Learning from the Loudest: Do the Most Active Information Providers Sway Investors' Equity Premium Expectations?

ABSTRACT

This paper studies how retail investors form - and continually reshape - their beliefs about future stock returns. Far from being anchored solely in economic fundamentals or rational expectations, investor beliefs emerge here as the outcome of an interplay among competing voices in the information ecosystem. Harnessing a uniquely comprehensive dataset of survey-based return expectations, coupled with text analyses of analyst reports and financial shows aired on 42 local news channels, I find that the sheer volume and prominence of certain analyst forecasts decisively shift investors' views on the equity risk premium. When widely visible outlets broadcast optimistic signals - particularly about earnings growth - retail expectations surge in a way that is both large and enduring, persisting for months. In contrast, less trumpeted insights from "quiet" experts are roundly ignored by the investing public, even though they contain predictive power for market returns. Remarkably, this attention-driven learning dynamic holds across almost all demographic segments, from high-net-worth investors to novices with modest portfolios. My findings present a new framework for understanding how pockets of information can powerfully amplify or dampen collective sentiment. By revealing how specific streams of market information tip the scales of investor belief, I illuminate a potent channel through which narratives, rather than strict fundamentals alone, shape price dynamics.

I. Introduction

Do subjective expectations of a retail investor about stock market risk premium vary over time? Does the information market affect the subjective expectations?

The studies by Greenwood and Shleifer, 2014 and Nagel and Xu, 2019 provide valuable insights into the formation of investors' expectations of future market returns. While Greenwood and Shleifer, 2014 find that investors tend to extrapolate recent market performance into the future, leading to time-varying expectations, Nagel and Xu, 2019 study suggests that investors' expectations of future market excess returns are "virtually constant." This apparent contradiction highlights the complexity of investors' subjective expectations and the need for further research to better understand the factors that influence these expectations.

I start with examining surveys of individual retail investors from three sources used in Greenwood and Shleifer, 2014 and Nagel and Xu, 2019 , Gallup, the Survey Research Center at the University of Michigan, and the Conference Board, and analyze subjective expectations of annual stock market returns and the risk-free rate from June 2002 to December 2019. I find that subjective expectations of market risk premium 12 months ahead is a non-stationary, persistent process.

To analyze this dynamic, I examine the market for financial information by collecting granular data from two major sources: (1) almost one million equity research reports from Investext, representing 545 distinct information providers, from investment banks to online platforms, and (2) thousands of episodes of financial television programs aired across 42 local channels, archived in the Moving Image Archive and the Internet Archive TV News. I find that the aggregated monthly sentiment from "loud" information providers and key financial TV shows rather than experts exhibits a cointegrated relationship with retail investors' subjective expectations of the equity risk premium, indicating a long-term equilibrium between "loud" information supply and belief formation.

I test four hypothesis

Hypothesis 1. If every report has equal weight or equivalently, the sentiment of infor-

mation providers is weighted by the reporting activity of information providers, there is an association between subjective expected excess return and sentiment about the growth of companies' earnings.

Hypothesis 2. The association between subjective expected excess return and sentiment about earnings growth is stronger for reports of information providers with higher reporting activity.

Hypothesis 3 Information from less active providers is not related to the temporal fluctuations of aggregate stock market return.

Hypothesis 4 There is no effect of income and stock investement amount on the learning pattern of retail investors.

To test the relationship between the aggregated monthly sentiment of information providers and subjective expectations of market risk premium, I utilize a vector error correction model. This model is based on the theory of cointegrated processes developed by Johansen, 1991 and Engle and Granger, 1987 and is commonly used in asset pricing research (Campbell and Shiller, 1988a, Campbell and Shiller, 1988c [1] and many others).

First, I show that the association between subjective expected excess return and sentiment about earnings growth is present if every report has equal weight or equivalently, the sentiment of information providers is weighted by the reporting activity of information providers.

Second, I show that the association between subjective expected excess return and sentiment about earnings growth is strong for reports written by information providers with higher reporting activity, and absent for reports written by information providers with lower reporting activity. Moreover, my findings show that shock to provider-weighted sentiment of earnings reports is a permanent disturbance that has a long-run effect on subjective expectations. A positive value for this shock raises subjective expectations to a new level in

[1] Campbell and Shiller, 1988a and Campbell and Shiller, 1988c use vector autoregression to forecast returns (or dividend growth) with other variables including the log dividend-price ratio. Since they calculated expected returns from an econometric forecasting model, they were estimating the discount rates that would be applied to cash flow by an investor with rational expectations.

three months and accounts for fifteen percent of its fluctuations (the rest of the impact is the impact of the previous month's subjective expectations that propagate shock even further).

Third, I also demonstrate that a trading strategy based on past-month information from less active information providers may provide valuable insights. To test this, I use two types of analysis. Firstly, I conduct a predictive regression as in Cochrane, 2006. The results show that the change in sentiment of active information providers is highly significant, while the change in sentiment of less active information providers is not significant. Secondly, I analyze the profitability of a sentiment-based trading strategy constructed using information from more and less active information providers, following the methodology of Jegadeesh and Titman, 1993 and Moskowitz and Grinblatt, 1999. The results show that the trading strategy based on information from less active providers is profitable and outperforms the returns of the S&P 500 Index. These findings suggest that information from less active information providers may not be noise, but rather a valuable addition to an investor's information set.

Fourth, I show that retail investors' learning patterns are shaped by their income level and the magnitude of their stock investments. Lower-income investors typically form their subjective expectations of stock market returns from insights gleaned from popular sources, like Jim Cramer's show, as well as provider-weghted sentiment. Conversely, higher-income investors with below-average stock investments seem to gravitate more towards the insights presented on 'Squawk on the Street' that are related to past stock market returns. Meanwhile, investors that boast both high incomes and considerable stock investments generally make their decisions based on information from provider-weghted sentiment of equity reports.

The paper points out on the importance of information interpretation and dissemination: the perception of events' sentiment can be strongly influenced by the reporting activity of information providers, leading to disparities in expected sentiment values. The straightforward example provided below illustrates the mechanism. Initially, without the influence of reporting activity, the expected sentiment of two contrasting events, one positive (+1) and the other negative (-1), is neutral or zero, as the positive and negative sentiments offset each

other. However, when I introduce information providers, as infomration intermediaries, with varying reporting activities into the equation, the expected sentiment score can change dramatically. Consider a scenario where one provider issues a single report for each event, while another issues two reports for the positive event and none for the negative. To calculate the new expected sentiment score, it's crucial to consider not only the sentiment scores themselves but also the frequency of reports by each provider. In our example, the denominator becomes the total number of reports (4), and the numerator is the sum of each sentiment score multiplied by its reporting frequency. This calculation yields a new expected sentiment score of 0.5, a substantial deviation from the original expectation of zero. This deviation can be attributed to differences in the Data Generating Process (DGP) of events and reported events. The DGP of events reflects the raw sentiment frequency, whereas the DGP of reported events captures the frequency of sentiment as shaped by reporting activity. As a result, the positive sentiment, which is overemphasized from the DGP of events perspective, appears accurately weighted from the DGP of reported events perspective. Whether in the financial market, media industry, or social behavior studies, such dynamics can influence public perception, decision-making, and overall sentiment toward certain events or phenomena.

The innovation of this paper is documenting that the reporting activity of information providers has a significant impact on investors' expectations. In particular, the study shows that more active information providers have a stronger influence on investor sentiment than less active providers. As a result, investors may have distorted expectations of the asset market. This perspective is an important contribution to the existing literature on subjective expectations, which has traditionally focused on the role of past market returns and other macroeconomic variables. By emphasizing the role of information providers in shaping investors' expectations, this study provides a more nuanced view of the mechanisms underlying the formation of subjective expectations. Additionally, this perspective sheds light on the potential biases that can arise in investors' expectations. If investors rely too heavily

on a subset of information providers, they may be more susceptible to biases and errors in their expectations. This finding has important implications for policymakers and market participants who seek to improve the accuracy and efficiency of asset prices.

Overall, linking the activity of information providers on the information market to investors' subjective expectations on the stock market provides a valuable contribution to the literature on financial markets and information processing. It highlights the importance of considering the role of information providers in shaping financial market outcomes and suggests new avenues for future research in this area.

Related Literature This paper makes contribution to three areas of literature. Firstly, this paper contributes to the literature examining subjective expectations by further investigating and refining our understanding of the temporal nature of retail investors' subjective expectations. While aligning with Greenwood and Shleifer, 2014 finding that these expectations are time-varying, my results demonstrate that these expectations are also a non-stationary and persistent process, nuanced by the aggregated monthly sentiment of information providers. This adds a new dimension to the studies of Malmendier and Nagel, 2011 and Nagel and Xu, 2019, revealing that the information market's dynamics and the sentiment of its participants are instrumental in shaping the trajectory of these time-varying subjective expectations.

Secondly, this paper has a strong connection to the literature that employs cointegration analysis to pinpoint the primary factors driving fluctuations in financial markets. Within the context of asset pricing, Bansal and Yaron, 2004 utilize cointegration to illustrate how long-run risks can contribute to risk premia. Similarly, Bansal et al., 2005 apply cointegration to analyze the term structure of interest rates and its relationship with macroeconomic variables. Other scholars like L. Hansen et al., 2005 and Bansal et al., 2007 have used cointegration to investigate the relationship between expected returns and dividend growth rates, and the relationship between asset prices, economic activity, and news shocks, respectively.

Engle and Granger, 1987 seminal work underscores the importance of cointegration in forecasting. They argue that if two time series are cointegrated, then the deviation of one series from the other, termed as the error-correction variable, is stationary and can be used to forecast future movements in both series. Within the context of this paper, focusing on subjective expectations and sentiment of earnings reports, this theorem implies that the deviation of the level of subjective expectations from the sentiment of earnings reports can significantly influence both the prediction of future subjective expectations growth rates and the innovations in subjective expectations. Therefore, the application of cointegration analysis in this paper enables a more profound understanding of the relationship between subjective expectations and sentiment of earnings reports.

My research augments this strand of literature that applies cointegration techniques in financial markets. I do this by applying a vector error correction model to demonstrate a cointegrated relationship between the aggregated monthly sentiment of information providers and retail investors' subjective expectations of stock market risk premium. This approach reveals a new facet of how the information market impacts investor behavior.

Finally, the paper contributes to the growing body of literature investigating the relationship between asset markets and information markets, particularly from the perspective of retail investors. Over the past few decades, an increasing number of studies have been dedicated to understanding the nuances of this relationship. The pioneering works of empirical finance, such as Roll, 1988, proposed that news is an exogenous process influencing asset prices. This theory suggests that asset markets are reactive, responding to the flow of information as it occurs. However, as the field evolved, it became evident that this relationship is more complex.

Barber and Odean, 2001 shifted the focus to the individual investor's perspective, examining how investors trade based on the information from their personal experiences and the impact of this on their portfolio performance. This work emphasizes the importance of personal interpretation and the individual's capacity to process information. Veldkamp,

2005 further advanced our understanding by highlighting the endogenous nature of news and its interaction with asset markets. Rather than simply reacting to information, asset markets, through the actions of their participants, play a role in shaping the flow and interpretation of news. The role of information providers or intermediaries has also been a topic of interest. Fishman and Hagerty, 2019 explored the influence of news on stock prices, illustrating that both the tone and content of news can have a significant impact on stock returns. This work underscores the importance of information quality and the role of media in the financial market ecosystem. Meanwhile, Tetlock, 2011 scrutinized the accuracy of financial analysts' predictions. The study found that analysts with more expertise are better able to predict future stock prices, highlighting the importance of expert knowledge and experience in the realm of financial forecasting. The advent of digital platforms has brought a new dimension to the information-asset market nexus. **eaton˙green˙roseman˙wu˙2021** illustrated how social media can increase the flow of information to retail investors while simultaneously leading to an increase in market "noise" from retail investor trading. This insight suggests that modern communication platforms can both enhance and distort the information environment for retail investors.

My work bridges several areas of existing scholarship and presents new insights into how the information market influences retail investors' subjective expectations. While Greenwood and Shleifer, 2014 and Nagel and Xu, 2019 provide critical starting points, highlighting the time-varying nature of retail investors' expectations and their tendency to extrapolate recent market performances, their analyses do not fully explain the mechanisms underlying these patterns. In response to this gap in the literature, my research delves deeper into the dynamics of these subjective expectations, revealing them as a non-stationary, persistent process.

Further, I extend our understanding of the information market by scrutinizing a large corpus of equity reports from InvesText. The analysis of this granular data demonstrates that the aggregated monthly sentiment of information providers follows a persistent pattern

over time and, importantly, shares a cointegrated relationship with retail investors' subjective expectations of stock market risk premiums. This finding expands the existing knowledge, providing empirical support for the endogeneity of news to asset markets as suggested by Veldkamp, 2005.

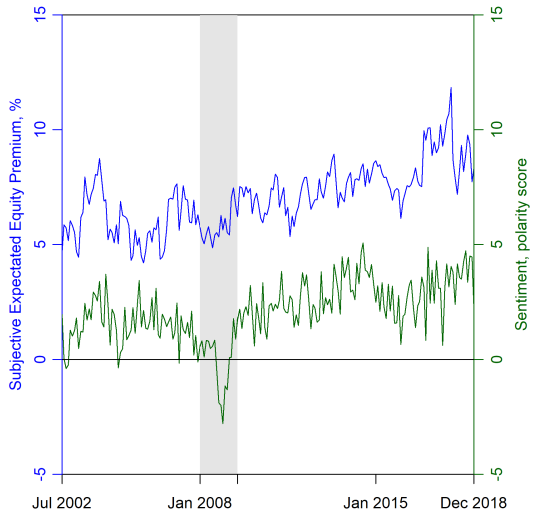
My research also advances our understanding of the differential impact of information providers based on their activity level, a topic not extensively addressed in existing literature. I show that retail investors' subjective expectations are more strongly influenced by information providers with higher reporting activity, underscoring the importance of information source prominence in shaping market perceptions. Moreover, my findings illuminate the potential value of less active information providers, suggesting that their contributions may not be mere noise but a valuable addition to an investor's information set, an insight that could have important implications for the design of investor strategies.

Moreover, I explore the socioeconomic dimensions of information processing in the asset market. My work reveals that income level plays a significant role in shaping the learning patterns of retail investors, with lower-income investors being more susceptible to the influences of highly active information providers. This finding sheds light on potential inequities in the access to and utilization of information in the asset market, prompting important questions about the distribution of resources and opportunities in financial markets.

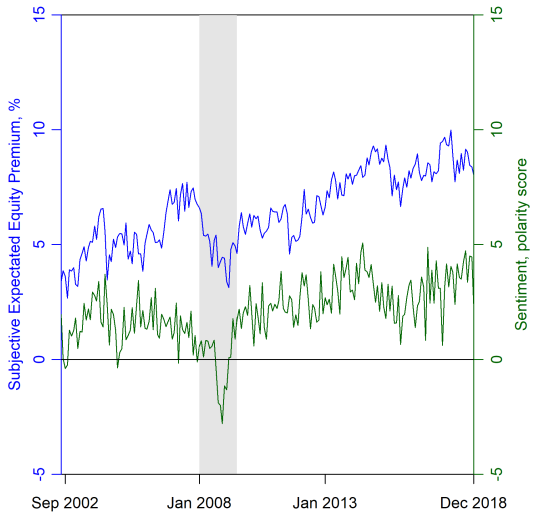
In sum, this paper extends our understanding of the interplay between the asset and information markets. It uncovers new dimensions of this relationship, and raises critical questions about equity and access in financial markets. This work paves the way for future research exploring the dynamics of investor behavior in the increasingly complex and digitized world of finance.

Figure 1. Subjective Equity Premium From Surveys & Provider-Weighted Aggregated Sentiment Of Analyst Reports

Green lines and green right y-axes on both plot correspond to aggregated sentiment of equity reports about 45 US blue-chip companies. Blue line and blue y-axis on the left plot show subjective equity premium from CCS survey. Blue line and blue y-axis on the right plot show subjective equity premium from MSC survey. Gray area is NBER recession.



(a) CCS Subjective Expected Equity Premium & s_t^w Sentiment



(b) MSC Subjective Expected Equity Premium & s_t^w Sentiment

II. Empirical Regularity

In this section, I illustrate an empirical regularity, a link between the retail investor's subjective expected equity premium and the aggregated sentiment reflected in equity reports about US blue-chip companies. The subsequent analysis brings into focus the potential role of sentiment expressed in these reports as an influencing factor in shaping retail investors' expectations about the equity premium.

Figure 1 shows the relationship between a retail investor's subjective expected equity premium and aggregated sentiment of 0.5 mln equity reports about 45 US blue-chip companies. The monthly subjective expected equity premium of a representative retail investor is obtained from CCS and MSC surveys, while the aggregated sentiment of equity reports is calculated from reports from 565 information providers.

The figure illustrates a strong co-movement between the monthly subjective expected equity premium of a representative retail investor and the aggregated sentiment of the equity

reports. It implies that the retail investor's expected equity premium could potentially be influenced by the collective sentiment of the equity reports.

This observation prompts further exploration and deeper understanding of the role of information providers in shaping retail investors' expectations and the mechanisms through which these sentiments are communicated and assimilated by retail investors.

III. Data And Measures

In this section, I present survey data on a retail investor's subjective expected equity premium, as well as a measure of expected equity premium from surveys. Additionally, I provide information on the reporting activity of information providers and outline measures of sentiment regarding earnings growth.

A. Asset Market: Subjective Expectations of Representative Retail Investor

In this subsection, I describe surveys employed to measure the subjective expected market risk premium. The primary resources for this purpose are two well-established surveys, the Michigan Survey of Consumers[2] and the Conference Board Consumer Confidence Survey[3]. These surveys are not only the most enduring monthly investigations in this field, but they are also widely recognized in economic and financial literature[4]. For additional context, I have added an Appendix that outlines eight other surveys, initiated by a range of central banks, the Federal Reserve system, and Vanguard. However, a notable limitation of these surveys is that they span a maximum period of ten years.

The Conference Board has been administering the Consumer Confidence Survey (CCS) consistently since 1967. Each month, roughly 5,000 individuals are selected to participate

[2] https://data.sca.isr.umich.edu/

[3] https://conference-board.org/data/consumerdata.cfm

[4] The surveys are used in Acemoglu and Scott, 1994; S. R. Baker and Fradkin, 2017; Barsky and Sims, 2012; Bram and Ludvigson, 1998; Carroll et al., 1994; Dees and Brinca, 2011; Lemmon and Portniaguina, 2006; Ludvigson, 2004; Matsusaka and Sbordone, 1995; Souleles, 2001, 2004; Throop, 2010 among amny others.

in the survey through mail. The response rate averages around 70%, yielding a substantial volume of completed questionnaires. The selection process for the sample adopts a balanced-quota design, where households are chosen based on distinct characteristics with the goal of making the sample representative of the broader population. The CCS primarily seeks respondents' opinions on expected changes in stock prices and interest rates for the upcoming twelve-month period. This set of questions has been incorporated since June 1987 and continues to this day. The collected responses to these questions are compiled in the Consumer Confidence Survey report, which can be accessed through a paid subscription.

The University of Michigan Surveys of Consumers (MSC) have been conducted consistently since 1958. It is monthly nationally representative survey based on approximately 500 telephone interviews of US households. The sample is randomly selected from a list of household telephone numbers, with about 70 percent of households responding. Responses are weighted based on Census strata to adjust for variation in the age and income distributions observed in monthly samples. Questionnaires focus on demographics, financial prospects for respondents, and the economy in general. I use questions covering the latter. To measure subjective expectations of market return, I focus on a survey question that asks respondents about the percent chance that a one thousand dollar investment in the stock market will increase in value in the year ahead. The question is available from June 2002 to current date. Answers are reported as a mean probability of increase in stock market in next year. The aggregated response data can be found in Table 20 "Probability of Increase in Stock Market in Next Year" in the Saving and Retirement section of the survey.[5] To measure subjective expectations of risk-free rate, I use a survey question that asks about expected direction of the change of interest rates for borrowing money a year ahead. The question is available from January 2008 to current date. The response data can be found in Table 31 "Expected Change in Interest Rates During The Next Year" in the Unemployment,

[5]Table 20 also contains and a coarse answers' distribution with brackets $\{0\%,\ 1-24\%,\ 25-49\%,\ 50\%,\ 51-74\%,\ 75-99\%,\ 100\%\}$ and number of "Do not Know" and "NA" responses. The SDA Customized Subset of Variables/Cases gives access to individual data and stratification weights. The name of the variable is $PSTK$ variable "Percent chance of investment increase in 1 year".

Interest Rates, Prices, Government Expectations section of the survey.[6]

The CCS survey aims to gauge expectations among the US population. Meanwhile, the MSC survey gathers responses from individuals who have invested over $10,000 in the stock market. In MSC samples, respondents with less than $10,000 in stock market investments or with no investments at all make up roughly 5% of the respondents each month.

To transform coarse probability estimates to point estimates of subjective expectation of market return, I follow Greenwood and Shleifer, 2014 and Nagel and Xu, 2019 methodology. They use UBS/Gallup surveys that contains point estimates of monthly subjective expectations of market return for the period between January 2000 and April 2003 and portfolio return expectations in percents from January 2000 to October 2007. The UBS/Gallup data is used to infer relationships between subjective expected probability of return growth and subjective return expectation on the UBS/Gallaup sample, and to impute the MSC and CCS percent return expectations in the longer sample from January 2000 to December 2019. The detailed description of the survey and the methodology is provided in Appendices A and B. In the case of risk-free rate, I use the Survey of Professional Forecasters[7], point estimates[8] as a benchmark for imputation of subjective expected risk-free rate[9] Appendix describes imputation procedure in details.

The subjective estimation of an equity risk premium $\tilde{r}_t$ a year ahead as the difference between a subjective expected return on the stock market a year ahead $E_t[\tilde{r}_{t\to t+12}]$ and a

[6]Table 20 also contains index, calculated as share of respondents who expect rate to go down minus share of respondents who expect rates to go up plus 100, and number of "Do not Know" and "NA" responses. The SDA Customized Subset of Variables/Cases gives access to individual data and stratification weights. The name of the variable is $RATEX$ variable "Interest Rates Up/Down Next Year".

[7]https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters

[8]For my analysis, I use the cross-sectional averages provided by the Federal Reserve Bank of Philadelphia. Specifically, I use the average of one, two, three and four quarter ahead forecast for the three-month Treasury bill rate. $(TBILL2+TBILL3+TBILL4+TBILL5)/4$ The respondents fill in these forecasts monthly. For the period from June 1987 to October 2020, average quarterly estimates are $\overline{TBILL2} = 3.03$, $\overline{TBILL3} = 3.10$, $\overline{TBILL4} = 3.17$, $\overline{TBILL5} = 3.28$ percent point.

[9]The correlation between SPF estimates and realized Market Yield on U.S. Treasury Securities at 1-Year Constant Maturity, Quoted on an Investment Basis, $GS1$, from Board of Governors of the Federal Reserve System (US), https://fred.stlouisfed.org/series/GS1 is 0.9935 in the sample from Jun 1987 to October 2020.

subjective expected risk-free rate over the next year $\tilde{y}_{t+1}$

$$\tilde{r}_t \equiv E_t[\tilde{r}_{t\to t+12} - \tilde{y}_{t\to t+12}] \tag{1}$$

In summary, this subsection outlins the approach employed for assessing the subjective expected market risk premium, focusing on two well-established surveys - the Michigan Survey of Consumers and the Conference Board Consumer Confidence Survey.

B. Market For Information: Monthly Sentiment About Earnings Growth

In this subsection, I provide a detailed overview of the data and the measurement techniques employed to gauge the sentiment of equity reports. I also show on how sentiment, captured from individual reports, are consolidated into an aggregated monthly time series.

I utilize the Thomson Reuters Embargoed Research Collection (InvesText) on the Mergent Online platform for this analysis. Mergent is part of the London Stock Exchange Group's Information Services Division. InvesText provides a vast, global repository of investment reports. This comprehensive collection includes both current and historical research reports on various companies, industries, products, and countries. These reports have been curated from over 1,700 brokerages, investment banks, and independent research firms. The database, starting from 1982, comprises more than 18 million reports.

I focus on equity reports[10] about 45 US blue-chip companies. These include diverse industry leaders such as Apple, recognized globally for its innovation in consumer electronics; JP Morgan Chase, a pillar of the financial sector; and the pharmaceutical giant Pfizer. Other renowned names on the list are IBM, a key player in technology services, and Exxon Mobil, one of the largest publicly traded oil and gas companies.[11]

[10]These reports are "COMPANY (EQUITY) REPORTS" type of reports in InvesText data base.

[11]The complete list of the companies is Alcoa, Apple, American International Group, Amgen, American Express, Boeing, Bank of America, Citigroup, Caterpillar, Salesforce, Cisco Systems, Chevron, Dow, DuPont, Walt Disney, General Electric, General Motors, Goldman Sachs, Home Depot, AlliedSignal (Honeywell), Hewlett Packard, IBM, Intel, International Paper Company, Johnson & Johnson, J.P. Morgan (JPMorgan Chase), Coca-Cola, Eastman Kodak, McDonald's, 3M Company (Minnesota Mining & Manufacturing),

Table I. InvesText. Meta Data Of Investment Report # BA_01012007-02282009

id	BA_01012007-02282009
Document Date	9/9/2008
Author	Suzanne H. Betts / Argus Research
information provider (Contributor)	Argus Research Corporation
Headline	BA: Machinist strike could cost Boeing $3 billion in sales
Language	English
Pages	6
Tickers	BA
Company Names	Boeing Co
Category	EQUITY
Countries	United States
Industries	Industrial Materials / Defense / Aerospace
Regions	North America
Subjects	
Report Styles	COMPANY (EQUITY) REPORTS

The choice of 45 companies is based on their presence in the Dow Jones Industrial Average Index from January 1, 2000, to December 31, 2019. The chosen 45 companies effectively represent the market. A synthetic market index, constructed as a weighted average of stock prices of these companies (weighted by their market capitalization), shares a three-year rolling correlation of about 80 percent with the S&P 500. This high correlation supports the assertion that these companies adequately represent the aggregate market dynamics.

My sample consists of meta data of 511,302 equity reports. Table I shows an example of a meta data of a report # BA_01012007-02282009. It consists of a document date, name of an analyst, name of the information provider, headline, language, number of pages, tickers, company names, category of report, countries, industries, report styles and a report ID in the InvesText.

Consider a report represented as $e_{d,f,c,ev}$, where $e \in \Theta$ symbolizes the text of a headline (editorial), $d \in D$ signifies a date, $f \in F$ denotes an information provider, $c \in C$ refers to a company listed in the Dow index, and ev stands for an event. Thus, the full set of reports,

Merck, Microsoft, Nike, Pfizer, Proctor & Gamble, Philip Morris, Raytheon (United Technologies), AT&T (SBC Communications), Travelers, UnitedHealth, Visa, Verizon, Walgreens Boots, Walmart, Exxon Mobil

Θ, can be defined as:

$$\Theta = \{e_{d,f,c,ev} = (d, f, c, ev) : d = \{\text{Dates}\}, \tag{2}$$

$$f = \{\text{Information Providers}\}, \tag{3}$$

$$c = \{\text{Companies}\}, \tag{4}$$

$$ev = \{\text{Events}\}\} \tag{5}$$

The set of dates D contains 6,945 dates d. There are 565 information providers f in the set F, 45 Dow companies c in the set C and up to 325,066 unique headlines (events) are in the set of event Ev. Table XX in the Appendix presents the top ten information providers. The list includes top investment banks with strong think tanks (for example, the Credit Suisse with the Credit Suisse Research Institute) and online platforms (for example, the Refinitiv StreetEvents that publishes verbatim representations of corporate and institutional events, the Trefis, an interactive financial online community structured around trends, forecasts and insights related to popular stocks in the US[12]).

Table II shows an example of set of reports Θ(Nike, Oct 2019) about Nike at October 2019. First three columns is the information form the data set, while Events section of the table illustrates a break down of events covered by information providers.

Table II show that panel of reports is sparse. No single information provider consistently publishes daily reports or reports about all listed events. Each provider, as shown in the example, exercises selectivity in their coverage. It's also noteworthy that the "10Q" event, which represents the issuance of a new 10-Q financial statement, is cumulative by nature. This means it potentially encompasses the other four events – the adoption of a new methodology for Return On Invested Capital ("ROIC"), the implications of the US-China trade dispute ("Tariffs"), Nike's acquisition of TraceMe ("M&A"), and the appointment of a new CEO ("CEO") – either in its current or subsequent reporting cycle. Therefore, the

[12]The start-up was founded in 2007 and joined the Thompson and Reuters platform later.

Table II. Example of Event Coverage: Headlines of Equity Reports About Nike Published at October 2019

The table presents headlines of equity reports regarding Nike that were published on weekdays in October 2019 and assigned non-zero sentiment scores. Five significant events are highlighted in the Events section of the table. These events include the issuance of a new 10-Q financial statement, the implementation of a new methodology for Return On Invested Capital (ROIC), the unfolding of the US-China trade dispute and associated tariffs, Nike's acquisition of TraceMe, and the appointment of a new CEO.

information provider, f	Date, d	Headline, e	Events, $Ev(Nike)$				
			10-Q	ROIC	Tariffs	M&A	CEO
Macquarie Research	Oct 7	Macquarie: NIKE (NKE US) (Outperform) - Oh I See... Your New ROIC		x			
	Oct 11	Macquarie: NIKE (NKE US) (Outperform) - 10Q: The Chessmaster	x				
	Oct 17	Macquarie: NIKE (NKE US) (Outperform) - Tracing the Digital Acquisitions				x	
	Oct 22	Macquarie: NIKE (NKE US) (Outperform) - Another "Digital Acquisition"					x
Oppenheimer	Oct 10	Product Innovation Still Fueling NKE	x				
	Oct 21	Trade Risks Abound, Leading Chains and Brands Still Managing Well			x		
Susquehanna Financial Group	Oct 8	Buy the Pullback; Adverse Impact on NKE from China's Battle with NBA Overblown			x		
Stock Traders Daily Research	Oct 15	Comprehensive Technical and Fundamental Analysis for NKE. This reports includes The Investment Rate, a macroeconomic leading indicator, and Market Analysis.	x				
Corporate Watchdog Reports	Oct 18	Watchdog Report: NKE - Red Flags and Warning Signs	x				
JPMorgan	Oct 22	NIKE, Inc. : "Win/Win" Hire w/ "Accelerate" ¿ Overhaul Opportunity; Mgmt Follow-Up Takes; Overweight					x
Piper Sandler Companies	Oct 23	Nike CEO Rotation Rounds Out Trifecta Of Athletic Leadership Changes This Week					x
Wedbush Securities	Oct 23	New CEO Has Large Sneakers to Fill as Company Dominates Consumer, Innovation					x

information landscape is not uniformly distributed, but instead varies significantly between different providers and across distinct events, emphasizing the complexity of interpreting and aggregating sentiment data.

Table II also shows that the headline of the report highlights the main event of the report and contains information about a information provider's sentiment on direction and strength of change in future earnings. For example, *"Macquarie: NIKE (NKE US) (Outperform) - Oh I See... Your New ROIC"* headline from Macquarie Research could be interpreted as positive, considering the optimistic language ('Outperform') and the mention of a new

ROIC methodology, while *"New CEO Has Large Sneakers to Fill as Company Dominates Consumer, Innovation"* headline from Wedbush Securities suggests mixed sentiments. It implies a challenge for the incoming CEO to meet high expectations but also emphasizes Nike's strong position in the market.

To measure the sentiment embedded in an editorial, I utilize the headline's textual sentiment. This involves the use of a polarity score to gauge the sentiment present in the headlines produced by information providers. Polarity, in this context, assesses the extent to which a text is negative or positive. For instance, a negative polarity score in an earnings report corresponds to declining earnings or negative earnings growth. Conversely, a positive polarity score in the same report signifies increasing earnings or positive earnings growth.

The usage of textual polarity is a common practice in behavioral finance, serving as a tool to examine the influence of sentiment on decision-making processes and market behaviors. Predominantly, two forms of sentiment are explored. The first is investor sentiment, defined as assumptions about future cash flows and investment risks unsupported by the current facts (M. Baker and Wurgler, 2007). The second is textual sentiment, which denotes the positivity or negativity level within texts. While investor sentiment encapsulates subjective judgments and behavioral attributes of investors, textual sentiment can encompass these aspects but also extends to the more objective reflections of circumstances within information providers and markets.

Measure of Sentiment of Headlines of Individual Reports I utilize the sentiment score calculation methodology outlined by Rinker, 2021, as described in their study. This method employs the Stanford Natural Language Processing (NLP) coreNLP annotation pipeline framework developed by Manning et al., 2014. The algorithm considers both the relative positions of words within a sentence and their context. This is achieved through a set of rules that were established based on a comprehensive training set of sentences, effectively making the algorithm pre-trained. To maintain objectivity in my analysis and

prevent any potential bias, I have opted not to train the algorithm on my own dataset. Instead, I rely on the initial pre-training to ensure the accurate quantification of sentiment scores in headlines[13]. Rinker, 2021 mephodology delivers fast, interpretable and accurate results. The algorithm $J : e_{d,f,c} \rightarrow s_{d,f,c}$ maps text of an editorial $e_{d,f,c}$ into a polarity score $s_{d,f,c} \in \mathbb{R}$

$$\forall e \in \Theta,\ J : e_{d,f,c,ev} \rightarrow s_{d,f,c,ev},\quad \text{such that}\ \ \forall e \in \Theta, \exists s \in \mathbb{R},\ J(s) = e \tag{6}$$

$J(\cdot)$ considers relations among words and accounts for relative position of words in a sentence.

$$s(e) = \frac{\sum_j f(w_j^+, w_j^-, \delta_j)}{\sqrt{n}} * 100 \tag{7}$$

where $f(\cdot)$ is a function of positions and scores of positive words w_j^+, negative words w_j^-, δ_j is a vector of polarity shifters (words-negators, words-amplifiers, words-deamplifiers and words that are adversative conjunctions), and n is a number of words and some punctuation signs in a headline.

Figure 2 shows an example of mapping from editorial $e_{d,f,c,ev}$ = *"Coca - Cola Co. : 2013 Should Be Better, But Not That Much Better; Lowering Estimates"* to a sentiment score $s_{d,f,c,ev} = -0.12$ using Rinker (2018) algorithm $J : e_{d,f,c,ev} \rightarrow s_{d,f,c,ev}$. In the first step, the algorithm finds positive and negative polarized words based on Loughran and McDonald, 2016 financial dictionary. The sentence contains two positive words "better" that weight

[13]Based on Wankhade et al., 2022, there are three approaches to measure polarity. The simplest one is a pre-trained rule-based model that uses a dictionary that assigns a predetermined sentiment score to each word and sums scores up. The rule-based solutions are fast, but has low accuracy. The naive bayes classifier uses conditional probabilities of each lexical feature occurring in either positive or negative text in the training data to arrive at the outcome. Naive Bayes model is used only if there is available complete data for training the model. As in my case, writing style is information provider type-specific, I have smaller training set for less active providers and bigger set for more active providers. The difference in size of the sets would affect the sentiment, so I don't use this method. A Deep Learning allows for processing data in a complex manner. A Long Short-Term Memory model, a type of Recurrent Neural Network, maps words positions in the sentence and polarity scores according to custom neural network (unsupervised learning) models. The method has high accuracy, but lacking interpretability, as the algorithm utilizes a set of hidden layers of cascading classification problems.

Figure 2. Simplified Example Of Polarity Score Calculation Based On Rinker (2018) Algorithm

Figure illustrates how Rinker, 2021 utilizes clusters of words around positive and negatives words to tune sentiment of a sentience. There are two polarized words in the sentence - "better" and "better". The polarized words are positive and have score $P = 1$ per word. Rinker, 2021 algorithm considers polarized words within clusters. A cluster consists of 5 words before a polarized word and two words after it. Underscored words shows words in clusters. A cluster score corrects P considering effect of amplifiers A, de-amplifiers D, negators neg and adversative conjunctions, b. A polarity score of a sentance is a sum of polarity scores of its clusters, $s_{d,f,c,ev}$

$$s_{d,f,c,ev} = cl_1 + cl_2, \quad \text{where} \quad cl_l = \frac{(1 + (A_l + D_l)) * P_l(-1)^{(2+\#\text{ of negators}_l)}}{\sqrt{n}}$$

and $A = 0.8 * (1 - \#\text{ of negators}_l) * \#\text{ of amplifires}_l + \mathbf{1}_{\#\text{ of adversative conjunctions}_l > 0}(1 + 0.25 * \#\text{ of adversative conjunctions}_l)$, $D = -0.8(\#\text{ of negators}_l * \#\text{ of amplifires}_l)$, n is a number of words in the sentence, including punctuation. So sentiment scores of the first and the second clusters are

$$s_{d,f,c,ev} = 0.28 - 0.40 = -0.12, \text{ where}$$

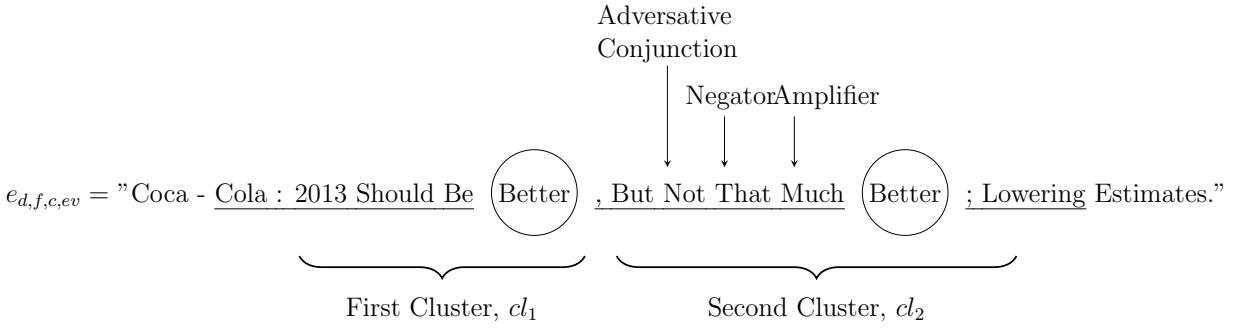
$$cl_1 = \frac{(1 + (0 + 0)) * P(-1)^{(2+0)}}{\sqrt{13}} = \frac{1}{\sqrt{13}} = 0.28, \text{ where } A_1 = 0.8 * (1 - 0) * 0 = 0$$

$$D_1 = -0.8 * (0 * 0 + 0) = 0$$

$$cl_2 = \frac{(1 + (1.25 - 0.8)) * P(-1)^{(2+1)}}{\sqrt{13}} = \frac{-1.45}{\sqrt{13}} = -0.40, \text{ where } A_2 = 0.8 * (1 - 1) * 1 + 1 + 0.25 * 1 = 1.25$$

$$D_2 = -0.8(1 * 1 + 0) = -0.8$$

The sentiment score for a sentence is a sum of scores of its clusters $s_{d,f,c,ev} = 0.28 - 0.40 = -0.12$



one point $P = 1$ each. Simplistic text processing stops here by assigning positive two as the sentiment score of the sentence. The Rinker, 2021 algorithm takes clusters that are formed around polarized words "better", cl_1 and cl_2, and utilises position and influence of polarity shifters (words-negators, words-amplifiers, words-deamplifiers and words that are adversative conjunctions) to tune the sentiment of polarized words. The first cluster cl_1, *"Co : Should Be Better"*, does not contain polarity shifters. Sentiment score of the first polarized words "better" is 1. The second cluster cl_2, *", But Not That Much Better; Lowering"*, contains

one adversative conjunction "but", one negator "not" and one amplifier "much" after the polarized word. Sentiment score of the second polarized words "better" is -1.45.

Figure 2 provides an illustration of how the editorial $e_{d,f,c,ev}$ = *"Coca - Cola Co. : 2013 Should Be Better, But Not That Much Better; Lowering Estimates"* is translated into a sentiment score $s_{d,f,c,ev} = -0.12$ using Rinker, 2021 algorithm $J : e_{d,f,c,ev} \rightarrow s_{d,f,c,ev}$. The initial step of the algorithm identifies words with positive and negative connotations, relying on Loughran and McDonald, 2016 financial dictionary. In this sentence, there are two positive words: "better," each contributing a point $P = 1$. A simple text analysis would conclude at this juncture, assigning a sentiment score of positive two to the sentence.

However, Rinker, 2021 algorithm delves deeper by identifying clusters formed around these polarized words, "better" - referred to as cl_1 and cl_2. It then adjusts the sentiment of these polarized words based on the position and impact of polarity shifters, which include negating words, amplifying words, de-amplifying words, and adversative conjunctions. The first cluster, cl_1 or *"Co : Should Be Better"*, does not have any polarity shifters, hence the sentiment score of the first occurrence of "better" remains as 1. The second cluster cl_2, *", But Not That Much Better; Lowering"*, includes an adversative conjunction "but", a negator "not", and an amplifier "much" following the polarized word. Therefore, the sentiment score of the second "better" is adjusted to -1.45.

When text is short and in form of sentences, Rinker, 2021 offers significant improvements over the more traditional "bag of words" approach[14] when mapping text to numerical values. On the sentence level, it provides a 4.6-fold increase in granularity of polarity score, with 2,281 unique polarity scores assigned using the Rinker's algorithm as compared to 495 unique scores using the "bag of words" approach. The detailed description of Rinker, 2021 algorithm is provided in the Appendix.

[14]The traditional "bag of words" approach calculates sentiment by summing the scores of positive and negative words, normalized by the square root of the total number of words and certain punctuation marks in a headline.

Measure of Expected Aggregate Market Earnings Growth. Given that the subjective expected equity premium of a representative investor is a monthly time series, and sentiment is a four-dimensional panel with day-information provider-company-event as the unit of observation, the sentiment data can be aggregated in various ways to test their association with the subjective equity premium of a retail investor.

I employ three strategies that use weighted averages to aggregate sentiments derived from individual reports. For the first strategy, the sentiment aggregation from individual reports is weighted according to each company's market capitalization. This means the sentiments linked to companies with larger market capitalizations carry more significance in the overall sentiment analysis. The second strategy involves placing equal emphasis on every information provider. This is achieved by calculating the average sentiment from each information provider and then deriving the arithmetic mean of these averages. Therefore, each information provider, regardless of their activity levels or the volume of reports they generate, is given equal weight. The third strategy involves calculating the arithmetic average of sentiments across all individual reports, thereby giving each report equal weight. This approach ensures that each report contributes equally to the overall sentiment, irrespective of the associated company or information provider. It is worth noting that this third aggregation strategy is mathematically equivalent to the process of weighting the average sentiment from each information provider by the number of reports they have produced. In other words, it is a weighted average where the weights correspond to the number of reports generated by each information provider.

For the first way of aggregation, I hypothesize that a signal about a company with a higher market capitalization, like Apple, will provide more information regarding the direction in which the aggregate stock market is moving than a signal about a company with a lower market capitalization, such as 3M Company. Indeed, the weighted average of the realized stock prices of 45 companies weighted by their market capitalization has 80 % three-year rolling correlation with the S&P 500. To aggregate the sentiment data according to this

hypothesis, I weight sentiment of by-company reports by number of report written about each company $\mathbf{e_{t,c}} \equiv \mathbf{s'_{t,c}n_{t,c}}$ for every company $c = \{c_1, c_2 \ldots, \}$ at month t. They form a vector $\mathbf{S_t^{w'}} = [e_{t,c_1},\ e_{t,c_2}, \ldots]$ at size $[1 \times m_t]$ of expected cross-news sentiment given a company, where m_t is a number of companies at time t. To aggregate the vector to one number, I use a vector of market capitalization of the companies $\mathbf{w}_t^C$ as weights

$$s_t^m \equiv \mathbf{S_t^{w'}w_t^C} = \begin{bmatrix} e_{t,c_1} & e_{t,c_2} & \ldots \end{bmatrix} \begin{bmatrix} w_t^1 \\ w_t^2 \\ \ldots \end{bmatrix} \tag{8}$$

to get monthly time series s_t^m of company-weighted aggregate sentiment scores, the measure that captures an economic structure of the market.

In the second way of aggregation, I hypothesize that if the market for information is populated by many atomistic information providers that deliver a homogeneous product, a random signal about the aggregate market performance, expectations over all by-information provider signals is equally important and should be informative. In this case, first, I look at the space of the reports through the lens of information providers, and next treat by-information provider signals as equally weighted. To aggregate the data according to this hypothesis, I weight sentiment of individual by-provider reports $\mathbf{e_{t,f}} \equiv \mathbf{s'_{t,f}n_{t,f}}$ for every information providers $f = \{f_1, f_2 \ldots, \}$ at month t, $\mathbf{S}'_t = [e_{t,f_1},\ e_{t,f_2}, \ldots]$, with n_t is a number of reporting providers at month t. To aggregate the vector to one number, I average the by-information provider sentiment scores,

$$s_t \equiv \frac{1}{n_t}\mathbf{S}'_t\mathbf{1} = \frac{1}{n_t}\begin{bmatrix} e_{t,f_1} & e_{t,f_2} & \ldots \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \ldots \end{bmatrix} \tag{9}$$

to get monthly time series s_t of equally information provider-weighted aggregate sentiment scores.

Table III. Toy Example of Sentiment Data
The table shows the number of reports per sentiment score, information provider and company per one month.

Company, c	information provider, f	Sentiment Score, s			Share of Reports
		-1	0	1	
Apple	J.P. Morgan	1			1/5
	Credit Suisse		2		2/5
Visa	J.P. Morgan		1		1/5
	Credit Suisse			1	1/5
Share of Reports		1/5	3/5	1/5	1

In the third way of aggregation, I hypothesize that if the market for information is populated by information providers with different activity levels that deliver heterogeneous products, a random signal about the aggregate market performance multiplied by an activity level of a information provider, expectations over aggregated by-information provider market signals weighted by the activity level should be informative. In this case, first, I look at the space of the reports through the lens of information providers, and then weight by-information provider signals by information providers' activity level. To aggregate the data according to this hypothesis, I weight sentiment of individual by-provider reports $\mathbf{S}_t^{'} = [e_{t,f_1},\ e_{t,f_2}, \ldots]$ for every $f = \{f_1, f_2 \ldots, \}$ at month t and weight these signals by the share of reports published by each information provider f in a given month t, $\mathbf{w_t^{f^{'}}} = [n_t^{f_1}\ \ n_t^{f_2}, \ldots]$

$$s_t^w = \mathbf{S}_t^{'}\mathbf{w_t^{f^{'}}} = \begin{bmatrix} e_{t,f_1} & e_{t,f_2} & \ldots \end{bmatrix} \begin{bmatrix} n_t^{f_1} \\ n_t^{f_2} \\ \ldots \end{bmatrix} \tag{10}$$

As shares of reports published are probabilities of a randomly picked report is published by a information provider f, $\mathbf{w_t^{f^{'}}} = [p(f_1)\ \ p(f_2), \ldots]$, the measure is equivalent to $E_F[E_{S|F}[p_{S|F}(s|f)]] = E[s]$ per month t, according to the data generating process on the market for information.

Table III offers a simplified example to explain how sentiment data is processed. This

"toy" dataset represents the number of reports per sentiment score (-1, 0, 1), information provider, and company, within a specific month. In this example, two information providers, J.P. Morgan and Credit Suisse, are considered, covering two companies, Apple and Visa. The specific breakdown of the reports is as follows: J.P. Morgan issues one report on Apple with a sentiment score of -1 and one on Visa with a sentiment score of 0. Credit Suisse, on the other hand, releases two reports on Apple with a sentiment score of 0 and one on Visa with a sentiment score of 1.

Before proceeding with the aggregation, let's calculate sentiment conditional on companies $e_{t,c}$ and sentiment conditional on information providers, $e_{t,f}$

$$e_{t,c_1}(\text{Apple}) = (-1) * \frac{1/5}{3/5} + 0 * \frac{2/5}{3/5} = -\frac{1}{3} \tag{11}$$

$$e_{t,c_2}(\text{Visa}) = 0 * \frac{1/5}{2/5} + 1 * \frac{1/5}{2/5} = \frac{1}{2} \tag{12}$$

$$e_{t,f_1}(\text{JPM}) = (-1) * \frac{1/5}{2/5} + 0 * \frac{1/5}{2/5} = -\frac{1}{2} \tag{13}$$

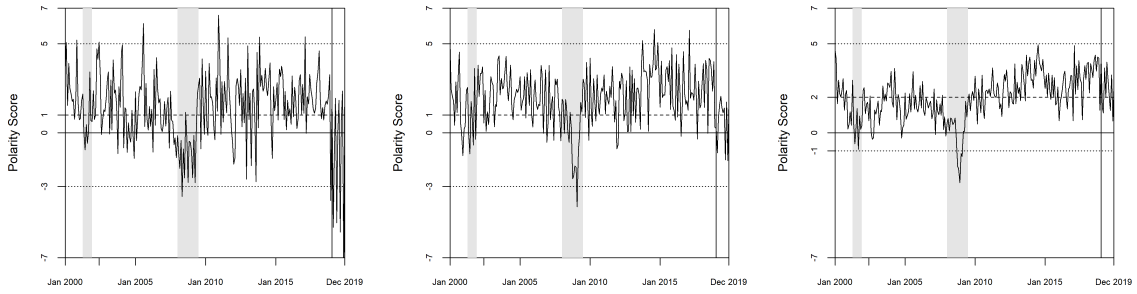
$$e_{t,f_2}(\text{CS}) = 0 * \frac{2/5}{3/5} + 1 * \frac{1/5}{3/5} = \frac{1}{3} \tag{14}$$

For the first aggregation approach, I calculate s_t^m as a weighted mean using the normalized market capitalization of companies as weights. Assuming that normalized market capitalization of Apple and Visa is $\mathbf{w}' = [w(\text{Apple}), w(\text{Visa})] = [2/3, 1/3]$ in a given month, $s_t^m = e_{t,c_1}(\text{Apple}) * w(\text{Apple}) + e_{t,c_2}(\text{Visa}) * w(\text{Visa})$,

$$s_t^m = \begin{bmatrix} e_{t,c_1}(\text{Apple}) & e_{t,c_2}(\text{Visa}) \end{bmatrix} \begin{bmatrix} w_t(\text{Apple}) \\ w_t(\text{Visa}) \end{bmatrix} = (-\frac{1}{3}) * \frac{2}{3} + \frac{1}{2} * \frac{1}{3} = -\frac{1}{18} \tag{15}$$

For the second aggregation approach, I calculate s_t, as ariphmetic average with of average

Figure 3. Aggregated Sentiment Score
Plot (a) shows company-weighted sentiment score, s_t^m. Plot (b) shows equally weighted by-information provider sentiment score, s_t . Plot (c) shows information provider-weighted by-information provider sentiment score, s_t^w per month. Dashed line on a plot is the mean of corresponding time series of sentiment scores, dotted lines are two standard deviations around the mean. Gray areas are NBER recessions.



(a) Company-Weighted **(b)** Equally-Weighted **(c)** Provider-Weighted

information providers' sentiment, $s_t = \frac{1}{2}\left(E[s|f{=}\text{JPM}] + E[s|f{=}\text{CS}]\right)$:

$$s_t = \frac{1}{n_t}\begin{bmatrix} e_{t,c_1}(JMP) & e_{t,c_2}(\text{CS}) \end{bmatrix} = \frac{1}{2}(-\frac{1}{2} + \frac{1}{3}) = -\frac{1}{12} \tag{16}$$

For the third aggregation approach, I calculate s_t^w as a weighted average using the percentage of published reports of an information provider as weights, $s_t^w = E\left[s|f{=}\text{JPM}\right] * p(f{=}\text{JMP}) + E\left[s|f{=}\text{CS}\right] * p(f{=}\text{CS})$

$$s_t^w \equiv \begin{bmatrix} e_{t,c_1}(\text{JMP}) & e_{t,c_1}(\text{CS}) \end{bmatrix} \begin{bmatrix} n_t^{\text{JMP}} \\ n_t^{\text{CS}} \end{bmatrix} = (-\frac{1}{2}) * \frac{2}{5} + \frac{1}{3} * \frac{3}{5} = 0 \tag{17}$$

Note that if partition $p(f = \text{JMP})$ and $p(f = \text{CS})$ is observed, the third case is equivalent to application of the law of total expectations and recovering expected value of sentiment[15].

Figure 3 shows the time series of the monthly aggregated sentiment score scores.

[15]Let the random variables X and Y, defined on the same probability space, assume a finite or countably infinite set of finite values. Assume that $E[X]$ is defined, that is, $\min(E[X_+], E[X_-]) < \infty$ $\min(E[X_+], E[X_-]) < \infty$. If $\{A_i\}$ is a partition of the probability space Ω, then

$$E(X) = \sum_i E(X \mid A_i)P(A_i)$$

Table IV. Descriptive Statistics

The table shows descriptive statistics of subjective expected equity premium over the next 12 months from MSC, $\tilde{r}_t^M$, and CCS, $\tilde{r}_t^C$, surveys, in percents. The realized and CCS data are from January 2000 to December 2019 and contain 240 observations. The MSC data is from June 2002 to January 2019 and contains 210 observations. The table also shows descriptive statistics of aggregate polarity score as company-weighted and information provider-weighted average with equal aggregation weights and number of published reports per information provider as an aggregation weight, scaled by 100, from January 2000 to December 2019. Sentiment scores are scaled by 100.

Variables	Mean	St. Dev.	Min	Max
Subjective Expected Equity Premium				
$\tilde{r}_t^M$	6.67	1.76	2.66	10.50
$\tilde{r}_t^C$	6.98	1.47	4.20	11.84
Sentiment Measures				
s_t^m	1.36	2.07	-8.05	6.60
s_t	1.93	1.60	-4.14	5.79
s_t^w	2.00	1.31	-2.78	4.94

C. Descriptive Statistics

Table (IV) provides descriptive statistics of subjective equity premium from the Michigan Survey of Consumers (MSC) and the Conference Board Consumer Survey (CCS). There is a sampling error of mean of subjective expected equity premium is 2.1 and 2.3 the MSC and CCS surveys. It indicates that mean of subjective expected annual equity premiums is statistically different from zero.

While mean subjective expected annualized equity premium is similar for MSC and CCS surveys, 6.67 % and 6.98 % per year, mean sentiment weighted by different weightings schemes differs. Mean sentiment weighted by number of providers reports is the highest, 2.00, while mean sentiment weighted by companies market capitalization s_t^m, is the lowest,

Proof.

$$E(E(X \mid Y)) = E\left[\sum_x x \cdot P(X{=}x \mid Y)\right] = \sum_y \left[\sum_x x \cdot P(X{=}x \mid Y{=}y)\right] \cdot P(Y{=}y)$$
$$= \sum_y \sum_x x \cdot P(X{=}x, Y{=}y) = E(X)$$

□

1.36.

Though mean subjective annualized equity premium is similar, subjective market price of risk in the sample, $\frac{\text{Mean}}{\text{St.Dev.}}$, differs. It is lower for MSC survey, 3.79, and higher for CCS survey expectations 4.51. Different standard deviation causes this difference.

I report two dispersion measures, standard deviation and values of maximum and minimum of variables. While standard deviation of subjective expectations time series is close, 1.76 versus 1.47, for MSC and CCS surveys, standard deviation of sentiment measures drops from 2.07 for sentiment weighted by companies market capitalization s_t^m to 1.31 for sentiment weighted by information providers activity s_t^m. Values of maximum and minimum values are olny positive for survey expectations. Sentiment weighted by companies market capitalization s_t^m has bigger in absolute value minimum, -8.05, than maximum, 6.60, while sentiment weighted by number of providers reports has the opposite, its minimum of -2.78 is almost two times lower in absolute value than maximum of 4.94.

IV. Empirical Strategy

In this section I describe results of stationarity tests, Elliott et al., 1996 and Zivot and Andrews, 1992 univariate unit root tests, as well as Johansen, 1988; Johansen, 1991 cointegration test. Considering test results, I introduce a vector error correction model that builds on the time series properties of the system of variables, aggregate sentiment scores of equity report headlines and subjective expectation of equity premium over the next 12 months.

A. Stationarity Tests

Figure (1) on page 10 shows that there is a visual upward drift in the monthly subjective expected equity premium and aggregated sentiment scores, so I start with stationarity tests.

As my sample includes one and a half business cycles[16] and contains noisy monthly data,

[16]Recession of 2008.

Table V. Elliott, Rothenberg and Stock (1996) and Zivot and Andrews (1992) Stationarity Tests

The Elliott et al., 1996 test tests level stationarity and has a regression specification

$$\Delta r_t^{e,d} = \pi r_{t-1}^{e,d} + \sum_1^p \phi_j \Delta r_{t-j}^{e,d} + \varepsilon_t$$

where $r_t^{e,d} = r_t^e - \beta_\phi^0 - \beta_\phi^1 t$ is GLS-detrended equity premium. The τ_τ is conventional t-statistics for the coefficient π testing the null hypothesis $H_0 : \pi = 0$ that series are non-stationary $I(1)$ with drift, versus an alternative $H_1 : |\pi| < 1$ is that time series are $I(0)$ with deterministic time trend. I use Schwert, 1988 to determine a number of lags.

The t_α statistics of Zivot and Andrews, 1992 tests trend stationarity with endogenous break λ and has a regression specification

$$r_t = \hat{\mu} + \hat{\alpha} r_{t-1} + \hat{\beta} t + \theta DT_t(\lambda) + \sum_i^p c_i \Delta r_{t-i} + \epsilon_t$$

where $DT_t(\lambda) = t - T(\lambda)$ if $t > T$ and 0 otherwise, and $p = 2$. The test has null hypothesis that the series has a unit root $I(1)$ without an exogenous structural break, against the alternative hypothesis that it is trend-stationary process with a one-time break in the trend occurring at an unknown point in time. The test statistics t_α estimates a break point that gives the most weight to the trend-stationary alternative.

	Test Statistics	
	τ_τ	t_α
Realized Return		
r_t^{1m}	-2.48	
Expected Subjective Equity Premium		
$\tilde{r}_t^M$	-2.38	
$\tilde{r}_t^C$	-2.39	
Sentiment Measures		
s_t^w	-3.14	-3.78
s_t^m	-2.62	
s_t	-3.16	-3.87

I use Elliott et al., 1996, ERS, stationarity test that has high asymptotic power for samples with a slow evolving trend and dominating random component. To account for a potential break in the trend function under the alternative hypothesis, I use Zivot and Andrews, 1992 , ZA, stationarity test.

Table (V) shows the values of the ERS statistics, τ_μ and the ZA statistics t_α for realized return and subjective expected premia from the MSC and CCS surveys, and sentiment time series. The τ_τ statistics of ERS test tests trend stationarity[17]. The statistics is more

[17]Critical values for ERS test are taken fromElliott et al., 1996 and equal to 3.48, -2.89 and -2.57 for 1%, 5% and 10% significance level. Decision rule is to reject $H0$ if tests statistics $<$ critical value.

than 5 % critical value of -2.89 for all time series except aggregated provider-weighted and equally-weighted sentiment score, so I cannot reject the null that they follow $I(1)$ process with drift[18] [19] For aggregated provider-weighted and equally-weighted sentiment score time series, I can reject the null and accept on an alternative that the variables are stationary $I(0)$ with deterministic trend[20]. The t_α statistics of ZA test test trend stationarity with an endogenous break in alternative hypothesis. The statistics is more than 5 % critical value of -4.42[21] for two sentiment measures. So, I cannot reject the null that the time series are $I(1)$ processes at 5% significance level. The potential break is at September 2000. As my MSC data starts at July 2002, I exclude the beginning of the sample from regressions, so the potential break does not affect the inference.

As I have persistent $I(1)$ variables in the sample, there might exist linearly independent vectors such that their linear combination is stationary, $I(0)$. To diagnose the presence of cointegrating relationships among variable, I run Johansen, 1991 test. The test shows that there exist one linearly independent vector.

In Tables VI, the results of the Johansen test for two set of variables, with subjective expectations from MSC survey $\{s_t^w,\ \tilde{r}_t^M\}$ and with subjective expectations from CCS survey

[18]The functional form of $I(1)$ time series with drift is

$$x_t = a + bt + x_{t-1} + \epsilon_t \tag{18}$$

[19]The persistence of realized return time series aligns with existing research and contributes to the ongoing debate about return predictability. Some scholars assert that expected returns contain a time-varying component, implying future return predictability (Campbell and Shiller, 1988b; Cochrane, 1991; Fama and French, 1988; Goetzmann and Jorion, 1993; Hodrick, 1992; Lettau and Ludvigson, 2001; Lewellen, 2004; Lustig and Van Nieuwerburgh, 2005; Menzly et al., 2004). However, others contend that such conclusions are debatable. They highlight that the relationship between financial ratios and future stock returns exhibits disconcerting features, including problematic inference due to extreme persistence of financial ratios (Ang and Bekaert, 2001; Ferson et al., 2003; Nelson and Kim, 1993; Stambaugh, 1999; Valkanov, 2003) and poor out-of-sample forecasting power (Bossaerts and Hillion, 1999; Goyal and Welch, 2003, 2004; Paye and Timmermann, 2003; Viceira, 1996).

[20]The functional form of $I(0)$ time series with deterministic time trend is

$$x_t = a + bt + \pi x_{t-1} + \epsilon_t, \quad \text{where} \quad |\pi| < 1 \tag{19}$$

[21]Critical values for a test specification with four lags in error correction term are $0.01 = -4.934$ $0.05 = -4.42$ $0.1 = -4.11$. Decision rule is to reject $H0$ if tests statistics $<$ critical value.

Table VI. Trace Statistics of Johansen Cointegration Test

The Trace test statistic is likelihood ratio test of the null hypothesis that the number of distinct cointegrating vectors is less than or equal to r against the alternative hypothesis of more than r cointegrating relations. The Trace test statistic, denoted $\lambda_{\text{trace}}(r)$, is given by

$$\lambda_{\text{trace}}(r) = -T \sum_{i=r+1}^{n} \ln(1 - \hat{\lambda}_i) \tag{20}$$

where T is the number of usable observations, $\hat{\lambda}_i$ is the ith largest canonical correlation between the $I(1)$ variables and their lagged first differences, and n is the number of variables in the system. The larger values of the test statistic provide stronger evidence against the null hypothesis.

	Variables	Test Statistics $\lambda_{\text{trace}}(r)$	
		$r = 0$	$r = 1$
CCS Survey	$(s_t^w, \tilde{r}_t^C)$	42.64	9.44
	$(s_t, \tilde{r}_t^C)$	64.21	10.88
	$(s_t^m, \tilde{r}_t^C)$	54.36	10.60
MSC Survey	$(s_t^w, \tilde{r}_t^M)$	37.36	5.56
	$(s_t, \tilde{r}_t^M)$	59.34	5.89
	$(s_t^m, \tilde{r}_t^M)$	47.06	5.88

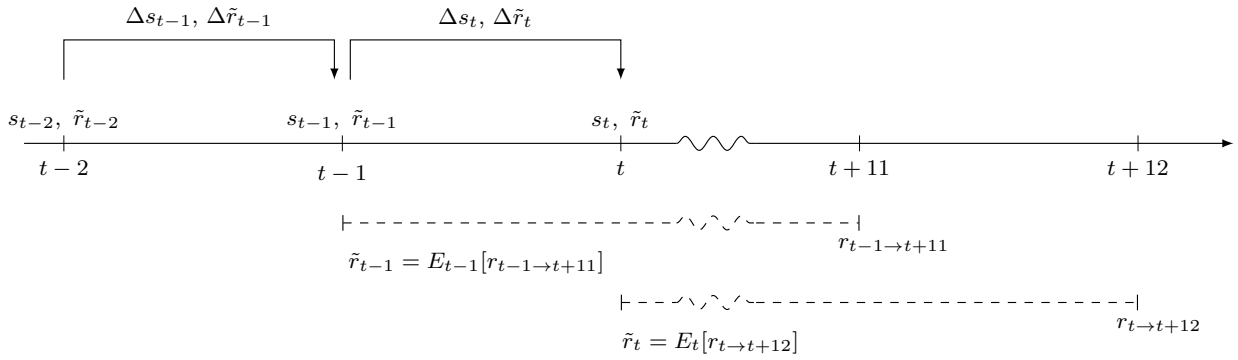
$\{s_t^w,\ \tilde{r}_t^C\}$, are given. Considering the statistic, the hypothesis of no cointegration can be rejected at the 1 % level for both sets. While the set with CCS survey expectations have one cointegration vector (tests statistics for rank $r = 1$, 9.44, is more that critical value of 8.18) for all sentiment measures at the 5 % level, tests for the set with MSC survey expectations yield contradictory conclusions about the cointegration rank. For this case, following Johansen and Juselius (1992), I study loading weights matrix of a cointegration vector, matrix α[22] The authors argued that if the values of $\alpha_{i,c}$ for $i = 1, 2$ at column c are close to zero, it is not significant. For systems with different surveys' expectations and aggregated sentiment scores, α matrices have elements in the second column (loading of the second cointegration vector) are close to zero (Table XXI with loading matrices is provided in Appendix). I conclude that there is one cointegration vector for systems with subjective

[22]Test is performed on an unrestricted VECM model of the form

$$\mathbf{\Delta Y_t} = \Gamma \mathbf{\Delta Y_{t-1}} + \alpha\beta^i \mathbf{\Delta Y_{t-2}} + \Phi D + \varepsilon_{\mathbf{t}} \tag{21}$$

where β is a matrix of coefficients of cointegration vectors and α is a matrix of the loading weights of cointegration vectors.

Figure 4. Timing of Variables
This figure shows the timing of variables at month t. s_t is an average sentiment score of equity reports published at month t. $\tilde{r}_t$ is a subjective expected equity premium from surveys sampled during month t about expected return 12 months ahead $\tilde{r}_t = E_t[r_{t\to t+12}]$.



expectations from both surveys.

B. Timing

Based on the survey methodology provided by the MSC and the CCS, subjective expectations issued at month t reflect expectations of respondents at month t. The MSC conducts its survey by phone throughout most of the month. Final figures for the full sample are subsequently made available at the end of the month and are not subject to further revision. The CCS mails questionnaires. The mailing is scheduled so that the questionnaires reach sample households on or about the first of each month. Returns flow in throughout the collection period, from first to last days of the month t, with the sample close-out for preliminary estimates occurring around the eighteenth of the month. Any returns received after then are used to produce the final estimates for the month, which are published with the release of the following month's data. The preliminary figures of subjective expectations from CCS are released on last Tuesday of month. Final figures are released with next month's release. I use final figures in my analysis.

As described in Data section, the sentiment s_t at month t is a weighted average of sentiments of individual reports published by information providers at month t. I estimate

that about 40 % of equity reports discuss financial statements[23] issued at month t, and about 60 % analyse impact of events on earnings. As events are randomly distributed within a month, the ones that occur at the end of the month might be reported at the beginning of the next month.

C. Misspecified Approach

This section presents a specification that ignore existence of cointegrating relationships between subjective expected equity premium,sentiment score and realized equity premium in the sample.

The OLS regression of subjective expected equity premium and sentiment score in differences is

$$\Delta\tilde{r}_t = a + b\Delta s_t + \epsilon_t \tag{22}$$

where $\Delta\tilde{r}_t$ is the difference in subjective expected equity premium calculated as $\tilde{r}_t - \tilde{r}_{t-1}$, and Δs_t is the difference in aggregated sentiment score calculated as $s_t - s_{t-1}$.

Table VII shows that no coefficient has statistical significance. Is this an indication that aggregate sentiment scores of equity report headlines are unrelated to subjective expectations of investors?

The Granger representation theorem of Engle and Granger, 1987 tells that if the levels are cointegrated then the data generation process has a representation as an error correction model (VECM). The VECM includes a lagged levels term but a regression VII in differences omits this term. This constrains the estimated coefficient on the lagged levels to be zero and also forces the estimated coefficients on the differenced regressors away from the values they would take if the model were correctly specified as VECM.

As the stationarity tests V and VI show that sentiment score s_t and realized equity

[23]The headlines that contain quarter number or a fiscal year.

Table VII. OLS Regression Specification
The regression specification is

$$\Delta\tilde{r}_t = a + b\Delta s_t + \epsilon_t \tag{23}$$

where $\Delta\tilde{r}_t$ is the difference in subjective expected equity premium calculated as $\tilde{r}_t - \tilde{r}_{t-1}$, and Δs_t is the difference in aggregated sentiment score calculated as $s_t - s_{t-1}$. I use three types of sentiment scores, equally weighted Δs_t, company-weighted Δs_t^m and provider-weighted Δs_t^w. I report standard errors in parentheses and the statistical significance is shown as *p<0.1, **p<0.05, and ***p<0.01.

	Dependent variable:					
	$\Delta\tilde{r}_t^M$	$\Delta\tilde{r}_t^C$	$\Delta\tilde{r}_t^M$	$\Delta\tilde{r}_t^C$	$\Delta\tilde{r}_t^M$	$\Delta\tilde{r}_t^C$
	(1)	(2)	(3)	(4)	(5)	(6)
Δs_t	0.01	−0.02				
	(0.02)	(0.03)				
Δs_t^m			0.004	0.01		
			(0.02)	(0.02)		
Δs_t^w					0.01	0.06
					(0.04)	(0.05)
Constant	0.02	0.01	0.02	0.02	0.02	0.01
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Observations	198	198	198	198	198	198
R^2	0.0003	0.002	0.0002	0.001	0.0005	0.01

premium r_t^e time series form cointegrating relationships, to account for the stochastic long-run equilibrium relation among variables, I use vector error correction model.

D. Vector Error Correction Model

In this section I describe mechanics of a vector error correction model. As my focus is not on particular contemporaneous coefficient estimates, but rather on how the variables respond to shocks dynamically, I also outline computing orthogonal impulse responses.

The VECM model specification is

$$\begin{cases} \Delta s_t = \underbrace{\alpha_1(s_{t-1} + \beta_1\tilde{r}_{t-1})}_{\text{long-term dynamics}} + \underbrace{\gamma_{ss}\Delta s_{t-1} + \gamma_{s\tilde{r}}\Delta\tilde{r}_{t-1}}_{\text{short-term dynamics}} + \phi_1 D + \varepsilon_{st} \\ \Delta\tilde{r}_t = \underbrace{\alpha_2(s_{t-1} + \beta_1\tilde{r}_{t-1})}_{\text{long-term dynamics}} + \underbrace{\gamma_{\tilde{r}s}\Delta s_{t-1} + \gamma_{\tilde{r}\tilde{r}}\Delta\tilde{r}_{t-1}}_{\text{short-term dynamics}} + \phi_2 D + \varepsilon_{\tilde{r}t} \end{cases} \tag{24}$$

To write the system in matrix form, let $\mathbf{\Delta Y}_t \equiv \begin{bmatrix} \Delta s_t & \Delta \tilde{r}_t \end{bmatrix}$, so

$$\mathbf{\Delta Y}_t = \alpha\beta^{'}\mathbf{Y}_{t-1} + \mathbf{\Gamma}\mathbf{\Delta Y}_{t-1} + \Phi D + \varepsilon_t \tag{25}$$

where $\alpha\beta^{'} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \begin{bmatrix} 1 & \beta_1 \end{bmatrix}$ captures long-run dynamics, while $\mathbf{\Gamma}$ is $\begin{bmatrix} \gamma_{ss} & \gamma_{s\tilde{r}} \\ \gamma_{\tilde{r}s} & \gamma_{\tilde{r}\tilde{r}} \end{bmatrix}$ matrix of coefficients that captures short-run effects. ΦD is a vector of constant terms, and $\varepsilon_t = [\varepsilon_{st}, \quad \varepsilon_{\tilde{r}t}]^{'}$ is a vector of stationary innovations, forecast errors of a variable conditional on observing its past values and the past values of sentiment and subjective expectations variables. $\varepsilon_{st} \sim N(0, \sigma_s^2)$ and $\varepsilon_{\tilde{r}t} \sim N(0, \sigma_{\tilde{r}}^2)$. As a general matter, ε_{st} and $\varepsilon_{\tilde{r}t}$ are correlated and its variance-covariance matrix of the system 25 Ω_ε has non-zero off-diagonal elements

$$\Omega_\varepsilon = E[\varepsilon_t \varepsilon_t^{'}] = \begin{bmatrix} \sigma_s^2 & \rho\sigma_s\sigma_{\tilde{r}} \\ \rho\sigma_s\sigma_{\tilde{r}} & \sigma_{\tilde{r}}^2 \end{bmatrix} \tag{26}$$

where ρ is a correlation between ε_{st} and $\varepsilon_{\tilde{r}t}$.

Whatever causes sentiment to rise (say a positive ε_{st}) would probably cause subjective expected equity premium to rise, too (so $\varepsilon_{\tilde{r}t}$ would also go up). Therefore, these innovations do not have a 'structural' interpretation[24].

Structural approach to the analysis presumes that the underlying driving forces of innovations are rooted in fundamental structural shocks. Let $u_t = [u_{st}, \quad u_{\tilde{r}t}]$ be "structural shocks", which are, by definition, uncorrelated with one another. Assume that there is a linear mapping between these structural shocks and the VECM system's 24 innovations:

$$\varepsilon_t = Bu_t \tag{27}$$

[24]Structural in the econometric sense means that they are mean zero and are uncorrelated with one another. Each is drawn from some distribution with known variance.

Taking the expectation of the outer product of the error vector with its transpose gives

$$E[\varepsilon_t \varepsilon_t^{'}] = BE[u_t u_t^{'}]B^{'} \tag{28}$$

Because structural shocks are uncorrelated, the off-diagonal elements of $E[u_t u_t^{'}]$ are zero. I can normalize the variance of each structural shock to be unity, which means that $E[u_t u_t^{'}] = I$. The above equation then becomes

$$\Omega_{\varepsilon} = BB^{'} \tag{29}$$

This is a system of equations that, without some assumptions, is under-determined. Indeed, there are four unique elements of B, but there are only three unique elements of Ω_{ε}, since a variance covariance matrix is symmetric. Hence, without imposing restrictions on B, it cannot be identified.

I use the most common restrictions - recursive restrictions, introduced by Sims, 1980. They impose timing assumptions - some shocks only affect some variables with a delay. Put differently, some of the elements of B are zero. Following macroeconomic literature, I employ a Cholesky decomposition of variance-covariance matrix as B.

Assume that to form subjective expectations, an investor reads reports published by information providers at time t about events that move the stock market.

$$s \rightarrow \tilde{r} \tag{30}$$

This would mean that the (1,2) element of B would be restricted to be zero. Given this restriction, the remaining elements of B is identified from the variance-covariance matrix of

residuals[25]

$$B = \begin{bmatrix} \sigma_s & 0 \\ \rho\sigma_{\tilde{r}} & \sigma_{\tilde{r}}\sqrt{1-\rho^2} \end{bmatrix} \tag{31}$$

where σ_s^2 is variance of ε_{st}, $\sigma_{\tilde{r}}^2$ is variance of $\varepsilon_{\tilde{r}t}$ and ρ is correlation of ε_{st} and $\varepsilon_{\tilde{r}t}$.

While $B(1,1)$ element of matrix B gives the standard deviation of the errors in equations explaining aggregated sentiment score dynamics, element $B(2,2)$ gives the conditional standard deviation of errors in the equation explaining subjective expected equity premium when the sentiment score errors are constant[26].

To calculate orthogonal responses of variables to shocks, I follow a standard two step procedure. First, I use Wold representation theorem, as in Wold, 1954, to write my VECM as vector moving average[27] model $MA(\infty)$. This step transform the model into a linear

[25]

$$\Omega_\varepsilon = \begin{bmatrix} \sigma_s^2 & \rho\sigma_s\sigma_{\tilde{r}} \\ \rho\sigma_s\sigma_{\tilde{r}} & \sigma_{\tilde{r}}^2 \end{bmatrix} = \begin{bmatrix} \sigma_s^2 & \rho\sigma_s\sigma_{\tilde{r}} \\ \rho\sigma_s\sigma_{\tilde{r}} & \rho^2\sigma_{\tilde{r}}^2 + (1-\rho^2)\sigma_{\tilde{r}}^2 \end{bmatrix} = \underbrace{\begin{bmatrix} \sigma_s & 0 \\ \rho\sigma_{\tilde{r}} & \sigma_{\tilde{r}}\sqrt{1-\rho^2} \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} \sigma_s & \rho\sigma_{\tilde{r}} \\ 0 & \sigma_{\tilde{r}}\sqrt{1-\rho^2} \end{bmatrix}}_{B'}$$

[26]What exactly the Cholesky decomposition does? if there is a one standard deviation shock to the sentiment, $\boldsymbol{\epsilon}_t = \begin{bmatrix} \sigma_s & 0 \end{bmatrix}'$, then $\mathbf{B}^{-1}$ will convert this shock into a vector

$$\mathbf{u_t} = \mathbf{B}^{-1}\boldsymbol{\epsilon}_t = \begin{bmatrix} \sigma_s^{-1} & 0 \\ -\rho(1-\rho^2)^{-\frac{1}{2}}\sigma_s^{-1} & (1-\rho^2)^{-\frac{1}{2}}\sigma_{\tilde{r}}^{-1} \end{bmatrix} \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -\rho\cdot(1-\rho^2)^{-\frac{1}{2}} \end{bmatrix}$$

The matrix $\mathbf{B}^{-1}$ rescales $\boldsymbol{\varepsilon}_t$ to have unit norm, $\mathrm{E}\big[\mathbf{B}^{-1}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top(\mathbf{B}^{-1})^\top\big] = \mathbf{I}$, and rotates the vector to account for the correlation ρ between ε_{st} and $\varepsilon_{\tilde{r}t}$. The rotation takes into account the correlation between ε_{st} and $\varepsilon_{\tilde{r}t}$, as matrix $\mathbf{B}^{-1}$ turns the shock $\boldsymbol{\varepsilon}_t = \begin{bmatrix} \sigma_s & 0 \end{bmatrix}^\top$ into a vector that is pointing 1 standard deviation in the s direction and $-\rho\cdot(1-\rho^2)^{-\frac{1}{2}}$ in the $\tilde{r}$ direction.

If there is a one standard deviation shock to the subjective expectaions, $\boldsymbol{\epsilon}_t = \begin{bmatrix} 0 & \sigma_{\tilde{r}} \end{bmatrix}'$, there is no response in the s direction and $(1-\rho^2)^{-\frac{1}{2}}$ response in $\tilde{r}$ direction

$$\mathbf{u_t} = \mathbf{B}^{-1}\boldsymbol{\epsilon}_t = \begin{bmatrix} \sigma_s^{-1} & 0 \\ -\rho(1-\rho^2)^{-\frac{1}{2}}\sigma_s^{-1} & (1-\rho^2)^{-\frac{1}{2}}\sigma_{\tilde{r}}^{-1} \end{bmatrix} \begin{bmatrix} 0 \\ \sigma_{\tilde{r}} \end{bmatrix} = \begin{bmatrix} 0 \\ (1-\rho^2)^{-\frac{1}{2}} \end{bmatrix}$$

[27]"Moving average" term is also used for the procedure of smoothing data with a running mean. A footnote in Pankratz, 1983, on page 48, says: "The label "moving average" is technically incorrect since the MA coefficients may be negative and may not sum to unity. This label is used by convention." Box and Jenkins, 1976 also says on page 10: "The name "moving average" is somewhat misleading because the weights ... need not total unity nor need that be positive. However, this nomenclature is in common use, and therefore we employ it."

combination of shocks.

$$\mathbf{Y_t} = \mathbf{D_0}\varepsilon_\mathbf{t} + \mathbf{D_1}\varepsilon_{\mathbf{t-1}} + \mathbf{D_2}\varepsilon_{\mathbf{t-2}} + \dots \tag{32}$$

where where $\mathbf{D_0}$, $\mathbf{D_1}$, $\mathbf{D_2}$ are coefficients of MA(∞). If I let $\mathbf{A}_1 = \mathbf{I} + \alpha\beta^{'} + \mathbf{\Gamma}$ and $\mathbf{A}_2 = -\mathbf{\Gamma}$,

$$\mathbf{D_0} = \mathbf{I} \tag{33}$$

$$\mathbf{D_1} = \mathbf{D_0}\mathbf{A_1} \tag{34}$$

$$\mathbf{D_2} = \mathbf{D_1}\mathbf{A_1} + \mathbf{D_0}\mathbf{A_2} \tag{35}$$

$$\dots \tag{36}$$

$$\mathbf{D_t} = \sum_{j=1}^{2} \mathbf{D_{t-j}}\mathbf{A_j} \quad for\ t = 1, 2, \dots \tag{37}$$

Second, following Sims, 1980 I use matrix B to orthgonalize the shocks in the MA(∞)[28]:

$$\mathbf{Y_t} = \mathbf{D_0}\mathbf{B}\mathbf{B}^{-\mathbf{1}}\varepsilon_\mathbf{t} + \mathbf{D_1}\mathbf{B}\mathbf{B}^{-\mathbf{1}}\varepsilon_{\mathbf{t-1}} + \mathbf{D_2}\mathbf{B}\mathbf{B}^{-\mathbf{1}}\varepsilon_{\mathbf{t-2}} + \dots \tag{39}$$

$$\equiv \mathbf{C_0}\mathbf{u_t} + \mathbf{C_1}\mathbf{u_{t-1}} + \mathbf{C_2}\mathbf{u_{t-2}} + \dots \tag{40}$$

where $u_t, u_{t-1}, u_{t-1}, \dots$ are i.i.d.

To find the impulse response function of $\tilde{r}_t$ to $u_{\tilde{s}t}$ over time, I set $u_{\tilde{s}t} = 1$. The impulse response on impact at $t = 0$ would be $C_0(2,1)$, the response after one period, at $t = 1$, would be $C_1(2,1)$, and so on. For example, orthonogalized impulse responses of $\tilde{r}_t$ to u_{st} in VECM with provider-weighted aggregated sentiment and CCS survey expectations is 0.14

[28]I denote a vector of variables as $\boldsymbol{Y}_t \equiv [s_t,\ \ \tilde{r}_t]$, and write the VECM system as VAR(2) model

$$\mathbf{\Delta Y}_t = \Gamma\mathbf{\Delta Y}_{t-1} + \alpha\beta^{'}\mathbf{Y}_{t-1} + \Phi D + \varepsilon_t \quad \Leftrightarrow \quad \mathbf{Y}_t = A_1\mathbf{Y}_{t-1} + A_2\mathbf{Y}_{t-2} + \Phi D + \varepsilon_t \tag{38}$$

where $A_1 \equiv \mathbf{I} + \Gamma + \alpha\beta^{'}$ and $A_2 \equiv -\Gamma$

at time $t = 0$, 0.17 at $t = 1$, 0.20 at $t = 2$

$$\mathbf{C_0} = \mathbf{D_0 B} = \begin{bmatrix} 0.92 & 0.00 \\ \textcircled{0.14} & 0.72 \end{bmatrix}$$

$$\mathbf{C_1} = \mathbf{D_1 B} = \begin{bmatrix} 0.30 & 0.11 \\ \textcircled{0.17} & 0.55 \end{bmatrix}$$

$$\mathbf{C_2} = \mathbf{D_2 B} = \begin{bmatrix} 0.44 & 0.21 \\ \textcircled{0.20} & 0.54 \end{bmatrix}$$

Mathematically $\mathbf{C_i}(2,1)$ elements of $\mathbf{C_i}$ matrices are:

$$\mathbf{C_0}(2,1) = \rho\sigma_{\tilde{r}} \tag{41}$$

$$\mathbf{C_1}(2,1) = \sigma_s(\alpha_2 + \gamma_{\tilde{r}s}) + \rho\sigma_{\tilde{r}}(\alpha_2\beta_1 + 1 + \gamma_{\tilde{r}\tilde{r}}) \tag{42}$$

$$\mathbf{C_2}(2,1) = \sigma_s\Big((\alpha_2 + \gamma_{\tilde{r}s})(\alpha_2\beta_1 + 1 + \gamma_{\tilde{r}\tilde{r}}) - \gamma_{\tilde{r}s} + (\alpha_2 + \gamma_{\tilde{r}s})(\alpha_1 + \gamma_{ss} + 1)\Big) + \tag{43}$$

$$+ \rho\sigma_{\tilde{r}}\Big((\alpha_2\beta_1 + 1 + \gamma_{\tilde{r}\tilde{r}})^2 - \gamma_{\tilde{r}\tilde{r}} + (\alpha_2 + \gamma_{\tilde{r}s})(\gamma_{s\tilde{r}} + \alpha_1\beta_1)\Big) \tag{44}$$

Dynamic orthogonal impusle responses $\mathbf{C_i}(2,1)$ are weighted sums of standard deviation of innovations in VECM subjective expectation equation, $\sigma_{\tilde{r}}$ and standard deviation of innovations in VECM sentiment equation, σ_s. Both, coefficeints that account for long-run dynamics, α_1, α_2, β_1 and short-term fluctuations, γ_{ss}, $\gamma_{s\tilde{r}}$, $\gamma_{\tilde{r}s}$, $\gamma_{\tilde{r}\tilde{r}}$, contributes to orphogonal impulse response of $\tilde{r}$ to shock in s_t.

E. Intuition

For illustarative purposes, following Stock and Watson, 1993, I can write the cointegrated system (24) with restrictions (31) in block triangular form. Following Stock and Watson, 1993, I focus only on long-term relationships betwenn variables, and ignore short-tern di-

namics,

$$\tilde{r}_t = \beta_1 s_t + e_{\tilde{r}t}, \quad \text{where } e_{\tilde{r}t} = \alpha_2 e_{st} + \varepsilon_{\tilde{r}t} \tag{45}$$

$$\Delta s_t = e_{st}, \quad \text{where } e_{st} = \sum_{j=0}^{\infty} \theta_0 \varepsilon_{st-1} + \theta_{t-2}\varepsilon_{st-2} + \theta_{t-3}\varepsilon_{st-3} + \dots \tag{46}$$

where innovations $\varepsilon_{\tilde{r}t} \sim I(0)$ and $\varepsilon_{st} \sim I(0)$ are from VECM model (24) without short-term dynamics, β_1 is a coefficient of conintegration vector in (24), and α_2 is a coeffitient of equilibrium adjustment vector in (24).

The first equation is $\tilde{r}_t = \beta_1 s_t + e_{\tilde{r}t}$ describes the long-term equilibrium relationship between $\tilde{r}_t$ and s_t. The parameter β_1 is the cointegration coefficient and measures how changes in s_t are associated with changes in $\tilde{r}_t$ in the long run. The term $e_{\tilde{r}t}$ is a disturbance term that captures short-term deviations from the long-term equilibrium. It is composed of two parts: $\alpha_2 e_{st}$ and $\varepsilon_{\tilde{r}t}$. The first part, $\alpha_2 e_{st}$, represents the error correction mechanism that adjusts $\tilde{r}$ towards its long-term equilibrium relationship with s_t. If e_{st} is positive, it means s_t is above its equilibrium value given $\tilde{r}$, and hence $\tilde{r}$ needs to increase to restore equilibrium. The adjustment is proportional to α_2. The second part, $\varepsilon_{\tilde{r}t}$ is an independent shock to $\tilde{r}$ that has nothing to do with s_t.

The second equation is $\Delta s_t = e_{st}$, which represents the change in s_t as a function of the disturbance term e_{st}. The disturbance term e_{st} is given by a sum of the past innovations ε_{st-j} weighted by the parameters θ_{t-j}. This equation captures the dynamics of s_t and how its changes are influenced by past shocks.

It's important to note that $\varepsilon_{\tilde{r}t}$ depends on β_1 and, if β_1 is not equal to zero, so does e_{st}. This means that the short-term dynamics and the long-term equilibrium of the system are interconnected. A change in sentiment s_t (for example, due to an innovation ε_{st}) not only directly affects s_t itself but also induces a change in the subjective expectations $\tilde{r}_t$ to restore the long-term equilibrium relationship.

The main economic intuition behind this model is that there are both long-term and

short-term forces at work in this economic system. The long-term forces maintain a stable relationship between $\tilde{r}_t$ and s_t, while the short-term forces cause temporary deviations from this equilibrium. The system constantly adjusts to these deviations, moving towards the long-term equilibrium.

V. Findings

In this section I discuss the tests conducted and the evidence garnered in support of the four hypotheses that are stated in the introduction section.

Hypothesis 1 If every report has equal weight or equivalently, the sentiment of information providers is weighted by the reporting activity of information providers, there is an association between subjective expected excess return and sentiment about the growth of companies' earnings.

Utilizing the findings from the VECM specifications presented in Tables XXII and XXIII, Figure 5 displays the orthogonal impulse response functions (OIRFs) along with confidence intervals[29]. These functions illustrate the effect of a one standard deviation shock in sentiment on the subjective expected equity premium. This is demonstrated across three sentiment score categories: information provider-weighted (s_t^w), equally-weighted (s_t), and company-weighted (s_t^m) sentiment scores.

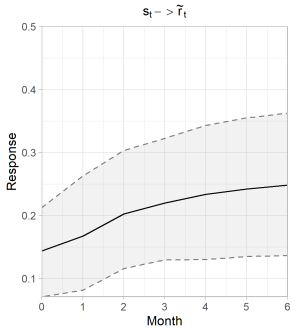
Figure 5 demonstrates that the information provider-weighted sentiment score has a significant impact on the subjective expected equity premium. The top two plots, (a) and (b), indicate that the adjustment of subjective expectations in response to a provider-weighted sentiment shock is both immediate and persistent. At the time of impact, $t = 0$, of one standard deviation shock in sentiment, subjective expectations grows 10 (MSC survey) to 15 basis points (CCS survey). The impact reaches 20-23 basis points in three months after

[29]Confidence intervals are represented as $CI_s = [s_{a/2}, s_{1-a/2}]$, where $s_{a/2}$ and $s_{1-a/2}$ correspond to the $a/2$ and $1 - a/2$ quantiles of the bootstrap distribution of orthogonalized coefficients C_τ in the MA(∞) representation of the VECM.

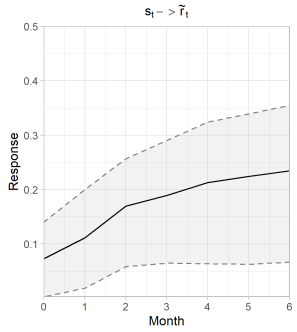
Figure 5. Hypothesis 1. Orthogonal Impulse Responses, in %

OIR of one standard deviation sentiment shock, s_t to subjective sxpectations, $\tilde{r}_t$ The sentiment score s_t is scaled by 100. 95 % confidence interval for the bootstrapped errors bands. 100 runs. Confidence intervals are represented as $CI_s = [s_{a/2}, s_{1-a/2}]$, where $s_{a/2}$ and $s_{1-a/2}$ correspond to the $a/2$ and $1-a/2$ quantiles of the bootstrap distribution of orthogonalized coefficients C_τ in the MA(∞) representation of the VECM.

Information Provider-weighted Aggregated Sentiment, $s_t^w \rightarrow \tilde{r}_t$

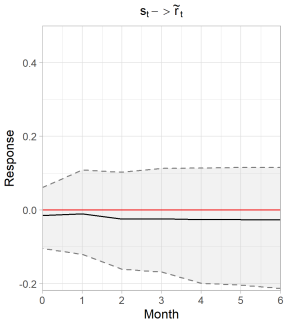


(a) CCS Survey

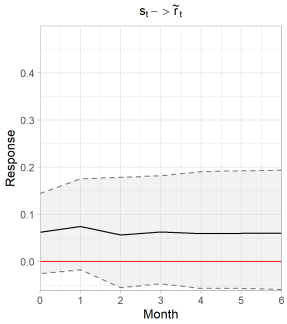


(b) MSC Survey

Equally-weighted Aggregated Sentiment, $s_t \rightarrow \tilde{r}_t$

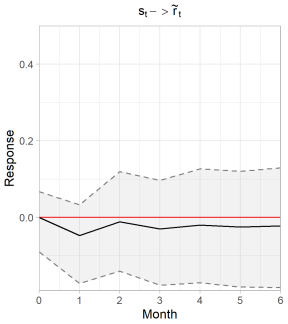


(c) CCS Survey

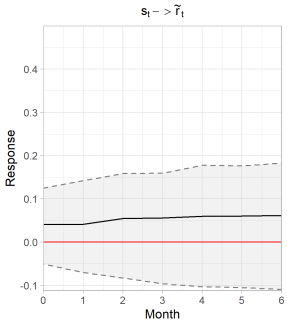


(d) MSC Survey

Company-weighted Aggregated Sentiment, $s_t^m \rightarrow \tilde{r}_t$



(e) CCS Survey



(f) MSC Survey

the shock.

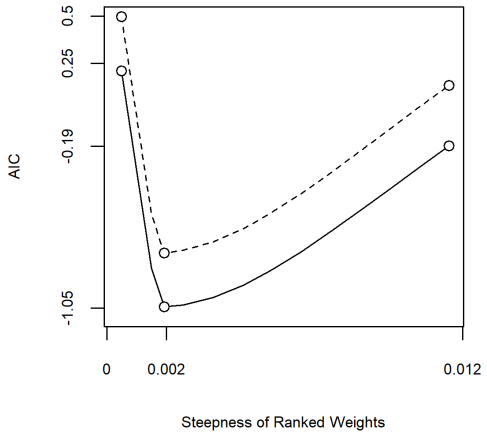
However, when subjective expectations are weighted by a company's market capitalization, there is no discernible reaction to the aggregated sentiment score shock, as evidenced by plots (e) and (f).

Figure 6. Steepness of Weighting $\mathbf{W_t^c}$ Vs. Lower Boundary of VECM AIC and Steepness of Weighting $\mathbf{W_t^c}$ Vs. Adjusted R^2 of Subjective Expectations Equation in VECM$(S_t, \tilde{r}_t)$

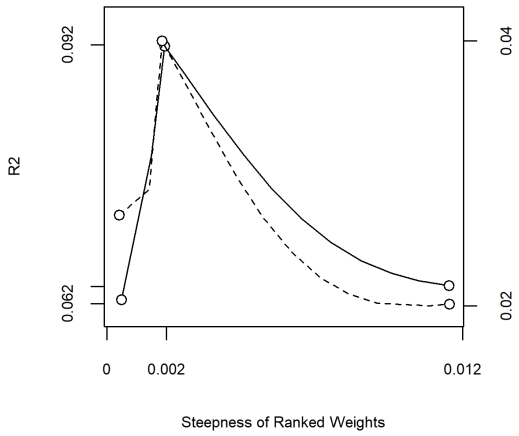
The X-axis represents the 'steepness' of a reparametrization scheme's weights. It's calculated as $\bar{sl} = \alpha \bar{sl(s_t)} + \beta \bar{sl(s_t^w)} + \gamma \bar{sl(s_t^e)}$, where $\bar{sl(s_t)}$, $\bar{sl(s_t^w)}$, and $\bar{sl(s_t^e)}$ are the slopes of individual weighting schemes. Each slope is determined by the formula $sl = \frac{1}{n_t}(\max w_t - \min w_t)$, where $\max w_t$ is the maximum weight in the scheme, $\min w_t$ is the minimum weight in the scheme, and n_t is the number of news items per month t.

The right plot visualizes the lowest values within the set of AIC (Akaike Information Criterion) of VECM (Vector Error Correction Model) models. In contrast, the left plot displays the highest values within the set of adjusted R^2 for the subjective expectations equation in VECM$(S_t^e, \tilde{r}_t)$. The solid black line represents specifications with subjective expectations from the MSC survey, while the dashed black line indicates those from the CBS survey. The left y-axis corresponds to the AIC of the subjective expectations equation from the MSC survey, whereas the right y-axis pertains to the AIC of the subjective expectations equation from the CBS survey.

A 'steepness' of 0 denotes an "equally-weighted" scheme. A 'steepness' of 0.002 signifies a weighting scheme that allocates 100% of the weight to provider-weighted sentiment. Lastly, a 'steepness' of 0.008 refers to a weighting scheme that gives 100% of the weight to exponentially-weighted sentiment.



(a) AIC



(b) Adj. R^2 of Subj.Exp. Equation

Are there mixed aggregation schemes that dominate by-information provider weighting? To answer the question, I construct 5,027 convex combinations of equally weighted $s_t = \frac{1}{n_t}\mathbf{1}\mathbf{S_t}'$, where $\mathbf{S_t}'$ is a vector of average sentiment of n_t information providers at month t, information provider-weighted $s_t^w = \mathbf{w_t^f}\mathbf{S_t}'$ and company-weighted $s_t^m = \mathbf{w_t^c}\mathbf{S_t}^{c,\prime} =$

$\mathbf{w_t^{c,adj}S_t}'$ sentiment measures, S_t^c,

$$S_t^c = \alpha s_t + \beta s_t^w + \gamma s_t^m = \mathbf{W_t^c S_t}' \tag{47}$$

$$\text{such that } \alpha + \beta + \gamma = 1, \tag{48}$$

$$\mathbf{W_t^c} = \alpha \frac{1}{n_t}\mathbf{1} + \beta \mathbf{w_t^f} + \gamma \mathbf{w_t^{c,adj}} \tag{49}$$

$$0 \leq \alpha \leq 1,\ 0 \leq \beta \leq 1,\ 0 \leq \gamma \leq 1 \tag{50}$$

where α, β and γ are shares that form the convex combination of sentiment weighting schemes[30] [31]

To characterize these big vectors of weights $\mathbf{W_t^c}$ for the ease of illustration of the results, I use "steepness of weights" $\bar{sl}$ of weighting schemes $\mathbf{W_t^c}$ to display results. "Steepness of weights" refers to the degree of disparity or differentiation between the weights assigned to different information providers within each weighting scheme $\mathbf{W_t^c}$. A steep weighting scheme assigns much higher weights to some information providers or companies and much lower weights to another ones. In contrast, a flat weighting scheme would assign similar weights to all information providers or companies. The steepness of a weighting scheme implicitly

[30] A total of 5,027 convex combinations are created using the following method. First, I generate three grids of weights, α, β, and γ, which range from 0 to 1 in increments of 0.01.

$$\alpha = 0, 0.01, 0.02, ..., 1\ \beta = 0, 0.01, 0.02, ..., 1\ \gamma = 0, 0.01, 0.02, ..., 1$$

Next, I create a data frame containing all possible combinations of these values and retain only those combinations where the sum of α, β, and γ is equal to one.

$$\alpha + \beta + \gamma = 1$$

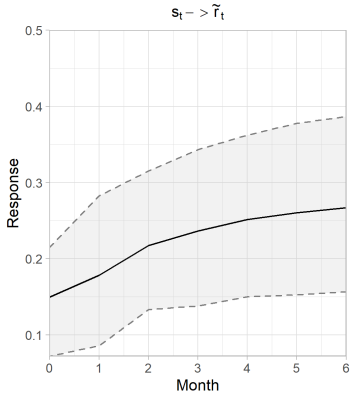
[31] As my data is in news-information provider-company-date granularity, I can map company-driven weights $\mathbf{w_t^c}$ to company-driven weights per information provider, $\mathbf{w_t^{c,adj}}$, $s_t^m = \mathbf{w_t^{c,adj}S_t}'$.

To illustrate the transformation, let's look at information providers A and B that both write reports news about companies 1 and 2. Information provider A writes three news articles about company 1 and four news acticles about company 2, $3s_1^A$, $4s_2^A$, while information provider B writes one news article about company1 and one news article about company 2, s_1^B, s_2^B. Market capitalization of company 1 is X and company 2 is Y, in percents of total market capitalization. Company-weighted sentiment is $s_t^c = \mathbf{w_t^c S_t}^{c,'} = (3s_1^A + s_1^B)\frac{1}{4}X + (4s_2^A + s_2^B)\frac{1}{5}Y$. If I want to map company-weighted weights per information provider, I can rearrange the equation as follows $s_t^m = \frac{7}{7}(3s_1^A\frac{1}{4}X + 4s_2^A\frac{1}{5}Y) + \frac{2}{2}(s_1^B\frac{1}{4}X + s_2^B\frac{1}{5}Y)$, where 7 is a number of reports written by information provider A and 2 is a number of reports written by information provider B.

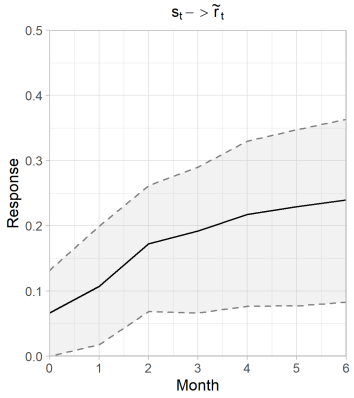
Figure 7. Orthogonal Impulse Response Functions of VECM With Sentiment of More and Less Active Providers

The sentiment score s_t is scaled by 100. 95 % confidence interval for the bootstrapped errors bands. 100 runs.

Information Provider-weighted Sentiment, More Active Providers, $s_t^{w,5} \rightarrow \tilde{r}_t$

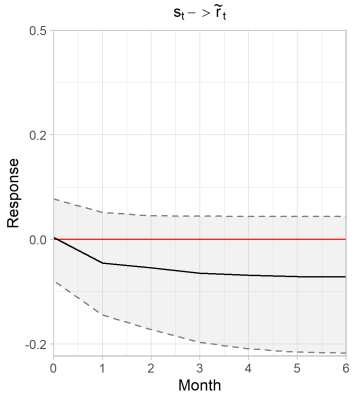


(a) CCS Survey

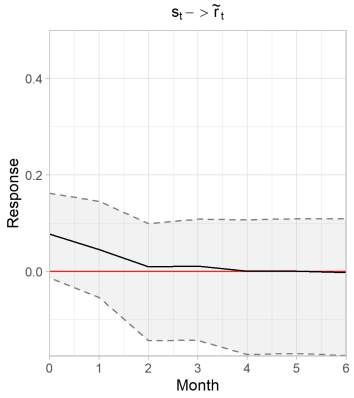


(b) MSC Survey

Information Provider-weighted Sentiment, Less Active Providers $s_t^{w,1:4} \rightarrow \tilde{r}_t$



(c) CCS Survey



(d) MSC Survey

tracks the degree of bias towards the observations that are given higher weights.

Figure 6 shows "steepness of weights" $\bar{sl}$ of $\mathbf{W_t^c}$ versus lower convex hull of AIC of VECM$(S_t^c, \tilde{r}_t)$ and upper boundary of the set of adjusted R^2 of subjective expectations equation in VECM$(S_t^c, \tilde{r}_t)$. Plots illustrate that specification with the minimum AIC information criterion and maximum adjusted R^2 of subjective expectation equation corresponds to the weighting scheme with steepness of weights $\bar{sl} = 0.002$. This steepness corresponds to the convex combination $\{\alpha, \beta, \gamma\} = \{0, 1, 0\}$ or 100 % provider-weighted sentiment.

Table VIII. Descriptive Statistics: Monthly Provider-Weighted Average Sentiment Score By Type of Provider

The table shows descriptive statistics of individual and aggregate polarity score, scaled by 100, from January 2000 to December 2019.

Variables	N	Mean	St. Dev.	Min	Max
Cross-News Panel					
Sentiment of Less Active Providers	21,120	0.35	19.40	-170.05	161.00
Sentiment of More Active Providers	489,360	1.99	20.10	-203.52	192.43
Aggregated Monthly Time Series					
Sentiment of Less Active Providers	240	0.58	3.3	-13.57	11.24
Sentiment of More Active Providers	240	2.09	1.37	-3.1	5.16

Hypothesis 2 The association between subjective expected excess return and sentiment about earnings growth is stronger for reports of information providers with higher reporting activity.

To test the hypothesis, I sort information providers by number of published reports over the sample period, use the top fifth quantile (top 20%) as more active information providers and first to forth quantiles as less active information providers. As information provider-weighted sentiment has the highest association with the subjective expected equity premium, I use information provider-weighted sentiment for building the two sentiment time series. The description of quantiles is provided in Appendix.

Table VIII shows descriptive statistics for individual and monthly time-series of aggregated provider-weighted sentiment for two types of information providers. The more active providers have 3.6 times higher monthly mean sentiment score and 2.4 lower standard deviatin.

I run the same VECM model using sentiment of more and less active providers weighted by information provider's activity, $s_t^{w,5}$ and $s_t^{w,1:4}$.

Tables XXV and XXIV show sentiment of less active providers $s_t^{w,1:4}$ has no association with subjective expected equity premium for both surveys. Sentiment of more active providers $s_t^{w,5}$ has highly significant long-term impact though cointegration vector and weakly significant short-term impact on subjective expected equity premium.

Table IX shows that adding lagged realized return increases AIC of VECM considerably.

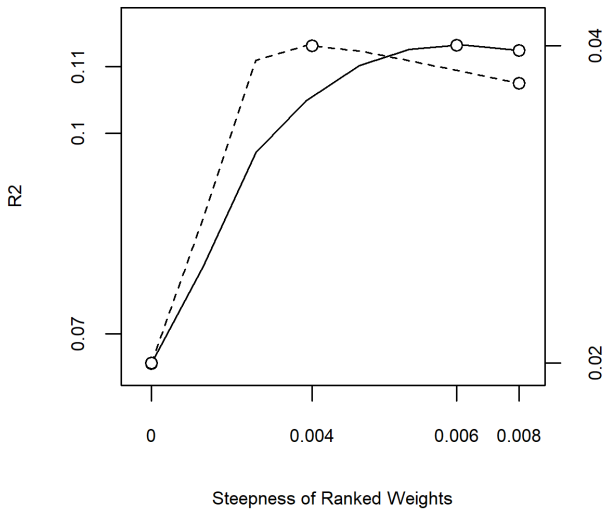
Table IX. AIC of VECM With Sentiment Of More and Less Active Information Providers and Adjusted R^2 of Equation for Subjective Expectations in VECM

Subjective Expectations	Sentiment	Models		
		VECM($s_t^w, \tilde{r}_t$)	VECM($r_{t-1}, s_t^w, \tilde{r}_t$)	VECM($r_{t-1}, \tilde{r}_t$)
AIC				
$\tilde{r}_t^M$	More Active Providers $s_t^{w,5}$	-0.72	1.57	1.62
	Less Active Providers $s_t^{w,1-4}$	1.48	3.92	
$\tilde{r}_t^C$	More Active Providers $s_t^{w,5}$	-1.01	1.87	1.91
	Less Active Providers $s_t^{w,1-4}$	1.16	4.22	
Adjusted R^2 of VECM Equation for Subjective Expectations				
$\tilde{r}_t^M$	More Active Providers $s_t^{w,5}$	0.10	0.08	0.07
	Less Active Providers $s_t^{w,1-4}$	0.07	0.07	
$\tilde{r}_t^C$	More Active Providers $s_t^{w,5}$	0.04	0.08	0.06
	Less Active Providers $s_t^{w,1-4}$	0.03	0.05	

What if I add even more weight to the sentiment of more active information providers? Before delving into the main question, note that VECM specifications XXII, XXIII, XXV, and XXIV indicate that the equation representing sentiment consistently exhibits a higher goodness of fit than the equation representing subjective expectations (approximately 30-40% compared to 4-10%). Furthermore, the coefficient of lagged sentiment in the sentiment equation is highly significant across all specifications. Given these observations, to answer the question I focus on the goodness of fit measures for the subjective expectations equation of VECM. Relying on the information criteria of the entire VECM system could lead to selection of a sentiment measure with the highest autoregressive component.

For test purposes, I introduce a new mesure, the average sentiment of the top 3 information providers. This metric allocates all weight to the average sentiment of the three most active information providers and assigns zero weight to all others. The choice of the top 3 providers serves as a straightforward, "rule of thumb" measure that can represent the concept of selective attention without necessitating any behavioral assumption about retail

Figure 8. Steepness of Weighting Scheme, $\bar{S}^e$, Vs. Upper Convex Hull of Set of Adjusted R^2 of Subjective Expectations Equation in VECM($S_t^e, \tilde{r}_t$)
X-axis shows "steepness" of weights of a reparametrization scheme, calculated as $\bar{sl} = \alpha \bar{sl}(s_t) + \beta \bar{sl}(s_t^w) + \gamma \bar{sl}(s_t^e)$, where $\bar{sl}(s_t)$, $\bar{sl}(s_t^w)$, $\bar{sl}(s_t^e)$ are slopes of invdividual weightening schemes with a slope $sl = \frac{1}{n_t}(\max w_t - \min w_t)$, where $\max w_t$ is the maximum weight in the scheme, $\min w_t$ is the minimum weight in the scheme and n_t is the number of news per month t. The plot shows upper convex hull of adjusted R^2 of subjective expectations equation in VECM($S_t^e, \tilde{r}_t$) with subjective expectations with MSC survey (solid black line) and CBS survey (dashed black line) versus the steepness of sentiment weighting scheme. Left y-axis corresponds to AIC of subjective expectations equation from MSC survey. Right y-axis corresponds to AIC of subjective expectations equation from CBS survey. 0 steepness corresponds to "equally-weighted" scheme; 0.042 steepness corresponds to weighting scheme that puts 100% of weight on $s_t^{w,5}$ sentiment, and 0.008 steepness corresponds to weighting scheme that puts 100% of weight on sentiment of top 3 information providers.



investors' limited attention span.

To construct the measure, every month I sort information providers from the most active to the least active ones using number of written reports in a given month. Let's denote the ordering as $1, 2, \ldots, n_t$, which sorts information providers from the most active one (1) to the least active one (n_t) within a month t. Next, I calculate average sentiment of top three information providers. So, weights in this "top-weighting" scheme are $\mathbf{w_t^e} = \{0.33, 0.33, 0.33, 0, 0 \ldots\}$.

Next, I construct 5,027 convex combinations of equally weighted $s_t = \frac{1}{n_t}\mathbf{1S_t}'$, information

provider-weighted $s_t^w = \mathbf{w_t^f S_t}'$ and top-weighted $s_t^e = \mathbf{w_t^e S_t}'$ sentiment measures

$$S_t^e = \mathbf{W_t^e S_t}' \tag{51}$$

$$\text{such that } \alpha + \beta + \gamma = 1, \tag{52}$$

$$\mathbf{W_t^e} = \alpha \frac{1}{n_t}\mathbf{1} + \beta \mathbf{w_t^f} + \gamma \mathbf{w_t^e} \tag{53}$$

$$0 \leq \alpha \leq 1,\ 0 \leq \beta \leq 1,\ 0 \leq \gamma \leq 1 \tag{54}$$

where α, β and γ are shares that form the convex combination of sentiment weighting schemes.

Figure 8 depicts the adjusted R^2 of the subjective expectation equation of the VECM in relation to the average steepness of the generated weighting scheme. The graph reveals that the peak adjusted R^2 for the subjective expectations equation in the VECM$(S_t^e, \tilde{r}_t)$, with expectations $\tilde{r}_t$ derived from the CBS survey (represented by the dashed line and right y-axis), tops out at $R^2 = 0.04$. This peak corresponds to a weighting scheme steepness of 0.004, which is made up of 73% provider-weighted sentiment and 27% top-weighted sentiment, hence $\alpha, \beta, \gamma = 0, 0.73, 0.27$. The sentiment measure $s_t^{w,5}$, which includes the 20% most active information providers, also exhibits an average weighting scheme steepness of 0.004 and a corresponding $R^2 = 0.04$. Therefore, for the CBS survey, the sentiment of the 20% most active providers, represented by $s_t^{w,5}$, could be the most representative of the "true" weighting scheme.

The graph also shows that the maximum adjusted $R^2 = 0.11$ for the subjective expectations equation of the VECM$(S_t^e, \tilde{r}_t)$, with expectations $\tilde{r}_t$ derived from the MSC survey (indicated by the solid line and left y-axis), reaches a peak at a steepness of 0.006. Since this steepness is higher than the 0.004 steepness of $s_t^{w,5}$ and higher than the $R^2 = 0.10$ of the subjective expectation equation in VECM$(s_t^{w,5}, \tilde{r}_t)$ with $s_t^{w,5}$, it can be inferred that investors from the MSC survey might lean towards learning from a set of providers with an even steeper weighting scheme. To avoid mechanical curve fitting, the precise determination

of weighting for MSC survey expectations should be informed by behavioral research studies. This would provide a more accurate reflection of how individuals weigh different sources of information, factoring in human cognitive biases and decision-making patterns.

Hypothesis 3 Information from less active providers is not related to the temporal fluctuations of aggregate stock market return.

First, I examine whether total annual return of the CRSP value-weighted portfolio[32] R_t is associated with the change in annual aggregated weighted sentiment of reports about 45 blue-chip companies. To study whether the sentiment of the reports contribute to time variations of aggregate stock market returns, or predictability of aggregate stock market return, I run a standard predictability regression as in Cochrane (2006) with and without the change in the sentiment and examine whether sentiment measures have statistical significance.

Table X presents specifications for the following predictability regressions

$$R_{t\to t+11} = \alpha + \beta_1 (D/P)_{t-11\to t} + \epsilon_t \tag{55}$$

$$R_{t\to t+11} = \alpha + \beta_1 (D/P)_{t-11\to t} + \beta_2 \Delta s^{w,1-4}_{t-12\to t-1} + \epsilon_t \tag{56}$$

$$R_{t\to t+11} = \alpha + \beta_1 (D/P)_{t-11\to t} + \beta_3 \Delta s^{w,5}_{t-12\to t-1} + \epsilon_t \tag{57}$$

$$R_{t\to t+11} = \alpha + \beta_1 (D/P)_{t-11\to t} + \beta_2 \Delta s^{w,1-4}_{t-12\to t-1} + \beta_3 \Delta s^{w,5}_{t-12\to t-1} + \epsilon_t \tag{58}$$

where $R_{t\to t+11}$ is the annual real total return of the CRSP value-weighted index at month t

[32] vwretd variable from CRSP

deflated by the CPI[33], $(D/P)_{t-11\to t}$ is the dividend-price ratio[34], $\Delta s^{w,5}_{t-12\to t-1}$ is the change in the average weighted lagged sentiment of more active information providers and $\Delta s^{w,1-4}_{t-12\to t-1}$ is the change in the lagged weighted sentiment of less active information providers.

As in Cochrane, 2006, I use L. P. Hansen and Hodrick, 1980 standard errors to account for serially correlated errors in the overlapping regression.

The first regression specification Table X replicates Cochrane, 2006 predictive regression. Second, third, and fourth regression specifications also include the average lagged weighted sentiment of more and less active information providers. It is shown in Table X that the change in the average sentiment of more active information providers ishighly statistically significant.

To explore non-overlapping predictive regression, I employ quarterly real return, quarterly dividend-price ratio, and quarterly change in sentiment to evaluate the statistical significance of sentiment measures. Table 9 shows that quartely sentiment of more active providers remains not only weakly statistically significant, but also increase adjusted R^2 from -0.01 to 0.03.

Second, I investigate the link between stock returns in a given month t and the sentiment change in reports in the previous month $t-1$. To achieve this, I use a cross-sectional return analysis. This process includes the development of a momentum trading strategy, which is

[33]The real monthly total return of CRSP value-weighted Index is calculated as

$$R_t = \frac{vwretd_t + 1}{cpi_t + 1} - 1 \tag{59}$$

where cpi_t is $CPIAUCSL_CH$ is the percent change in Consumer Price Index for All Urban Consumers divided by 100 from U.S. Bureau of Labor Statistics, https://fred.stlouisfed.org/series/CPIAUCSL and annual return $R_{t\to t+11}$ is equal to

$$R_{t\to t+11} = \Pi_t^{t+11}(R_t + 1) - 1 \tag{60}$$

where R_t is the annual real total return of the CRSP value-weighted index at month t.

[34]The dividend-price ratio is calculated as

$$(D/P)_{t-11\to t} = \Pi_{t-11}^{t}\frac{vwretd_t + 1}{vwretx_t + 1} - 1 \tag{61}$$

where $vwretd_t$ is the annual real total return on the CRSP value weighted portfolio and $vwretx_t$ is the return on the CRSP value weighted portfolio excluding dividends at month t

Table X. Predictive Regression: Overlapping Regression Specification
The regression specifications are

$$R_{t\to t+11} = \alpha + \beta_1 (D/P)_{t-11\to t} + \epsilon_t$$
$$R_{t\to t+11} = \alpha + \beta_1 (D/P)_{t-11\to t} + \beta_2 \Delta s^{w,1-4}_{t-12\to t-1} + \epsilon_t$$
$$R_{t\to t+11} = \alpha + \beta_1 (D/P)_{t-11\to t} + \beta_3 \Delta s^{w,5}_{t-12\to t-1} + \epsilon_t$$
$$R_{t\to t+11} = \alpha + \beta_1 (D/P)_{t-11\to t} + \beta_2 \Delta s^{w,1-4}_{t-12\to t-1} + \beta_3 \Delta s^{w,5}_{t-12\to t-1} + \epsilon_t$$

where $R_{t\to t+11}$ is an annual return of value-weighted market index with dividends $r_{t\to t+11} = \Pi_t^{t+11}(vwretd_t + 1)/(\pi_t + 1) - 1$ adjusted for monthly inflation π_t, $(D/P)_{t-11\to t}$ is annual dividend-price ratio calculated from monthly $vwretd_t$ and $vwretx_t$ as $\Pi_{t-11}^{t}(vwretd_t + 1)/(vwretx_t + 1) - 1$. $s_t^{w,1-4}$, $\Delta s^{w,1-4}_{t-12\to t-1}$ and $\Delta s^{w,5}_{t-12\to t-1}$ are average annual sentiment of less and more active information providers. I report Hansen-Hodrick (1980) standard errors with 12 month window in parentheses and the statistical significance is shown as *p<0.1, **p<0.05, and ***p<0.01.

	Dependent variable:			
	$R_{t\to t+11}$			
	(1)	(2)	(3)	(4)
$(D/P)_{t-11\to t}$	17.98***	17.55***	17.43***	17.14***
	(0.01)	(0.01)	(0.01)	(0.01)
$\Delta s^{w,1-4}_{t-12\to t-1}$		0.03		0.03
		(0.06)		(0.06)
$\Delta s^{w,5}_{t-12\to t-1}$			0.10***	0.09***
			(0.02)	(0.02)
Constant	0.71	0.72	0.72*	0.72*
	(0.45)	(0.44)	(0.43)	(0.43)
Observations	197	197	197	197
R^2	0.09	0.09	0.09	0.10
Adjusted R^2	0.08	0.08	0.08	0.08

based on sentiments conveyed in reports from both highly active and less active information providers. By juxtaposing the cumulative returns from strategie, I assess the impact that change in sentiment of active and less active information providers has on stock returns.

The momentum strategy I utilize is based on the sentiment of report headlines from information providers in the previous month. Every month, this strategy ranks 45 stocks according to the sentiment about the company in the preceding month's reports and then assigns these stocks to portfolios. The portfolios are held for one month.

In particular, I use the sentiment-based momentum strategy as in Jegadeesh and Titman,

Table XI. Predictive Regression Specification: Non-overlapping Regression Specification

The regression specifications are

$$R_q = \alpha + \beta_1 (D/P)_{q-1} + \epsilon_q$$
$$R_q = \alpha + \beta_1 (D/P)_{q-1} + \beta_2 \Delta s_{q-1}^{w,1-4} + \epsilon_q$$
$$R_q = \alpha + \beta_1 (D/P)_{q-1} + \beta_3 \Delta s_{q-1}^{w,5} + \epsilon_q$$
$$R_q = \alpha + \beta_1 (D/P)_{q-1} + \beta_2 \Delta s_{q-1}^{w,1-4} + \beta_3 \Delta s_{q-1}^{w,5} + \epsilon_q$$

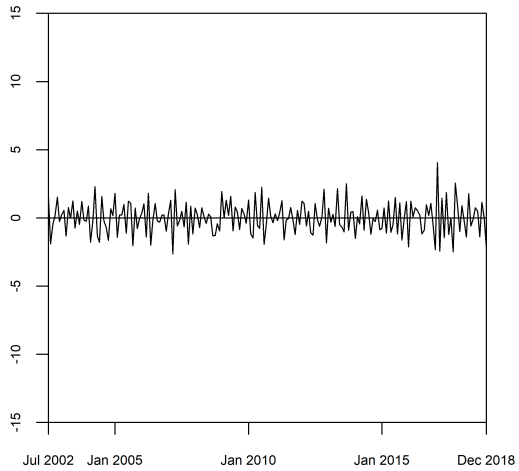
where R_q is a real return of CRSP value-weighted portfolio, $(D/P)_{q-1}$ is quarterly dividend-price ratio calculated from monthly $vwretd_t$ and $vwretx_t$ as $\Pi_t^{t+2}(vwretd_t + 1)/(vwretx_t + 1) - 1$. I use quarterly change in sentiment $\Delta s_q^{w,1-4}$ and $\Delta s_q^{w,5}$. I report heteroscedasticity-consistent standard errors (with HC3 adjustment for small sample size) in parentheses and the statistical significance is shown as *p<0.1, **p<0.05, and ***p<0.01.

	Dependent variable:			
	sprtrn.x			
	(1)	(2)	(3)	(4)
$(D/P)_{q-1}$	3.26	4.37	9.93	9.73
	(17.84)	(18.08)	(13.79)	(13.84)
$\Delta s_{q-1}^{w,1-4}$		−0.14		−0.08
		(0.43)		(0.44)
$\Delta s_{q-1}^{w,5}$			2.15*	2.14*
			(1.21)	(1.24)
Constant	0.003	−0.004	−0.03	−0.03
	(0.09)	(0.09)	(0.07)	(0.07)
Observations	65	64	64	64
R^2	0.001	0.004	0.06	0.06
Adjusted R^2	−0.01	−0.03	0.03	0.02

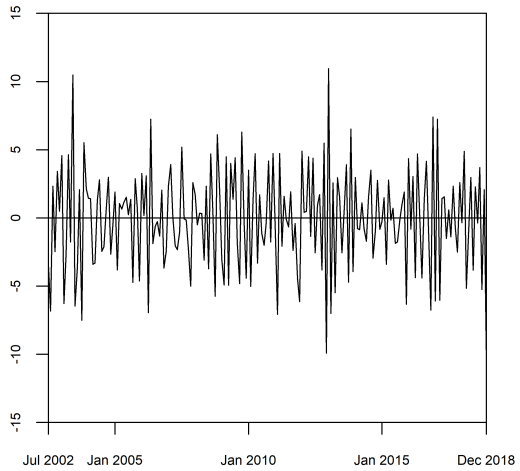
1993 and Moskowitz and Grinblatt, 1999. However, instead of using past returns to sort stocks into portfolios, I use sentiment from the previous month's reports. The strategy involves investing equally in the stocks with the most and least pessimistic sentiment.

I use two datasets of monthly reports — one for active providers and another for less active providers. Both datasets are structured with the month, PERMNO of the company, monthly sentiment, and monthly return. In each dataset, every month, companies are sorted based on their sentiment from the previous month and assigned a quintile number (from 1 to 5) for the current month. This creates a sentiment-driven quintile. The average return

Figure 9. Time Series of Change in Sentiment of More and Less Active Information Providers



(a) Change in Sentiment of More Active Providers, $\Delta s_t^{w,5}$



(b) Change in Sentiment of Less Active Providers,$\Delta s_t^{w,1:4}$

for each month-quintile is then calculated. To ensure equal numbers of companies in each quintile, only the top 8 companies are retained for each month-quintile. The new dataset now contains the month, sentiment-driven quintile, and average return at time t. Next, a long-short portfolio is created by going long on the 5th quintile and shorting the 1st quintile. This is the equivalent of subtracting the return of the 1st quintile from the return of the 5th quintile at time 't', yielding a new dataset with the month and portfolio return for that month. Finally, I calculate the average return and standard deviation of the portfolio returns, providing the sentiment-based trading strategy's average return and standard deviation for portfolios formed based on more active and less active information providers.

Table XII indicates that a trading strategy that relies on the sentiment of less active information providers generates a higher expected return than a strategy based on the sentiment of more active providers. Interestingly, these results from the cross-sectional analysis seems to contradict the outcomes of the predictive time-series regression.

Table XII indicates that the sentiment from less active information providers in the previous month (month $t-1$) could be a better predictor of stock returns in the current month (month t) compared to the sentiment from more active information providers. This

Table XII. Monthly Return of Return-based and Sentiment-based Trading Strategies

Sorting based on previous month aggregated sentiment of information provider's type. One month strategy formation period. Holding period is 1 month. Trading strategy based on information from less active information providers is to long stocks with the most pessimistic report headlines and short stocks with the most optimistic report headlines. Trading strategy based on information from more active information providers is to long stocks with the most optimistic report headlines and short stocks with the most pessimistic report headlines. 240 months.

Strategy	Average Return	St.Dev.	Sharpe Ratio
Return-Based Strategies			
S&P 500	0.28	0.49	0.31
	(0.03)		
Momentum based on r_{t-1}	-0.11	0.58	-0.40
	(0.04)		
Sentiment-Based Strategies			
Based on less active providers $s_{t-1}^{w,1:4}$	0.30	0.36	0.50
	(0.02)		
Based on more active providers $s_{t-1}^{w,5}$	0.16	0.46	0.08
	(0.03)		

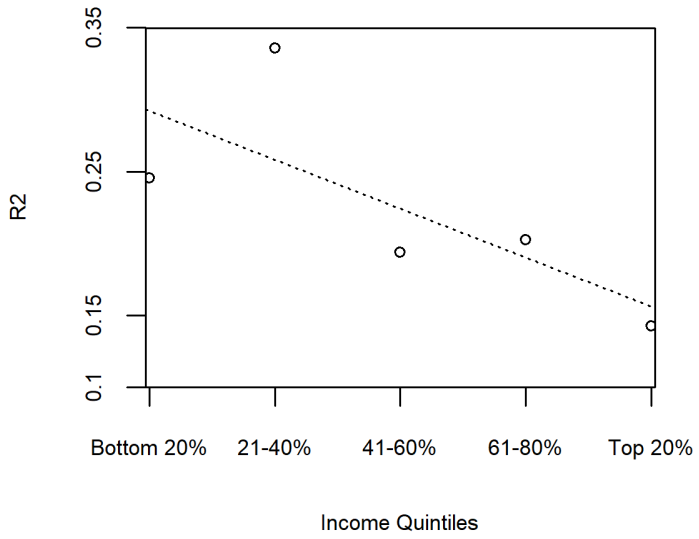
observation emphasizes the potential importance of recent data from less active providers in analyzing the stock market.

Note that the trading strategies based on information from less active and more active providers follow opposite directions. A strategy that utilizes data from less active providers would involve buying stocks with the most pessimistic report headlines from the previous month and selling those with the most optimistic headlines. Conversely, a strategy that relies on data from more active providers would involve buying stocks with the most optimistic report headlines from the previous month and selling those with the most pessimistic headlines. This trend aligns with the negative sign of information from less active providers in the non-overlapping predictability regression (XI), as well as with the descriptive statistics of sentiments from both types of providers.

Descriptive statistics presented in Table VIII shows that less active providers tend to be more conservative in their reports (mean sentiment of individual reports is 0.35) and publish less frequently (21,120 reports in sample), compared to more active providers who generally

Figure 10. Average Goodness of Fit of Expectations Equation Of VECM($s_t^w, \tilde{r}_t$) With Subjective Expectations of Investors With Different Income and Stock Investment Amount Quintiles

The plot displays the average adjusted R^2 values of the subjective expectations equation in the VECM($s_t^w, \tilde{r}_t$), where $\tilde{r}_t$ represents the subjective expectations of investors from different income brackets across quintiles of stock investment amounts within each income quintile, according to data from the MSC. The X-axis represents income quintiles, while the Y-axis represents the adjusted R^2 of the subjective expectations equation in the VECM. A dotted line is included to depict the regression of the R^2 values on income quintiles for visual reference. These computations are based on monthly data from June 2002 through December 2018.



express more optimism (mean sentiment of individual reports is 1.99) and publish more often (489,360 reports in sample). Hence, when less active providers issue positive equity reports, they are typically more pessimistic and published relatively later than those from more active providers. Consequently, the stock market return tends to decrease in response to positive news from less active providers. For instance, if more active providers issue 10 reports predicting a 20% growth in Apple's stock, a subsequent single report from a less active provider forecasting a 10% growth may be associated with a decrease in the stock market return.

Hypothesis 4 There is no effect of income on the learning pattern of retail investors.

I examine whether investors' income and stock investment amount affect the relationship between the aggregated polarity of earnings reports and investors' subjective expecta-

tions. As the MSC provides microdata on income and stock investment of all respondents, I construct subjective expectations for investors within MSC quintiles of income[35] and MSC quintiles of stock investment amount[36] . For my analysis, I use 18 combinations of income and stock investment quintiles formed from individual surveys conducted the University of Michigan and accumulated in the University of Michigan's Surveys of Consumers from June 2008 through December 2012.

I employ the data as provided by the University of Michigan, which is adjusted based on random sampling and population adjustment procedures. There are no additional statistical adjustments made on my part. This approach aligns with the standard practices within economics and finance literature that work with the MSC to analyse subjective expectations of individuals regarding the economy (M. Baker and Wurgler, 2007; Barsky and Sims, 2012; Brunnermeier et al., 2014; Ludvigson, 2004; Souleles, 2004[37] among others).

I run VECM($s_t^w, \tilde{r}_t$) with subjective expectations of investors within income quintiles. Figure 10 displays the adjusted R^2 of a subjective expectaions equation of the VECM($s_t^w, \tilde{r}_t$). The VECM($s_t^w, \tilde{r}_t$) with the highest R^2 (25- 35%) corresponds to models with subjective expectations of investors with bottom 20 percentile incomes and 21-40 percentile incomes. Those respondents who belong to the bottom 20% income quintile and the bottom 20% investment quintile earn an average of $18,753 a year and invest an average of $6,919 in stocks. Their stock investments account for 37 percent of their annual income. Those in the

[35]Variable YTL5 in MSC database.

[36]Variable STL5 in MSC database.

[37]Ludvigson, 2004 made extensive use of the Michigan Survey of Consumers to investigate the predictive power of consumer confidence for consumption growth. The paper concludes that consumer sentiment has significant power in forecasting future consumption growth, especially for non-durable goods and services. Souleles, 2004 used the Michigan Survey of Consumers to examine how changes in consumer sentiment affect consumer spending. The study found that changes in consumer confidence have significant and substantial effects on household consumption, with these effects being larger for households that are more likely to be liquidity constrained. Barsky and Sims, 2012 used the Michigan Survey of Consumers to study news shocks. They relied on the survey's data about consumer expectations regarding personal and macroeconomic conditions to identify news shocks. They then analyzed how these shocks propagate into macroeconomic quantities. M. Baker and Wurgler, 2007 used the Index of Consumer Sentiment from the Michigan Survey in their study on investor sentiment and its effects on the cross-section of stock returns. Their analysis provides evidence that investor sentiment may indeed affect asset prices. Brunnermeier et al., 2014 utilized the survey data to investigate the role of belief disagreements in financial markets. They used the data on consumer expectations to measure disagreement and its impact on stock price volatility.

21-40% income quintile with stock investments up to 60% earn on average $35,972 per year, with an average stock investment of $35,573. Their stock investments account for 99 percent of their annual income.

VECM with the lowest R^2 corresponds to models with subjective expectations of investors in the top 20% of the income distribution. On average, they earn $189,888 per year and hold $600,127 in stocks. There is 316% of annual income attributed to stock holdings.[38]

The data shows that retail investors with lower income, in 0-20% and 21-40% income quintiles, tend to follow provider-weghted equity reports, whereas investors with top 20% incomes are less susceptible to this learning pattern. First, this finding suggest that income level is a significant factor in learning patterns. Second, it suggests simultaneously that lower-income retail investors tend to follow more active providers (and may be influenced more by the popularity of certain sources) and/or lower income investor read equity reports.

There is a consensus in the literature that higher-income retail investors have a greater tendency to read and interpret financial reports. Souleles, 2004, found that it is the case as higher-income households have the larger financial stakes involved and the capacity to afford professional financial advice. Similarly, D'Acunto et al., 2019 argued that wealthier individuals are more likely to consume financial news because they have more resources at stake in financial markets. As such, they are more incentivized to stay informed and make optimal decisions. Their study also suggested that wealthy investors have better access to quality information, which might explain their higher consumption of financial reports. Lower-income investors are generally found to be less engaged with financial reports. Hastings and Tejeda-Ashton, 2008 found that lower-income investors tend to be less financially literate and are, therefore, less likely to read and understand financial reports. Their study also highlighted that lower-income individuals often face barriers such as lack of time or expertise, which prevent them from effectively utilizing financial reports. Finally, studies like Lusardi and Mitchell, 2014 have raised concerns about financial literacy among lower-income individuals,

[38]The respondents have an annual income ranging from $1,000 to $500,000.

suggesting that they may not consume or interpret financial reports effectively. This could contribute to suboptimal financial decision-making among this demographic. To summarize, literature suggests that higher-income investors are generally more active consumers of financial reports due to larger financial stakes and access to resources, whereas lower-income investors may face barriers preventing them from effectively consuming and understanding financial reports.

Considering the findings in the light of the literature, I propose a hypothesis that lower income investors' behavior may be more driven by their preference for popular information sources, rather than the equity reports these sources publish. To test this hypothesis, I include data from a well-known information source and re-examine the learning patterns observed earlier.

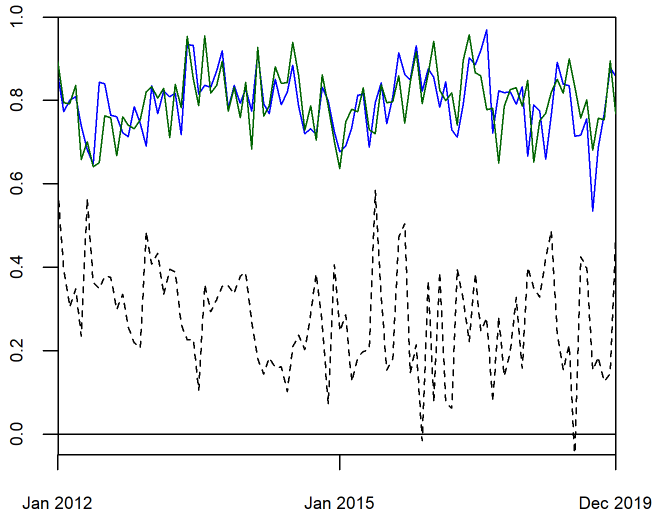
What can help to explain the direction of heterogeneity of learning patterns? To further investigate the potential factors that may affect the observed diversity in learning patterns, I will examine whether other prevalent information sources affect retail investors in heterogenious way.

Television shows could be a viable source of information given that several studies, including those by Gershuny & Robinson (1998) and Robinson & Godbey (1997), have identified a negative correlation between television viewing and income levels. Therefore, I hypothesize that popular television shows focusing on stock investing might have different impacts on investors across varying income levels. However, since I lack granular data regarding the television viewing habits of retail investors, this section is speculative.

Using the Moving Image Archive and the Internet Archive TV News at the Internet Archive[39], a non-profit free digital library of Internet sites and cultural artifacts, I collected transcripts of 8,128 episodes of Mad Money show, 2,329 episodes of the Squawk on the Street show and 1,710 episodes on 60 Minutes show that were aired from June 2, 2009 to December 31, 2019 on CNBC, American basic cable business news channel, and 42 local channels,

[39]https://archive.org/about/

Figure 11. Polarity of Mad Money (Dark Green Solid Line), Squawk on the Street (Blue Solid Line) and 60 Minutes (Back Dashed Line) Shows, Monthly
Plot presents monthly polarity of Mad Money (dark green solid line), Squawk on the Street (blue solid line) and 60 Minutes (back dashed line) shows. Data is from June 2002 to December 2018.



such as KPIX, television station licensed to San Francisco, California, a WUSA, a television station in Washington, D.C., and WBAL, a television station in Baltimore, Maryland. The Internet Archive misses year 2011 for all three shows, so I use only episodes from January 1, 2012 to December 31, 2019 for my analysis. That is 3,668 episodes of Mad Money show, 2,134 episodes of Squawk on the Street show and 1,710 episodes on the 60 Minutes show.

I picked Mad Money and Squawk on the Street shows as they are both about stock investing, but have different narrative style. I use 60 Minutes show as a control, to check whether a natural text processing algorithm does produce different polarity score for different shows.

Mad Money is a financial television program that airs weeknights on the CNBC network 6PM ET and 11PM ET. The show began airing on March 14, 2005. It is a 60 minutes show that is hosted by Jim Cramer, a former hedge fund manager and stockbroker. The program is focused on providing stock market analysis, investment advice, and stock recommendations to viewers [40]. The show typically features Cramer discussing the day's top stock market

[40] According to Cramer, the term "mad money" refers to the funds that are available for investing in stocks,

news and events, as well as interviewing business leaders and market analysts. Cramer also takes calls from viewers, answering their questions about stocks and providing investment advice.

Mad Money is known for its fast-paced and entertaining style, with Cramer often using humorous sound effects and props to illustrate his points. The show also features a "Lightning Round" segment, in which Cramer rapidly gives his opinion on various stocks that viewers ask about. Overall, Mad Money is a popular program in the financial media landscape, providing retail investors with insights and analysis on the stock market.

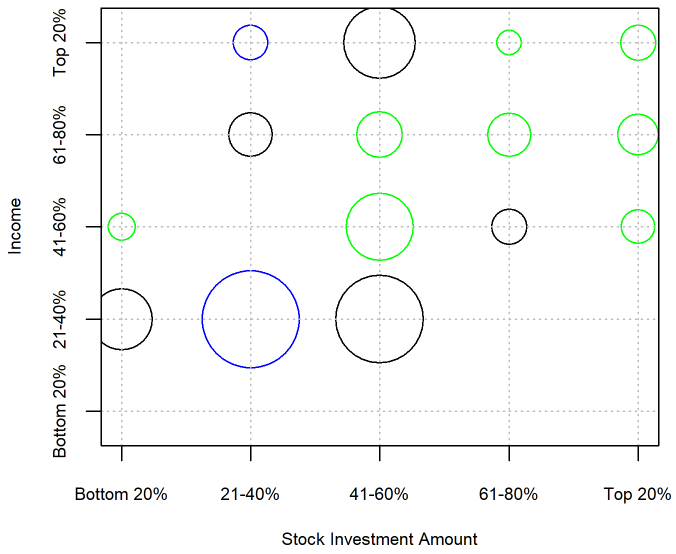
Squawk on the Street is a financial news television program that airs weekdays on the CNBC network. The show debuted on December 19, 2005. It is two hours show, from 9 am to 11 am ET, that is co-hosted by Carl Quintanilla, David Faber, and Morgan Brennan. Squawk on the Street is known for its coverage of the stock market, business news, and analysis of the day's top stories. The program is broadcast live from the floor of the New York Stock Exchange, and features frequent updates on the markets and trading activity. The hosts interview industry experts and business leaders, providing viewers with insights into the latest trends and developments in the financial world.

Squawk on the Street also covers breaking news and events that can impact the stock market and global economy. The program features analysis of corporate earnings reports, economic data releases, and other key indicators that affect the markets. Overall, Squawk on the Street is a viable source of information and analysis for investors and anyone interested in the financial markets.

As a control, I also use 60 Minutes show, an American television news magazine program that has been airing on CBS since 1968. The show has generally kept the Sunday evening format, and starts at 7:00 p.m. ET. It is known for its in-depth investigative journalism and hard-hitting interviews with news makers and public figures. Each episode of 60 Minutes

but not for retirement purposes, as retirement savings should be placed in more conservative investment options such as a 401K, individual retirement account (IRA), savings account, bonds, or stocks that pay dividends. Source: Cramer, James; Mason, Cliff (2006). Mad Money: Watch TV, Get Rich. New York: Simon & Schuster. p. 45. ISBN 978-1-4165-3790-8.

Figure 12. "Best" VECM Specifications (Color) and Adjuster R^2 of VECM's Subjective Expectations Equations (Size of Circle)
The bubble plot shows adjusted R^2 of the "best" VECM with subjective expectations $\tilde{r}_t$ of investors with different level of investment in stocks and income from MSC survey. X-axis shows Stock Investment Amount quintile. Y-axis shows Income quintiles. The size of a circle is proportional to adjusted R^2 of VECM's subjective expectations equation. The biggest circle is equivalent adjusted $R^2 = 39\%$, the smallest to the $R^2 = 8\%$. Black color corresponds to VECM$(jc_t, s_t, \tilde{r}_t)$ with the following ordering of variables $jc \to s \to \tilde{r}$, where jc is polarity of the Mad Money episodes and s is a sentiment of equity reports. Blue color corresponds to VECM$(r_{t-1}, s_t, \tilde{r}_t)$ or VECM$(r_{t-1}, q_t, \tilde{r}_t)$ with ordering $r \to q \to \tilde{r}$, where q is polarity of the Squawk on the Street episodes. Green color corresponds to VECM$(s_t, \tilde{r}_t)$ with ordering $s \to \tilde{r}$ The calculations are based on monthly data from January 2012 to December 2018.



typically consists of several segments, each covering a different news story or topic. The show covers a wide range of issues, including politics, business, science, technology, and entertainment. The segments are usually around 12-15 minutes long, and are presented by veteran correspondents who specialize in the topic being covered. I use this show to check that its polarity differ from episodes of Mad Money and Squawk on the Street.

The Internet Archive transcribes speech into text. As with many transcription algorithms, letter capitalization and punctuation is often lost in the text. For this reason, I employ the "bag of words" text processing method that sums up Loughran and McDonald, 2016 scores of positive w_j^+ and negative w_j^- words in every episode's text

$$B(e) = \frac{\sum_{j=1}^{m} h(w_j^+) + \sum_{j=1}^{g} h(w_j^-)}{\sqrt{n}} * 100, \qquad m + g \leq n \tag{62}$$

where Loughran and McDonald, 2016 scores $h(\cdot)$ of positive words w_j^+, negative words w_j^-, m is a number of positive words in a headline, g is a number of negative words in a sentence, and n is a number of words and some punctuation signs in a headline.

On average, an episode's transcript has 11,705 words, so this method picks the polarity of transcripts well. Taking an average of the polarity of episodes aired within a month, I aggregate the polarity of individual episodes into monthly time series.

Figure 11 shows time series of aggregated monthly polarity of three shows. Plot shows that Mad Money, jc_t and Squawk on the Street, q_t time series are highly correlated. Pearson correlation of 0.60^{***} is high and highly significant. As expected, Monthly 60 Minutes polarity, nm_t, is negatively correlated with both financial shows. It has -0.14 correlation with Mad Money and -0.03 with Squawk on the Street polarity scores.

Next, as empirical tests indicate that the time series of TV show polarity has a unit root[41] and are cointegrated with the aggregate subjective expectations and the polarity of equity reports, I proceed with VECM model. I run 42 Johansen, 1991 estimation procedures and corresponding VECM or VAR models per quintile expecttaions with two, v_i, v_j, and three variables, v_i, v_j, v_k. The v_i, v_j and v_i, v_j, v_k are combinations of variables from a vector that includes quintile subjective expectations, sentiment measures, stock market return, and TV polarity measures.

If in previous sections I rely on VECM AIC and VECM ordering to compare models, in this section, I use VECM or VAR AIC, models' ordering and, in addition, Granger-type causality tests that pin down impact of polarity of reports and polarity of TV shows on subjective expectations.

I follow Toda and Phillips, 1991 to evaluate Granger causality in Johansen-type error correction models. The approach uses Wald statistics with null hypothesis of non-causality and tests whether coefficients in differences and coefficients in error correction vector in

[41] Elliott et al., 1996 test statistics for the specification with a constant is -2.21 for Mad Money monthly polarity, jc_t, and -2.53 for Squawk on the Street monthly polarity. Zivot and Andrews, 1992 test statistics is -3.28 for Mad Money monthly polarity, jc_t, and -3.38 for for the Squawk on the Street monthly polarity.

VECM equation (24) of explanatory variable of interest are zero.

I start with selecting models with p-value of Wald statistics less than 0.125 for every VECM or VAR model of subjective expectations of retail investors within income-stock investment amount quintiles. Next, I select the models with the lowest AIC within the subsets. The detailed table with Wald statsics, AIC and adjusted R^2 of subjective expectations equation within all selected VECMs is provided in Table XXVIII.

Figure 23 shows adjusted R^2 of subjective expectations equation of VECM of selected model (size of a bubble) on the grid of retail investors' income quintiles and quintiles of stock investment amount. Colors correspond to different combinations of variables in VECM or VAR.

Black color bubbles lay at the left bottom of the grid, accounts for subjective expectations of retail investors with bottom 20% and 21-40% income quintile and corresponds to VECM with the following ordering of variables

$$jc \to s^w \to \tilde{r} \tag{63}$$

$$jc \to s^c \to \tilde{r} \tag{64}$$

where jc is monthly polarity of Mad Money show, s^w is monthly polarity of provider-weighted equity reports, s^c is monthly polarity of company-weighted equity reports, and $\tilde{r}$ is retail investor's subjective expectations of stock market return.

Blue color bubbles lay on the left side of the grid and accounts for subjective expectations of retail investors with income in 21-40%, and top 20% quintiles, who has stock investment amount in 21-40% quintile. Blue color corresponds to VECM models with variables ordering that include past realized stock market return r_{t-1} and the Squawk on the Street sentiment

$$r \to q \to \tilde{r} \tag{65}$$

where q is monthly polarity of Squawk on the Street show and r is past stock market return.

Green color bubbles lay at the center of the grid, accounts for subjective expectations of retail investors within 41-60% and 61-80% and top 20% income quintiles and 41-60% and 61-80% and top 20% investment amount quintiles. Green color corresponds to VECMs with ordering

$$\begin{bmatrix} s^{w} \\ s^{w,1-4} \\ s^{w,5} \end{bmatrix} \rightarrow \tilde{r} \tag{66}$$

where s^{w}, $s^{w,1-4}$, $s^{w,5}$ are provider-weighted sentiment, sentiment of 20% more active providers and sentiment of less active providers correspondingly.

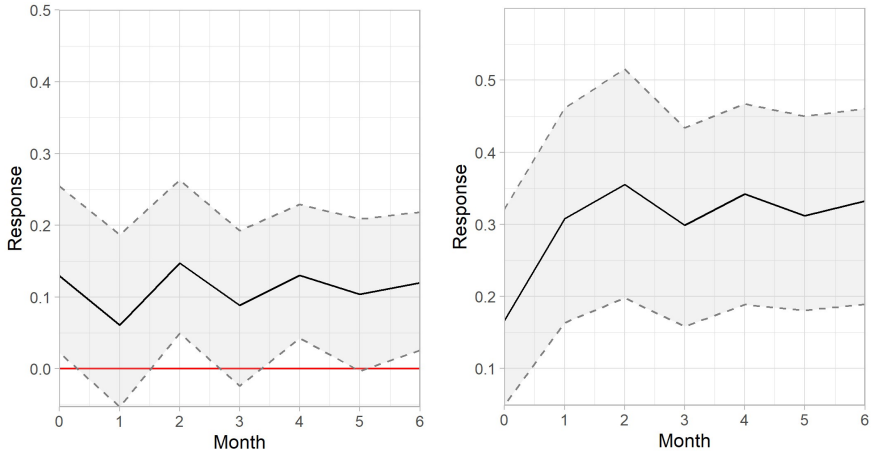
In the Appendix, I included a grid based on VAR models incorporating changes in the sentiment of both reports and TV shows. The objective is to ascertain the factors that influence the rate of change in subjective expectations.

The analysis reveals that the learning patterns of retail investors are influenced by their income level and the size of their stock investments. Investors with lower incomes tend to form their subjective expectations of stock market returns based on insights from Jim Cramer's show. On the other hand, investors with higher incomes but below-average stock investments might lean more towards insights from 'Squawk on the Street'. Meanwhile, investors with both high incomes and substantial stock investments tend to base their decisions on information from equity reports.

The dynamics of the system can be understood through Figure 13, which presents examples of orthogonal impulse responses of subjective expectations from investors with low and high incomes, but within the same stock investment amount quintile. The left plot depicts how a sentiment from Jim Cramer's show influences the subjective expectations of retail investors who fall within the 21-40% income quintile. A shock equivalent to one standard deviation that increases the sentiment leads to a direct positive effect on subjective expectations. Furthermore, these expectations remain positive, although they fluctuate around

Figure 13. Orthogonal Impulse Response Functions With Expecttaions of Investors In Low Income Vs High Income Quintiles Within Same Stock Investment Quintile

Orthogonal Impulse Response Functions sentiment to expectations for "best" models for low and high income. 95 % confidence interval for the bootstrapped errors bands. 100 runs.



(a) $jc \to \tilde{r}$, Expectaions of investors in 21-40% Income Quintile and 41-60% Stocks Investment Quintile

(b) $s^w \to \tilde{r}$, Expecttaions of investors in top 20% Income Quintile and 41-60% Stocks Investment Quintile

the level of the initial impact. On the right, the plot demonstrates the impact of a positive shock, equivalent to one standard deviation, from provider-weighted reports on the subjective expectations of a retail investor in the top 20% income bracket. Upon impact, subjective expectations surge, they continue to grow in the first and second months, and finally stabilize at the level they achieved after the first month.

To sum up, the sentiment derived from a popular television show and provider-weighted reports both positively influence the subjective expectations of retail investors. However, the intensity and duration of this impact vary contingent upon the income levels of the investors.

A. Mechanism

A.1. Process in Hand

Suppose there are two events: one is positive with a score of +1, and the other is negative with a score of -1. The expected sentiment of these events is zero. However, when I add information providers with different reporting activity, I get a different result. Let one

provider publish one report about the positive event and one report about the negative event, while another provider publish two reports about the positive event and zero reports about the negative event.

To calculate the expected sentiment with this additional information, I need to weight the sentiment scores by the frequency of each event and the frequency of reporting by each provider. So the denominator is $2 + 1 + 1 = 4$. The numerator is the sum of the products of the sentiment scores and the frequency of reporting by each provider. For the positive event, the numerator is $1 * (1 + 2)$, since one provider reports it twice and the other reports it once. For the negative event, the numerator is $-1 * (1 + 0)$, since one provider reports it once and the other reports it zero times. Adding these two products and dividing by the denominator gives us:

$$\frac{1 * (1 + 2) + (-1) * (1 + 0)}{2 + 1 + 1} = \frac{1}{2} \tag{67}$$

This means that the expected sentiment score is now 0.5, which is different from the expected sentiment score of zero without the addition of information providers.

The reason for this difference is that the data generating process (DGP) of events and the DGP of reported events differ. The frequency of sentiment from the DGP of events may not necessarily match the frequency of reported sentiment from the DGP of reported events. In this case, the positive sentiment is overweighted from the perspective of the DGP of events, but the sentiment is correctly weighted from the perspective of the DGP of reported events.

A.2. Why would a retail investor ignore information from less active providers?

Firstly, it could be a rational. If there are many information providers available, the weight of sentiment provided by less active providers would be relatively small. Therefore, it may be rational to assign a zero weight to less active providers, as the law of large numbers would suggest that their impact on the overall sentiment calculation would be negligible.

Second, it might be a friction from supply side. Investors may be influenced by analysts' upward earnings forecast bias, driven by compensation and investment-banking incentives, as in Lin and McNichols (1993) and Dugar and Nathan (1993), to believe that fewer neutral or pessimistic reports from less active information providers are less trustworthy than a constant flow of positive reports from more active information providers. Also, reports from more active information providers may be easier to access than those from less active ones. Although another possibility is that high reporting activity may be too overwhelming and may not allow an investor to read all reports of active information provider, so he might prefer to wait for and read an informative report from less active information providers.

Third, it might be a friction from demand side. Excessive coverage of active information providers may increase investor overconfidence by overstating precision of the informational content in analyst reports. Overconfidence is expected to feed investors' illusion of knowledge resulting in disregarding reports of less active information providers. Investors' judgment biases, as in Barberis, Shleifer, and Vishny (1998), Daniel, Hirshleifer, and Subramanyam (1998), and Hirshleifer and Teoh (2003) can also influence investors' choices.

The test of behavioral supply and demand stories would require granular data on subjective expectations, including data on when and which reports investors read, as well as when and what investor watch.

VI. Robustness Check

A. Exclusion of 2019

My company-weighted sentiment measure is highly volatile in 2019, putting it at a disadvantage compared to other more homoskedastic sentiment measures. I exclude 2019 from VECM specifications. Plot 24 in Appendix A shows effect of exclusion of one year, from 2002 to 2019, on AIC criterion of the VECM $(s_t^w,\ \tilde{r}_t)$. Higher line is AIC of the MSC expectations regression. Lower line is AIC of CCS expectations regression. The plot shows that exclusion

of 2019 does not change AIC of VECM dramatically.

B. Strategy Formation Window

Figure 25 in Appendix A examines average annualized return and standard deviation of sentiment-based momentum strategies with one-month to twelve months formation period and holding period of one month. Blue line corresponds to the strategy that is built on information of less active information providers, while green line - on information of more active information providers. To eliminated selection, every quantile's portfolio is restricted to contain top eight stocks. Dashed line corresponds to standard momentum strategy restricted to top eight stocks in each standard momentum quantile's portfolios as.

Plots (a) and (c) on Figure 25 shows that sentiment-based momentum strategies based on information from less active information providers earn higher average return than ones that are based on information from less active information providers over July 2002 to December 2018 sample period. Plots (a) - (d) show that when the strategy formation period is one month (this corresponds to the point $t-1$ on x-axis of every plot), the average return of sentiment-based strategy based on information from more active providers is 1.95% per year with a standard deviation of 15.98%. The corresponding returns of strategy based on information from less active providers earns an average return of 3.66% with standard deviation of 12.38%. The dotted line shows return of standard momentum strategy implemented on the set of 45 blue-chip stocks with the maximum eight stocks in each portfolio.

VII. Conclusion

This paper takes a deeper dive into the temporal variability retail investors' subjective expectations regarding stock market risk premium. It establishes a connection between these expectations and the market for information, by examining an extensive dataset of equity reports. The aggregated monthly sentiment derived from these reports illustrates a

persistently fluctuating pattern over time and holds a cointegrated relationship with retail investors' subjective expectations of stock market risk premium.

The investigation further strengthens the link between the sentiment about earnings growth, derived from the reports, and the subjective expected excess returns. This association is particularly robust for reports generated by information providers with a higher reporting activity, while it is relatively absent for those written by less active providers. Moreover, it is demonstrated that the sentiment shock from provider-weighted reports has a lasting impact on subjective expectations, adjusting them to a new level in a span of three months, while accounting for a substantial proportion of their fluctuations.

Meanwhile, the paper also highlights the potential value of information from less active providers. While their change in sentiment might not show statistical significance in a overlapping ammual predictive regression, a sentiment-based trading strategy constructed using their monthly information can outperform the returns of the S&P 500 Index. This suggests that such information should not be readily dismissed as noise, but rather considered a valuable component of an investor's informational set.

Furthermore, the paper shows that learning patterns of retail investors seem to be significantly influenced by their income levels and their stock investment amounts. Lower-income investors tend to derive their expectations from popular sources and provider-weighted sentiment, while higher-income investors with below-average stock investments rely more on insights related to past stock market returns. On the other hand, high-income investors with substantial stock investments base their decisions on provider-weighted sentiment from equity reports. This suggests a potential vulnerability for lower-income retail investors who may be more susceptible to biased or incomplete information.

This paper underscores the impact of information providers' reporting activity on shaping investor expectations, thereby contributing significantly to the existing literature on subjective expectations. It provides a more nuanced perspective on the mechanisms underlying the formation of these expectations and potential biases that can arise therein. Moreover, it

highlights the role of information providers in shaping financial market outcomes, suggesting new research directions in this area.

By integrating insights from various existing studies and introducing new findings, this research contributes to the literature examining subjective expectations, as well as the relationship between asset and information markets. It highlights the significance of considering the role of information providers in shaping financial market outcomes and paves the way for future research in this area. It calls attention to potential inequities in the access and utilization of information in financial markets, raising questions about the distribution of resources and opportunities therein.

Appendix A. Data on Subjective Expectations

Appendix A. Conference Board. Consumer Confidence Survey

The Conference Board[42] runs Consumer Confidence Survey (CCS).

Sampling frame. The monthly survey uses an address-based mail sample design. The sampling frame is derived from the files created by the U.S. Postal Service, which represent near-universal coverage of all residential households in the United States. The CCS frame is updated monthly to ensure up-to-date coverage of U.S. households.

Sampling. The CCS uses a probability sample design to select each month's random sample from the household universe frame. The frame is first stratified geographically within the census division to provide a proportionate geographic distribution, after which a systematic sample of household addresses is selected. The sample addresses are then used for the mailing.

Sample size. About 3,500 surveys are completed each month. About 2,500 for end-of-month release; 3,500 for later revision.

Field period. The CCS mailing is scheduled so that the questionnaires reach sample households on or about the first of each month. Returns flow in throughout the collection period, with the sample close-out for preliminary estimates occurring around the eighteenth of the month. Any returns received after then are used to produce the final estimates for the month, which are published with the release of the following month's data. Completed questionnaires are checked in as they are received and then scheduled for data entry. Data fields are edited for invalid entries and, if necessary, are flagged for review. As part of the ongoing quality control process, a random sample of questionnaires is selected for independent

[42] https://conference-board.org/data/consumerconfidence.cfm

review/validation by a senior member of the data collection staff. The targeted responding sample size - approximately 3,000 completed questionnaires - has remained essentially unchanged throughout the history of the CCS.

Fieldwork. The Nielsen Company[43].

Weighting. To improve the accuracy of the estimates and ensure the proportionate representation of key categories in the estimates, the CCS uses a post-stratification weighting structure covering the following categories: Census Division (9 Census divisions), Age of Head of Household (<30, 30-39, 40-49, 50-59, 60+), Gender of Head of Household (Male/Female), Income of Household (<15,000; 15,000-24,999; 25,000-34,999; 35,000-49,999; 50,000-74,999; 75,000-99,999; 100,000-124,999; 125,000+).

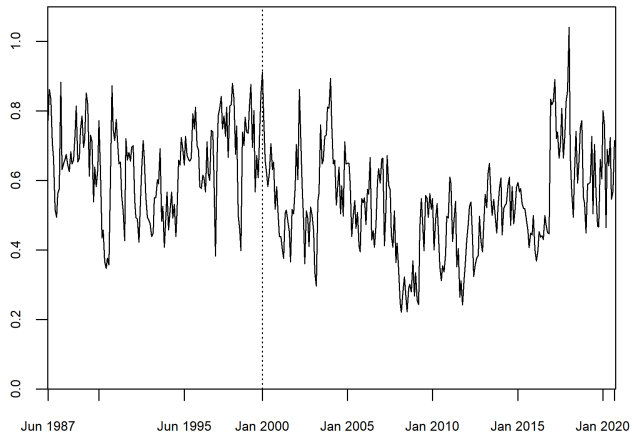
The post-stratification weighting uses an iterative proportional fitting technique for simultaneously balancing sample weights across several different population control groups. This technique ensures that sample-based estimates of the household population categories match the independent census population controls within +/- 1 percent.

Seasonal Adjustment. Data as of January 2011 use the Census X-12 seasonal adjustment software for the publication series where needed. Seasonal adjustment helps remove periodic seasonal fluctuations in the series due to events such as weather, holidays, and the beginning and end of the school year. While the CCS series are typically not highly seasonal, the X-12 software helps reduce any residual seasonality in the various data series.

[43]As of February 2011, The Conference Board has changed survey providers from TNS to The Nielsen Company for ongoing CCS operational support. Nielsen uses a mail survey specifically designed for the Consumer Confidence Survey. The new design uses a probability-design random sample, poststratification weights (for gender, income, geography, and age), and the U.S. Census X-12 seasonal adjustment. The CCS concepts, questions and mail survey collection method remain unchanged.

From September 2010 to January 2011, a five-month pilot test of the new sample design was conducted in parallel with the existing design. Three months of previously published data (November 2010 to January 2011) have been restated to smooth the transition, which makes November 2010 the effective changeover month.

Figure 14. Conference Board Index IND_t^{CB}, monthly: June 1987 - September 2020



Release. Preliminary figures are released on last Tuesday of month. Final figures are released with next month's release.

Questions about return expectation. The surveys elicit respondents simple categorical beliefs about whether the stock prices will likely increase, decrease, or stay the same (or whether they are undecided, which we include in the same category).

As per Nagel & Xu (2019), I construct the Index IND_t^{CB} as the ratio of those who respond with an increase to the sum of those who respond with a decrease or the same:

$$IND_t^{CB} = \frac{n_t^{increase}}{n_t^{decrease} + n_t^{same}} \tag{A1}$$

Appendix B. Michigan Survey of Consumers

Sampling frame. The Michigan Survey of Consumers[44] is a monthly nationally representative survey based on approximately 500 telephone interviews with adult men and women living in households in the coterminous United States (48 States plus the District of Columbia). The sample is designed as a rotating panel. For each monthly sample, an independent cross-section sample of households is drawn. The respondents chosen in this

[44]https://data.sca.isr.umich.edu/

drawing are then reinterviewed six months later. A rotating panel design results, and the total sample for any one survey is normally made up of 60% new respondents, and 40% being interviewed for the second time. The MSC provides access to panel of individual monthly responses.

The MSC uses random digit dialing (RDD) telephone sampling to draw the monthly national probability sample. The specific RDD procedure used at the Survey Research Center (SRC) is a one-stage list-assisted design. The list-assisted sampling frame consists of all hundred series[45] which have at least one listed household number. The frame is produced by aggregating all directory-listed household telephone numbers to the hundred series level. These listed hundred series form a subset of approximately 40 percent of the total possible hundred series which can be formed from all Area Code/Exchanges in the Bellcore system. Each hundred series is associated with 100 possible phone numbers - which can be listed household, unlisted household, nonresidential, non-working or unassigned. Because of the way telephone numbers are assigned, a hundred series which has at least one listed household number is more likely to have other residential telephone numbers. Business numbers are often segregated in reserved hundred series and other hundred series are not used. While the incidence of working household numbers is about 22 percent in the set of all possible hundred series from the Bellcore Area Code/Exchanges, the incidence of working household numbers is about 50 percent in the set of listed hundred series.

Household telephone samples fail to include the approximately 6% of U.S. households that are not telephone subscribers, although the percentage of nonsubscribers is declining over time. Past analysis suggests that nonsubscribers are disproportionately poor, live in the rural areas, and are more likely to rent and live alone than the rest of the population. Current studies of the bias which results from the exclusion of non telephone subscribers indicate that it is not severe and probably is within the accuracy requirements for most, but

[45]The term "hundred series" refers to the first eight digits of a phone number - the area code, exchange, and the first two digits of the remaining four numbers. One hundred possible phone numbers can be formed from each hundred series by adding the set of numbers "00" to "99" to create 10-digit phone numbers.

not all, survey research projects.

Sampling. The monthly Survey of Consumers sample, which are selected from a list-assisted RDD frame using the GENESYS Sampling System, are stratified, one-stage, equal probability samples of telephone households in the contiguous United States (48 states and the District of Columbia). GENESYS uses the Donnelly Quality Index Database (100% Phone File) as the basis for its RDD sampling frame along with auxiliary files including the Bellcore file of valid area codes and exchanges.

The GENESYS list-assisted frame is stratified by geography and urbanicity. Explicit strata are formed by crossing Census Division by MSA/non-MSA status[46]. Within each MSA stratum, there is an ordering by size of MSA and within MSA by exchanges serving the county containing the central city, followed by those serving remaining non-central city counties; within non-MSA strata, exchanges are ordered geographically in a serpentine fashion within each Census Division. The GENESYS sampling frame is updated twice yearly. Area code changes are incorporated as needed between the semi-annual updates.

List-assisted RDD sample designs for telephone surveys differ from those for personal interview surveys in that selection probabilities are assigned on the basis of the number of possible phone numbers which can be formed from the set of listed hundred series in a defined group of area codes/ exchange codes rather than on population totals for geographic areas such as counties, cities, and blocks.

The list-assisted RDD design provides for an equal probability sample of all telephone households; within each household, probability methods are also used to select one adult as the designated respondent. At the time of the initial contact with the household, a listing is taken of all household members that are 18 or older. From this list of eligible respondents, a specific member of the household is selected by the interviewer using the "respondent selection table" assigned to that household's coversheet. These selection tables are assigned

[46]MSA is Metropolitan Statistical Areas. Definition can be found in https://www2.census.gov/geo/pdfs/reference/GARM/Ch13GARM.pdf

to households so that each adult has a known selection probability, across households of all sizes, as well as differences in age and sex composition. Giving each selected respondent a weight equal to the number of adults in the household would then transform the sample of households to a sample of the adult population.

Sample size. 250-300 for mid-month release. 500 for end-of-month revision.

Field period. Michigan conducts its survey by phone throughout most of the month. Final figures for the full sample are subsequently made available at the end of the month and are not subject to further revision.

Fieldwork. Michigan Survey Research Center.

Weighting. Household head weight is used in the monthly expectation surveys. The household weights are designed to yield a representative sample of all U.S. households.

Data from the Current Population Surveys conducted by the Census are used to adjust for variations in the age and income distributions observed in the monthly samples. In practice, the post stratification weights do not yield "weighted" response distributions that differ significantly from the "unweighted" results - that is, the differences are within the margin of the expected sampling error.

The RDD and reinterview portions of the sample are post-stratified separately. This permits the construction of weights designed for analyses based solely on cases in either portion of the sample, and allows the pooling of cases when the analyses are based on the full sample. The separate post-stratification also explicitly recognizes the underlying differences between initial refusals and panel attrition. The potential non response bias in the RDD portion of the sample relate to several factors:

a) establishing contact with the selected households - for example, some phones may never be answered as the occupants are away for an extended period of time, or because

answering machines are used to screen and avoid calls;

b) establishing contact with the selected respondent - interviews are conducted only with the designated respondent, no substitutions are allowed even if the designated respondent is unavailable for the entire study period due to work schedules, travel, and so forth;

c) the willingness of the selected respondent to be interviewed.

For the reinterview portion of the sample, there are additional sources of non response bias related to our ability to recontact respondents that have moved, changed phone numbers, or discontinued phone service. Willingness to be interviewed a second time may reflect different considerations on the part of the respondent, especially given their knowledge about the content of the interview. Before the weights for the RDD and the reinterview portions of the sample are integrated one further adjustment is made, based on the strengths of the rotating panel design of the monthly surveys.

The rotating panel design offers important statistical advantages for the measurement of change over time. The statistical advantage stems from the reduction in the standard errors of the observed differences in observed means between two overlapping samples as compared with two independent samples. The variances of the estimated differences over time are reduced to the extent that the repeated measures in the reinterview portion of the sample are positively correlated. Due to the correlation, each case in the reinterview portion of the sample contributes less to the variance (by one minus the correlation coefficient) than cases from the RDD sample. To take advantage of this variance reduction feature, the weights given to the RDD cases are decreased relative to the reinterview cases so as to achieve estimates of differences with minimum variance.

Seasonal Adjustment. No information on seasonal adjustment.

Release. Preliminary figures are released mid-month. Final figures are released at end of the month.

Figure 15. MSC: Expected Percent Chance of Increase of $1,000 Investment in Diversified Stock Mutual Fund in the Year Ahead, $\bar{p}_t^{MSC}$, monthly
Data is from June 2002 to October 2020

Questions about return expectation. The MSC reports the perceived probability that an investment in a well-diversified stock fund will increase in value over a one-year horizon.

The question is: "What do you think the percent chance that this one thousand dollar investment will increase in value in the year ahead, so that it is worth more than one thousand dollars one year from now?" The question is available from June 2002 to current date. The aggregated response data can be found in Table 20 "Probability of Increase in Stock Market in Next Year" in the Saving and Retirement section of the survey. Answers are reported as a mean probability of increase in stock market in next year $\bar{p}$ and a coarse answers' distribution in $\{0\%,\ 1-24\%,\ 25-49\%,\ 50\%,\ 51-74\%,\ 75-99\%,\ 100\%\}$ with "Do not Know" and "NA" options. The MSC also gives access to individual responses with post-stratification weights.

Using law of iterated expectations, mean probability of increase in stock market in next year, $\bar{p}$, %, can be calculated from aggregated data as

$$\bar{p}_t^{MSC} = \frac{1}{\sum_{i=1}^{N} w_{i,t}} \sum_{i=1}^{N} w_{i,t} p_{i,t} \tag{A2}$$

where $N = 7$ is a number of probability partitions in $\{0\%,\ 1-24\%,\ 25-49\%,\ 50\%,\ 51-74\%,\ 75-99\%,\ 100\%\}$, p_i average probability in a partition i and w_i is a number of respondents estimated probability of increase in value in partition i. It is equivalent to weighted sum of individual responses.

The MSC expectations index $\bar{p}_t^{MSC}$ is

As the UBS/Gallup's benchmark surveys contain responses of investors with minimum of $10,000 invested in stock market, I check MSC statistics for respondents reported to invest more and less $ 10,000. Based on MSC sampling methodology, 5 % of people interviewed

Figure 16. MSC: Share of Interviewed People With More (black line) and Less $10,000 (blue line) Investment in Stock Market, Monthly
Data is from June 2002 to November 2020.

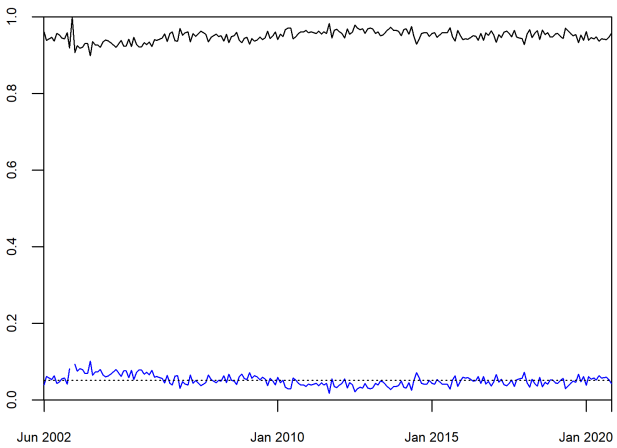


Table XIII. Monthly MSC Expectation Index (Equally-Weighted Average) for Investors Investing More and Less $ 10,000: June 2002 - November 2020

Statistic	Mean	St. Dev.	Min	Max
Invested less $10,000	49.70	9.13	24.42	72.14
Invested more $10,000	49.36	6.13	33.44	61.44

have investment in stock market less than $10,000.

Summary statistics of average monthly responses for the two groups of investors shows that expectations of respondents invested less $ 10,000 is mean-preserving spread of expectations of respondents invested more that $ 10,000.

Given the stability of the sampling and that time series of expectations of investors invested less $ 10,000 is a mean-preserving spread of the expectation index of investors who invested more than $ 10,000, I take full MSC sample for the analysis.

Appendix C. UBS/Gallup Survey

The Roper Center for Public Opinion Research [47] at the Cornell University provides access to UBS/Gallup US Investor Optimism Index surveys. The monthly data ranges from February 1999 to October 2007 and profiles individual investors. As metrics in 1999 are

[47] https://ropercenter.cornell.edu/

volatile, I use a sample from January 2000 through October 2007. It constitutes 93 monthly polls.

Sampling frame. The survey is conducted on a nationally representative sample of respondents holding stocks, bonds, or mutual funds worth at least $10,000.[48] Gallup screens for U.S. investors using a nationally representative sample of U.S. adults aged 18 and older living in all 50 states and the District of Columbia.

Sampling frame includes a listing of all possible household telephone numbers in the continental United States. It's created from all telephone exchanges in the U.S. and estimates of the number of residential households for each exchange.

Sampling. Gallup samples phone numbers using random-digit-dial (RDD) methods[49]. The RDD procedure utilizes random generation of phone numbers from the sample frame. Participants change from survey to survey.

Traditionally, the Gallup implements the following stratification scheme. The United States is divided into seven size of-community strata: cities of population 1,000,000 and over, 250,000 to 999,999, and 50,000 to 249,999, with the urbanized areas of all these cities forming a single stratum; cities of 2,500 to 49,999; rural villages; and farm or open country rural areas. Within each of these strata, the population is further divided into seven regions: New England, Middle Atlantic, East Central, West Central, South, Mountain, and Pacific Coast. Within each size-of-community and regional stratum the population is arrayed in geographic order and zoned into equal size groups of sampling units. Pairs of localities in each zone are selected with probability of selection proportional to the size of each locality's population-producing two replicated samples of localities.

[48]Information on Gallup survey sampling procedures was excerpted from George H. Gallup, The Gallup Poll, Public Opinion 1934-1971, Vol. 1, 1935-1948 (New York: Random House, 1972), pp. vi-viii; George H. Gallup, The Gallup Opinion Index, Report No. 162 (Princeton, NJ: The Gallup Poll, January 1979), pp. 29, 30; George Gallup, The Sophisticated Poll Watcher's Guide (Princeton, NJ: Princeton Opinion Press, 1976), p. 102; and from information provided by The Gallup Organization, Inc.

[49]https://www.albany.edu/sourcebook/pdf/app5.pdf

The stratification by regions is routinely supplemented by fitting each obtained sample to the latest available U.S. Census Bureau estimates of the regional distribution of the population. Also, minor adjustments of the sample are made by educational attainment (for males and females separately), based on the annual estimates of the U.S. Census Bureau derived from their Current Population Survey. The sample procedure described is designed to produce an approximation of the adult civilian population living in the United States, except for those persons in institutions such as prisons or hospitals.

Systematic procedures are in place to maintain the integrity of the sample. If there is no answer or the line is busy, the number is stored in the computer and redialed a few hours later or on subsequent nights of the survey period. Procedures are utilized to assure that the within-household selection process is random in households that include more than one adult. One method involves asking for the adult with the latest birthday; if that adult is not home the number is stored for a call back. These procedures are standard methods for reducing the sample bias that would otherwise result from under representation of persons who are difficult to find at home.

Sample size. There are about 1000 observations per month.

Field period. The UBS/Gallup conducts interviews of investors during the first two weeks of every month.

Fieldwork. The Gallup company[50]

Weighting. Individual responses are weighted by UBS/Gallup's Weighting Variable for Aggregation WTFCTR, $w_{i,t}$. Gallup weights samples to correct for unequal selection probability, nonresponse in the sampling frame. Gallup also weights its final samples to match the

[50]https://www.gallup.com/178685/methodology-center.aspx

U.S. population according to gender, age, race, ethnicity[51], education, region[52], population density[53], and phone status (cellphone only, landline only, both, and cellphone mostly). Demographic weighting targets for the U.S. are based on the most recent Current Population Survey figures for the aged 18 and older U.S. population. Population density targets are based on the most recent U.S. Census.

Seasonal Adjustment. No information on seasonal adjustment.

Release. The UBS/Gallup reports the results on the last Monday of the month.

Questions about return expectation. Following Nagel and Xu (2018), I use data from two survey questions about expected returns. The first question[54] asks about the expected rate of return the respondent expects to receive from investing in the stock market over the next 12 months:

"What overall rate of return do you expect to get on your portfolio in the next twelve months?"

This question is available until April 2003.

[51]Race, ethnicity - Nonwhite is comprised of individuals who report themselves as any combination of the following classifications: Hispanic, American Indian, other Indian, Asian, and black. Black and Hispanic are subcategories of nonwhite. However, due to variation in respondent reporting, the category white may also include some Hispanics.

[52]The four regions of the country as reported in Gallup public opinion survey results are

- East - Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Pennsylvania, Maryland, Delaware, West Virginia, District of Columbia;
- Midwest - Ohio, Michigan, Indiana, Illinois, Wisconsin, Minnesota, Iowa, Missouri, North Dakota, South Dakota, Nebraska, Kansas;
- South - Virginia, North Carolina, South Carolina, Georgia, Florida, Kentucky, Tennessee, Alabama, Mississippi, Arkansas, Louisiana, Oklahoma, Texas; and
- West- Montana, Arizona, Colorado, Idaho, Wyoming, Utah, Nevada, New Mexico, California, Oregon, Washington, Hawaii, Alaska.

[53]Urbanization - Central cities have populations of 50,000 and above. Suburbs constitute the fringe and include populations of 2,500 to 49,999. Rural areas are those that have populations of under 2,500.

[54]It is Question # 15 from February 1999 to December 2001, and Question # 10 form January 2002 to April 2003

The second question[55] asks participants about the return they expect on their own portfolio: "Thinking about the stock market more generally, what overall rate of return do you think the stock market will provide investors during the coming twelve months? (Open ended and code actual percent)"

This question was in the survey until October 2007.

For both questions, possible answers are

Score	Answer
0-99	Code actual percent, %
997	997+
998	(Do not know)
999	(Refused)

After every respondent's answer, an interviewer codes in a separate entry whether the expected rate of return number is positive or negative, $sgn_{i,t}$.[56] The UBS/Gallup interviewer's instructions states that "if you are unsure whether the number is positive or negative, then ask the respondent. As a general rule, you should assume it to be positive, unless the respondent explicitly says "Minus"; or in some other way indicates the number is negative."

I use micro data from 93 polls, to calculate aggregate expected market $\tilde{r}^{(12)}_{m,t}$ and portfolio $\tilde{r}^{(12)}_{p,t}$ returns for the next 12 months as

$$\tilde{r}^{(12)}_{m,t} = \sum_i w_{i,t} sgn_{m,i,t} \tilde{r}^{(12)}_{m,i,t} \tag{A3}$$

$$\tilde{r}^{(12)}_{p,t} = \sum_i w_{i,t} sgn_{m,i,t} \tilde{r}^{(12)}_{m,i,t} \tag{A4}$$

where $\sum_{i,t} w_{i,t} = 1$. I use UBS/Gallup aggregated means[57] to cross verify the expectations

[55]It is Question # 16 from February 1999 to December 2001, and Question # 12 form January 2002 to October 2007

[56]It is Question # 16A from February 1999 to December 2001, and Question # 13 from January 2002 to October 2007 .

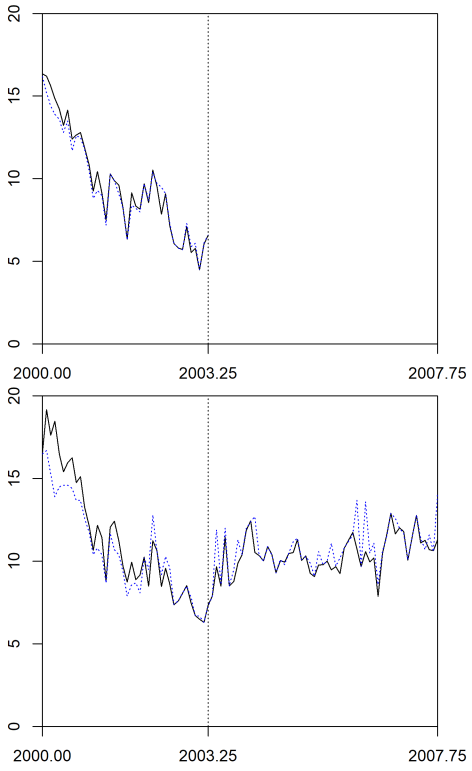
[57]April 2003 report contains monthly expected stock market returns aggregated by UBS/Gallup and can be found in the Reports, Data Tables & Other Materials section of "Gallup/UBS

that I get form microdata. Weighted mean of expected stock return from microdata is 99.1% correlated with UBS/Gallup aggregated mean of expected stock market return. Equally weighted mean of expected stock return from mictodata is 98.2% correlated with UBS/Gallup aggregated mean of expected stock market return. Weighted mean of expected portfolio return from microdata is 90.4 % correlated with UBS/Gallup aggregated mean of expected portfolio return. Equally weighted - 89.0 %.

The time series of corresponding aggregated weighted expected return proxies are provided below.

Figure 17. Aggregated UBS Gallup Investors' Expectations from Microdata, 12 month Ahead

Stock Market Return Expectations (top plot), Percent, Portfolio Return Expectations (dottom plot), Annual Percent Monthly: January 2000 - October 2007.



Poll # 2003-INVEST04: April, 2003 US Investor Optimism Index [Roper # 31089585]" poll, https://doi.roper.center/?doi=10.25940/ROPER-31089585. October 2007 report contains history of monthly expected portfolio returns aggregated by UBS/Gallup and can be found in the Reports, Data Tables & Other Materials section of "Gallup/UBS Poll # 2007-INVEST10: October, 2007 US Investor Optimism Index [Roper # 31089639]" poll. https://doi.roper.center/?doi=10.25940/ROPER-31089639. December 2001 report contains monthly expected interest rates https://doi.roper.center/?doi=10.25940/ROPER-31089569

Table XIV. Summary Statistics
UBS/Gallup Expected Annual Stock Market $\tilde{r}^{UBS/Gallup}_{m,t\to t+12}$ and Portfolio Return $r^{UBS/Gallup}_{port,t\to t+12}$ A Year From Now, Annual Percent Monthly: January 2000 - October 2007

	N	Mean	St. Dev.	Min	Max
From microdata. $\tilde{r}^{UBS/Gallup}_{m,t\to t+12}$	40	9.68	3.21	4.49	16.34
Aggregate. $\tilde{r}^{UBS/Gallup}_{m,t\to t+12}$	40	9.50	2.95	4.50	16.20
From microdata. $r^{UBS/Gallup}_{port,t\to t+12}$	93	10.77	2.53	6.31	19.18
Aggregate. $r^{UBS/Gallup}_{port,t\to t+12}$	93	10.78	2.14	6.30	16.70

Sampling error. All sample surveys are subject to sampling error, that is, the extent to which the results may differ from those that would be obtained if the entire population surveyed had been interviewed. The size of sampling errors depends largely on the number of interviews.

The following table may be used in estimating sampling error in the Gallup surveys. The computed allowances have taken into account the effect of the sample design upon sampling error. They may be interpreted as indicating the range (plus or minus figure shown) within which the results of repeated samplings in the same time period could be expected to vary, 95% of the time, assuming the same sampling procedure, the same interviewers, and the same questionnaire.

The table would be used in the following manner: Assume a reported percentage is 33 for a group that includes 1,000 respondents. Proceed to row "Percentages near 30" in the table and then to the column headed, "1,000." Figure in this cell is four, which means that at the 95% confidence level, the 33% result obtained in the sample is subject to a sampling error of plus or minus four points.

Table XV.

Gallup. Recommended Allowance For Sampling Error (Plus or Minus) at 95% Confidence Level, Points

Percentages near	Sample size					
	1,000	750	600	400	200	100
10	2	3	3	4	5	7
20	3	4	4	5	7	9
30	4	4	4	6	8	10
40	4	4	5	6	8	11
50	4	4	5	6	8	11
60	4	4	5	6	8	11
70	4	4	4	6	8	10
80	3	4	4	5	7	9
90	2	3	3	4	5	7

Appendix D. Survey of Professional Forecasters

The Survey of Professional Forecasters[58] is the oldest quarterly survey of macroeconomic forecasts in the United States. The survey began in 1968 and was conducted by the American Statistical Association and the National Bureau of Economic Research. The Federal Reserve Bank of Philadelphia took over the survey in 1990.

Sampling frame. The monthly survey uses an address-based mail sample design. The sampling frame is derived from the files created by the U.S. Postal Service, which represent near-universal coverage of all residential households in the United States. The CCS frame is updated monthly to ensure up-to-date coverage of U.S. households.

Sampling. The CCS uses a probability sample design to select each month's random sample from the household universe frame. The frame is first stratified geographically within the census division to provide a proportionate geographic distribution, after which a systematic sample of household addresses is selected. The sample addresses are then used for the

[58]https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters

mailing.

Sample size. About 3,500 surveys are completed each month. About 2,500 for end-of-month release; 3,500 for later revision.

Field period. The CCS mailing is scheduled so that the questionnaires reach sample households on or about the first of each month. Returns flow in throughout the collection period, with the sample close-out for preliminary estimates occurring around the eighteenth of the month. Any returns received after then are used to produce the final estimates for the month, which are published with the release of the following month's data. Completed questionnaires are checked in as they are received and then scheduled for data entry. Data fields are edited for invalid entries and, if necessary, are flagged for review. As part of the on-going quality control process, a random sample of questionnaires is selected for independent review/validation by a senior member of the data collection staff. The targeted responding sample size - approximately 3,000 completed questionnaires - has remained essentially unchanged throughout the history of the CCS.

Fieldwork. The Nielsen Company[59].

Weighting. To improve the accuracy of the estimates and ensure the proportionate representation of key categories in the estimates, the CCS uses a post-stratification weighting structure covering the following categories: Census Division (9 Census divisions), Age of Head of Household (<30, 30-39, 40-49, 50-59, 60+), Gender of Head of Household

[59]As of February 2011, The Conference Board has changed survey providers from TNS to The Nielsen Company for ongoing CCS operational support. Nielsen uses a mail survey specifically designed for the Consumer Confidence Survey. The new design uses a probability-design random sample, poststratification weights (for gender, income, geography, and age), and the U.S. Census X-12 seasonal adjustment. The CCS concepts, questions and mail survey collection method remain unchanged.

From September 2010 to January 2011, a five-month pilot test of the new sample design was conducted in parallel with the existing design. Three months of previously published data (November 2010 to January 2011) have been restated to smooth the transition, which makes November 2010 the effective changeover month.

(Male/Female), Income of Household (<15,000; 15,000-24,999; 25,000-34,999; 35,000-49,999; 50,000-74,999; 75,000-99,999; 100,000-124,999; 125,000+).

The post-stratification weighting uses an iterative proportional fitting technique for simultaneously balancing sample weights across several different population control groups. This technique ensures that sample-based estimates of the household population categories match the independent census population controls within +/- 1 percent.

Seasonal Adjustment. Data as of January 2011 use the Census X-12 seasonal adjustment software for the publication series where needed. Seasonal adjustment helps remove periodic seasonal fluctuations in the series due to events such as weather, holidays, and the beginning and end of the school year. While the CCS series are typically not highly seasonal, the X-12 software helps reduce any residual seasonality in the various data series.

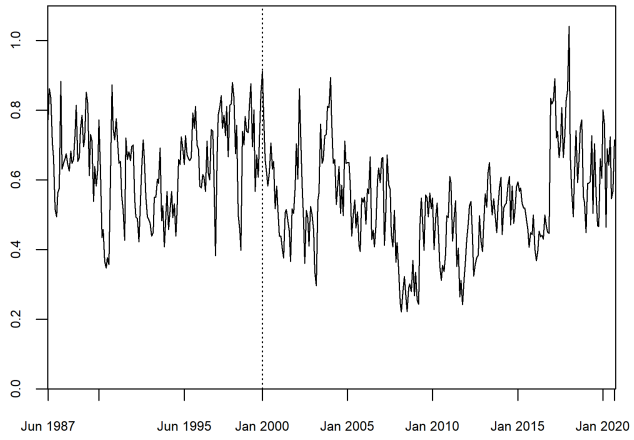
Release. Preliminary figures are released on last Tuesday of month. Final figures are released with next month's release.

Questions about return expectation. The surveys elicit respondents simple categorical beliefs about whether the stock prices will likely increase, decrease, or stay the same (or whether they are undecided, which we include in the same category).

As per Nagel & Xu (2019), I construct the Index IND_t^{CB} as the ratio of those who respond with an increase to the sum of those who respond with a decrease or the same:

$$IND_t^{CB} = \frac{n_t^{increase}}{n_t^{decrease} + n_t^{same}} \tag{A5}$$

Figure 18. Conference Board Index IND_t^{CB}, Monthly
Data is from June 1987 to September 2020



Appendix E. Panel on Household Finances (PHF) by Bundesbank since 2011

The German Panel on Household Finances (PHF)[60] is a panel survey on household finance and wealth in Germany, covering the balance sheet, pension, income, work life and other demographic characteristics of private households living in Germany. The panel survey is conducted by the Research Centre of the Deutsche Bundesbank.

The first two waves were carried out in 2010/2011 and 2014, respectively, in cooperation with infas Institut für angewandte Sozialwissenschaften, Bonn. Net samples of 3,565 (wave 1) and 4,461 (wave 2) randomly selected households were collected. The collection of the data of the third wave ended in November 2017. The data are currently in the preparation stage. First results and a scientific use file are expected to be published in early 2019. Around 5,000 households participated in the third wave.

Wealthy households are oversampled on the basis of microgeographic indicators in order to better match the distribution of wealth across households and to shed light on the composition of wealth. A strong attempt is being made to select households from all economic strata. Participation is strictly voluntary.

The survey is designed to be a full panel, i.e. all households are re-contacted. The

[60]https://www.bundesbank.de/en/bundesbank/research/panel-on-household-finances

intended survey frequency is three years. Almost half of the 4,461 households in wave two took part for the second time.

The results of the first waves of our study were published in several Bundesbank monthly bulletin articles, reports and papers. The micro data from wave one and two are available for scientific research projects through the Bundesbank's Research Data and Service Centre.

Aside from being an encompassing survey on household finance in Germany, PHF is an integral part of the Household Finance and Consumption Survey (HFCS). This system of wealth surveys collects ex ante harmonised micro data in every country of the euro area.

Appendix F. European Community Household Panel by ECB and member national banks

The European Community Household Panel (ECHP) [61] is an eight-year, longitudinal household survey covering 14 EU member states from 1994 to 2001. For more recent, comparable, panel data, are in EU Statistics on Income and Living Conditions (EU-SILC) and described below.

Subject interviews in ECHP covered: overall financial situation, income data, working life, housing, social relations, health and biographical observations. EU Member States included in ECHP were Austria, Belgium, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden and the United Kingdom.

These interviews cover a wide range of topics concerning living conditions. They include detailed income information, financial situation in a wider sense, working life, housing situation, social relations, health and biographical information of the interviewed.

The total duration of the ECHP was 8 years, running from 1994 to 2001 (8 waves). The then Member States involved were Belgium, Denmark, Germany, Ireland, Greece, Spain, France, Italy, Luxembourg, the Netherlands, Austria, Portugal, Sweden and the United Kingdom. As from 2003/2004, the EU-SILC survey covers most of the above-mentioned

[61] https://ec.europa.eu/eurostat/web/microdata/european-community-household-panel

topics.

The ECHP consists of panel data, meaning that the same respondents within each country have answered the survey year after year. All EU households and citizens of 16 years of age or more are in the target population. Common sampling requirements and standards are employed in all countries – probability sampling procedures are used. In the first wave, 60,500 households and 130,000 individuals were interviewed.

The data is collected through face-to-face interviews

Plenty of the datasets under the Income and living conditions (ILC) domain under theme "Population and social conditions" contain ECHP based data for the above mentioned periods. This includes several indicators on monetary poverty and distribution of income, which are analysed in different ways (eg. different cut-off thresholds, by age, gender, activity status, tenure status...).

There is also a selection of indicators on non-monetary deprivation derived from ECHP, notably on housing conditions.

Some indicators in the health care collections of the public health domain are derived from ECHP as well.

Appendix G. EU Statistics on Income and Living Conditions (EU-SILC)

EU-SILC is a cross-sectional and longitudinal sample survey, coordinated by Eurostat, based on data from the European Union member states. EU-SILC provides data on income, poverty, social exclusion and living conditions in the European Union. EU-SILC stands for 'European Union Statistics on Income and Living Conditions.' There are two data scopes:

- Cross-sectional data pertaining to fixed time periods, with variables on income, poverty, social exclusion and living conditions, and
- Longitudinal data pertaining to individual-level changes over time, usually observed over four years.

Details of the database are on the Eurostat EU-SILC resource page. The 2019 EU-SILC data coverage table is at this link.

Social exclusion and housing-condition observations are collected at household level. Income data is collected at personal level, with some components included in the 'Household' section. Labour, education and health observations only apply to persons aged 16 or older.

EU-SILC was established to provide data on structural indicators of social cohesion (at-risk-of-poverty rate, S80/S20 and gender pay gap) and to provide relevant data for the two 'open methods of coordination' in the field of social inclusion and pensions in Europe.

The EU-SILC 2019 release extended data coverage from Junem 2015 to Novemberm 2019. Eurostat periodically issues revisions of earlier waves. The data dossier is structured as follows:

- Data: Cross-sectional
- Data: Longitudinal
- Documentation
- Metadata (for all waves)

Appendix H. Survey of Consumer Expectations by New York Fed since 2013

The New York Fed's Survey of Consumer Expectations (SCE)[62] gathers information on consumer expectations regarding inflation, household finance, the labor and housing markets, and other economic issues. Its overall goal is to fill the gaps in existing data sources (such as the University of Michigan Survey of Consumers, the Federal Reserve Board's Survey of Consumer Finances, and the Bureau of Labor Statistics' Consumer Expenditure Survey) pertaining to household expectations and behavior by providing a more integrated data approach.

The SCE started in June 2013, after a six-month initial testing phase. It is a nationally representative, internet-based survey of a rotating panel of about 1,300 household heads,

[62]https://www.newyorkfed.org/microeconomics/sce

where household head is defined as the person in the household who owns, is buying, or rents the home. The survey is conducted monthly. New respondents are drawn each month to match various demographic targets from the American Community Survey (ACS), and they stay on the panel for up to twelve months before rotating out. The survey instrument is fielded on an internet platform designed by the Demand Institute, a nonprofit organization jointly operated by the Conference Board and Nielsen. The respondents for the SCE come from the sample of respondents to the Consumer Confidence Survey (CCS), a mail survey conducted by the Conference Board. In turn, the respondents for the CCS are selected from the universe of U.S. Postal Service addresses. From that universe, a new random sample is drawn each month, stratified only by Census division.

The SCE has several components. First, it includes a core monthly module on expectations about a number of macroeconomic and household-level variables. In this module, respondents are asked about their inflation expectations, as well as their expectations regarding changes in home prices and the prices of various specific spending items, such as gasoline, food, rent, medical care, and college education. The core survey also asks for expectations about unemployment, interest rates, the stock market, credit availability, taxes, and government debt. In addition, respondents are asked to report their expectations about several labor market outcomes that pertain to them, including changes in their earnings, the perceived probability of losing their current job (or leaving their job voluntarily), and the perceived probability of finding a job. Finally, the core survey asks about the expected change in respondent households' overall income and spending. As described in more detail below, these questions about expectations are fielded at various time horizons and with various formats, including both point and density forecasts. Second, each month, the SCE contains a supplementary "ad hoc" module on special topics. Three such modules are repeated every four months, leaving three "floating" supplements per year on topics that are determined as the need arises. The three repeating supplements are on credit access, labor market, and spending. Topics covered so far in the "floating" supplement include (but are

not limited to) the Affordable Care Act, student loans, workplace benefits such as childcare and family leave, and the use of insurance products.

Together, the core monthly module and the monthly supplement take about fifteen minutes to complete. Finally, SCE respondents also fill out longer surveys (up to thirty minutes in length, and separate from the monthly survey) each quarter on various topics. Most of these surveys are repeated at a yearly frequency. Since each SCE panelist stays in the panel for up to twelve months, these annual surveys can be used as independent repeated cross sections, although they obviously can be linked to the monthly core survey panel responses. The SCE currently contains quarterly surveys on the housing market, the labor market, informal work participation, and consumption, saving, and assets. A subset of these surveys is designed in part or wholly by other Federal Reserve Banks.

Appendix I. Survey of Household Economics and Decision making by the Federal Reserve Board since 2013

The Federal Reserve Board has conducted the Survey of Household Economics and Decisionmaking (SHED)[63], which measures the economic well-being of U.S. households and identifies potential risks to their finances. The survey includes modules on a range of topics of current relevance to financial well-being including credit access and behaviors, savings, retirement, economic fragility, and education and student loans. [64]

The survey is designed with three primary motivations

1. Monitor trends in consumer behavior and sentiment particularly among low- and moderate-income populations
2. Cast light on current issues affecting financial well-being

[63] https://www.federalreserve.gov/consumerscommunities/shed.htm

[64] Also, the SHED asks about informal income-earning activities that happen outside of formal work. Many types of arrangements that are included in other studies—such as temp-agency work or subcontracted work—are unlikely to be included, whereas activities that are excluded by other studies—such as working under the table and selling goods—are included. Informal and independent work overlap, but are not synonymous.

3. Fill data gaps and provide insights into questions for which there may not be other reliable data sources.

It's conducted annually in the fourth quarter of each year since 2013.

Ipsos, a private consumer research firm, administers the survey using its KnowledgePanel, a nationally representative probability-based online panel. Ipsos selects respondents for the KnowledgePanel based on address-based sampling (ABS)[65] SHED sample is made up of three components:

- New respondents randomly selected (3,054 adults),
- Oversample of adults with household income under $40,000 (1,556 adults),
- Reinterviewed respondents from 2015 SHED survey (2,033 adults).

Appendix J. Canadian Survey of Consumer Expectations since 2015

The Canadian Survey of Consumer Expectations (CSCE) [66] is a quarterly survey aimed at measuring household views of inflation, the labour market and household finances, as well as topical issues of interest to the Bank of Canada. The CSCE also provides data by age, geography, income and education.

The Canadian Survey of Consumer Expectations is a nationally representative, internet-based quarterly survey of a rotating panel of approximately 2,000 heads of households.2 It is administered by a large polling firm on behalf of the Bank of Canada. Respondents participate in the panel for up to a year, with a roughly equal number joining and leaving the panel each quarter. This reduces variability caused by changes in composition, allowing for greater stability and precision in the estimates. The survey's target population is adult residents of Canada aged 18 or older. The survey is conducted in February, May, August and November and is offered in both English and French. Respondents answer questions

[65]Prior to 2009, respondents were also recruited using random-digit dialing.

[66]https://www.bankofcanada.ca/publications/canadian-survey-of-consumer-expectations/ https://www.bankofcanada.ca/publications/canadian-survey-of-consumer-expectations/canadian-survey-of-consumer-expectations-references/

about inflation, the labour market and household finances and demographic questions about themselves and their household.

Appendix K. Online Survey of Consumer Expectations by Bundesbank in 2019

The Bundesbank is currently undertaking a pilot study to investigate whether a regular consumers expectation survey can provide information that is useful for policy-making[67]. There are chiefly two questions that are of relevance in this context:

Would such a study be able to supply the Bundesbank and policymakers with a broad and up-to-date picture of consumers' economic expectations in Germany? To what extent can the obtained data assist the Bundesbank's and other institutions' researchers in analysing current economic developments? The survey and the questionnaire were designed and prepared by the Deutsche Bundesbank's Research Centre in cooperation with external experts. The public opinion research company forsa has been commissioned with conducting the survey. The pilot survey will initially comprise three waves containing both recurring and wave-specific questions. For each wave of the survey, around 2,000 representative members of the general public will be asked to respond. Some of the respondents will be asked multiple times. Participation in the study is voluntary and will take about 20 minutes.

The collected data will be used exclusively for the production of statistics, for monetary and financial stability purposes, as well as for study and research. There will be no commercial use. The collected data will always be stored separately from personal data and identification of individual persons will not be possible, even for the researchers at the Bundesbank.

[67] https://www.bundesbank.de/en/bundesbank/research/pilot-survey-on-consumer-expectations/bundesbank-online-pilot-survey-on-consumer-expectations-794568

Appendix L. Ifo Business Tendency Survey

The Ifo Business Climate Survey [68] is a leading indicator of German economic activity, compiled by the Munich-based Ifo Institute for Economic Research.

The Ifo Business Climate Survey is based on approximately 9,000 monthly survey responses from German firms in manufacturing, construction, the service sector, and trade. The companies surveyed are asked to provide feedback on whether their current business situation is good, satisfactory, or poor, as well as assess their expectations for the next six months as either more favorable, unchanged, or more unfavorable.

The responses of the firms are weighted according to the economic importance of each industry, and a net balance is calculated for each assessment: good/poor for the current situation, and more favorable/more unfavorable for the outlook—the "satisfactory" and "unchanged" responses are regarded as neutral and thus not included.

The business climate itself, the main subject of the survey, is then calculated as the mean of these two balances. The outcome is constructed to yield outcomes between –100, assuming every firm gives a negative response to both questions, and +100, meaning every firm gives a positive response to both questions.

The headline survey number that is released is, however, recalculated in the form of an index, which will be set to 100 in a base year. The base year currently in use is 2005.

Appendix M. Atlanta Fed Survey of Business Uncertainty since 2015

In partnership with Steven Davis of the University of Chicago Booth School of Business and Nicholas Bloom of Stanford University, the Federal Reserve Bank of Atlanta has created the Atlanta Fed/Chicago Booth/Stanford Survey of Business Uncertainty (SBU). This innovative panel survey measures the one-year-ahead expectations and uncertainties that firms have about their own employment, capital investment, and sales. The sample covers all regions of the U.S. economy, every industry sector except agriculture and government,

[68] https://www.ifo.de/en

and a broad range of firm sizes.

The SBU elicits a 5-point probability distribution over 12-month-ahead sales, employment, and capital expenditures for each firm. It also elicits current values of these quantities. The survey's innovative design allows the calculation of each firm's expected growth rate over the next year and its degree of uncertainty about its expectations. Policy makers and researchers can use SBU data to help forecast economic activity and better understand how business expectations and uncertainty affect employment, sales, investment, and other economic outcomes.

Each survey form below goes to about one-third of the panel members each month. A given panel member will receive each of the three forms over the course of three months. In addition to the core survey questions posed in the forms below, we typically ask at least one special question each month.

Appendix N. Vanguard Research Initiative

The Vanguard Research Initiative (VRI)[69] is a collaboration of the University of Michigan, New York University, and Vanguard.

VRI surveys are administered via the internet to a panel of Vanguard clients to gather complementary information to Vanguard's administrative data. The panel was chosen by inviting Vanguard account holders fulfilling the following criteria: over 55 years old, have a domestic address, no immediate record of a Vanguard annuity purchase, hold between $10,000 and $5 million in assets with Vanguard, have a valid email registered with Vanguard, have logged on in the past six months.

The sample was stratified such that each age group above 55 will be adequately represented, as well as singles. The sample is also divided between individual accounts and employer-sponsored accounts.

The initial cohort of 9,000 respondents joined the VRI in 2013 with Survey 1. A new

[69] https://ebp-projects.isr.umich.edu/VRI/index.html

cohort of 3,700 respondents joined the VRI in 2016 with Survey 5. The original cohort was also given Survey 5.

The project employs data on a panel of savers that includes detailed wealth, health, and demographic information. This panel data set comprises over 9,000 Vanguard clients. By the joint use of administrative account data and surveys, the project employs an innovative infrastructure for understanding the decisionmaking and well-being of older Americans. The project also innovates by combining this distinctive measurement infrastructure with survey questions, modeling, and estimation that can yield precise quantification of the considerations that affect decisionmaking and well-being leading up to retirement and during retirement. A key innovation is to pose strategic survey questions (SSQs), a form of contingent stated preference question.

Appendix B. Imputation of Point Estimates

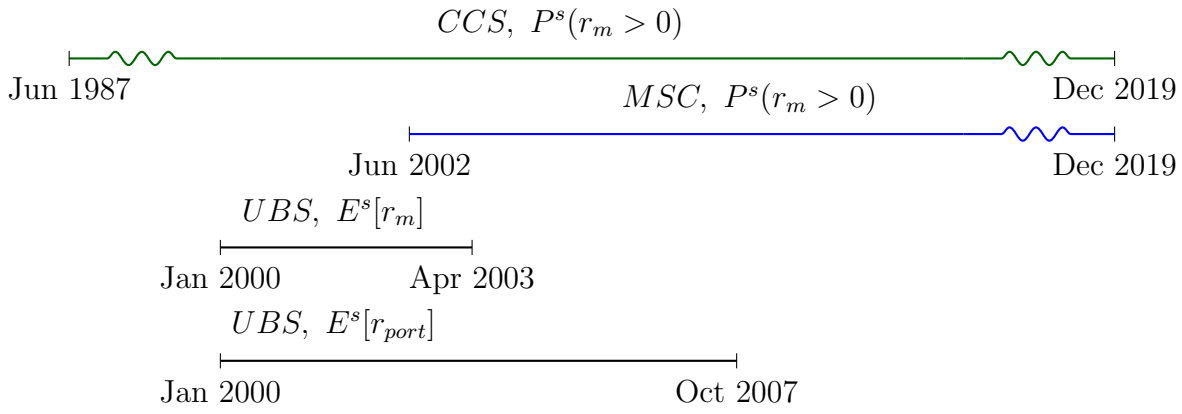
Appendix A. Imputation of Percent Return Expectations

As the Consumer Confidence Survey (CCS) from the Conference Board[70], and the Michigan Survey of Consumers[71] (MSC) surveys contain coarse probability estimations of stock market return, and UBS/Gallup surveys contain percent return expectations, I use UBS/Gallup[72] data to construct benchmark source of expectations.

Figure 19 shows timing of the CCS, the MSC and the UBS/Gallup surveys. The top green line shows timing of the CCS survey that starts on June 1987 and ends on December 2019. The blue line shows timing of the MSC survey that starts on June 2002 and ends on December 2019. The UBS/Gallup surveys overlap the CSS and the MSC surveys. However, it is clear that regressing subjective market growth probabilities on UBS/Gallup subjective market returns directly would not work for the MSC data, as the UBS/Gallup

[70] https://conference-board.org/data/consumerdata.cfm
[71] https://data.sca.isr.umich.edu/
[72] https://ropercenter.cornell.edu/

Figure 19. Timing and Output of Surveys on Expected Subjective Market Return Over the Next 12 Months
Two top lines, green and blue, show timing of the CCS Conference Board survey and the Michigan Surveys of Consumers. The surveys provide subjective probability that market return over the next 12 months will be positive, $P^s(r_m > 0)$. Bottom two lines show timing of UBS/Gallup surveys on expected market return. The upper of the two lines shows timing of UBS survey question asking about subjective expected stock market return in percent, $E^s[r_m]$. The lower of the two lines shows timing of UBS survey question about expected portfolio return.



subjective market return expectations overlaps with UBS/Gallup market returns only in ten points, from June 2002 to April 2003. The UBS/Gallup data allows to extend subjective expectations in percents by utilizing a UBS/Gallup survey on subjective portfolio return expectations. $E^s[r_{port}]$ that overlaps with the MSC survey over five and a half years, from June 2002 to October 2007.

As a result, following Nagel & Xu (2019) methodology, I fit the UBS/Gallup return expectations to MSC and Conference Board probability estimates and impute the MSC and CCS percent return expectations in three steps.

First, I expand UBS/Gallup stock market expectations. Form January 2000 to April 2003 UBS/Gallup reports both expectations of stock market return and portfolio return. From May 2003 to October 2007, the UBS/Gallup survey respondents report only the return that they expect on their own portfolio. Following Nagel & Xu (2019) I impute market return expectations by regressing subjective expected market returns $\tilde{r}^{UBS,\ 01-2000\ to\ 04-2003}_{m,t\to t+12}$

Table XVI. OLS Regression Specification

The regression is used for imputation of Market Return Expectations,
$\tilde{r}^{UBS,\ 01-2000\ to\ 04-2003}_{m,t\to t+12} = a_p + b_p \tilde{r}^{UBS,\ 01-2000\ to\ 04-2003}_{port,t\to t+12} + \epsilon_t$

	Dependent variable:
	$\tilde{r}^{UBS,\ 01-2000\ to\ 04-2003}_{m,t\to t+12}$, %
$\tilde{r}^{UBS,\ 01-2000\ to\ 04-2003}_{port,t\to t+12}$, %	0.867^{***}
	(0.033)
Constant	-0.073
	(0.394)
Observations	40
R^2	0.947
Adjusted R^2	0.945

on individual subjective expected portfolio returns $\tilde{r}^{UBS,\ 01-2000\ to\ 04-2003}_{port,t\to t+12}$ using the sample segment where both variables are provided and employing the fitted value from that regression $\hat{r}^{USB}_{m,t\to t+12}$ when market return expectations are not provided.

$$\tilde{r}^{UBS,\ 01-2000\ to\ 04-2003}_{m,t\to t+12} = a_p + b_p \tilde{r}^{UBS,\ 01-2000\ to\ 04-2003}_{port,t\to t+12} + \epsilon_t \tag{B1}$$

$$\hat{r}^{USB}_{m,t\to t+12} = a_p + b_p \tilde{r}^{UBS}_{port,t\to t+12} \tag{B2}$$

Because during this overlap period the movements in the expectations of returns are highly correlated, the overlap allows me to map the expected portfolio returns into expected market return over the entire period upto October 2007.

A blue line represents the fitted expected stock market return $\hat{r}^{USB}_{m,t\to t+12}$ in the plot below.

Second, to impute percentage expectations from MSC and Conference Board estimates, I regress the fitted percentage expectations $\hat{r}^{UBS}_{m,t\to t+12}$ from the UBS/Gallup on the MSC probability $\bar{p}^{MSC}_t$ and on IND^{CB}_t ratio. Since the Conference Board surveys ask about stock price increases, I subtract the current dividend yield[73] of the CRSP value weighted index from the dependent variable in this regression and add it back to the fitted value.

Particularly, for the interval from January 2000 to October 2007 I run the following

[73]CRSP: "Dividend Yield is another name for Income Return. It is the ratio of the ordinary dividends of a security or index to the previous price."

regressions to get coefficients for predictive regression $\{a_{MSC}, b_{MSC}; a_{BC}, b_{CB}\}$:

$$\hat{r}^{UBS}_{m,t\to t+12} - y^{div}_t = a_{MSC} + b_{MSC}\bar{p}^{MSC}_t + \epsilon_t \tag{B3}$$

$$\hat{r}^{UBS}_{m,t\to t+12} - y^{div}_t = a_{BC} + b_{CB}IND^{CB}_t + \epsilon_t \tag{B4}$$

where $\hat{r}^{UBS}_{m,t\to t+12}$ is expanded UBS/Gallup subjective expected market return and y^{div}_t is dividend yield calculated calculated from the CRSP as

$$y^{div}_t = \frac{vwretd_t + 1}{vwretx_t + 1} - 1 \tag{B5}$$

where $vwretd_t$ is value-weighted return including distributions and $vwretx_t$ is value-weighted return. Next, I use the coefficients $\{a_{MSC}, b_{MSC}; a_{BC}, b_{CB}\}$ and $\{\bar{p}^{MSC}_t, IND^{CB}_t\}$ coarse subjective probability estimates from the MSC and CB to impute subjective expected market returns in percents for the period from January 2000 to December 2020:

$$\tilde{r}^{MSC,no\ div}_{m,t\to t+12} = a_{MSC} + b_{MSC}\bar{p}^{MSC}_t \tag{B6}$$

$$\tilde{r}^{CB,no\ div}_{m,t\to t+12} = a_{CB} + b_{CB}IND^{CB}_t \tag{B7}$$

and add back dividend yield to come up to subjective expected market return with dividends.

$$\tilde{r}^{MSC}_{m,t\to t+12} = \tilde{r}^{MSC,no_div}_{m,t\to t+12} + y^{div}_t \tag{B8}$$

$$\tilde{r}^{CB}_{m,t\to t+12} = \tilde{r}^{CB,no_div}_{m,t\to t+12} + y^{div}_t \tag{B9}$$

As a results, I have long time series of monthly subjective expected market returns in percent $r^{MSC}_{m,t\to t+12}$ and $r^{CB}_{m,t\to t+12}$ that I will use in my analysis.

Figure 20 shows the imputed subjective expected stock market returns in the next twelve months from the MSC (dashed blue line), $\tilde{r}^{MSC}_{m,t\to t+12}$, and from the CCS (solid green line), $\tilde{r}^{CB}_{m,t\to t+12}$. Gray areas are recessions. We see that subjective expected market returns for a

Figure 20. Subjective Expected Market Return A Year Ahead, %
The green solid line corresponds to subjective market return imputed from the Conference Board, $\tilde{r}^{CB}_{m,t\to t+12} - \tilde{r}^{CB}_{f,t\to t+12}$. The blue dashed line is subjective market return imputed from the MSC $\tilde{r}^{MSC}_{m,t\to t+12} - \tilde{r}^{MSC}_{f,t\to t+12}$. The monthly time series are in percent and span from January 2000 to December 2019. Gray areas are NBER recessions.

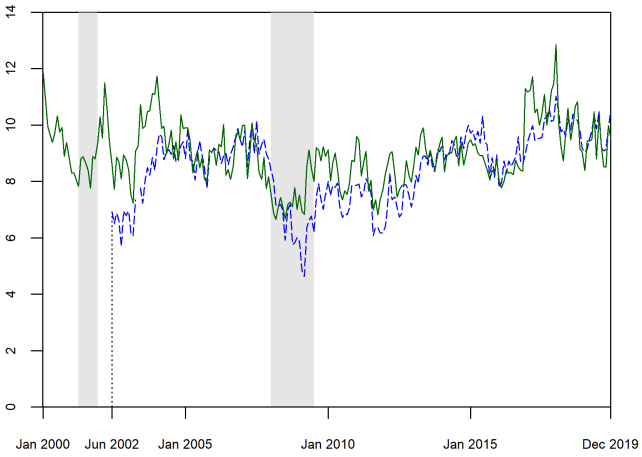
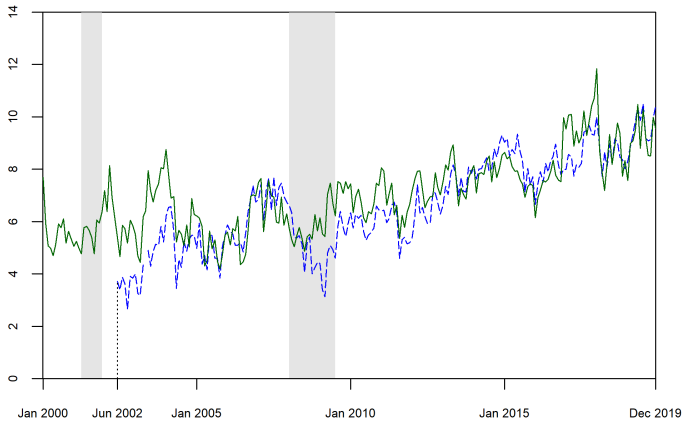


Figure 21. Subjective Expected Excess Market Return A Year Ahead, %
The green line corresponds to subjective excess market return imputed from the Conference Board, $\tilde{r}^{CB}_{m,t\to t+12} - \tilde{r}^{CB}_{f,t\to t+12}$. The blue line is subjective excess market return imputed from the MSC $\tilde{r}^{MSC}_{m,t\to t+12} - \tilde{r}^{MSC}_{f,t\to t+12}$. The monthly time series are in percent and span from January 2000 to December 2019. Gray areas are NBER recessions.



year ahead are positive and highly correlated. The correlation is 82.6 %. Both $\tilde{r}^{MSC}_{m,t\to t+12}$ and $\tilde{r}^{CB}_{m,t\to t+12}$ are time-varying and respond to recession of 2008-2009. They decrease beforehand and start moderate growth during the recession.

Table XVII. Moments of Imputed Subjective Expected Market Return

Variables	N	Mean	St. Dev.	Min	Max
Expected Subjective Stock Market Return					
$\tilde{r}_t^M$	210	8.24	1.30	4.26	11.01
$\tilde{r}_t^C$	240	8.98	1.09	6.66	12.84

Appendix B. Imputation of Percent Interest Rate Expectations

I use the quarterly average of the daily levels of 3-month treasury bill rate expected over next four quarters from the Survey of Professional Forecasters[74], SPF, as a benchmark.

$$\bar{r}^{SPF}_{f,q\to q+4} = \frac{1}{4}\sum_{j=0}^{4} \tilde{r}^{SPR}_{q+j\to q+j+1|q} \tag{B10}$$

where q is a quarter.

As the SPF forecasters include non-random, self-selected representatives from academia, government, labor, consulting and banking, I cannot use the SPF forecast as is. I use the SPF forecasts as a benchmark to impute investors expectations from the MSC and Conference Board surveys.

The MSC and Conference Board surveys contains investors' coarse probability estimation of an interest rate change. As the MSC and the Conference Board interest rate questions do not mention Treasury rate, the expected interest rate might include risk premium. Given extensive empirical evidence, I assume that interest rate expectations are associated with Treasury yield expectations and impute expected one-year Treasury yield using Nagel & Xu methodology.

[74]Survey of Professional Forecasters

Table XVIII. Moments of Imputed Subjective Expected Risk-Free Rate

Variables	N	Mean	St. Dev.	Min	Max
Expected Subjective Risk-Free Rate					
$\tilde{r}^{M}_{f,t}$	240	2.02	1.27	0.00	5.61
$\tilde{r}^{C}_{f,t}$	240	2.00	1.17	0.00	5.00

To account for trend stationary, I add trend component t to the imputation regression:

$$\bar{r}^{SPF}_{f,q\to q+4} = a_{MSC} + b_{MSC}t + c_{MSC}\bar{p}^{MSC}_{t} + d_{MSC}t\bar{p}^{MSC}_{t} + \epsilon_t \tag{B11}$$

$$\bar{r}^{SPF}_{f,q\to q+4} = a_{CB} + b_{CB}t + c_{CB}IND^{CB}_{t} + d_{CB}t\bar{p}^{MSC}_{t} + \epsilon_t \tag{B12}$$

As the quarterly rates are daily averages within a quarter, I treat quarterly rates r_q as monthly rates within a quarter. So quarterly rates within a year $\{q_1, q_2, q_3, q_4\}$ are mapped to monthly rates as

$\{q_1, q_1, q_1, q_2, q_2, q_2, q_3, q_3, q_3, q_4, q_4, q_4\}$.

I use fitted values from the imputation regressions as expected subjective risk-free rates:

$$\tilde{r}^{MSC}_{f,t\to t+12} = a_{MSC} + b_{MSC}t + c_{MSC}\bar{p}^{MSC}_{t} + d_{MSC}t\bar{p}^{MSC}_{t} \tag{B13}$$

$$\tilde{r}^{CB}_{f,t\to t+12} = a_{CB} + b_{CB}t + c_{CB}IND^{CB}_{t} + d_{CB}t\bar{p}^{MSC}_{t} \tag{B14}$$

I use long monthly series $\tilde{r}^{MSC}_{f,t\to t+12}$ and $\tilde{r}^{CB}_{f,t\to t+12}$ as subjective expected risk-free rates.

The black solid line on the plot below shows the SPF forecast of three-month treasury bill rate four quarters ahead. It is highly correlated with actual one-year Treasury yield (black dotted line). Plot also shows imputed subjective expectations of interest rate form MSC (blue line) and Conference Board (green line) surveys.

Table XIX. Top 10 InvesText's Information Providers by Number of Published Equity Reports About the Dow Companies

Tables show top 10 information providers by number of reports published, percent of number of reports published, number of covered companies, period when a information provider is in the InvesText and tickers of covered companies. There are 46 tickers in InvesTest out of 47 the Dow companies that were in the Dow in the sample period. Sample is from January 1, 2000 to December 31, 2019.

information provider	Percent of Reports, %	# Companies Covered	In InvesText		Tickers Covered
			From	To	
Credit Suisse	14.9	46	2000-01	2019-12	All
Oppenheimer & Co., Inc.	6.2	45	2000-01	2019-12	All except DOW
Deutsche Bank	6.0	46	2000-01	2019-12	All
JPMorgan	4.3	46	2000-01	2019-12	All
Cowen and Company	4.3	43	2000-01	2019-12	All except AIG, GM, TRV
RBC Capital Markets	4.1	46	2000-01	2019-12	All
Wells Fargo Securities, LLC	3.8	43	2000-01	2019-12	All except DOW, KODK, MMM
Bear Stearns & Co. Inc.	3.4	41	2000-01	2008-05	All except CRM, DOW, GM, PM, V
Refinitiv StreetEvents	3.1	46	2002-01	2019-12	All
Piper Sandler Companies	2.8	44	2000-01	2019-12	All except DOW, RTX

Appendix C. Information Providers

Appendix A. Ton 10 information providers

Appendix B. Ton 10 Information Providers in Each Group

Table XX. Top 10 InvesText's Information Providers by Number of Published Equity Reports About the Dow Companies

Tables show top 10 information providers in 5th and 1-4 quantiles, percent of number of reports published, number of covered companies, period when a information provider is in the InvesText and tickers of covered companies. There are 46 tickers in InvesTest out of 47 the Dow companies that were in the Dow in the sample period. Sample is from January 1, 2000 to December 31, 2019.

information provider	Percent of Reports In Group, %	# Companies Covered	In InvesText		Tickers Covered
			From	To	
Top 10 in 5th Quantile					
Credit Suisse	14.9	46	2000-01	2019-12	All
Oppenheimer & Co., Inc.	6.2	45	2000-01	2019-12	All except DOW
Deutsche Bank	6.0	46	2000-01	2019-12	All
JPMorgan	4.3	46	2000-01	2019-12	All
Cowen and Company	4.3	43	2000-01	2019-12	All except AIG, GM, TRV
RBC Capital Markets	4.1	46	2000-01	2019-12	All
Wells Fargo Securities, LLC	3.8	43	2000-01	2019-12	All except DOW, KODK, MMM
Bear Stearns & Co. Inc.	3.4	41	2000-01	2008-05	All except CRM, DOW, GM, PM, V
Refinitiv StreetEvents	3.1	46	2002-01	2019-12	All
Piper Sandler Companies	2.8	44	2000-01	2019-12	All except DOW, RTX
Top 10 in 1-4th Quantile					
Miller Tabak & Co.	1.7	31	2010-06	2015-06	AA, AAPL, AIG, AMGN, AXP, BA, BAC, C, CAT, CRM, CSCO, CVX, DD, DIS, GE, GS, INTC, JNJ, JPM, KO, MCD, MRK, MSFT, PFE, RTX, T, UNH, VZ, WBA, WMT, XOM
Summit Insights Group	1.7	12	2012-10	2019-10	AAPL, CRM, CSCO, DD, HPQ, IBM, INTC, MRK, MSFT, NKE, T, VZ
SEENSCO	1.7	34	2014-09	2019-10	AAPL, AMGN, AXP, BA, BAC, CAT, CSCO, CVX, DD, DIS, GE, GS, HD, IBM, INTC, JNJ, JPM, KO, MCD, MMM, MO, MRK, MSFT, NKE, PFE, PG, RTX, T, TRV, UNH, V, VZ, WMT, XOM
Crispidea	1.7	31	2014-04	2019-12	AAPL, AMGN, BA, BAC, C, CAT, CRM, CSCO, CVX, DD, DIS, GE, GS, HON, HPQ, IBM, INTC, JNJ, JPM, KO, MCD, MMM, MRK, MSFT, PFE, RTX, T, UNH, VZ, WMT, XOM
Rosenblatt Securities, Inc.	1.6	8	2014-10	2019-12	AA, AAPL, CRM, CSCO, DIS, INTC, MSFT, T
Desjardins Securities	1.5	10	2004-02	2016-08	AA, CAT, CSCO, GM, HPQ, INTC, IP, PM, T, WMT
Yuanta Research	1.5	9	2007-07	2019-10	AAPL, BA, C, CSCO, HPQ, IBM, INTC, MSFT, NKE
Tucker Anthony Sutro Capital Markets	1.5	27	2000-01	2001-10	AXP, BA, BAC, C, CSCO, CVX, DD, DIS, GE, HD, HON, HPQ, IBM, INTC, JPM, MMM, MRK, MSFT, NKE, PFE, PG, RTX, T, VZ, WBA, WMT, XOM
Berenberg	1.5	23	2003-01	2019-12	AA, AAPL, BA, BAC, C, CAT, CRM, CSCO, CVX, GE, GM, GS, HON, IBM, JPM, MO, MRK, MSFT, NKE, PFE, PG, PM, XOM
Acquisdata	1.5	25	2014-05	2019-12	AA, AAPL, AMGN, BA, BAC, C, CAT, CRM, CSCO, CVX, DIS, GM, GS, HPQ, IBM, INTC, JNJ, JPM, MRK, MSFT, PFE, RTX, T, VZ, XOM

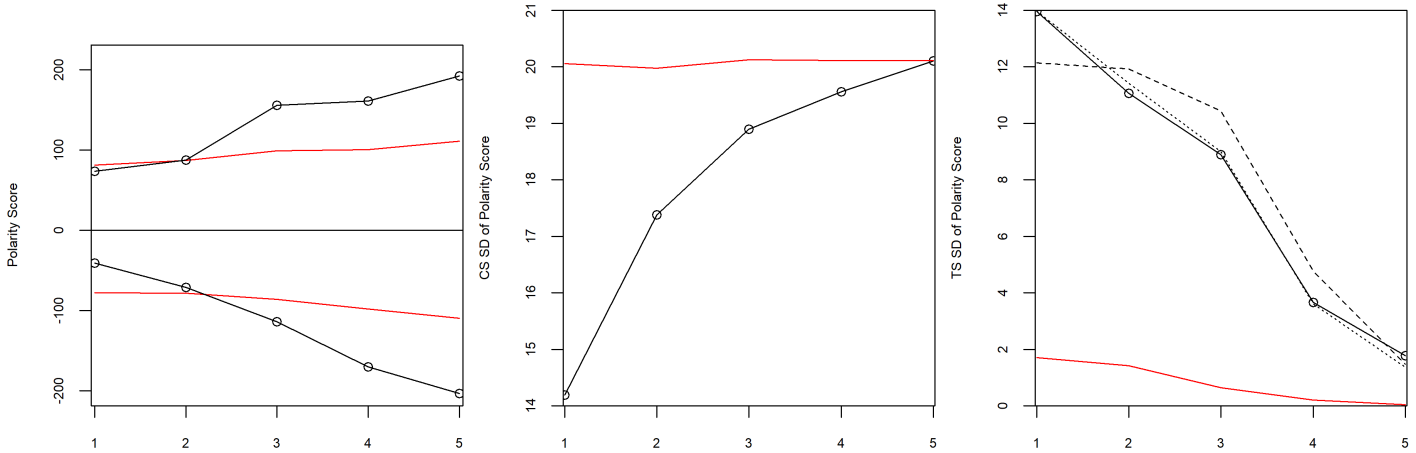
Appendix D. Types of Providers

Figure 22. Measures of Cross-News and Time-Series Dispersion of Sentiment About Earnings Growth Per information provider Cluster

The left plot shows minimum, $\min(s^{cl}) == \min s_{e,d,f,c}|cl$, and maximum, $\max(s^{cl}) = \max s_{e,d,f,c}|cl$, of sentiments about earnings growth of individual reports per cluster cl per editorial e, day d, information provider f and company c, for $cl \in \{1, \ldots, 5\}$. Red lines are random minimum and maximum of a simulated sample of the size n^{cl} from Normal distribution with mean and standard deviation of a cross-news cluster number five $N(1.99, 20.10)$.

Right plot shows cross-news sample standard deviation of sentiment about earnings growth per information provider cluster. It is calculated as $\bar{\sigma}^{cl} = \sqrt{\frac{1}{n^{cl}-1}\sum(s^{cl}_{e,d,f,c} - \bar{s}^{cl})^2}$, for $cl \in \{1, \ldots, 5\}$, e is an event type, d is a day , f is an information provider and c is a company, $\bar{s}^{cl} = E[p(s_{e,d,f,c})|cl]$ is mean of sentiment about earnings growth per information provider cluster. Red line is a mean of a simulated sample of the size n^{cl} from Normal distribution with cross-sectional mean and standard deviation of a cluster number five $N(1.99, 20.10)$.

Bottom plot shows a sample standard deviation of monthly average sentiment about earnings growth per information provider's cluster. For every cluster, the sample standard deviation of monthly average sentiment is calculated as $\bar{\sigma} = \sqrt{\frac{1}{n-1}\sum(s_t - \bar{s})^2}$, where $s_t = \{s^m_t, s_t, s^w_t\}$ and $\bar{s} = \{\bar{s}^m, \bar{s}, \bar{s}^w\}$. Solid black line is a standard deviation of s^f_t per cluster. Dashed line is a standard deviation of s^m_t per cluster. Dotted line is a standard deviation of s^w_t per cluster. Red line a standard deviation of a simulated sample of the size a cluster from Normal distribution with mean and standard deviation of a cross-news cluster number five $N(1.99, 20.10)$. The sample is from July 1, 2002 to December 31, 2019.



(a) Cross-news: Range of Individual Sentiment Per Cluster

(b) Cross-news: SD of Individual Sentiment Per Cluster

(c) Time-series: SD of Monthly Sentiment Per Cluster

Appendix E. Rinker (2018) Sentiment Algorithm

Each sentence s is broken into an ordered words

$$s = \{w_1, w_2, \ldots, w_k, \ldots, w_n\} \tag{E1}$$

where w_k are the words within sentences. Punctuation is removed with the exception of pause punctuation (commas, colons, semicolons) which are considered a word within the sentence. Denote pause words as cw.

First, the words in each sentence w_k are compared to a dictionary of polarized words[75] and weighted based on the sentiment dictionary. Denote polarized words as pw_k.

Second, each polarized word forms a polarized context cluster c_l which is a subset of the a sentence $c_l \subseteq s$. The polarized context cluster c_l of words is pulled from around the polarized word pw_k and defaults to two words before and five words after pw_k. The cluster can be represented as

$$c_l = \{w_{l,k-2}, \ldots, pw_{l,k}, \ldots, w_{l,k+5}\} \tag{E2}$$

The words in this polarized context cluster l are tagged as neutral $w^0_{l,k}$ or as valence-shifters, such as negators $w^n_{l,k}$, amplifiers (intensifiers) $w^a_{l,k}$, or de-amplifiers (downtoners) $w^d_{l,k}$. Neutral words hold no value in the equation but affect word count n.

The polarized word $pw_{l,k}$ is then weighted based on words sentiment dictionary and then further weighted by the number of the valence shifters surrounding the positive or negative word $pw_{l,k}$ in cluster l.

Amplifiers (intensifiers) increase the polarity of the polarized word $pw_{l,k}$. Amplifiers $w^a_{l,k}$ become de-amplifiers if the context cluster contains an odd number of negators $w^n_{l,k}$. De-amplifiers (downtoners) work to decrease the polarity (deamplifier weight is constrained to

[75] I used Loughran & McDonald's (2016) positive/negative financial word list as sentiment lookup values that assigns +1 to positive polarity word and −1 to negative polarity word.

-1 lower bound). Negation $w^n_{l,k}$ acts on amplifiers/de-amplifiers as discussed but also flip the sign of the polarized word. Negation is determined by raising -1 to the power of the number of negators $w^n_{l,k} + 2$. Simply, this is a result of a belief that two negatives equal a positive, 3 negatives a negative and so on.

The adversative conjunctions (i.e., 'but', 'however', and 'although') also weight the context cluster. Denote a number of adversative conjunctions within the cluster before the polarized word as $n^b_{l,ad}$ and after polarized word as $n^a_{l,ad}$. Adversative conjunctions before polarized word up-weight the cluster by

$$1 + 0.85 * n^b_{l,ad} \tag{E3}$$

while adversative conjunctions after the polarized word down-weight the cluster by

$$1 - 0.85 * n^a_{l,ad} \tag{E4}$$

This corresponds to the belief that an adversative conjunction makes the next clause of greater values while lowering the value placed on the prior clause.

Last, these weighted context clusters c_l are summed c and divided by the square root of the word count $\sqrt{n}$ yielding an unbounded polarity score δ for each sentence.

$$\delta = \frac{c}{\sqrt{n}} \tag{E5}$$

where

$$c = \sum_{l}((1 + w_{l,amp} + w_{l,deamp}) * pw_{l,k}(-1)^{2+w_{l,neg}}) \tag{E6}$$

$$w_{l,amp} = w_{l,b}\mathbb{1}_{w_{l,b}>1} + \sum(w_{l,neg} * (0.85 * w^{a}_{l,k}) \tag{E7}$$

$$w_{l,b} = 1 + 0.85(n^{b}_{l,ad} - n^{a}_{l,ad}) \tag{E8}$$

$$w_{l,neg} = \left(\sum w^{n}_{l,k}\right)(mod2) \tag{E9}$$

$$w_{l,deamp} = \max\{w_{l,deamp'}, 1\} \tag{E10}$$

$$w_{l,deamp'} = w_{b}\mathbb{1}_{w_{l,b}<1} + \sum(0.85 * (-w_{l,neg} * w^{a}_{l,k} + w^{d}_{l,k})) \tag{E11}$$

Pause $cw_{l,k}$ locations are indexed and considered in calculating the upper and lower bounds in the polarized context cluster, as these marks indicate a change in thought and words prior are not necessarily connected with words after these punctuation marks.

The lower bound of the polarized context cluster is constrained to $\max\{pw_{l,k-4}, 1, \max\{cw_{l,k} < pw_{l,k}\}\}$ and the upper bound is constrained to $\min\{pw_{l,k+4}, n, \min\{cw_{l,k} > pw_{l,k}\}\}$ where n is the number of words in the sentence.

Figure 22 shows two measures of dispersion, the range and the standard deviation of sentiment scores per information providers' quantile. The plot on the left shows minimum and maximum cross-news sentiment scores per quantile. The range, as a difference between maximum and minimum cross-news sentiment score is monotonically increasing. The right graph shows the standard deviation of cross-news sentiment scores. It increases monotonically from 12 to 20.

To show that the increase in dispersion is far from being mechanical consequence of increase in the sample size, I include a realization of Monte Carlo simulation of a range and a standard deviation of sentiment scores generated by a similar information structure with $\{132, 103, 111, 112, 113\}$ information providers that draw correspondingly one, two, ten, seventy-five and 2,039 random normal numbers from $N(1.99, 20.10)$ with a mean and stan-

dard deviation of the largest, fifth quantile 240 times (number of months in the sample). The red lines show the results of a simulation. Although the simulated range widens slightly for the fifth quantile, the magnitude of the change is significantly smaller than the change in the range in the data. The standard deviation of the simulated sentiment scores remains roughly the same, in contrast with the 1.4 times increase in cross-news standard deviation in the data.

Appendix F. Johansen (1988, 1991) Maximum Likelihood Estimation

An $(n \times 1)$ vector $\mathbf{y}$, was said to exhibit h cointegrating relations if there exist h linearly independent vectors $\mathbf{a_1}$,$\mathbf{a_2}$, ..., $\mathbf{a_h}$ such that $\mathbf{a_i^{'}y_t}$, is stationary. To uniquely identify the vectors, the normalization condition such as $a_{11} = 1$ is imposed. For this normalization we would put y_{1t}, on the left side of a regression and the other elements of y, on the right side.

Let $\mathbf{y}$ denote An $(n \times 1)$ vector. The maintained hypothesis is that $\mathbf{y}$ follows a $VAR(p)$ in levels. As any pth-order VAR can be written i the form

$$\mathbf{\Delta y_t} = \xi_\mathbf{1}\mathbf{\Delta y_{t-1}} + \xi_\mathbf{2}\mathbf{\Delta y_{t-3}} + \ldots + \xi_\mathbf{p-1}\mathbf{\Delta y_{t-p+1}} + \alpha + \xi_\mathbf{0}\mathbf{y_{t-1}} + \varepsilon_\mathbf{t} \tag{F1}$$

with

$$E(\varepsilon_\mathbf{t}) = 0 \tag{F2}$$

$$E(\varepsilon_\mathbf{t}\varepsilon_\tau^{'}) = \begin{cases} \mathbf{\Omega}, & \text{for} \quad t = \tau \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{F3}$$

Johansen (1991) describes his procedure using $\xi_\mathbf{0}\mathbf{y_{t-p}}$ instead of $\xi_\mathbf{0}\mathbf{y_{t-1}}$. Since $\mathbf{y_{t-p}} = \mathbf{y_{t-1}} - \mathbf{\Delta y_{t-1}} - \mathbf{\Delta y_{t-2}}, \ldots - \mathbf{\Delta y_{t-p+1}}$, the residuals are numerically identical to ones described

in the text (of Hamilton (1994)).

Suppose that each individual variable $\mathbf{y_{i,t}}$ is $I(1)$, although h linear combinations of $\mathbf{y_t}$ are stationary. This implies that $\xi_\mathbf{0}$ can be written in the form

$$\xi_\mathbf{0} = -\mathbf{B}\mathbf{A}^{'} \tag{F4}$$

for $\mathbf{B}$ an $(n \times k)$ matrix and $A^{'}$ an $(h \times n)$ matrix/ That is, under the hypothesis of h cointegrating relations, only h separate linear combinations of the level of $\mathbf{y_{t-1}}$ (the h elements of $\mathbf{z_{t-1}} = \mathbf{A}^{'}\mathbf{y_{t-1}}$ appears in the equation above.

Consider a sample of $T + p$ observations on $\mathbf{y}$ denoted $(\mathbf{y_{-p+1}}, \mathbf{y_{-p+2}}, \ldots, \mathbf{y_T})$. If the disturbances $\varepsilon_\mathbf{t}$ are Gaussian, then the log likelihood of $(\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_T})$ conditional on $(\mathbf{y_{-p+1}}, \mathbf{y_{-p+2}}, \ldots, \mathbf{y_T})$ is given by

$$L(\mathbf{\Omega}, \xi_\mathbf{1}, \xi_\mathbf{2}, \ldots, \xi_\mathbf{p-1}, \alpha, \xi_\mathbf{0}) = (-Tn/2)\log(2\pi) - (T/2)\log|\mathbf{\Omega}| \tag{F5}$$

$$-\frac{1}{2}\sum_{t=1}^{T}\Big[(\mathbf{\Delta y_t} - \xi_\mathbf{1}\mathbf{\Delta y_{t-1}} \ldots - \xi_\mathbf{p-1}\mathbf{\Delta y_{t-p+1}} - \alpha - \xi_\mathbf{0}\mathbf{y_{t-1}})^{'} \tag{F6}$$

$$\times\mathbf{\Omega}^{-1}(\mathbf{\Delta y_t} - \xi_\mathbf{1}\mathbf{\Delta y_{t-1}} \ldots - \xi_\mathbf{p-1}\mathbf{\Delta y_{t-p+1}} - \alpha - \xi_\mathbf{0}\mathbf{y_{t-1}})\Big] \tag{F7}$$

The goal is to chose $\mathbf{\Omega}, \xi_\mathbf{1}, \xi_\mathbf{2}, \ldots, \xi_\mathbf{p-1}, \alpha, \xi_\mathbf{0}$ to maximize the likelihood function above subject to the constraint that $\xi_\mathbf{0}$ can be written in the form $\xi_\mathbf{0} = -\mathbf{B}\mathbf{A}^{'}$.

The next sections summarize Johansen's algorithm.

Appendix .1. Step 1. Calculate Auxiliary Regressions

The first step is to estimate a $(p-1)$th-order VAR for $\mathbf{\Delta y_t}$; that is, regress the scalar $\Delta y_{i,t}$ on a constant and all the elements of the vectors $\mathbf{y_1}$, $\mathbf{y_2}$, ... , $\mathbf{y_{t-p+1}}$ by OLS. Collect the $i = 1, 2, \ldots, n$ OLS regressions in vector form as

$$\mathbf{\Delta y_t} = \hat{\pi_\mathbf{0}} + \mathbf{\Pi_1}\hat{\Delta} y_{t-1} + \mathbf{\Pi_2}\hat{\Delta} y_{t-2} + \ldots + \mathbf{\Pi_{p-1}}\hat{\Delta} y_{t-p+1} + \hat{\mathbf{u_t}} \tag{F8}$$

where $\mathbf{\Pi_i}$ denotes an $(n \times n)$ matrix of OLS coefficients estimates and $\mathbf{\hat{u}_t}$ denotes $(n \times 1)$ vector of OLS residuals.

We also estimate a second battery of regressions, regressing the scalar $\Delta y_{i,t-1}$ on a constant and $\mathbf{y_{t-1}}$, $\mathbf{y_{t-2}}$, ... , $\mathbf{y_{t-p+1}}$ for $i = 1, 2, \ldots, n$. Write this second set of OLS regressions as

$$\mathbf{\Delta y_{t-1}} = \hat{\theta} + \mathbf{\Psi_1}\hat{\Delta} y_{t-1} + \mathbf{\Psi_2}\hat{\Delta} y_{t-2} + \ldots + \mathbf{\Psi_{p-1}}\hat{\Delta} y_{t-p+1} + \mathbf{\hat{v}_t} \tag{F9}$$

with $\mathbf{\hat{v}_t}$ the $(n \times 1)$ vector of residuals from this second battery of regressions.

Johansen (1991) paper's procedure calculates $\mathbf{v_t}$ instead of $\mathbf{\hat{v}_t}$, where $\mathbf{v_t}$ is OLS residual from regression of $\mathbf{y_{t-p}}$ on a constant and $\mathbf{y_{t-1}}$, $\mathbf{y_{t-2}}$, ... , $\mathbf{y_{t-p+1}}$.

Appendix .2. Step 2. Calculate Canonical Correlations

Next calculate the sample variance-covariance matrices of the OLS residuals $\mathbf{\hat{u}_t}$ and $\mathbf{\hat{v}_t}$:

$$\hat{\Sigma_{VV}} \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbf{\hat{v}_t}\hat{\mathbf{v_t}'} \tag{F10}$$

$$\hat{\Sigma_{UU}} \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbf{\hat{u}_t}\hat{\mathbf{u_t}'} \tag{F11}$$

$$\hat{\Sigma_{UV}} \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbf{\hat{u}_t}\hat{\mathbf{v_t}'} \tag{F12}$$

$$\hat{\Sigma_{VU}} \equiv \hat{\Sigma_{UV}'} \tag{F13}$$

From these, find the eigenvalues of the matrix

$$\hat{\Sigma_{VV}}^{-1}\hat{\Sigma_{VU}}\hat{\Sigma_{UU}}^{-1}\hat{\Sigma_{UV}} \tag{F14}$$

with eigenvalues ordered $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_n$. The maximum value attained by the log likelihood function subject to the constraint that there are h cointegrating relations is given

by

$$L^* = -\frac{Tn}{2}\log(2\pi) - \frac{Tn}{2} - \frac{T}{2}\log|\hat{\Sigma_{UU}}| - \frac{Tn}{2}\sum_{i=1}^{h}\log(1-\hat{\lambda}_i) \tag{F15}$$

Appendix .3. Step 3. Calculate Maximum Likelihood Estimates of Parameters

If we are interested only in a likelihood ratio test of the number of cointegrating relations, step 2 provides all the information needed. If maximum likelihood estimates of parameters are also desired, these can be calculated as follows.

Let $\hat{\mathbf{a_1}}$, $\hat{\mathbf{a_2}}$, ..., $\hat{\mathbf{a_h}}$ denote the $(n \times 1)$ eigenvectors of $\hat{\Sigma_{VV}}^{-1}\hat{\Sigma_{VU}}\hat{\Sigma_{UU}}^{-1}\hat{\Sigma_{UV}}$ associated with the h largest eigenvalues. These provide a basis for the space of cointegrating relations; that is, the maximum likelihood estimate is that any cointegrating vector can be written in the form

$$\mathbf{a_1} = b_1\hat{\mathbf{a_1}} + b_2\hat{\mathbf{a_2}} + \ldots + b_h\hat{\mathbf{a_h}} \tag{F16}$$

for some choice of scalars $(b_1, b_2, \ldots, b_n)$. Johansen suggested normalizing these vectors $\hat{\mathbf{a_i}}$ so that $\hat{\mathbf{a_i}}'\hat{\Sigma_{VV}}\hat{\mathbf{a_i}} = 1$. For example, if the eigenvectors $\hat{\mathbf{a_i}}$ of $\hat{\Sigma_{VV}}^{-1}\hat{\Sigma_{VU}}\hat{\Sigma_{UU}}^{-1}\hat{\Sigma_{UV}}$ are calculated from a standard computer program that normalizes $\tilde{\mathbf{a_i}}'\tilde{\mathbf{a_i}} = 1$, Johansen's estimate is $\hat{\mathbf{a_i}} = \tilde{\mathbf{a_i}} + \sqrt{\tilde{\mathbf{a_i}}'\hat{\Sigma_{VV}}\tilde{\mathbf{a_i}}}$. Collect the first h normalized vectors in an $(n \times h)$ matrix $\hat{\mathbf{A}}$:

$$\hat{\mathbf{A}} = [\hat{\mathbf{a_1}}, \hat{\mathbf{a_2}}, \ldots, \hat{\mathbf{a_h}}] \tag{F17}$$

Then the MLE of $\hat{\xi}_0$ is given by

$$\hat{\xi_0} = \hat{\Sigma_{UV}}\hat{\mathbf{A}}\hat{\mathbf{A}}' \tag{F18}$$

The MLE of $\hat{\xi}_{\mathbf{i}}$ for $i = 1, 2, \ldots, p-1$ is

$$\hat{\xi}_{\mathbf{i}} = \hat{\mathbf{\Pi}}_{\mathbf{i}} - \hat{\xi}_{\mathbf{0}} \hat{\mathbf{\Psi}}_{\mathbf{p-1}} \tag{F19}$$

and the MLE of α is

$$\hat{\alpha} = \hat{\pi_{\mathbf{0}}} - \hat{\xi}_{\mathbf{0}} \hat{\theta} \tag{F20}$$

The MLE of $\mathbf{\Omega}$ is

$$\hat{\mathbf{\Omega}} = \frac{1}{T} \sum_{t=1}^{T} [(\hat{\mathbf{u_t}} - \hat{\xi}_{\mathbf{0}} \hat{\mathbf{v_t}})(\hat{\mathbf{u_t}} - \hat{\xi}_{\mathbf{0}} \hat{\mathbf{v_t}})'] \tag{F21}$$

Appendix G. Loading Matrices

Table XXI. Matrix α
Matrix α of loading weights of potential cointegration vectors in the test procedure.

	$\tilde{r}_t^C$	$\tilde{r}_t^M$
s_t^w	$\begin{bmatrix} -0.369 & 0.027 \\ 0.096 & -0.044 \end{bmatrix}$	$\begin{bmatrix} -0.365 & 0.013 \\ 0.126 & -0.014 \end{bmatrix}$
s_t	$\begin{bmatrix} -0.687 & 0.001 \\ -0.008 & -0.007 \end{bmatrix}$	$\begin{bmatrix} -0.707 & 0.001 \\ 0.007 & 0.002 \end{bmatrix}$
s_t^m	$\begin{bmatrix} -0.652 & 0.001 \\ -0.011 & -0.006 \end{bmatrix}$	$\begin{bmatrix} -0.623 & 0.001 \\ 0.014 & 0.002 \end{bmatrix}$

Appendix H. VECM. Hypothesis 1

Table XXII. Hypothesis 1. VECM With Subjective Expectaions From MSC Survey

	Dependent variable:					
	VECM With Provider-weighted Sentiment		VECM With Equally-weighted Sentiment		VECM With Company-weighted Sentiment	
	$\Delta\tilde{r}_t^M$	Δs_t^w	$\Delta\tilde{r}_t^M$	Δs_t	$\Delta\tilde{r}_t^M$	Δs_t^m
	(1)	(2)	(3)	(4)	(5)	(6)
v_{t-1}	0.126^{**}	-0.365^{***}	0.007	-0.707^{***}	0.014	-0.623^{***}
	(0.050)	(0.074)	(0.040)	(0.092)	(0.034)	(0.094)
$\Delta\tilde{r}_{t-1}^M$	-0.317^{***}	0.174^{*}	-0.277^{***}	0.349^{**}	-0.277^{***}	0.332^{*}
	(0.070)	(0.105)	(0.070)	(0.160)	(0.069)	(0.191)
Δs_{t-1}^w	0.066	-0.684^{***}	0.020	-0.830^{***}	0.007	-0.920^{***}
	(0.047)	(0.070)	(0.031)	(0.072)	(0.026)	(0.071)
Constant	0.234^{**}	-0.585^{***}	0.028	-0.033	0.031	-0.145
	(0.093)	(0.139)	(0.045)	(0.104)	(0.046)	(0.127)
Observations	196	196	196	196	196	196
R^2	0.106	0.333	0.078	0.412	0.077	0.480
Adj. R^2	0.092	0.323	0.064	0.403	0.062	0.472

Table XXIII. Hypothesis 1. VECM With Subjective Expectaions From CSS Survey

	Dependent variable:					
	VECM With Provider-weighted Sentiment		VECM With Equally-weighted Sentiment		VECM With Company-weighted Sentiment	
	$\Delta\tilde{r}_t^C$	Δs_t^w	$\Delta\tilde{r}_t^C$	Δs_t	$\Delta\tilde{r}_t^C$	Δs_t^m
	(1)	(2)	(3)	(4)	(5)	(6)
v_{t-1}	0.096^{*}	-0.369^{***}	-0.008	-0.687^{***}	-0.011	-0.652^{***}
	(0.056)	(0.070)	(0.045)	(0.089)	(0.040)	(0.094)
$\Delta\tilde{r}_{t-1}^C$	-0.228^{***}	0.155^{*}	-0.194^{***}	0.224	-0.192^{***}	0.294^{*}
	(0.073)	(0.092)	(0.070)	(0.138)	(0.070)	(0.164)
Δs_{t-1}^w	0.062	-0.695^{***}	0.001	-0.823^{***}	-0.028	-0.941^{***}
	(0.055)	(0.070)	(0.037)	(0.072)	(0.030)	(0.071)
Constant	0.327^{*}	-1.180^{***}	0.011	-0.376^{***}	0.002	-0.863^{***}
	(0.188)	(0.236)	(0.059)	(0.115)	(0.074)	(0.173)
Observations	196	196	196	196	196	196
R^2	0.053	0.343	0.039	0.413	0.044	0.491
Adj. R^2	0.039	0.333	0.024	0.404	0.029	0.483

Appendix I. VECM. Hypothesis 1.1

Table XXIV. Hypothesis 1.1. VECM With Subjective Expectaions From MSC Survey

	Dependent variable:			
	VECM With Sentiment of More Active Providers		VECM With Sentiment of Less Active Providers	
	$\Delta\tilde{r}_t^M$	Δs_t^w	$\Delta\tilde{r}_t^M$	Δs_t^w
	(1)	(2)	(3)	(4)
v_{t-1}	0.131^{***}	-0.358^{***}	-0.020	-0.668^{***}
	(0.048)	(0.073)	(0.022)	(0.095)
$\Delta\tilde{r}_{t-1}^M$	-0.322^{***}	0.157	-0.276^{***}	0.354
	(0.070)	(0.107)	(0.069)	(0.305)
Δs_{t-1}^w	0.066	-0.662^{***}	-0.004	-0.876^{***}
	(0.046)	(0.070)	(0.017)	(0.073)
Constant	0.263^{***}	-0.622^{***}	0.018	-0.405^{**}
	(0.096)	(0.147)	(0.047)	(0.205)
Observations	196	196	196	196
R^2	0.111	0.320	0.081	0.430
Adj. R^2	0.097	0.310	0.067	0.421

Table XXV. Hypothesis 1.1. VECM With Subjective Expectaions From CSS Survey

	Dependent variable:			
	VECM With Sentiment of More Active Providers		VECM With Sentiment of Less Active Providers	
	$\Delta\tilde{r}_t^C$	Δs_t^w	$\Delta\tilde{r}_t^C$	Δs_t^w
	(1)	(2)	(3)	(4)
v_{t-1}	0.106^{**}	-0.357^{***}	-0.028	-0.678^{***}
	(0.054)	(0.069)	(0.025)	(0.094)
$\Delta\tilde{r}_{t-1}^C$	-0.236^{***}	0.137	-0.195^{***}	0.140
	(0.073)	(0.094)	(0.070)	(0.264)
Δs_{t-1}^w	0.068	-0.668^{***}	-0.022	-0.880^{***}
	(0.054)	(0.069)	(0.019)	(0.073)
Constant	0.384^{**}	-1.220^{***}	-0.022	-0.972^{***}
	(0.194)	(0.250)	(0.063)	(0.236)
Observations	196	196	196	196
R^2	0.058	0.328	0.046	0.433
Adj. R^2	0.043	0.317	0.031	0.424

Appendix J. VECM. Hypothesis 1.1. With r_{t-1}

Table XXVI. Hypothesis 1.1. VECM With Subjective Expectaions From MSC Survey and Past Market Return

	Dependent variable:					
	VECM With Sentiment of More Active Providers			VECM With Sentiment of Less Active Providers		
	$\Delta\tilde{r}_t^C$	Δs_t^w	Δr_{t-1}	$\Delta\tilde{r}_t^C$	Δs_t^w	Δr_{t-1}
	(1)	(2)	(3)	(4)	(5)	(6)
v_{t-1}	-0.030^{*}	0.074^{***}	-1.113^{***}	-0.009	0.196^{***}	-0.992^{***}
	(0.016)	(0.026)	(0.097)	(0.016)	(0.075)	(0.094)
$\Delta\tilde{r}_{t-1}^M$	-0.292^{***}	-0.018	1.359^{***}	-0.298^{***}	0.261	1.674^{***}
	(0.071)	(0.113)	(0.421)	(0.073)	(0.353)	(0.442)
Δs_{t-1}^w	0.024	-0.548^{***}	1.172^{***}	0.008	-0.566^{***}	0.190^{**}
	(0.042)	(0.066)	(0.245)	(0.013)	(0.062)	(0.077)
Δr_{t-2}	-0.003	0.049^{**}	-1.018^{***}	0.008	0.049	-0.960^{***}
	(0.012)	(0.019)	(0.072)	(0.012)	(0.058)	(0.073)
Constant	0.044	-0.019	0.574^{**}	0.022	0.093	-0.735^{**}
	(0.046)	(0.072)	(0.269)	(0.047)	(0.227)	(0.284)
Observations	196	196	196	196	196	196
R^2	0.101	0.269	0.520	0.088	0.311	0.483
Adj. R^2	0.082	0.254	0.510	0.068	0.297	0.472

Table XXVII. Hypothesis 1.1. VECM With Subjective Expectaions From CSS Survey and Past Market Return

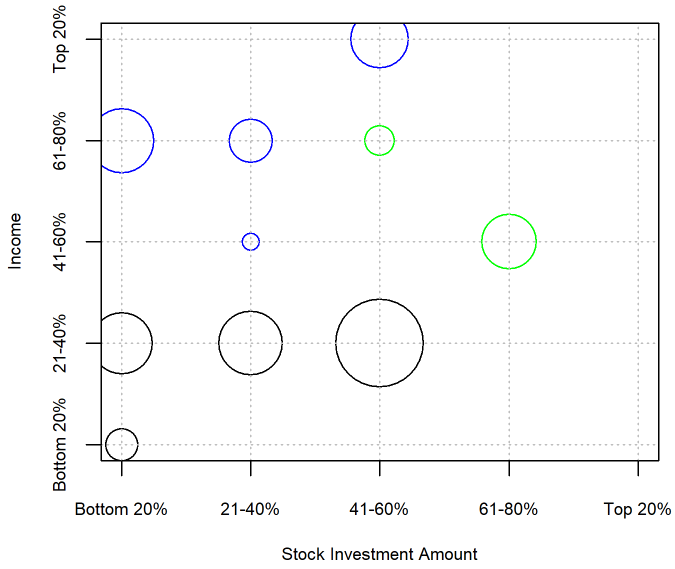
	Dependent variable:					
	VECM With Sentiment of More Active Providers			VECM With Sentiment of Less Active Providers		
	$\Delta\tilde{r}_t^C$	Δs_t^w	Δr_{t-1}	$\Delta\tilde{r}_t^C$	Δs_t^w	Δr_{t-1}
	(1)	(2)	(3)	(4)	(5)	(6)
v_{t-1}	-0.058^{***}	0.074^{***}	-1.053^{***}	-0.039^{**}	0.234^{***}	-0.974^{***}
	(0.019)	(0.025)	(0.098)	(0.018)	(0.076)	(0.098)
$\Delta\tilde{r}_{t-1}^C$	-0.217^{***}	-0.043	0.736^{*}	-0.201^{***}	0.185	1.145^{***}
	(0.074)	(0.102)	(0.395)	(0.076)	(0.316)	(0.408)
Δs_{t-1}^w	0.061	-0.545^{***}	1.065^{***}	0.0001	-0.582^{***}	0.252^{***}
	(0.048)	(0.065)	(0.254)	(0.015)	(0.062)	(0.080)
Δr_{t-2}	-0.017	0.050^{**}	-0.981^{***}	-0.007	0.065	-0.949^{***}
	(0.014)	(0.020)	(0.076)	(0.014)	(0.060)	(0.077)
Constant	0.038	-0.011	0.444	-0.060	0.397	-1.829^{***}
	(0.052)	(0.071)	(0.277)	(0.063)	(0.262)	(0.338)
Observations	196	196	196	196	196	196
R^2	0.094	0.271	0.485	0.070	0.321	0.457
Adj. R^2	0.075	0.255	0.474	0.051	0.306	0.446

Appendix K. VAR with Change in Sentiment

Figure 23. "Best" VAR Specifications (Color) and Adjuster R^2 of VAR's Subjective Expectations Equations (Size of Circle)

The bubble plot shows adjusted R^2 of the "best" VAR with subjective expectations $\Delta\tilde{r}_t$ of investors with different level of investment in stocks and income from MSC survey. X-axis shows Stock Investment Amount quintile. Y-axis shows Income quintiles. The size of a circle is proportional to adjusted R^2 of VECM's subjective expectations equation. The biggest circle is equivalent adjusted $R^2 = 39\%$, the smallest to the $R^2 = 8\%$.

Gray color corresponds to VAR$(\Delta^2 jc_t, \Delta^2 s_t, \Delta\tilde{r}_t)$ with the following ordering of variables $\Delta^2 jc \to \Delta^2 s \to \Delta\tilde{r}$, where $\Delta^2 jc$ is second difference of polarity of the Mad Money episodes and $\Delta^2 s$ is a second difference of sentiment of equity reports. Blue color corresponds to VAR$(\Delta r_{t-1}, \Delta^2 s_t, \Delta\tilde{r}_t)$ or VAR$(\Delta r_{t-1}, \Delta q_t, \Delta\tilde{r}_t)$ with ordering $\Delta r \to \Delta^2 q \to \Delta\tilde{r}$, where $\Delta^2 q$ is polarity of the Squawk on the Street episodes. Green color corresponds to VAR$(\Delta^2 s_t, \Delta\tilde{r}_t)$ with ordering $\Delta^2 s \to \Delta\tilde{r}$ The calculations are based on monthly data from January 2012 to December 2018.



Appendix L. "Best" VECM or VAR Models With Subjective Expectations of Investors Within Income and Stock Invested Amount Quintiles

Table XXVIII. "Best" VECM or VAR Models With Subjective Expectations of Investors Within Income and Stock Invested Amount Quintiles
First three columns of table show VECM(v_1, v_2, v_3) variables v_1, v_2 and v_3. The first "Wald Statistics" column shows Wald test statistics with null hypothesis that lags of v_1 and $\alpha\beta^{'}$ are zero in v_2 equation of VECM. The second "Wald Statistics" column shows Wald test statistics with null hypothesis that lags of v_2 and $\alpha\beta^{'}$ are zero in v_3 equation of VECM. The first "Measures of Fit" column shows AIC, while the second - adjusted R^2 of subjective expectations equation. Monthly data is from January 2012 to December 2018.

VECM or VAR			Wald Statistics		Measure of Fit	
v_1	v_2	v_3	H_0 : $v_1 \nrightarrow v_2$	H_0 : $v_2 \nrightarrow v_3$	AIC	R_a^2
jc	s^w	$\tilde{r}(I2, S1)$	0.01	0.04	-15.91	0.27
r	q	$\tilde{r}(I2, S2)$	0.07	0.05	-13.80	0.43
jc	s^w	$\tilde{r}(I2, S3)$	0.01	0.13	-15.58	0.39
	$s_t^{w,5}$	$\tilde{r}(I3, S1)$		0.06	-9.88	0.12
	$s_t^{w,1:4}$	$\tilde{r}(I3, S3)$		0.10	-7.87	0.30
jc	s^w	$\tilde{r}(I3, S4)$	0.02	0.12	-14.83	0.16
	$s_t^{w,1:4}$	$\tilde{r}(I3, S5)$		0.07	-8.42	0.15
jc	s^c	$\tilde{r}(I4, S2)$	0.00	0.11	-14.66	0.20
	$s_t^{w,5}$	$\tilde{r}(I4, S3)$		0.06	-10.04	0.20
	$s_t^{w,1:4}$	$\tilde{r}(I4, S4)$		0.10	-7.84	0.19
	s^w	$\tilde{r}(I4, S5)$		0.04	-10.08	0.18
r	q	$\tilde{r}(I5, S2)$	0.13	0.06	-12.77	0.16
jc	s^c	$\tilde{r}(I5, S3)$	0.00	0.01	-13.86	0.32
	s^w	$\tilde{r}(I5, S4)$		0.13	-9.82	0.11
	s^w	$\tilde{r}(I5, S5)$		0.07	-9.56	0.16

Appendix M. Robustness Check

Figure 24. AIC of VECM (s_t, $\tilde{r}_t$) Without One Year

Plot shows AIC of VECM regression explaining subjective equity premium without year on X-axis. Monthly data is from July 2002 to December 2019. The higher dashed line is the average AIC of the VECM with MSC expectations. The lower dashed line is AIC of VECM with the CCS expectations regression. Blue dots corresponds to AIC of VECM with subjective expectations from the CCS survey; white dots -AIC of VECM with subjective expectations from the MSC survey.

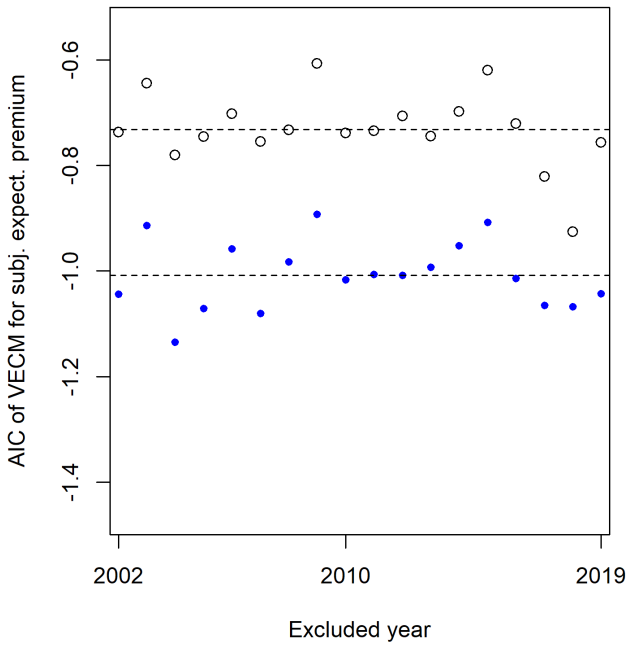
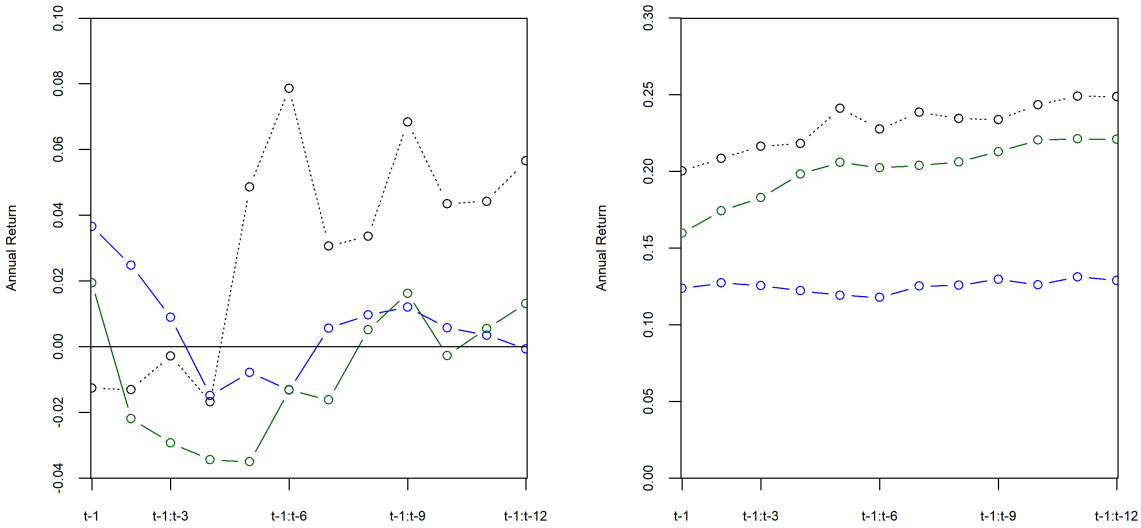


Figure 25. Annualized Mean Return And Standard Deviation of Return Of Standard Momentum Strategy (Dashed Line) And Sentiment-Based Momentum Strategy Based on Information From Less Active (Blue Line) and More Active (Green Line) Information Providers Vs. Strategy Formation Period From $t-1$ to $t-1 \rightarrow t-12$

The calculations are based on the sample that consists of 45 US blue-chip companies and spans from July 2002 to December 2018.



(a) Annualized Mean Return Vs. Strategy Formation Period

(b) Annualized St.Dev. Of Return Vs. Strategy Formation Period

References

Acemoglu, D., & Scott, A. (1994). Consumer confidence and rational expectations: Are agents' beliefs consistent with the theory? *The Economic Journal*, *104*(422), 1–19.

Ang, A., & Bekaert, G. (2001). Stock return predictability: Is it there? *Review of Financial Studies*, *20*(3), 651–707.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*(2), 129–152.

Baker, S. R., & Fradkin, A. (2017). The impact of unemployment insurance on job search: Evidence from google search data. *The Review of Economics and Statistics*, *99*(5), 756–768.

Bansal, R., Dittmar, R., & Kiku, D. (2007). *Cointegration and consumption risks in asset returns* (Working Paper). National Bureau of Economic Research.

Bansal, R., Dittmar, R. F., & Lundblad, C. (2005). Consumption, dividends, and the cross-section of equity returns. *Journal of Finance*, *60*, 1639–1672.

Bansal, R., & Yaron, A. (2004). Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles. *Journal of Finance*, *59*(4), 1481–1509. https://doi.org/10.1111/J.1540-6261.2004.00670.X

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, *116*(1), 261–292.

Barsky, R. B., & Sims, E. R. (2012). Information, animal spirits, and the meaning of innovations in consumer confidence. *American Economic Review*, *102*(4), 1343–77.

Bossaerts, P., & Hillion, P. (1999). Implementing statistical criteria to select return forecasting models: What do we learn? *Review of Financial Studies*, *12*(2), 405–428.

Box, G. E., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. Holden-Day.

Bram, J., & Ludvigson, S. (1998). Does consumer confidence forecast household expenditure? a sentiment index horse race. *Federal Reserve Bank of New York Economic Policy Review*, *4*(2), 59–78.

Brunnermeier, M. K., Simsek, A., & Xiong, W. (2014). A welfare criterion for models with distorted beliefs. *The Quarterly Journal of Economics*, *129*(4), 1753–1797.

Campbell, J. Y., & Shiller, R. J. (1988a). The dividend-price ratio and expectations of future dividends and discount factors. *The Review of Financial Studies*, *1*, 195–228.

Campbell, J. Y., & Shiller, R. J. (1988b). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies*, *1*(3), 195–228.

Campbell, J. Y., & Shiller, R. J. (1988c). Stock prices, earnings, and expected dividends. *The Journal of Finance*, *43*, 661–676.

Carroll, C. D., Fuhrer, J. C., & Wilcox, D. W. (1994). Does consumer sentiment forecast household spending? if so, why? *American Economic Review*, *84*(5), 1397–1408.

Cochrane, J. H. (1991). Production-based asset pricing and the link between stock returns and economic fluctuations. *Journal of Finance*, *46*(1), 209–237.

Cochrane, J. H. (2006). The dog that did not bark: A defense of return predictability. *Review of Financial Studies*, *21*(4), 1533–1575.

D'Acunto, F., Weber, M., & Jin, C. (2019). The effect of large investors on asset quality: Evidence from subprime mortgage securities. *The Journal of Finance*.

Dees, S., & Brinca, P. S. (2011). Consumer confidence as a predictor of consumption spending: Evidence for the united states and the euro area. *International Economics*, *128*(1), 1–14.

Elliott, G., Graham, C., Rothenberg, T. J., & Stock, J. H. (1996). Efficient tests for an autoregressive unit root. *Econometrica*, *64*(4), 813–836.

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, *55*(2), 251–276.

Fama, E. F., & French, K. R. (1988). Dividend yields and expected stock returns. *Journal of Financial Economics*, *22*(1), 3–25.

Ferson, W. E., Sarkissian, S., & Simin, T. T. (2003). Spurious regressions in financial economics? *Journal of Finance*, *58*(4), 1393–1413.

Fishman, A., & Hagerty, K. M. (2019). The impact of news on stock prices.

Goetzmann, W. N., & Jorion, P. (1993). Testing the predictive power of dividend yields. *Journal of Finance*, *48*(2), 663–679.

Goyal, A., & Welch, I. (2003). Predicting the equity premium with dividend ratios. *Management Science*, *49*(5), 639–654.

Goyal, A., & Welch, I. (2004). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, *21*(4), 1455–1508.

Greenwood, R., & Shleifer, A. (2014). Expectations of returns and expected returns. *The Review of Financial Studies*, *27*, 714–746.

Hansen, L., Heaton, J., & Li, N. (2005). *Consumption strikes back?* [Working paper, University of Chicago].

Hansen, L. P., & Hodrick, R. J. (1980). Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. *Journal of Political Economy*, *96*, 829–853.

Hastings, J. S., & Tejeda-Ashton, L. (2008). Financial literacy, information, and demand elasticity: Survey and experimental evidence from mexico. *The Journal of Consumer Affairs*.

Hodrick, R. J. (1992). Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *Review of Financial Studies*, *5*(3), 357–386.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*, 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, *12*(2-3), 231–254.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: journal of the Econometric Society*, 1551–1580.

Lemmon, M., & Portniaguina, E. (2006). Consumer confidence and asset prices: Some empirical evidence. *The Review of Financial Studies*, *19*(4), 1499–1529.

Lettau, M., & Ludvigson, S. (2001). Resurrecting the (c)capm: A cross-sectional test when risk premia are time-varying. *Journal of Political Economy*, *109*(6), 1238–1287.

Lewellen, J. (2004). Predicting returns with financial ratios. *Journal of Financial Economics*, *74*(2), 209–235.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey [Available at SSRN: https://ssrn.com/abstract=2504147 or http://dx.doi.org/10.2139/ssrn.2504147].

Ludvigson, S. C. (2004). Consumer confidence and consumer spending. *Journal of Economic Perspectives*, *18*(2), 29–50.

Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, *52*(1), 5–44.

Lustig, H., & Van Nieuwerburgh, S. (2005). Housing collateral, consumption insurance, and risk premia: An empirical perspective. *Journal of Finance*, *60*(3), 1167–1219.

Malmendier, U., & Nagel, S. (2011). Title of the paper. *Journal Name*, Page numbers.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. https://doi.org/10.3115/v1/P14-5010

Matsusaka, J. G., & Sbordone, A. M. (1995). Consumer confidence and economic fluctuations. *Economic Inquiry*, *33*(2), 296–318.

Menzly, L., Santos, T., & Veronesi, P. (2004). Understanding predictability. *Journal of Political Economy*, *112*(1), 1–47.

Moskowitz, T. J., & Grinblatt, M. (1999). Do industries explain momentum? *The Journal of Finance*, *54*(4), 1249–1290.

Nagel, S., & Xu, Z. (2019). Asset pricing with fading memory. *The Review of Financial Studies*, *32*, 4273–4313.

Nelson, C. R., & Kim, M. J. (1993). Predictable stock returns: The role of small sample bias. *Journal of Finance*, *48*(2), 641–661.

Pankratz, A. (1983). *Forecasting with univariate box-jenkins models: Concepts and cases*. Wiley. https://doi.org/10.1002/9780470316566.index

Paye, B. S., & Timmermann, A. (2003). Instability of return prediction models. *Unpublished Manuscript*.

Rinker, T. W. (2021). sentimentr: Calculate text polarity sentiment [version 2.9.0]. https://github.com/trinker/sentimentr

Roll, R. (1988). R2. *Journal of Finance*, *43*(3), 541–566.

Schwert, G. W. (1988). *Tests for unit roots: A monte carlo investigation* (Working Paper No. 73). National Bureau of Economic Research. https://doi.org/10.3386/t0073

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, *48*(1), 1–48. https://ideas.repec.org/a/ecm/emetrp/v48y1980i1p1-48.html

Souleles, N. S. (2001). Consumer sentiment and household expenditure. *Journal of the American Statistical Association*, *96*(456), 1359–1367.

Souleles, N. S. (2004). Expectations, heterogeneous forecast errors, and consumption: Micro evidence from the michigan consumer sentiment surveys. *Journal of Money, Credit, and Banking*, *36*(1), 39–72.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, *54*(3), 375–421.

Stock, J. H., & Watson, M. W. (1993). A simple estimator of cointegrating vectors in higher order integrated system. *Econometrica*, *61*, 783–820. https://doi.org/10.2307/2951763

Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information?

Throop, A. W. (2010). *Consumer sentiment: Its rationality and usefulness in forecasting expenditure—evidence from the michigan micro data* (tech. rep.). Federal Reserve Bank of San Francisco.

Toda, H. Y., & Phillips, P. C. (1991). *Vector autoregression and causality: A theoretical overview and simulation study* (tech. rep. No. 1244). Cowles Foundation Discussion Papers. https://elischolar.library.yale.edu/cowles-discussion-paper-series/1244

Valkanov, R. (2003). Long-horizon regressions: Theoretical results and applications. *Journal of Financial Economics*, *68*(2), 201–232.

Veldkamp, L. L. (2005). Information markets and the comovement of asset prices. *The Review of Economic Studies*, *72*(3), 823–845.

Viceira, L. M. (1996). Testing for structural change in the predictability of asset returns. *Unpublished Manuscript*.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780.

Wold, H. (1954). *A study in the analysis of stationary time series* (2nd ed.) [With an Appendix on "Recent Developments in Time Series Analysis" by Peter Whittle]. Almqvist; Wiksell Book Co.

Zivot, E., & Andrews, D. W. (1992). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics*, *10*(3), 251–270.