily excess market returns. Daily excess market returns are obtained from Kenneth French's website.

**Table 10: ICAPM Tests: Predictive regressions for SVAR**
This table presents the results of single and multiple long-horizon regressions for the stock market variance (SVAR), at horizons of 1, 3, 13, 26, 39, 52, and 65 four-week periods ahead. The independent variables are quadriweekly factors (HKR, SMB, HML). SVAR is computed as the sum of squared daily excess market returns over each four-week period. $^{*}$, $^{**}$, and $^{***}$ denote statistical significance at the 10%, 5%, and 1% levels, respectively.

**(a)** Panel A: Single predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| HKR | -0.0222 | -0.0218 | -0.1515 | -0.2249 | -0.1550 | -0.2944** | -0.2785 |
| | (0.0454) | (0.0930) | (0.1603) | (0.1618) | (0.1659) | (0.1478) | (0.1720) |
| R-squared | 0.0058 | 0.0010 | 0.0053 | 0.0057 | 0.0020 | 0.0061 | 0.0046 |
| R-squared Adj. | 0.0005 | -0.0045 | -0.0004 | -0.0005 | -0.0047 | -0.0012 | -0.0035 |

**(b)** Panel B: Multivariate predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| SMB | -0.0278 | -0.0381 | 0.0734 | 0.1879* | 0.0877 | 0.2004 | 0.0080 |
| | (0.0210) | (0.0400) | (0.1558) | (0.1054) | (0.1444) | (0.1950) | (0.1346) |
| HML | -0.0598 | -0.0529 | -0.1885* | -0.4643** | -0.3624* | -0.3409* | -0.3269** |
| | (0.0420) | (0.0974) | (0.1040) | (0.2191) | (0.2040) | (0.1863) | (0.1437) |
| HKR | -0.0856* | -0.0788 | -0.3102* | -0.6130** | -0.4530* | -0.5746** | -0.5310** |
| | (0.0481) | (0.0928) | (0.1818) | (0.2570) | (0.2644) | (0.2698) | (0.2354) |
| R-squared | 0.0815 | 0.0146 | 0.0125 | 0.0266 | 0.0122 | 0.0131 | 0.0121 |
| R-squared Adj. | 0.0666 | -0.0015 | -0.0046 | 0.0083 | -0.0080 | -0.0090 | -0.0124 |

I then conduct single and multivariate long-horizon regressions on SVAR:

$$\text{SVAR}_{t,t+q} = a_q + b_q z_t + v_{t,t+q} \tag{8}$$

Above, $SVAR_{t,t+q} = SVAR_{t+1} + \cdots + SVAR_{t+q}$. In both single-factor (Panel A) and multivariate (Panel B) models, HKR consistently shows negative coefficients across all horizons. HKR shows significance with $q = 52$ in the univariate case and for almost all horizons when conditional to SMB and HML. Thus, when future investment opportunities are driven by market volatility, the expanded FF3 model meets the ICAPM criteria.

Coefficients become statistically significant at longer horizons, particularly in the multivariate model. The negative slopes observed for HKR in the predictive regressions of

market volatility are consistent with the positive risk price found in the cross-sectional regression, which provides support for the factor's relevance within the ICAPM.

## 5.2 Second constraint: Consistent sign between return forecasting and cross-sectional pricing of state variables

As the second criterion outlined in Maio and Santa-Clara (2012) for considering a multifactor model as an ICAPM variant, the sign of a state variable's forecast for changes in investment opportunities (specifically, the expected market return) should align with the sign of its innovation's pricing in cross-sectional analyses. For instance, if a particular state variable predicts an increase in expected returns, it should correspondingly command a positive risk premium in the cross-sectional evaluation of the associated multifactor model.

Examining Table 9, I observe that higher values of HKR consistently predict higher future market returns across various time horizons. The coefficients for HKR are positive for all $q$ values, with statistical significance at shorter horizons. This predictive power is further supported by the cross-sectional pricing evidence in Tables 7 and 11. In Table 7, HKR exhibits a positive and statistically significant coefficient across all three models (Mkt, FF3, and FF5). Similarly, in Table 6, HKR shows positive coefficients for both the market model and the FF3 model, with strong statistical significance in the market model.

These results demonstrate that portfolios more exposed to HKR command a higher risk premium, aligning with the positive predictive relationship observed in the ICAPM tests. The consistency between HKR's positive impact on future market returns and its positive pricing in cross-sectional tests fulfills the second constraint for interpreting this model as an ICAPM variant.

## 5.3 Third constraint: Reasonable relative risk aversion (RRA) estimates derived from market risk premium

The ICAPM's third constraint requires that the market risk premium, derived from cross-sectional analyses, should serve as a reasonable proxy for the representative investor's relative risk aversion (RRA). According to Mehra and Prescott (1985), the estimate for the RRA coefficient should fall between one and ten. To estimate the market risk premium, I employ efficient Generalized Method of Moments (GMM), following the methodology outlined by Bryzgalova (2015). The results are presented in Table 11. The analysis yields a plausible and statistically significant value for RRA, indicating that the HKR factor sat-

isfies the ICAPM restrictions.

From the tests conducted in this section, we demonstrate that the HKR factor satisfies the key constraints imposed by the Intertemporal Capital Asset Pricing Model. HKR shows significant predictive power for future market returns and volatility. The positive sign of HKR's risk price in cross-sectional tests aligns with its positive impact on future market returns. Additionally, the market risk premium yields a plausible estimate for the representative investor's relative risk aversion. These results support the interpretation of HKR as a legitimate risk factor capturing time-varying investment opportunities within the ICAPM framework.

**Table 11: Cross-sectional asset pricing tests of the HKR factor**
This table presents the results of cross-sectional asset pricing tests examining the pricing of the HKR factor. The tests are conducted using efficient Generalized Method of Moments (GMM) estimation as described in Bryzgalova (2015). Three model specifications are considered: market model (Mkt), Fama-French three-factor model (FF3), and Fama-French five-factor model (FF5). The test assets are 27 portfolios sorted on size, value, and KKR. Reported coefficients represent factor risk prices, with standard errors in parentheses. The sample period is 2006-2020. The positive and statistically significant HKR factor risk prices across all specifications, coupled with economically plausible estimates for the market price of risk, suggest that HKR satisfies key Intertemporal CAPM (ICAPM) restrictions.

| | Mkt | FF3 | FF5 |
|---|---|---|---|
| | (1) | (2) | (3) |
| HKR | 0.571*** | 0.246* | 0.248* |
| | (0.141) | (0.134) | (0.146) |
| Mkt.RF | 1.296*** | 1.145*** | 1.032*** |
| | (0.288) | (0.293) | (0.296) |
| SMB | | -0.056 | -0.123 |
| | | (0.156) | (0.166) |
| HML | | -0.871*** | -0.925*** |
| | | (0.209) | (0.212) |
| RMW | | | 0.316* |
| | | | (0.177) |
| CMA | | | -0.858*** |
| | | | (0.182) |
| Observations | 204 | 204 | 204 |
| $R^2$ | 0.789 | 0.869 | 0.889 |

*p<0.1; **p<0.05; ***p<0.01

# 6 Conclusion

This study introduces a novel approach to measuring and pricing knowledge capital risk in financial markets. By leveraging textual analysis of firms' 10-K filings, I construct a High-Knowledge Risk (HKR) factor that captures the unique risk profile associated with

R&D-intensive firms. I demonstrate that this factor offers significant explanatory power for cross-sectional equity returns, even after controlling for well-established Fama-French factors.

The HKR factor exhibits several noteworthy characteristics. Portfolios formed on high HKR consistently outperform those with low HKR over a 14-year period, suggesting a substantial risk premium associated with knowledge capital intensity. The factor boasts the highest Sharpe ratio among compared factors, indicating an efficient trade-off between risk and return. The persistent significance of HKR in asset pricing tests suggests that it captures a unique dimension of risk not fully accounted for by traditional factors. Additionally, high HKR equities generally exhibit higher return volatility, particularly during market stress periods, reflecting the uncertain nature of R&D outcomes.

These results have important implications for both academic research and practical applications in finance and suggest that incorporating knowledge capital risk can enhance the explanatory power of asset pricing models, particularly for innovation-driven sectors. The HKR factor offers a new dimension for portfolio construction and risk management, potentially improving diversification and return optimization strategies. The pricing of knowledge capital risk provides insights for more accurate firm valuation, especially for R&D-intensive companies where traditional metrics may fall short. Furthermore, understanding the market pricing of knowledge capital risk can inform corporate decisions on R&D investments and disclosure strategies.

In conclusion, this paper contributes to the growing literature on intangible capital valuation and highlights the critical role of innovation-related risk in modern financial markets. As the knowledge economy continues to evolve, the ability to accurately measure and price knowledge capital risk will become increasingly vital for investors, managers, and policymakers alike.

# References


Aboody, David and Baruch Lev (2000) "Information asymmetry, R&D, and insider gains," *The Journal of Finance*, 55 (6), 2747–2766.

Ai, Hengjie, Jun E Li, Kai Li, and Christian Schlag (2020) "The collateralizability premium," *The Review of Financial Studies*, 33 (12), 5821–5855.

Amore, Mario Daniele (2020) "Innovation disclosure in times of uncertainty," *Journal of Economics & Management Strategy*, 29 (4), 792–815.

Andrei, Daniel, William Mann, and Nathalie Moyen (2019) "Why did the q theory of investment start working?" *Journal of Financial Economics*, 133 (2), 251–272.

Atkeson, Andrew and Ariel Burstein (2019) "Aggregate implications of innovation policy," *Journal of Political Economy*, 127 (6), 2625–2683.

Belo, Frederico, Vito D Gala, Juliana Salomao, and Maria Ana Vitorino (2022) "Decomposing firm value," *Journal of Financial Economics*, 143 (2), 619–639.

Belo, Frederico, Jun Li, Xiaoji Lin, and Xiaofei Zhao (2017) "Labor-force heterogeneity and asset prices: The importance of skilled labor," *The Review of Financial Studies*, 30 (10), 3669–3709.

Belo, Frederico, Chen Xue, and Lu Zhang (2013) "A supply approach to valuation," *The Review of Financial Studies*, 26 (12), 3029–3067.

Blei, David M, Andrei Ng, and Michael Jordan (2003) "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*

Bloom, Nicholas and John Van Reenen (2007) "Measuring and explaining management practices across firms and countries," *The Quarterly Journal of Economics*, 122 (4), 1351–1408.

Brown, James R, Steven M Fazzari, and Bruce C Petersen (2009) "Financing innovation and growth: Cash flow, external equity, and the 1990s R&D boom," *The Journal of Finance*, 64 (1), 151–185.

Bryzgalova, Svetlana (2015) "Spurious factors in linear asset pricing models," *LSE manuscript*, 1 (3), 6.

Cochrane, John (2009) *Asset pricing: Revised edition*: Princeton University Press.

Corrado, Carol, John Haltiwanger, and Daniel Sichel (2009a) *Measuring Capital in the New Economy*: University of Chicago Press.

Corrado, Carol, Charles Hulten, and Daniel Sichel (2009b) "Intangible capital and US economic growth," *Review of income and wealth*, 55 (3), 661–685.

Crouzet, Nicolas and Janice Eberly (2023) "Rents and intangible capital: A q+ framework," *The Journal of Finance*, 78 (4), 1873–1916.

Dolphin, Rian, Barry Smyth, and Ruihai Dong (2022) "Stock embeddings: Learning distributed representations for financial assets," *arXiv preprint arXiv:2202.08968*.

Du, Xin and Kumiko Tanaka-Ishii (2020) "Stock embeddings acquired from news articles and price history, and an application to portfolio optimization," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 3353–3363.

Eisfeldt, Andrea L, Antonio Falato, and Mindy Z Xiaolan (2023) "Human capitalists," *NBER Macroeconomics Annual*, 37 (1), 1–61.

Eisfeldt, Andrea L and Dimitris Papanikolaou (2013) "Organization capital and the cross-section of expected returns," *The Journal of Finance*, 68 (4), 1365–1406.

Fama, Eugene F and Kenneth R French (2015) "A five-factor asset pricing model," *Journal of Financial Economics*, 116 (1), 1–22, https://doi.org/10.1016/j.jfineco.2014.10.010.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019) "Text as data," *Journal of Economic Literature*, 57 (3), 535–574.

Golubov, Andrey and Theodosia Konstantinidi (2019) "Where is the risk in value? Evidence from a market-to-book decomposition," *The Journal of Finance*, 74 (6), 3135–3186.

Grossman, Gene M and Elhanan Helpman (1991) "Quality ladders in the theory of growth," *The review of economic studies*, 58 (1), 43–61.

Hansen, Lars Peter, John C Heaton, and Nan Li (2005) "Intangible risk," in *Measuring Capital in the New Economy*, 111–152: University of Chicago Press.

Kalamara, Eleni, Arthur Turrell, Chris Redl, George Kapetanios, and Sujit Kapadia (2022) "Making text count: economic forecasting using newspaper text," *Journal of Applied Econometrics*, 37 (5), 896–919.

Kogan, Leonid and Dimitris Papanikolaou (2019) "Technological innovation, intangible capital, and asset prices," *Annual Review of Financial Economics*, 11 (1), 221–242.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman (2017) "Technological innovation, resource allocation, and growth," *The Quarterly Journal of Economics*, 132 (2), 665–712.

Koh, Ping-Sheng and David M. Reeb (2015) "Missing R&D," *Journal of Accounting and Economics*, 60 (1), 73–94, https://doi.org/10.1016/j.jacceco.2015.03.004.

Koh, Ping-Sheng, David M Reeb, and Wanli Zhao (2018) "CEO confidence and unreported R&D," *Management Science*, 64 (12), 5725–5747.

Kusupati, Aditya, Gantavya Bhatt, Aniket Rege et al. (2022) "Matryoshka representation learning," *Advances in Neural Information Processing Systems*, 35, 30233–30249.

Leung, Woon Sau, Kevin P Evans, and Khelifa Mazouz (2020) "The R&D anomaly: Risk or mispricing?" *Journal of Banking & Finance*, 115, 105815.

Li, Dongmei (2011) "Financial constraints, R&D investment, and stock returns," *The Review of Financial Studies*, 24 (9), 2974–3007.

Li, Wendy CY and Bronwyn H Hall (2020) "Depreciation of business R&D capital," *Review of Income and Wealth*, 66 (1), 161–180.

Lima, Luiz Renato and Lucas Lúcio Godeiro (2023) "Equity-premium prediction: Attention is all you need," *Journal of Applied Econometrics*, 38 (1), 105–122.

Maio, Paulo and Pedro Santa-Clara (2012) "Multifactor models and their consistency with the ICAPM," *Journal of Financial Economics*, 106 (3), 586–613.

Mehra, Rajnish and Edward C Prescott (1985) "The equity premium: A puzzle," *Journal of Monetary Economics*, 15 (2), 145–161.

Merton, Robert C (1973) "An intertemporal capital asset pricing model," *Econometrica: Journal of the Econometric Society*, 867–887.

Mezzanotti, Filippo and Timothy Simcoe (2023) "Innovation and Appropriability: Revisiting the Role of Intellectual Property," Working Paper 31428, National Bureau of Economic Research, 10.3386/w31428.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013) "Distributed Representations of Words and Phrases and their Compositionality," in Burges, C.J., L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger eds. *Advances in Neural Information Processing Systems*, 26: Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Newey, Whitney K and Kenneth D West (1987) "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 777–787.

Peri, Giovanni, Kevin Shih, and Chad Sparber (2015) "STEM workers, H-1B visas, and productivity in US cities," *Journal of Labor Economics*, 33 (S1), S225–S255.

Peters, Ryan H. and Lucian A. Taylor (2017) "Intangible capital and the investment-q relation," *Journal of Financial Economics*, 123 (2), 251–272, https://doi.org/10.1016/j.jfineco.2016.03.011.

Rahimikia, Eghbal, Stefan Zohren, and Ser-Huang Poon (2021) "Realised Volatility Forecasting: Machine Learning via Financial Word Embedding," *SSRN Electronic Journal*, 10.2139/ssrn.3895272.

Romer, Paul M. (1990) "Endogenous Technological Change," *Journal of Political Economy*, 98 (5, Part 2), S71–S102, 10.1086/261725.

Sehrawat, Saurabh (2019) "Learning Word Embeddings from 10-K Filings Using PyTorch," *SSRN Electronic Journal*.

Stambaugh, Robert F. and Yu Yuan (2016) "Mispricing Factors," *The Review of Financial Studies*, 30 (4), 1270–1315, 10.1093/rfs/hhw107.

Zheng, Tingguo, Xinyue Fan, Wei Jin, and Kuangnan Fang (2024) "Words or numbers? Macroeconomic nowcasting with textual and macroeconomic data," *International Journal of Forecasting*, 40 (2), 746–761.

# Appendices

## A Firm count by year

Table 12 presents the annual count of firms with available 10-K filings in EDGAR, alongside the count of firms that meet the filtering criteria outlined in section 3.1.

**Table 12:** Annual Firm Counts: Total and Filtered

| Year | Total Firms | Filtered Firms |
|---|---|---|
| 2006 | 5,685 | 2,466 |
| 2007 | 6,445 | 2,714 |
| 2008 | 6,931 | 2,305 |
| 2009 | 8,244 | 2,190 |
| 2010 | 8,122 | 2,290 |
| 2011 | 8,019 | 2,356 |
| 2012 | 7,797 | 2,316 |
| 2013 | 7,560 | 2,401 |
| 2014 | 7,560 | 2,518 |
| 2015 | 7,531 | 2,528 |
| 2016 | 7,196 | 2,431 |
| 2017 | 6,896 | 2,394 |
| 2018 | 6,804 | 2,418 |
| 2019 | 6,683 | 2,404 |
| 2020 | 6,531 | 2,332 |
| 2021 | 6,936 | 2,308 |
| 2022 | 6,899 | 1,885 |

*Note:* Filtered firms meet criteria for NYSE, AMEX, and NASDAQ membership; minimum stock price; and 10-K filing requirement, as defined in section 3.1.

# B Robustness test: defining HKR via LDA (Latent Dirichlet Allocation)

As an alternative method to define KKR, I employ Latent Dirichlet Allocation (LDA) to identify textual topics within Item 1A (Risk Factors) of firms' 10-K filings.

LDA is a generative statistical model widely used for topic modeling in natural language processing. It assumes that each document in a corpus can be represented as a mixture of latent topics, with each topic having a particular weight. Formally, for $K$ topics, each document $i$ has topic weights $\omega_{i1}, ..., \omega_{iK}$. Each topic $k$ is associated with a word probability vector $\theta_k$, defining the likelihood of words appearing under that topic (Blei et al., 2003).

Given a document's word count vector $X_i$ of length $n_i$, the word distribution is modeled as a multinomial distribution:

$$X_i \sim \text{Multinomial}(n_i, \omega_{i1}\theta_1 + ... + \omega_{iK}\theta_K) \tag{9}$$

Prior to applying LDA, I perform several preprocessing steps on the filtered firm texts. I use the `spacy` Python library for lemmatization, which transforms words into their base form (e.g., "take", "took", and "taken" are standardized to "take"). This process ensures consistent semantics across the corpus. `spacy` leverages WordNet, a comprehensive English lexical database curated by Princeton University, for this task.

To capture richer semantic insights, I extract significant collocations (e.g., "patent application") using the method outlined in Mikolov et al. (2013). This approach yields pertinent bigrams and trigrams with a minimum occurrence threshold of 5, ensuring only statistically relevant combinations are included. I then compile all discovered words, bigrams, and trigrams into a comprehensive dictionary.

Finally, I transform the texts into a bag-of-words model using both the dictionary and the n-gram processed texts. In this representation, each word's frequency in a document (denoted as $c_{ij}$) is preserved, but the sequence of words is omitted, resulting in the final depiction of the corpus.

LDA output consists of a list of topics, each represented by a collection of words, and a weight distribution across these topics for each document. In my implementation, I specify a parameter $k$ for the number of topics and provide my previously created dictionary. The choice of $k$ is typically made ad hoc, prioritizing interpretability (Gentzkow et al., 2019)

It is important to note that LDA is an unsupervised learning method, operating with-

out predefined labels. This can pose challenges in interpretation, as generated topics may not align with intuitive descriptions.

Several challenges arise in using LDA for KKR extraction. First, the extracted KKR assumes a bimodal distribution, with over 80% of firms having KKR close to 0. Therefore, firms are divided into four bins of KKR, divided by percentiles 80, 90, and 95.

Second, high KKR is heavily concentrated in Health firms, potentially creating a measure that primarily clusters Health firms. Finally, as an unsupervised method, LDA-generated topics cannot be directly guided to represent specific intended concepts.

Despite these challenges, firms identified as having high knowledge capital risk through this method do exhibit higher levels of kurtosis in their returns, which is consistent with firms engaged in significant innovation processes.

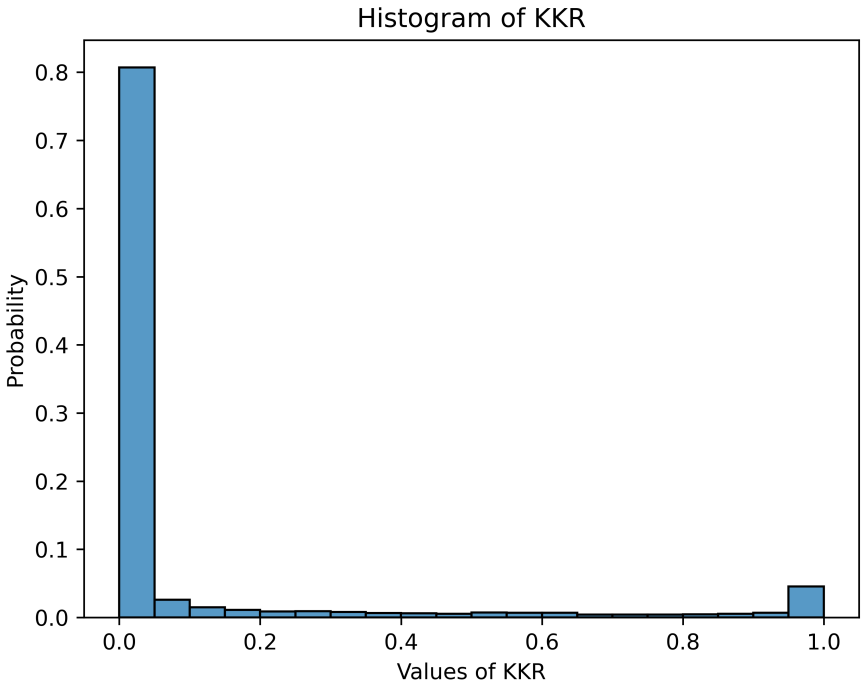



**Figure 12:** Distribution of *KKR* across the entire set of firms (LDA)

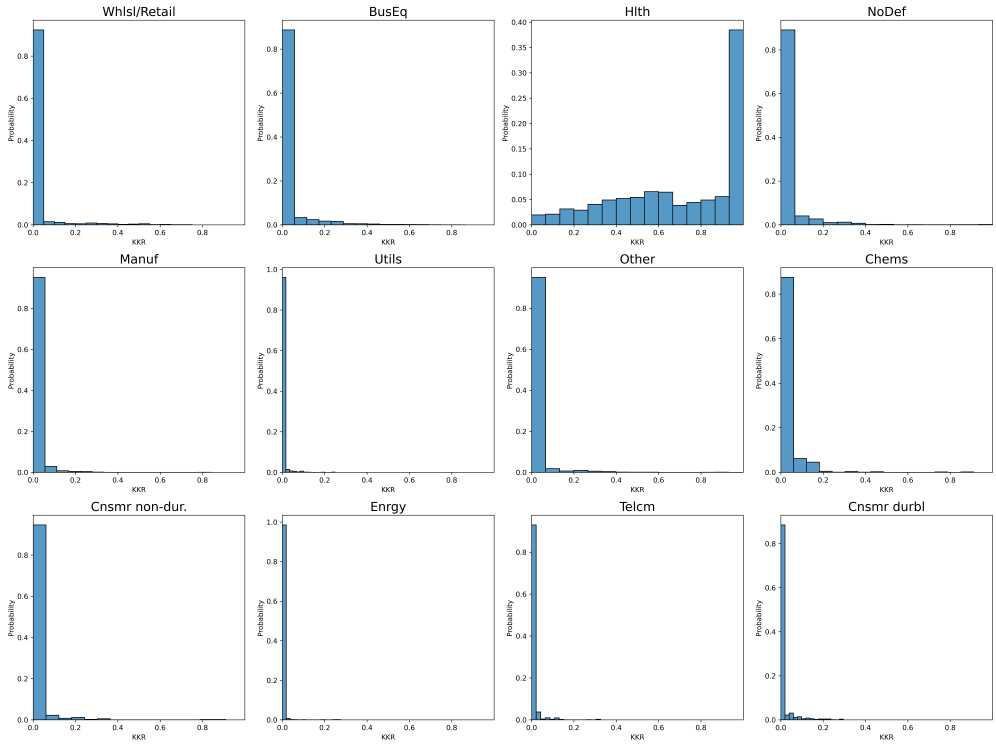



**Figure 13:** Distribution of KKR by industry, as defined by Fama and French's 12-industry model

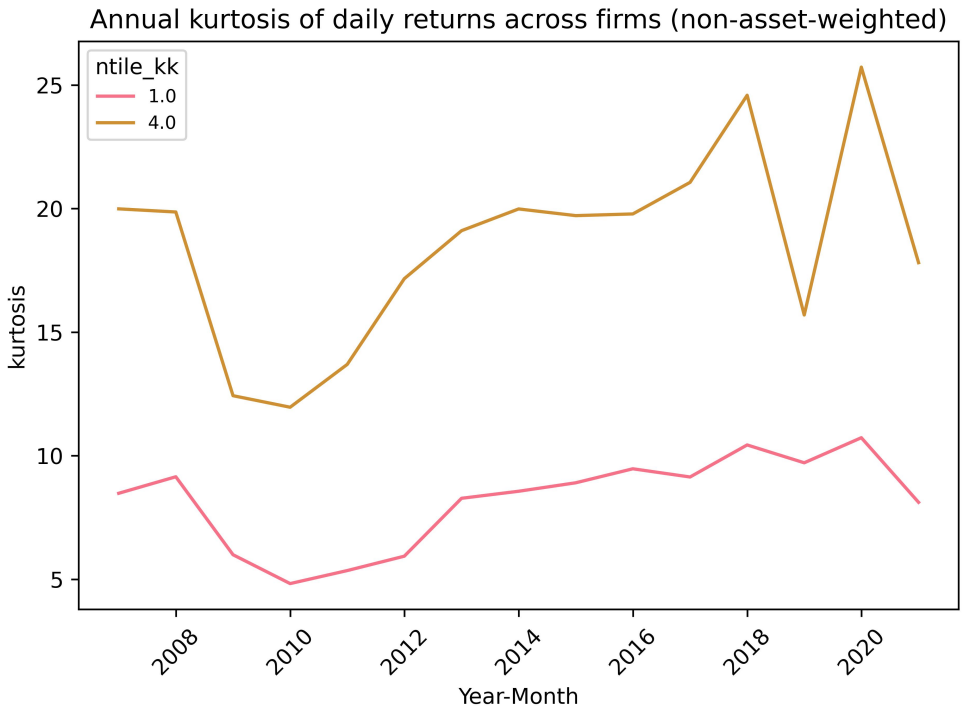



**Figure 14:** Kurtosis of Returns for different firms (LDA), grouped by KKR. The yellow line represents firms above the 95th percentile. The red line represents firms below the 80th percentile.

# C Robustness test: defining HKR via HDP (Hierarchical Dirichlet Processing)

As an additional robustness check, I employ Hierarchical Dirichlet Process (HDP) to identify textual topics within Item 1A (Risk Factors) of firms' 10-K filings. HDP extends the Latent Dirichlet Allocation (LDA) model by allowing for a potentially infinite number of topics, providing more flexibility in topic modeling.

HDP is a nonparametric Bayesian approach to clustering grouped data. In this context, each firm's 10-K filing is a group of words, and I aim to discover topics (clusters of words) that are shared across these documents. Formally, HDP can be described as:

$$G_0 \sim \text{DP}(\gamma, H) \tag{10}$$

$$G_j \sim \text{DP}(\alpha, G_0) \quad \text{for each group } j \tag{11}$$

$$\theta_{ji} \sim G_j \quad \text{for each word } i \text{ in group } j \tag{12}$$

$$x_{ji} \sim F(\theta_{ji}) \quad \text{for each word } i \text{ in group } j \tag{13}$$

Where DP denotes a Dirichlet Process, $H$ is the base distribution, $\gamma$ and $\alpha$ are concentration parameters, $G_0$ is the global distribution over topics, $G_j$ is the topic distribution for document $j$, $\theta_{ji}$ is the topic for word $i$ in document $j$, and $F$ is the distribution of words given a topic.

I apply the same preprocessing steps as described in Appendix A, including lemmatization using the `spacy` Python library, extraction of significant collocations, and transformation into a bag-of-words model.

The key advantage of HDP over LDA is that it automatically infers the number of topics from the data, potentially capturing more nuanced aspects of KKR. The output of HDP includes a list of topics (each represented by a collection of words) and the topic proportions for each document.

However, HDP faces similar challenges to LDA in KKR extraction. As an unsupervised method, HDP-generated topics may not directly align with predefined concepts of KKR, making interpretation challenging. The resulting KKR measure may exhibit a non-uniform distribution across firms, potentially complicating comparative analyses. Additionally, high KKR values may cluster in specific industries, potentially biasing the measure towards certain sectors. These challenges underscore the need for careful validation and interpretation of the HDP-derived KKR measure.

Despite these challenges, HDP offers a more flexible approach to topic modeling that

may capture subtle variations in KKR across firms. I find that firms identified as having high knowledge capital risk through this method exhibit higher levels of return kurtosis, consistent with firms engaged in significant innovation processes.

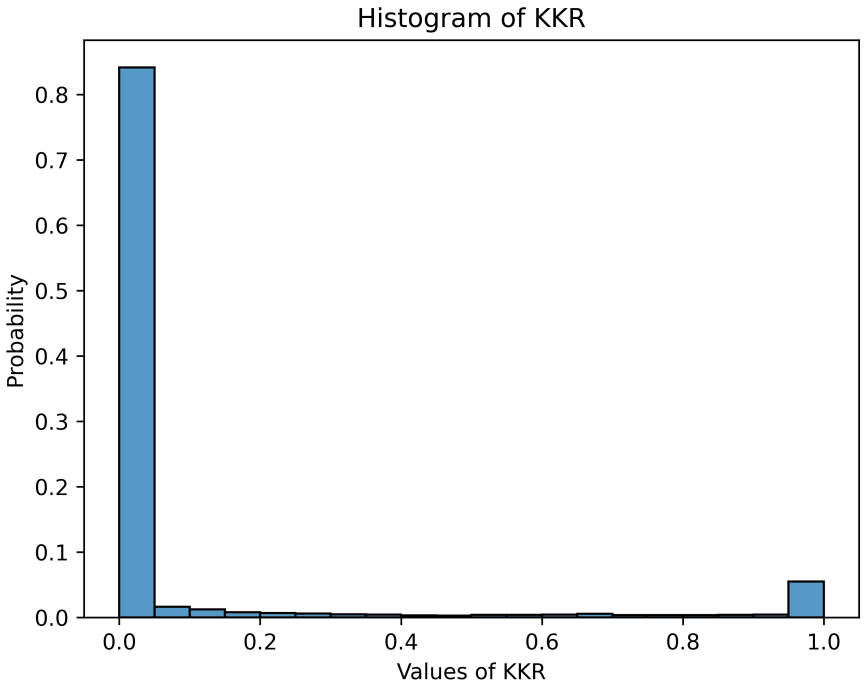



**Figure 15:** Distribution of *KKR* across the entire set of firms (HDP)

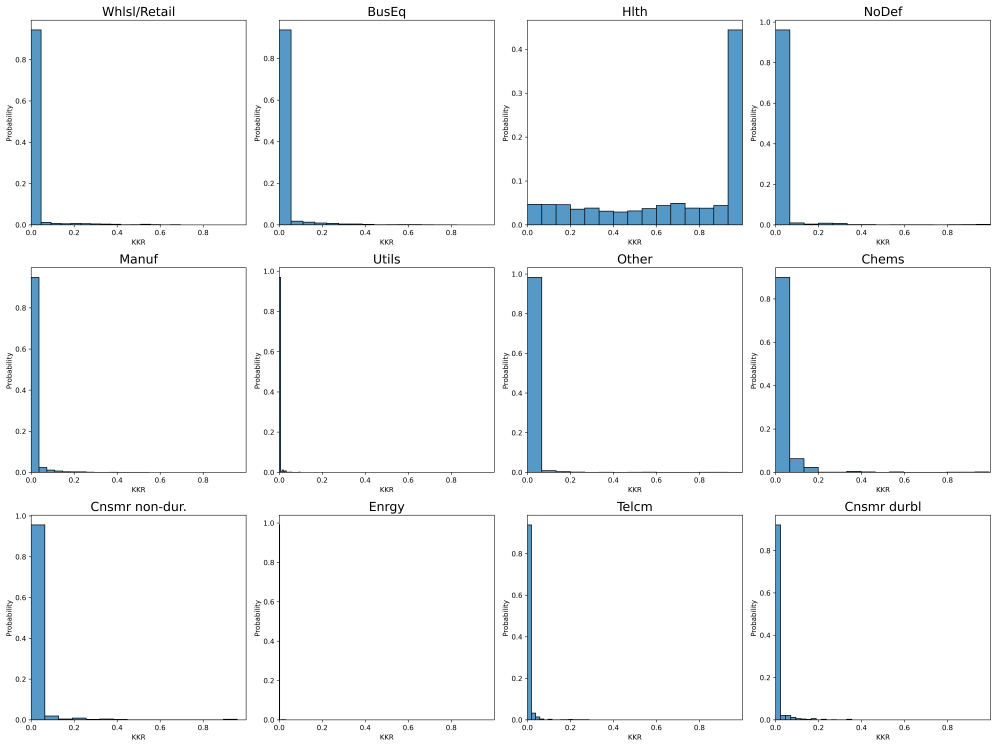



**Figure 16:** Distribution of KKR by industry, as defined by Fama and French's 12-industry model (HDP)

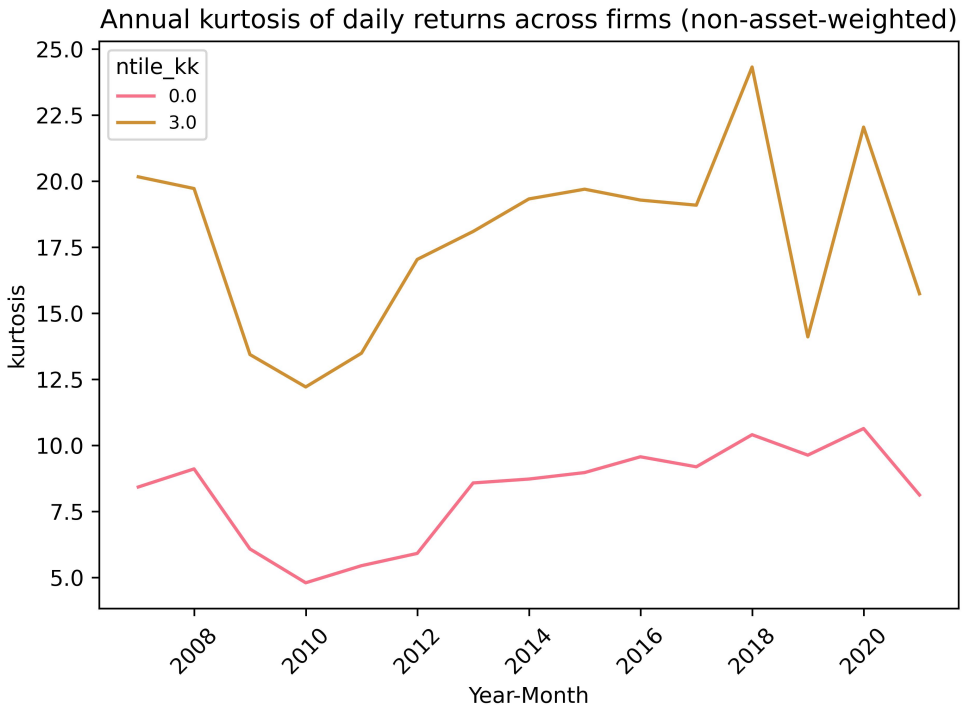
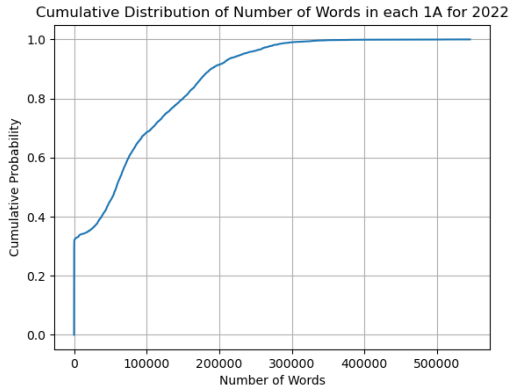
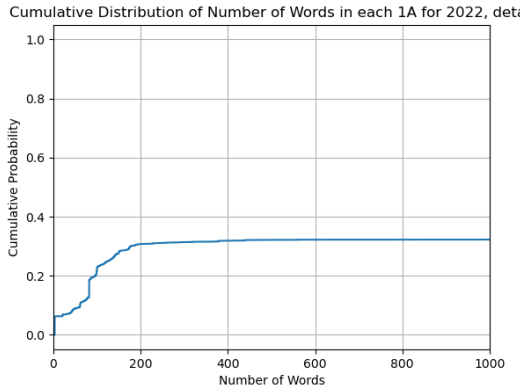



**Figure 17:** Kurtosis of Returns for different firms (HDP), grouped by KKR. The yellow line represents firms above the 95th percentile. The red line represents firms below the 80th percentile.

# D Word distribution by document (1A)




**(a)** Cumulative Distribution of Number of Words




**(b)** Cumulative Distribution of Number of Words, Zoom

**Figure 18:** Cumulative Distribution of Number of Words in 2022

# E Robustness test: Out-of-sample prediction power of HKR on portfolio returns

**Table 13:** Predictive power: assigning 30% of the set as train set (to compute HKR); 70% of the set as test set (to assign portfolios)

| | Mkt | Mkt+ | FF3 | FF3+ | FF5 | FF5+ |
|---|---|---|---|---|---|---|
| Intercept | -0.0004 | -0.0004 | -0.0003 | -0.0003 | -0.0002 | -0.0002 |
| | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Predicted | 1.0544 | 1.0523 | 1.0061 | 1.0042 | 1.0092 | 1.0084 |
| | (0.0089) | (0.0080) | (0.0057) | (0.0055) | (0.0051) | (0.0049) |
| R-squared | 0.7596 | 0.7927 | 0.8777 | 0.8825 | 0.8918 | 0.8958 |
| R-squared Adj. | 0.7596 | 0.7926 | 0.8777 | 0.8825 | 0.8918 | 0.8958 |

**Table 14:** Predictive power: assigning 70% of the set as train set (to compute HKR); 30% of the set as test set (to assign portfolios)

| | Mkt | Mkt+ | FF3 | FF3+ | FF5 | FF5+ |
|---|---|---|---|---|---|---|
| Intercept | -0.0006 | -0.0005 | -0.0004 | -0.0004 | -0.0003 | -0.0003 |
| | (0.0002) | (0.0002) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Predicted | 1.0655 | 1.0525 | 1.0128 | 1.0101 | 1.0187 | 1.0168 |
| | (0.0114) | (0.0098) | (0.0082) | (0.0075) | (0.0075) | (0.0069) |
| R-squared | 0.6936 | 0.7348 | 0.8095 | 0.8175 | 0.8254 | 0.8326 |
| R-squared Adj. | 0.6935 | 0.7348 | 0.8095 | 0.8175 | 0.8254 | 0.8326 |

**Table 15:** Predictive power: assigning 80% of the set as train set (to compute HKR); 20% of the set as test set (to assign portfolios)

| | Mkt | Mkt+ | FF3 | FF3+ | FF5 | FF5+ |
|---|---|---|---|---|---|---|
| Intercept | -0.0006 | -0.0006 | -0.0004 | -0.0005 | -0.0002 | -0.0003 |
| | (0.0002) | (0.0002) | (0.0002) | (0.0002) | (0.0001) | (0.0001) |
| Predicted | 1.0654 | 1.0535 | 1.0108 | 1.0095 | 1.0170 | 1.0169 |
| | (0.0131) | (0.0115) | (0.0095) | (0.0089) | (0.0088) | (0.0083) |
| R-squared | 0.6437 | 0.6815 | 0.7606 | 0.7696 | 0.7782 | 0.7863 |
| R-squared Adj. | 0.6437 | 0.6815 | 0.7606 | 0.7695 | 0.7782 | 0.7862 |

# F Annual skewness of daily returns across firms

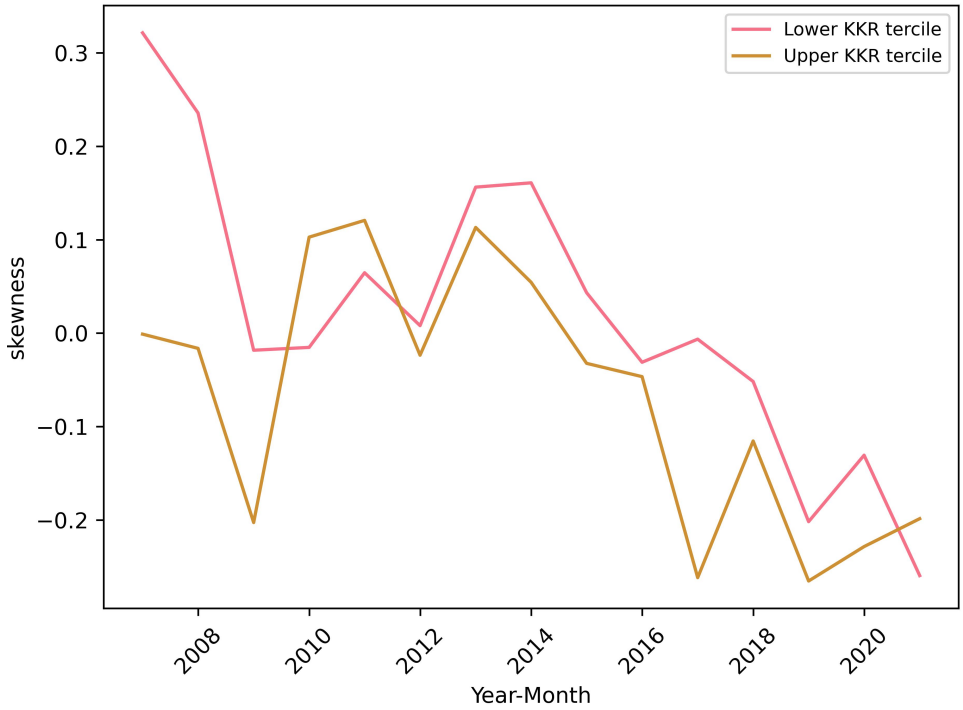



**Figure 19:** Skewness of daily stock returns by KKR quintile. The figure shows the annual skewness of daily returns for firms in different KKR quintiles from 2006 to 2020. Higher KKR quintiles generally exhibit somewhat more positive skewness, indicating a higher probability of extreme positive returns for firms with greater knowledge capital risk.

# G Robustness test: 2009-2020 period

**Table 16:** ICAPM Tests: Predictive regressions for market returns

**(a)** Panel A: Single predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| HKR | 0.4858 | 0.6215* | 0.6448 | 0.8305 | 0.6763 | 0.7922 | 0.9497 |
| | (0.3149) | (0.3743) | (0.6446) | (0.8136) | (0.7731) | (0.6867) | (0.7417) |
| R-squared | 0.0398 | 0.0273 | 0.0069 | 0.0060 | 0.0050 | 0.0065 | 0.0079 |
| R-squared Adj. | 0.0346 | 0.0220 | 0.0012 | -0.0001 | -0.0017 | -0.0008 | -0.0002 |

**(b)** Panel B: Multivariate predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| SMB | -0.2559 | -0.0234 | 0.1526 | -0.7061 | -0.0423 | -0.0595 | -0.1839 |
| | (0.2277) | (0.2664) | (0.3457) | (0.4335) | (0.1718) | (0.1856) | (0.2439) |
| HML | 0.3926 | -0.1739 | 0.1880 | 0.3784 | -0.3474 | -0.2768 | 0.2297 |
| | (0.3135) | (0.3455) | (0.3199) | (0.6218) | (0.5181) | (0.6331) | (0.5490) |
| HKR | 0.8207 | 0.3095 | 0.2964 | 0.8572 | -0.1047 | 0.3243 | 1.3854 |
| | (0.5814) | (0.6693) | (1.0971) | (1.4901) | (1.2057) | (1.3305) | (1.5801) |
| R-squared | 0.0432 | 0.0274 | 0.0049 | 0.0208 | 0.0091 | 0.0193 | 0.0300 |
| R-squared Adj. | 0.0243 | 0.0079 | -0.0164 | -0.0023 | -0.0170 | -0.0099 | -0.0031 |

**Table 17:** ICAPM Tests: Predictive regressions for SVAR

**(a)** Panel A: Single predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| HKR | -0.0203 | -0.0078 | -0.1177 | -0.2078 | -0.1408 | -0.3011** | -0.3034* |
| | (0.0420) | (0.0837) | (0.1566) | (0.1643) | (0.1703) | (0.1410) | (0.1743) |
| R-squared | 0.0047 | 0.0001 | 0.0031 | 0.0048 | 0.0016 | 0.0062 | 0.0054 |
| R-squared Adj. | -0.0006 | -0.0053 | -0.0026 | -0.0014 | -0.0051 | -0.0011 | -0.0027 |

**(b)** Panel B: Multivariate predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| SMB | -0.0086 | -0.0266 | -0.0727 | 0.0489 | 0.1084*** | 0.2604** | -0.0267 |
| | (0.0127) | (0.0220) | (0.0906) | (0.1012) | (0.0420) | (0.1289) | (0.1252) |
| HML | -0.0672 | -0.1132 | -0.0457 | -0.2000 | -0.0348 | 0.0164 | -0.0050 |
| | (0.0482) | (0.0859) | (0.0823) | (0.1690) | (0.1673) | (0.0648) | (0.1140) |
| HKR | -0.0459 | -0.0507 | 0.1141 | -0.0379 | 0.2833 | 0.1243 | 0.3358* |
| | (0.0647) | (0.1110) | (0.1101) | (0.2302) | (0.3020) | (0.3838) | (0.1924) |
| R-squared | 0.0970 | 0.0770 | 0.0213 | 0.0258 | 0.0371 | 0.0394 | 0.0381 |
| R-squared Adj. | 0.0792 | 0.0585 | 0.0003 | 0.0028 | 0.0118 | 0.0109 | 0.0053 |

**Table 18:** Alphas, $\beta_{HKR}$, and weekly excess returns for equities sorted into 27 portfolios, during 2009-2020 period

| Portfolio | $\alpha$ | t($\alpha$) | $\beta_{\text{HKR}}$ | t($\beta_{\text{HKR}}$) | RET |
|---|---|---|---|---|---|
| S/L/0 | 0.01% | 1.056 | -0.301 | -25.609 | 0.08% |
| S/M/0 | 0.04% | 8.727 | -0.316 | -38.535 | 0.14% |
| S/H/0 | 0.11% | 24.123 | -0.237 | -23.058 | 0.27% |
| M/L/0 | -0.06% | -14.083 | -0.730 | -59.475 | 0.02% |
| M/M/0 | 0.07% | 29.377 | -0.426 | -139.870 | 0.18% |
| M/H/0 | 0.17% | 50.649 | -0.289 | -46.793 | 0.41% |
| B/L/0 | -0.04% | -8.984 | -0.447 | -36.091 | 0.05% |
| B/M/0 | 0.03% | 7.700 | -0.831 | -132.824 | 0.1% |
| B/H/0 | 0.09% | 40.984 | -0.343 | -60.158 | 0.3% |
| S/L/1 | 0.0% | 0.635 | -0.264 | -24.896 | 0.04% |
| S/M/1 | -0.0% | -1.320 | -0.109 | -10.009 | 0.15% |
| S/H/1 | 0.08% | 11.486 | -0.006 | -0.521 | 0.27% |
| M/L/1 | -0.09% | -28.964 | -0.245 | -40.615 | 0.06% |
| M/M/1 | 0.04% | 13.857 | -0.024 | -5.910 | 0.2% |
| M/H/1 | 0.16% | 59.811 | 0.283 | 61.041 | 0.42% |
| B/L/1 | -0.01% | -2.559 | -0.146 | -11.608 | 0.12% |
| B/M/1 | 0.01% | 2.959 | -0.345 | -45.087 | 0.15% |
| B/H/1 | 0.08% | 37.905 | 0.013 | 2.697 | 0.28% |
| S/L/2 | -0.11% | -23.871 | 0.114 | 8.509 | 0.09% |
| S/M/2 | -0.04% | -14.664 | 0.327 | 40.882 | 0.17% |
| S/H/2 | 0.08% | 8.870 | 0.350 | 23.460 | 0.37% |
| M/L/2 | -0.15% | -38.594 | 0.472 | 63.445 | 0.05% |
| M/M/2 | 0.01% | 2.742 | 0.522 | 59.016 | 0.2% |
| M/H/2 | 0.14% | 50.325 | 0.404 | 53.998 | 0.41% |
| B/L/2 | -0.02% | -5.739 | 0.111 | 8.788 | 0.13% |
| B/M/2 | -0.03% | -12.884 | 0.241 | 37.484 | 0.17% |
| B/H/2 | 0.09% | 39.267 | 0.486 | 97.372 | 0.34% |

# H Robustness test: dividing HKR into quintiles

**Table 19:** Fama-MacBeth Regressions of Portfolio Weekly Excess Returns

| | Mkt | FF3 | FF5 |
|---|---|---|---|
| | (1) | (2) | (3) |
| HKR | 0.00190*** | 0.00139** | 0.00131** |
| | (0.00071) | (0.00062) | (0.00059) |
| MktRF | -0.00104 | 0.00025 | -0.00021 |
| | (0.00159) | (0.00128) | (0.00130) |
| SMB | | 0.00013 | -0.00012 |
| | | (0.00049) | (0.00053) |
| HML | | -0.00241*** | -0.00229*** |
| | | (0.00074) | (0.00073) |
| RMW | | | -0.00006 |
| | | | (0.00042) |
| CMA | | | -0.00112*** |
| | | | (0.00039) |
| Intercept | 0.00309*** | 0.00191*** | 0.00243*** |
| | (0.00083) | (0.00068) | (0.00073) |
| Observations | 29430 | 29430 | 29430 |
| N. of groups | 45 | 45 | 45 |
| $R^2$ | 0.54154 | 0.66299 | 0.65060 |

*p<0.1; **p<0.05; ***p<0.01

**Table 20:** Summary Statistics of Factor Returns: HKR divided into quintiles

| Factor | Count | Mean | SD | Sharpe | 25% | 50% | 75% |
|---|---|---|---|---|---|---|---|
| Mkt.RF | 757 | 0.75% | 11.48% | 0.065 | -3.80% | 1.37% | 6.80% |
| SMB | 757 | -0.01% | 5.65% | -0.002 | -3.40% | 0.07% | 3.25% |
| HML | 757 | -0.43% | 7.49% | -0.057 | -3.55% | -0.64% | 2.61% |
| CMA | 757 | -0.04% | 2.96% | -0.015 | -2.00% | -0.13% | 1.69% |
| RMW | 757 | 0.20% | 3.65% | 0.055 | -1.87% | 0.05% | 2.22% |
| HKR | 757 | 0.44% | 6.04% | 0.073 | -2.74% | 0.51% | 3.74% |

**Table 21:** ICAPM Tests: Predictive regressions for market returns (HKR in quintiles)

**(a)** Panel A: Single predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| HKR | 0.5205* | 0.5672* | 0.2520 | 0.4900 | 0.3076 | 0.4541 | 0.4023 |
| | (0.3008) | (0.3434) | (0.4645) | (0.6376) | (0.6526) | (0.4724) | (0.5646) |
| R-squared | 0.0627 | 0.0313 | 0.0014 | 0.0028 | 0.0014 | 0.0029 | 0.0019 |
| R-squared Adj. | 0.0577 | 0.0261 | -0.0043 | -0.0033 | -0.0053 | -0.0044 | -0.0062 |

**(b)** Panel B: Multivariate predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| SMB | -0.2589 | -0.1268 | -0.1015 | -0.6893* | 0.2207 | 0.1561 | 0.0764 |
| | (0.2029) | (0.2645) | (0.4577) | (0.3649) | (0.4150) | (0.5056) | (0.6400) |
| HML | 0.1664 | -0.3072 | -0.0125 | 0.2653 | -0.3053 | -0.1307 | -0.2859 |
| | (0.2695) | (0.3671) | (0.4965) | (0.6505) | (0.4906) | (0.5135) | (0.4350) |
| HKR | 0.6361* | 0.3243 | 0.2492 | 0.7039 | 0.1022 | 0.3600 | 0.2282 |
| | (0.3705) | (0.4232) | (0.4454) | (0.6009) | (0.6628) | (0.5653) | (0.4722) |
| R-squared | 0.0731 | 0.0449 | 0.0017 | 0.0064 | 0.0024 | 0.0031 | 0.0027 |
| R-squared Adj. | 0.0580 | 0.0292 | -0.0156 | -0.0122 | -0.0180 | -0.0192 | -0.0220 |

**Table 22:** ICAPM Tests: Predictive regressions for SVAR (HKR in quintiles)

**(a)** Panel A: Single predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| HKR | -0.0396 | -0.0549 | -0.0909 | -0.1414 | -0.0740 | -0.1730* | -0.1856 |
| | (0.0408) | (0.0802) | (0.1138) | (0.1178) | (0.1288) | (0.1002) | (0.1186) |
| R-squared | 0.0246 | 0.0081 | 0.0026 | 0.0030 | 0.0006 | 0.0028 | 0.0027 |
| R-squared Adj. | 0.0194 | 0.0027 | -0.0031 | -0.0032 | -0.0061 | -0.0046 | -0.0054 |

**(b)** Panel B: Multivariate predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| SMB | -0.0189 | -0.0197 | 0.0500 | 0.1489 | 0.0484 | 0.1584 | -0.0297 |
| | (0.0208) | (0.0362) | (0.1468) | (0.1027) | (0.1383) | (0.1938) | (0.1378) |
| HML | -0.0690 | -0.0797 | -0.1212 | -0.3433** | -0.2547* | -0.2162* | -0.2228* |
| | (0.0431) | (0.0983) | (0.0885) | (0.1694) | (0.1480) | (0.1275) | (0.1149) |
| HKR | -0.0945** | -0.1174 | -0.1718 | -0.3682** | -0.2369 | -0.3199** | -0.3117** |
| | (0.0457) | (0.0836) | (0.1135) | (0.1500) | (0.1612) | (0.1508) | (0.1394) |
| R-squared | 0.1177 | 0.0290 | 0.0059 | 0.0159 | 0.0065 | 0.0059 | 0.0072 |
| R-squared Adj. | 0.1034 | 0.0131 | -0.0113 | -0.0025 | -0.0138 | -0.0163 | -0.0174 |

# I Robustness test: dividing HKR into deciles

| | Mkt+ | FF3+ | FF5+ |
|---|---|---|---|
| | (1) | (2) | (3) |
| HKR | 0.00196** | 0.00147** | 0.00143** |
| | (0.00080) | (0.00072) | (0.00067) |
| MktRF | -0.00093 | -0.00013 | -0.00044 |
| | (0.00151) | (0.00128) | (0.00127) |
| SMB | | 0.00010 | -0.00012 |
| | | (0.00049) | (0.00054) |
| HML | | -0.00215*** | -0.00208*** |
| | | (0.00071) | (0.00070) |
| RMW | | | -0.00010 |
| | | | (0.00039) |
| CMA | | | -0.00073** |
| | | | (0.00034) |
| Intercept | 0.00295*** | 0.00223*** | 0.00257*** |
| | (0.00076) | (0.00068) | (0.00065) |
| Observations | 58860 | 58860 | 58860 |
| N. of groups | 90 | 90 | 90 |
| $R^2$ | 0.38906 | 0.53746 | 0.51662 |

**Table 23:** Summary Statistics of Factor Returns: HKR divided into deciles

| Factor | Count | Mean | SD | Sharpe | 25% | 50% | 75% |
|---|---|---|---|---|---|---|---|
| HKR | 757 | 7.70% | 11.85% | 0.65 | -31.4% | 9.8% | 67.2% |
| Mkt.RF | 757 | 9.39% | 19.02% | 0.49 | -36.5% | 17.8% | 125.5% |
| SMB | 757 | -0.15% | 9.37% | -0.02 | -33.4% | 0.8% | 47.4% |
| HML | 757 | -5.00% | 12.42% | -0.40 | -34.5% | -7.4% | 36.5% |
| RMW | 757 | 2.41% | 6.04% | 0.40 | -20.1% | 0.6% | 30.3% |
| CMA | 757 | -0.51% | 4.90% | -0.10 | -21.3% | -1.6% | 22.4% |

**Table 24:** ICAPM Tests: Predictive regressions for market returns (HKR in deciles)

**(a)** Panel A: Single predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| HKR | 0.3814* | 0.4379 | 0.1742 | 0.4026 | 0.3331 | 0.3573 | 0.3578 |
| | (0.2304) | (0.2957) | (0.3895) | (0.5363) | (0.5647) | (0.3666) | (0.4962) |
| R-squared | 0.0452 | 0.0253 | 0.0009 | 0.0024 | 0.0020 | 0.0022 | 0.0019 |
| R-squared Adj. | 0.0401 | 0.0200 | -0.0048 | -0.0038 | -0.0047 | -0.0051 | -0.0062 |

**(b)** Panel B: Multivariate predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| SMB | -0.2318 | -0.1074 | -0.0783 | -0.6978* | 0.1659 | 0.1766 | 0.0369 |
| | (0.1961) | (0.2517) | (0.4423) | (0.3715) | (0.3710) | (0.4976) | (0.6320) |
| HML | 0.0731 | -0.3599 | -0.0686 | 0.2002 | -0.2500 | -0.1943 | -0.2803 |
| | (0.2323) | (0.3434) | (0.4757) | (0.6459) | (0.4748) | (0.4937) | (0.4371) |
| HKR | 0.4232* | 0.2077 | 0.1460 | 0.5785 | 0.1927 | 0.2414 | 0.2294 |
| | (0.2347) | (0.3403) | (0.3406) | (0.4895) | (0.5410) | (0.4029) | (0.3895) |
| R-squared | 0.0534 | 0.0425 | 0.0012 | 0.0061 | 0.0027 | 0.0027 | 0.0028 |
| R-squared Adj. | 0.0381 | 0.0268 | -0.0161 | -0.0126 | -0.0176 | -0.0197 | -0.0219 |

**Table 25:** ICAPM Tests: Predictive regressions for SVAR (HKR in deciles)

**(a)** Panel A: Single predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| HKR | -0.0209 | -0.0150 | -0.0079 | -0.0835 | -0.0338 | -0.1114 | -0.1132 |
| | (0.0343) | (0.0733) | (0.0926) | (0.0903) | (0.1016) | (0.0823) | (0.1061) |
| R-squared | 0.0092 | 0.0008 | 0.0000 | 0.0013 | 0.0002 | 0.0014 | 0.0013 |
| R-squared Adj. | 0.0039 | -0.0046 | -0.0057 | -0.0049 | -0.0066 | -0.0059 | -0.0069 |

**(b)** Panel B: Multivariate predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| SMB | -0.0265 | -0.0411 | 0.0049 | 0.1216 | 0.0216 | 0.1332 | -0.0672 |
| | (0.0209) | (0.0395) | (0.1292) | (0.0984) | (0.1323) | (0.2074) | (0.1318) |
| HML | -0.0482 | -0.0354 | -0.0409 | -0.2667 | -0.1947 | -0.1519 | -0.1444 |
| | (0.0376) | (0.0831) | (0.0993) | (0.1650) | (0.1422) | (0.1250) | (0.1171) |
| HKR | -0.0524 | -0.0380 | -0.0292 | -0.2285** | -0.1311 | -0.2013 | -0.1638 |
| | (0.0335) | (0.0666) | (0.0980) | (0.1071) | (0.1127) | (0.1246) | (0.1099) |
| R-squared | 0.0737 | 0.0117 | 0.0005 | 0.0099 | 0.0042 | 0.0032 | 0.0041 |
| R-squared Adj. | 0.0587 | -0.0045 | -0.0168 | -0.0087 | -0.0161 | -0.0191 | -0.0206 |

**Table 26:** Alphas, $\beta_{HKR}$, and weekly excess returns for equities sorted into 60 portfolios. KKR is divided in deciles

| Portfolio | $\alpha$ | t($\alpha$) | $\beta_{\text{HKR}}$ | t($\beta_{\text{HKR}}$) | RET |
|---|---|---|---|---|---|
| S/L/0 | -0.01% | -1.236 | -0.197 | -27.521 | 0.0% |
| S/M/0 | 0.07% | 9.917 | -0.111 | -16.724 | 0.17% |
| S/H/0 | -0.14% | -18.103 | -0.196 | -21.183 | 0.02% |
| M/L/0 | -0.07% | -14.729 | -0.555 | -72.288 | 0.0% |
| M/M/0 | 0.13% | 27.518 | -0.334 | -78.474 | 0.21% |
| M/H/0 | 0.13% | 20.580 | -0.288 | -37.756 | 0.33% |
| B/L/0 | -0.07% | -6.270 | -0.542 | -53.558 | -0.0% |
| B/M/0 | 0.1% | 18.423 | -0.623 | -120.729 | 0.15% |
| B/H/0 | 0.1% | 17.615 | -0.274 | -35.633 | 0.36% |
| S/L/5 | -0.02% | -3.705 | -0.203 | -28.883 | 0.01% |
| S/M/5 | -0.08% | -15.499 | -0.017 | -3.557 | 0.08% |
| S/H/5 | 0.09% | 8.077 | -0.012 | -1.494 | 0.33% |
| M/L/5 | -0.07% | -10.503 | -0.128 | -17.501 | 0.09% |
| M/M/5 | 0.07% | 19.778 | 0.062 | 7.904 | 0.23% |
| M/H/5 | 0.12% | 14.980 | 0.109 | 17.348 | 0.38% |
| B/L/5 | -0.05% | -10.243 | 0.057 | 5.093 | 0.07% |
| B/M/5 | -0.08% | -9.362 | 0.009 | 1.086 | 0.11% |
| B/H/5 | 0.06% | 18.448 | -0.061 | -17.273 | 0.25% |
| S/L/9 | -0.13% | -35.226 | 0.224 | 37.494 | 0.07% |
| S/M/9 | -0.12% | -14.488 | 0.309 | 33.625 | 0.06% |
| S/H/9 | -0.0% | -0.209 | 0.507 | 57.356 | 0.32% |
| M/L/9 | -0.15% | -28.584 | 0.353 | 56.268 | 0.1% |
| M/M/9 | -0.05% | -11.157 | 0.487 | 80.256 | 0.16% |
| M/H/9 | 0.1% | 16.352 | 0.583 | 79.769 | 0.38% |
| B/L/9 | -0.01% | -3.136 | 0.215 | 26.278 | 0.2% |
| B/M/9 | -0.08% | -26.683 | 0.382 | 62.206 | 0.12% |
| B/H/9 | 0.06% | 26.889 | 0.365 | 90.429 | 0.32% |

# J Robustness test: defining HKR as cosine similarity to term "patents"

**Table 27:** Fama-MacBeth Regressions of Portfolio Weekly Excess Returns

| | Mkt | FF3 | FF5 |
|---|---|---|---|
| | (1) | (2) | (3) |
| HKR | 0.00146*** | 0.00098* | 0.00087* |
| | (0.00055) | (0.00051) | (0.00050) |
| MktRF | -0.00103 | 0.00069 | 0.00063 |
| | (0.00174) | (0.00134) | (0.00139) |
| SMB | | 0.00010 | -0.00007 |
| | | (0.00051) | (0.00054) |
| HML | | -0.00248*** | -0.00237*** |
| | | (0.00075) | (0.00072) |
| RMW | | | 0.00043 |
| | | | (0.00050) |
| CMA | | | -0.00140*** |
| | | | (0.00045) |
| Intercept | 0.00293*** | 0.00154* | 0.00162* |
| | (0.00094) | (0.00086) | (0.00083) |
| Observations | 17658 | 17658 | 17658 |
| N. of groups | 27 | 27 | 27 |
| $R^2$ | 0.48441 | 0.66451 | 0.67963 |

*p<0.1; **p<0.05; ***p<0.01

**Table 28:** Summary Statistics of Factor Returns (HKR in terciles)

| Factor | Count | Mean | SD | Sharpe | 25% | 50% | 75% |
|---|---|---|---|---|---|---|---|
| Mkt.RF | 757 | 0.75% | 11.48% | 0.065 | -3.80% | 1.37% | 6.80% |
| SMB | 757 | -0.01% | 5.65% | -0.002 | -3.40% | 0.07% | 3.25% |
| HML | 757 | -0.43% | 7.49% | -0.057 | -3.55% | -0.64% | 2.61% |
| CMA | 757 | -0.04% | 2.96% | -0.015 | -2.00% | -0.13% | 1.69% |
| RMW | 757 | 0.20% | 3.65% | 0.055 | -1.87% | 0.05% | 2.22% |
| HKR | 757 | 0.45% | 5.23% | 0.086 | -2.35% | 0.48% | 3.17% |

**Table 29:** ICAPM Tests: Predictive regressions for market returns (HKR in terciles)

**(a)** Panel A: Single predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| HKR | 0.3691 | 0.4204 | 0.1868 | 0.0056 | -0.1941 | -0.0456 | -0.1167 |
| | (0.3051) | (0.3886) | (0.5421) | (0.7466) | (0.6825) | (0.4768) | (0.6246) |
| R-squared | 0.0237 | 0.0130 | 0.0005 | 0.0000 | 0.0004 | 0.0000 | 0.0001 |
| R-squared Adj. | 0.0184 | 0.0076 | -0.0052 | -0.0062 | -0.0063 | -0.0073 | -0.0080 |

**(b)** Panel B: Multivariate predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| SMB | -0.1567 | 0.0506 | -0.0406 | -0.3911 | 0.6612 | 0.6050 | 0.5625 |
| | (0.1820) | (0.2532) | (0.4917) | (0.4751) | (0.5459) | (0.6563) | (0.7632) |
| HML | -0.0105 | -0.6967 | -0.1096 | -0.2505 | -1.0663* | -0.8312 | -1.0228* |
| | (0.2162) | (0.4348) | (0.5602) | (0.7027) | (0.5713) | (0.5985) | (0.5555) |
| HKR | 0.3352 | -0.3260 | 0.0827 | -0.2226 | -1.2442* | -0.8945 | -1.1409** |
| | (0.2571) | (0.5377) | (0.5262) | (0.5937) | (0.6587) | (0.6446) | (0.4772) |
| R-squared | 0.0288 | 0.0422 | 0.0008 | 0.0030 | 0.0090 | 0.0051 | 0.0066 |
| R-squared Adj. | 0.0130 | 0.0265 | -0.0165 | -0.0157 | -0.0113 | -0.0171 | -0.0180 |

**Table 30:** ICAPM Tests: Predictive regressions for SVAR (HKR in terciles)

**(a)** Panel A: Single predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| HKR | -0.0069 | 0.0137 | -0.1060 | -0.0813 | -0.0068 | -0.1333 | -0.1554 |
| | (0.0476) | (0.1077) | (0.1564) | (0.1381) | (0.1515) | (0.1161) | (0.1167) |
| R-squared | 0.0006 | 0.0004 | 0.0023 | 0.0006 | 0.0000 | 0.0010 | 0.0012 |
| R-squared Adj. | -0.0048 | -0.0050 | -0.0034 | -0.0055 | -0.0067 | -0.0063 | -0.0069 |

**(b)** Panel B: Multivariate predictive regressions

| | $q = 1$ | $q = 3$ | $q = 13$ | $q = 26$ | $q = 39$ | $q = 52$ | $q = 65$ |
|---|---|---|---|---|---|---|---|
| SMB | -0.0293 | -0.0533 | 0.0624 | 0.1412 | 0.0428 | 0.1543 | -0.0253 |
| | (0.0214) | (0.0484) | (0.1566) | (0.1222) | (0.1834) | (0.2548) | (0.2011) |
| HML | -0.0559 | -0.0161 | -0.1649 | -0.3704** | -0.2651 | -0.2442 | -0.2756 |
| | (0.0374) | (0.0951) | (0.1152) | (0.1521) | (0.1818) | (0.1741) | (0.2029) |
| HKR | -0.0719 | -0.0108 | -0.2638 | -0.4390** | -0.2636 | -0.3814 | -0.4206* |
| | (0.0494) | (0.1205) | (0.2024) | (0.1820) | (0.2578) | (0.2427) | (0.2375) |
| R-squared | 0.0636 | 0.0088 | 0.0072 | 0.0124 | 0.0049 | 0.0039 | 0.0064 |
| R-squared Adj. | 0.0484 | -0.0075 | -0.0100 | -0.0061 | -0.0154 | -0.0184 | -0.0183 |

**Table 31:** Alphas, $\beta_{HKR}$, and weekly excess returns for equities sorted into 27 portfolios. KKR is defined as cosine similarity to term "patents"

| Portfolio | $\alpha$ | t($\alpha$) | $\beta_{\text{HKR}}$ | t($\beta_{\text{HKR}}$) | RET |
|---|---|---|---|---|---|
| S/L/0 | -0.04% | -6.388 | -0.339 | -38.962 | 0.06% |
| S/M/0 | 0.03% | 7.294 | -0.313 | -40.642 | 0.16% |
| S/H/0 | 0.12% | 39.290 | -0.183 | -19.359 | 0.29% |
| M/L/0 | -0.02% | -7.305 | -0.570 | -54.061 | 0.08% |
| M/M/0 | 0.05% | 32.229 | -0.452 | -163.419 | 0.19% |
| M/H/0 | 0.14% | 42.199 | -0.342 | -70.546 | 0.37% |
| B/L/0 | -0.08% | -21.065 | 0.022 | 3.180 | 0.03% |
| B/M/0 | 0.05% | 12.823 | -0.519 | -86.527 | 0.15% |
| B/H/0 | 0.09% | 31.859 | -0.362 | -99.410 | 0.28% |
| S/L/1 | -0.03% | -6.510 | -0.286 | -35.071 | 0.03% |
| S/M/1 | 0.03% | 7.820 | -0.085 | -9.954 | 0.16% |
| S/H/1 | 0.08% | 15.528 | 0.179 | 16.880 | 0.28% |
| M/L/1 | -0.09% | -25.228 | -0.338 | -39.580 | 0.04% |
| M/M/1 | 0.04% | 13.555 | -0.084 | -20.641 | 0.19% |
| M/H/1 | 0.16% | 73.815 | 0.148 | 28.852 | 0.41% |
| B/L/1 | 0.04% | 9.139 | -0.096 | -8.603 | 0.15% |
| B/M/1 | -0.04% | -12.332 | -0.201 | -24.098 | 0.1% |
| B/H/1 | 0.13% | 81.113 | 0.031 | 4.307 | 0.38% |
| S/L/2 | -0.06% | -11.468 | 0.092 | 7.584 | 0.13% |
| S/M/2 | -0.08% | -27.085 | 0.454 | 37.976 | 0.14% |
| S/H/2 | 0.04% | 3.740 | 0.804 | 72.332 | 0.3% |
| M/L/2 | -0.15% | -35.210 | 0.381 | 50.015 | 0.04% |
| M/M/2 | 0.01% | 2.780 | 0.470 | 67.540 | 0.19% |
| M/H/2 | 0.17% | 51.074 | 0.897 | 80.997 | 0.46% |
| B/L/2 | -0.03% | -9.056 | 0.073 | 7.735 | 0.12% |
| B/M/2 | -0.0% | -1.923 | 0.270 | 46.026 | 0.18% |
| B/H/2 | 0.05% | 19.238 | 0.563 | 142.554 | 0.28% |