Similarity in Financial Markets [*]

Constantin Charles[†]
Pengfei Sui[‡]

This version: January 20, 2026

Abstract

We propose a framework in which investors draw on similar past periods when forming their beliefs or deciding on their actions. Motivated by this framework, we construct a measure that assigns weights to past months based on their similarity to today. Using our measure, we construct similarity-based beliefs and show that they can explain return expectations from surveys as well as higher-moment beliefs implied by stock options and the VIX. We also show that similarity can explain the repurchasing decisions of individual investors. Our results show that similarity is a useful principle for understanding beliefs and actions in financial markets.

[*] We thank Peter Andre, Kai Barron, Francesca Bastianello, Stefano Cassella, Zhi Da, Joey Engelberg, Cary Frydman, Stefano Giglio, Katrin Goedker, Huseyin Gulen, Hongye Guo, Tony He, Lawrence Jin, Eben Lazarus, Tse-Chun Lin, Jun Liu, Ian Martin, Thomas Maurer, Daniel Metzger, Jun Pan, Cameron Peng, Tim Pleskac, Suproteem Sarkar, Rui Shen, Andrei Shleifer, Kelly Shue, Johannes Wohlfart, Zhengyang Xu, and Hongjun Yan, as well as participants at the Chicago Booth Conference in Behavioral Finance and Decision-Making, China International Conference in Finance, Early Career in Behavioral Economics Conference, London Behavioral Finance Group Meeting, NBER Behavioral Finance Fall Meeting, City University of Hong Kong, CUHK-Shenzhen, Fudan University, Helsinki Finance Seminar, London School of Economics, Nantes University & Audencia Business School, Shanghai Jiao Tong University, Stockholm School of Economics, Tsinghua University, University College London, University of Cologne, University of Hong Kong, University of Zurich, Warwick Business School, and WHU for helpful comments. We thank Terrance Odean for providing the trading data. This paper supersedes our previous work titled “Marketwide Memory.”

[†] London School of Economics. Contact: c.charles@lse.ac.uk

[‡] The Chinese University of Hong Kong, Shenzhen. Contact: psui@cuhk.edu.cn

1 Introduction

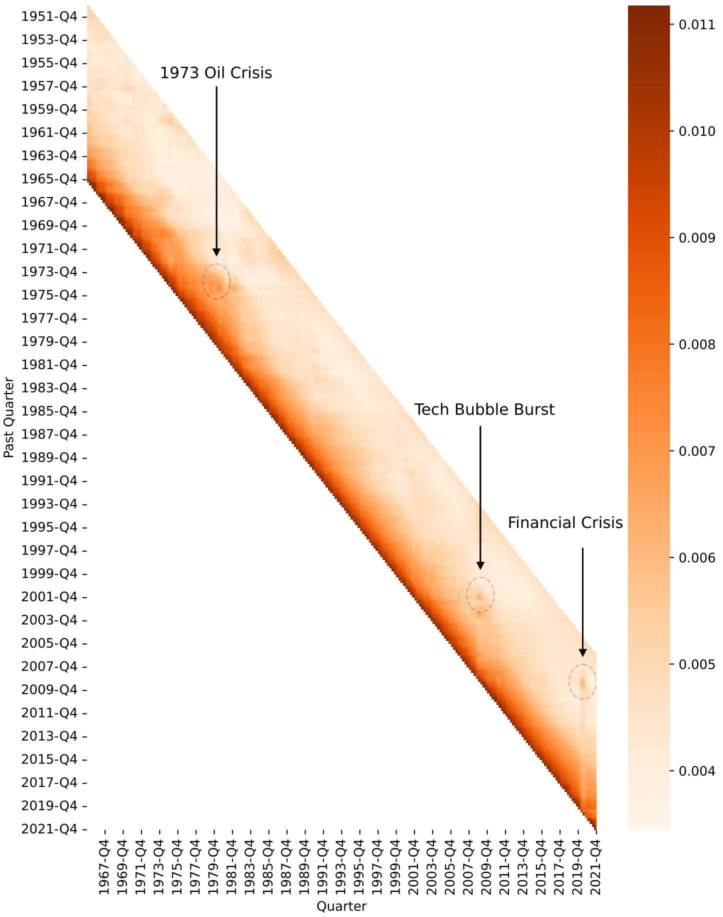


Figure 1. This figure displays the weight that our measure assigns to the past 60 quarters (= 15 years) for each quarter from 1966 to 2021. The x-axis indicates the quarter in which the hypothetical cue occurs, while the y-axis indicates the weight assigned to past quarters. A darker shade of red indicates a higher weight.

A large and growing literature in finance documents that investors' beliefs and actions are shaped by their past experiences. A key finding in this literature is that the perceived similarity between the past and the present is a crucial filter for how investors process information (Malmendier (2021a,b); Malmendier and Wachter (2024)). This aligns with long-standing insights from psychology, which recognizes similarity as a fundamental principle guiding human cognition. A growing body of work in economics and finance builds on this idea, and assumes that people draw on similar past events or experiences when forming their beliefs or deciding on an action. For instance, in case-based theory, people rely on analogies to past cases when making decisions (Gilboa and Schmeidler (1995)). In addition to such active reasoning by similarity, recent theories that incorporate aspects of human memory into economics and finance are built around the idea that people are more likely to recall similar past experiences (Mullainathan (2002); Bordalo et al. (2020); Bordalo et al. (2023); Wachter and Kahana (2024)).

Beyond such cognitive foundations, similarity is also a key ingredient in non-linear state-dependent forecasting, such as ‘nearest neighbor’ forecasting. The focus on similarity, in both rational and psychological approaches, is based on the idea that in economic environments with state-dependent dynamics, data from historically similar conditions are often more informative than an unconditional average of all past data.

In this paper, we propose a framework that uses similarity as its organizing principle. In this framework, when people try to form beliefs about the future or decide on an action, they draw on past periods when things were similar to what they are today, and use data from those prior periods to make their forecasts and to choose their actions. Motivated by our framework, we construct a similarity-based measure that specifies which historical periods investors draw on at a given point in time. We show that our measure (i) matches the historical periods that managers and analysts discuss during corporate events, (ii) is useful in explaining return expectations from investor surveys as well as higher moment beliefs implied by stock options and the VIX, and (iii) can explain repurchasing decisions of individual investors. Overall, our results suggest that similarity is a principle that can be broadly applied to understand beliefs and actions in financial markets.

To build intuition for our approach, consider an investor who evaluates a stock at the height of the 2020 Covid-19 pandemic. This evaluation of the stock, in the context of a global pandemic, serves as a cue that leads the investor to draw on similar past periods (e.g., the 2008 financial crisis). For each stock, we represent each period as a high-dimensional vector that comprises the stock’s features in that period, such as its price, return, trading volume, profitability, capital structure, etc., as well as broader contextual features of the economic environment, like GDP growth and the inflation rate.[1]

Having represented each period as a vector of features, we calculate the cosine similarity of the current period’s vector with historical vectors. We assume that, all other things equal, the investor is more likely to

[1] As an alternative approach, we use the narratives and topical themes extracted from business news coverage, sourced from Bybee et al. (2024), to represent each period. We find very similar results with this alternative approach.

draw on past periods for which the historical vector is more similar to the current vector. This captures the force of similarity, the organizing principle of our framework.

It is possible that many historical vectors are similar to the current period's vector. These historical periods all compete for retrieval, making it less likely that the investor draws on any particular historical period. Thus we assume that, all other things equal, the investor is less likely to draw on a particular historical period if many historical vectors are similar to the current period's vector. This offsetting force directly results from a normalization, which we implement to ensure that the probability distribution of weights on past periods integrates to 1. Viewed through the lens of associative memory theory, this normalization captures the force of interference (Kahana (2012)).

By combining these two offsetting forces, our measure yields the probability with which a representative investor draws on a particular historical period when cued with the current period's vector. We make two assumptions when calculating this probability. First, we calculate similarity stock-by-stock, effectively assuming that investors narrowly frame each stock. This assumption is supported by previous work showing that investors' thinking is quite "narrow" (Benartzi and Thaler (1995); Barberis (2013)). Second, we use a rolling look-back window over the past 180 months (=15 years) for each stock. Thus, for each stock in each month, our measure yields a distribution of weights over the previous 180 months.

Figure 1 above gives a visual representation of the probabilities generated by our measure from 1966-2021, aggregated to the quarterly and market levels, respectively. The x-axis indicates the quarter in which the hypothetical cue occurs, while the y-axis indicates the weight that our measure puts on past quarters (averaged across stocks). These probability distributions display a strong recency effect, where recent months receive a higher weight than distant months, consistent with the recency effect discussed in prior studies (Barberis et al. (2015); Cassella and Gulen (2018); Jin and Sui (2022); Nagel and Xu (2022)). In our framework, we do not impose any direct assumptions related to recency. Rather, the principle of similarity naturally generates the recency effect.

In addition to generating a strong recency effect, our measure also places high weights on past crises if the cue occurs during turbulent economic times. For instance, if cue occurs during the 2020 Covid-19 pandemic, our measure assigns a high weight to the 2008 financial crisis. These spikes in the weights on past crises are a key distinguishing characteristic of our measure and cannot be captured by approaches that focus only on recency. Another key strength of our approach is that it generates weights specific to each stock, allowing both the strength of the recency effect and the spikes in weights on past crises to vary across different stocks.

We validate our measure by demonstrating that it generates patterns that closely align with historical periods discussed during corporate events. We collect transcripts of corporate events, such as earnings calls, and extract the historical periods mentioned during these events. Our measure strongly matches the historical periods that are mentioned. For example, the 2008 financial crisis is heavily discussed during corporate events occurring in the 2020 Covid-19 pandemic. A simple extrapolation model, where the weights on past periods decay exponentially, does not capture these patterns, but our measure does. More generally, when we run a horserace, our measure continues to explain which historical periods are mentioned while the coefficient on this simple extrapolation model is insignificant (though positive).

Having validated our measure, we apply it to three different financial market settings. In our first setting, we show that similarity-based return expectations explain survey-based return expectations. We construct similarity-based return expectations for the aggregate market for each month by weighting historical monthly returns of the S&P 500 index with the weights assigned to these past months by our measure. This approach effectively assumes that investors are sampling historical returns based on similarity in order to construct their return expectations. We find that these similarity-based return expectations are strongly correlated (corr = 0.70; rank corr = 0.69) with return expectations elicited from the Gallup investor survey (Greenwood and Shleifer (2014)). In regression tests, we estimate that a one standard deviation increase in the similarity-based return expectation (corresponding to a 3.8 percentage point higher annualized return) is associated with a 15 percentage point increase in the net bullishness of investors (percentage point difference between the share of bullish and bearish investors) about the stock market over the next 12 months.

We also show that our approach captures more than mere extrapolation of the recent past. To implement these tests, we decompose the similarity-based return expectation into a recent-period component (most recent 12 months) and a distant-period component (more than 12 months in the past), and show that both components independently explain survey expectations. We find similar results when we use a five-year cutoff to distinguish between recent and distant periods. Importantly, expectations from distant periods continue to explain survey expectations when we directly benchmark our measure against a simple extrapolation model, either by controlling for cumulative returns over the past 12 months or by controlling for an exponentially-weighted average of returns over the past five years. By showing that our measure captures more than a simple model of extrapolation, these results also address the potential concern that our similarity-based measure might just be a proxy for recent returns.

In a direct test of the key distinguishing feature of our measure, we find that expectations from distant periods have the most explanatory power during recessions. Recessions are exactly the kinds of periods in which the current environment cues distant but similar crises, such as the tech bubble or the 2008 financial crisis. Further, since such crises occur relatively infrequently, they stand out more. These results show how the principle of similarity – the underlying driver of our measure – allows our measure to draw on distant periods precisely when the current economic environment closely resembles these distant periods.

In our second setting, we show that similarity can serve as a unifying framework to explain not only first-moment beliefs (i.e., return expectations), but also higher-moment beliefs in financial markets. We implement these tests by leveraging the rich cross-sectional variation inherent in our measure, and construct similarity-based volatility for each stock in each month. Specifically, we define similarity-based volatility as the standard deviation of monthly returns over the past 180 months, where each historical return is weighted with its associated weight. We find that similarity-based volatility derived from our measure can explain cross-sectional variation in implied volatility from stock options, establishing a link between similarity and higher-moment beliefs in financial markets. The explanatory power of similarity-based volatility survives even when we control for a large set of alternative predictors of option-implied volatility.

Moving to our third setting, we show that the weights captured by our measure can explain trading behavior of individual investors. In our tests, we revisit previously documented evidence that investors are more likely to repurchase a stock if they previously sold it for a gain (Strahilevitz et al. (2011)). We show that similarity is an important mechanism driving this effect: holding fixed the previously realized gain (loss), investors are more (less) likely to repurchase the stock if the weight on the month in which the gain (loss) was realized is higher. A key strength of this setting is that we can use each investor's history of portfolio holdings and trades to construct *investor-specific* weights, which we show dominate the weights of a representative investor in explaining investor behavior.

Our paper contributes to the literature on beliefs in asset pricing (Greenwood and Shleifer (2014); Barberis et al. (2015); Giglio et al. (2021); Da et al. (2021); Brunnermeier et al. (2021); Adam and Nagel (2023)). We show that similarity-based beliefs generated by our measure can explain return expectations from surveys as well as higher moments implied by stock options and the VIX. In a related paper, Chen et al. (2025) use the idea of similarity to construct a predictor of *realized* aggregate US stock returns from a large corpus of newspaper articles. By contrast, we show the direct link between similarity and *ex-ante* beliefs. Further, our measure is available for the entire cross-section of stocks, not just for the aggregate market.

We also contribute to the literature on experience effects (Malmendier and Nagel (2011, 2016); Malmendier (2021a,b)). Our measure generates some of the key features documented by this literature, including a strong recency effect and a high weight on past crises when the current environment cues such crises. Our measure also matches the historical periods that experts – specifically financial analysts and managers – discuss during corporate events. In terms of actions, we show that similarity-based associations can help explain how past trading experiences affect investors' future repurchasing decisions (Kaustia and Knüpfer (2008); Choi et al. (2009); Strahilevitz et al. (2011)). Taken together, these results support the current push in the literature towards incorporating similarity as a principle to better understand and theoretically model experience effects (Malmendier and Wachter (2024)).

Finally, our paper contributes to a growing literature arguing that similarity-based thinking has broad applications in finance. Theoretical work has emphasized the role of similarity in decision theory (Gilboa and Schmeidler (1995)) and in models of human memory (Mullainathan (2002); Gennaioli and Shleifer (2010); Bordalo et al. (2020); Bodoh-Creed (2020); Bordalo et al. (2023); Wachter and Kahana (2024); Voigt (2023); Bordalo et al. (2025)). Several experimental and empirical studies test the predictions of these models, with a focus on the memory mechanism (Charles (2022); Goetzmann et al. (2022); Colonnelli et al. (2024); Jiang et al. (2025); Enke et al. (2024); Gödker et al. (2025); Graeber et al. (2024); Charles (2025)). In a related study, Chen and Huang (2025) use machine learning to back out latent memory associations from analyst forecasts. In contrast, we start with similarity as a defined organizing principle to construct belief measures ex-ante. Importantly, while our measure can be microfounded by theories of memory, it is also consistent with other interpretations, such as explicit look-up of similar past data.

2 An Organizing Theoretical Framework

This section presents the conceptual framework that guides our empirical approach. In our framework, investors have access to a database of historical events but filter this information through the lens of similarity. The events in the database comprise events that the investor personally experienced as well as events that the investor did not personally experience but can look up (e.g., by reading the financial press). We proceed in two steps. First, we define a weighting mechanism (*Assumption 1*) that determines the probability of drawing on a specific past event based on its similarity to the current environment. Second, we define similarity-based beliefs (*Assumption 2*), where investors use these weights to construct subjective belief distributions.

2.1 Similarity-Based Weighting

Assume that a representative investor has access to a database E, which contains $M > 1$ historical events. Each event can be represented as a high-dimensional vector of $F > 1$ features. Following Graeber et al. (2024), we assume that these features are rich enough to capture both the quantitative as well as qualitative content of each event. For example, an event associated with a stock includes quantitative features such as the stock's return, as well as qualitative features like the stories and narratives about the stock. In addition to these stock-level features, there are also contextual features, like the state of the economy during the event. Just as the stock-level features, contextual features can be quantitative or qualitative in nature. For example, the growth rate of the economy is a quantitative contextual feature, while a story about the economy is a qualitative contextual feature.

Suppose the investor faces a question Q about a stock at time t.[2] The question brings to mind both stock-level features as well as broader, contextual features. Together, these features comprise the cue κ_t,

[2] We intentionally keep the nature of the question abstract here, since this section focuses on the mechanics of similarity-based weighting. In the next section, when we discuss similarity-based beliefs, we consider more specific questions.

which causes the investor to draw on past events. We formalize how investors draw on past events in our first assumption:

Assumption 1. The probability that the investor draws on event e when faced with cue κ_t is given by:

$$r(e, \kappa_t) = \frac{S(e, \kappa_t)}{\sum_{e' \in E} S(e', \kappa_t)}, \tag{1}$$

where S is a symmetric function.

The numerator in the above expression captures the force of similarity: the investor is more likely to draw on event e if it is more similar to the cue κ_t. The denominator contains the sum of similarities between the cue κ_t and all events $e' \in E$, thereby ensuring that the probability distribution integrates to 1. When interpreted through the lens of associative memory theory, the denominator captures the force of interference: the investor cannot fully control the recall process, and all events $e' \in E$ compete for retrieval. Thus, if there are many events that are similar to the cue, the weight on the focal event e is lower (Kahana (2012);Bordalo et al. (2020); Bordalo et al. (2023)).

In the above example, where the investor faces a question Q about a stock at time t, the vector κ_t contains both stock-level as well as contextual features. However, an investor might be cued by a broader theme, such as "inflation fears" or "tech sector uncertainty." By adjusting the cue vector κ_t such that all stock-level entries are null, our framework can accommodate such types of cues. We touch on these possibilities in Section 3.6, where we use market-level narrative themes as cues.

2.2 Similarity-Based Beliefs

Suppose now that Q is a probabilistic question, such as *"What is the expected return of stock l?"*. This question requires the investor to construct an underlying distribution of returns, which leads to our second assumption:

Assumption 2. When evaluating a probabilistic question Q, the investor samples past events e from the database E following Equation (1) and assigns a sampling-based probability to each retrieved event.

This sampling yields a subjective probability measure. Using the sampling-based probability and the historical return of each event, the investor can calculate the similarity-based return expectation as well as higher-moment beliefs for stock l.[3]

3 An Empirical Measure of Similarity-Based Weighting in Financial Markets

In this section, we apply the insights from the conceptual framework to construct an empirical measure of similarity-based weighting in financial markets. Our main measure is based on quantitative features, such as a stock's return and macroeconomic indicators like the GDP growth, that can be easily sourced from well-known databases. However, we also present an alternative measure based on qualitative features, namely the narratives and topical themes extracted from business news.

3.1 General Approach

In our setting, events e and cues κ_t are vectors in F dimensions. We calculate the similarity $S(e, \kappa_t)$ between an event e and a cue κ_t using the cosine similarity measure[4]:

$$S(e, \kappa_t) = 0.5 \times \frac{e \cdot \kappa_t}{\|e\| \cdot \|\kappa_t\|} + 0.5 \tag{2}$$

Here, $e \cdot \kappa_t$ represents the dot product of vectors e and κ_t, $\|e\|$ represents the magnitude (length) of vector e, and $\|\kappa_t\|$ represents the magnitude (length) of vector κ_t. The cosine similarity value ranges from -1 to 1. A value closer to 1 indicates a higher similarity between the vectors, a value closer to -1 indicates a lower similarity. We normalize the cosine similarity into the range $[0, 1]$ by scaling it with a factor of 0.5 and then shifting it by 0.5. For the normalized measure, the similarity is equal to 1 if an event e is identical to the cue. Conversely, the similarity value is 0 if e is the opposite of κ_t.[5]

[3] This approach is similar in spirit to Gennaioli et al. (2024).

[4] The cosine similarity of two vectors, A and B, is calculated by taking the dot product of the vectors and dividing it by the product of their magnitudes: $\frac{A \cdot B}{\|A\| \cdot \|B\|}$

[5] Due to the normalization, two randomly chosen vectors have an expected similarity of 0.5. While at first blush this may appear high, this is not an issue for our empirical approach. The reason is that when we apply Equation (1) to construct our measure, it is relative similarity that matters, not absolute similarity.

As in our conceptual framework, the vectors e and κ_t consist of stock-level features as well as broader, contextual features. While the potential list of features is near endless, we focus on a rich set of features that captures the high-dimensional nature of stock market events. These features can be sorted into three groups: (i) stock-level variables, (ii) firm-level financial ratios, and (iii) broad macroeconomic variables.[6]

At the stock-level, we include stock returns, stock prices, dollar volume, and trading volume. Following Van Binsbergen et al. (2023), we also include a large set of financial ratios, such as the book-to-market ratio and the dividend yield. At the macro-level, we consider consumption growth, GDP growth, growth of industrial production, the unemployment rate, and the inflation rate. These variables can affect firms' earnings and are thus potentially relevant to investors. In total, we consider 72 variables, which we describe in detail in Table IA.1 in the Appendix. All variables are at the monthly level and as of the end of each month.

The above variables exhibit variation at different scales, which can lead to undesired scale-driven weighting effects when using the cosine similarity method. We address this issue by standardizing each variable to have a mean of zero and a standard deviation of one. Using the standardized variables, we compute similarity on a stock-by-stock basis over a rolling 180-month (=15-year) window. This choice reflects a balance of trade-offs. A 15-year horizon provides a sufficiently long look-back period to capture past crises. While longer windows are possible, they would require all stocks in the sample to have been listed for more than 15 years, thereby increasingly selecting for a narrow subset of long-lived firms.[7] Shorter windows, in contrast, would allow for broader stock coverage but may miss important, salient events from the distant past. We adopt the 15-year window as a compromise.

Calculating similarity stock-by-stock implies an assumption of narrow framing: when cued with stock l, investors only draw on past events of stock l. In principle, our empirical approach could allow for cross-stock similarity, where an investor might also draw on past events of other stocks. However, implementing this would be computationally very intensive. Moreover, the assumption of narrow framing is consistent with previous work showing that investors' thinking is quite "narrow" (Benartzi and Thaler (1995); Barberis

[6] In Section 3.4, we analyze which types of features are most important for our results.

[7] In Appendix Figure IA.1, we show our main figure using a 50-year look-back window.

(2013)). We therefore adopt stock-level narrow framing throughout the paper.

Having calculated similarity using Equation 2, we construct the stock-specific weight by applying Equation (1):

$$\pi_{l,t,t-h} \equiv \frac{S(e_{l,t-h}, \kappa_{l,t})}{\sum_{h'=1}^{180} S(e_{l,t-h'}, \kappa_{l,t})} \tag{3}$$

where $e_{l,t-h}$ is an event of stock l in month $t-h$, and $\kappa_{l,t}$ is a stock-specific cue for stock l in month t. Conceptually, $\pi_{l,t,t-h}$ is our empirical proxy for the weight that a representative investor assigns to event $e_{l,t-h}$ when cued with $\kappa_{l,t}$. At this point, it is useful to briefly discuss our choice of notation. In our theoretical framework, the investor can draw on past events not only of stock l, but also of other stocks. In contrast, in all of our empirical applications, the narrow framing assumption implies that the investor only draws on past events of the focal stock l. To distinguish between the two, we reserve $r(e, \kappa_t)$ for the theory and use $\pi_{l,t,t-h}$ as its empirical counterpart. Since events in our tests are at the stock-month level, $\pi_{l,t,t-h}$ can be interpreted as the stock-l-specific probability that a representative investor draws on month $t-h$ when cued in month t.

3.1.1 Data

Our sample consists of stocks (share code 10 or 11) listed on the NYSE, Amex, and Nasdaq (exchange code 1, 2, or 3) that have been listed for at least 15 years.[8] We source stock-level variables from CRSP and firm-level variables from the Financial Ratios Suite offered by Wharton Research Data Services.[9] The macroeconomic variables are obtained from the real-time dataset provided by the Federal Reserve Bank of Philadelphia (Croushore and Stark (2001)).[10] Our monthly sample starts in January 1966, as the firm-level

[8] In Figure IA.1 in the Appendix, we focus on even longer time windows by constructing our measure for firms that have been listed for at least 50 years.

[9] When constructing the firm-level variables, we assume that they are publicly available two months after the end of a fiscal period. Suppose a firm's fiscal year-end is on December 31. We assume that the information from its Form 10-K becomes publicly available at the end of February. These assumptions are consistent with the WRDS Financial Ratios Suite, which assumes that information becomes publicly available two months after the end of a fiscal period.

[10] Macroeconomic variables are often revised after their initial release. We use the initially released values, as this is the most up-to-date information that is available at the time of release.

variables are only available after 1951 and we require a rolling window of 15 years. Our sample ends in December 2021.

We standardize each variable stock-by-stock, using data from the past 15 years to calculate the mean and standard deviation. Specifically, we subtract the mean from each data point and then divide it by the standard deviation. If firm-level variables are missing, we follow Van Binsbergen et al. (2023) and fill them with the monthly industry-level median of the firm's Fama-French 49 industry. After filling, approximately 3% of our observations still have at least one variable with missing data. The majority of these missing values are due to missing trading volume. We set these values to zero when calculating the cosine similarity.

3.2 Characterizing the Measure

We begin by visualizing the resulting probability distributions. For the purposes of visualization, we construct the market-level probability $\bar{\pi}_{t,t-h}$ by taking the equally-weighted average of $\pi_{l,t,t-h}$ across all stocks in our sample in month t[11]:

$$\bar{\pi}_{t,t-h} = \frac{1}{|L_t|} \sum_{l \in L_t} \pi_{l,t,t-h} \tag{4}$$

where L_t is the set of stocks in our sample in month t, and $|L_t|$ is the number of stocks in L_t.

In Figure 1, we further aggregate $\bar{\pi}_{t,t-h}$ to the quarterly level and show the full probability distribution for each quarter from 1967 to 2021. The x-axis represents the quarter in which the hypothetical cue occurs, while the y-axis represents the weights on past quarters. The weights stacked vertically above each quarter constitute one full probability distribution for which the weights sum to one. A darker shade of red indicates a higher weight.

Figure 1 displays a striking recency effect: the weight on recent quarters is much larger than the weight on distant quarters, as evidenced by the dark red area near the lower diagonal line. The recency effect has previously been studied in the finance literature (Malmendier and Nagel (2011); Greenwood and Shleifer (2014); Barberis et al. (2015); Nagel and Xu (2022); Jin and Sui (2022)). However, these studies typically resort to ad-hoc configurations to generate a recency effect. For example, Barberis et al. (2015) and Jin

[11] When calculating this average, we only include stocks for which we have data on the previous 180 months, to ensure that we have a valid weight on each of the previous 180 months.

and Sui (2022) directly assume that investors use exponentially-decaying weights. Similarly, Nagel and Xu (2022) generate the recency effect by assuming that investors use a constant-gain learning rule. In contrast, we do not impose any direct assumptions related to recency. Instead, the principle of similarity naturally generates the recency effect.[12]

A key feature of experience effects is their long-lasting nature (Malmendier (2021a,b)). For instance, Malmendier and Nagel (2011) show that disastrous experiences, like the Great Depression, can have long-lasting effects on financial risk-taking and return expectations. In a more recent study with a memory focus, Jiang et al. (2025) show that distant dramatic events are more likely to be recalled when investors are cued with drastic economic dynamics. Thus, an empirical approach that only focuses on recency (e.g., by imposing exponentially-decaying weights) misses these important patterns.

Here, we show that our measure naturally generates high weights on past crises if the cue occurs during extreme episodes. In Figure 1, we highlight three dramatic episodes: the 2020 Covid-19 pandemic, the 2008 financial crisis, and the 1979 oil crisis. For each of these episodes, our measure assigns high weights to past crises. For instance, during the 2020 Covid-19 pandemic, our measure assigns a high weight to the 2008 financial crisis. Similarly, during the 2008 financial crisis, the tech bubble burst receives a high weight. Finally, during the 1979 oil crisis, the 1973 oil crisis is weighted more heavily.[13]

These high weights on distant but similar events are a key distinguishing characteristic of our measure, one that a simple extrapolation model would miss. To make this point visually salient, in Figure 2, we show the weights implied by the exponential-decay model of Greenwood and Shleifer (2014), using the average decay parameter $\lambda = 0.56$ from their study.[14] As Figure 2 shows, this model generates a strong recency

[12] The recency effect in our measure is driven by autocorrelated or sticky variables. There are many possible microfoundations for recency, including both psychological microfoundations (e.g., memory and attention) and rational microfoundations (e.g., if the economy switches between different, sticky states). Since our measure cannot distinguish these different possibilities, we remain agnostic about the exact forces that are generating recency.

[13] In Figure IA.1 in the Appendix, we recreate Figure 1 using firms that existed for at least 50 years. Focusing on these firms allows us to extend the look-back window from 15 to 50 years. The resulting figure shows that past crises–going all the way back to the 1973 oil crisis–receive high weights during the 2008 financial crisis or the 2020 Covid-19 pandemic.

[14] Throughout the paper, we refer to the λ parameter from Greenwood and Shleifer (2014) several times. We use the Gallup-specific

effect with a fast decay, but it does not generate the high weights on past crises generated by our measure.

In Figures 3 and 4, we dive into the dynamics generated by our measure, by providing snapshots of the weights immediately before and then during a dramatic event. The x-axis in these figures indicates the number of months between the cueing month and the past month, while the y-axis indicates the weights assigned to past months.

Take Figure 3 as an example. This figure illustrates the weighting dynamics before and during the 2008 financial crisis. The upper panel shows (in red) the weights on past months if the cue occurs in June 2007, just before the outbreak of the crisis. The weights gradually decay, consistent with the average weights in our sample (displayed in green). In contrast, the lower panel shows that if the cue occurs in December 2008, in the middle of the financial crisis, our measure assigns higher weights to the months around the tech bubble burst.

The patterns in in Figure 4 are perhaps even more striking. In this figure, we present the weights just before and then during the 2020 Covid-19 pandemic. In the upper panel, we show that if the cue occurs in December 2019, immediately before the outbreak of the pandemic, the weights gradually decay, aligning with the average weights in our sample. In the lower panel, we show that six months later, when the Covid-19 pandemic is in full swing, the weight on the 2008 financial crisis is dramatically higher.

Our measure naturally generates these patterns through the principle of similarity. Intuitively, as a dramatic event like the 2020 Covid-19 pandemic unfolds, our rich set of stock-level, firm-level, and macroeconomic features captures the high-dimensional nature of events during that period and places higher weights on past periods with similar features. Further, since crises occur relatively infrequently, these past events ‘stand out more’ (are subject to less interference), which also leads to higher weights.

3.3 Validating the Measure

Perhaps the most direct way of validating our measure is to show that it generates weighting patterns that match the historical periods that market participants actually talk about. To this end, we show that our

$\lambda = 0.77$ whenever we work with the Gallup survey data, and we use $\lambda = 0.56$, which is the average λ across several surveys, for all other tests.

measure generates patterns that closely align with historical periods discussed during corporate events, like earnings calls.[15]

We collect transcripts of corporate events from Refinitiv StreetEvents for January 2001 to December 2021. These transcripts are verbatim representations of Earnings Calls, M&A Calls, Sales Calls, Analyst Meetings, and Corporate Conference Presentations. Q&A sessions of these calls are also included in the transcripts. In Section A in the Appendix, we describe in detail how we construct the sample and how we process the data to extract past episodes mentioned during each event.

To get a feel for the extracted patterns, we begin by aggregating the data to the monthly level. This aggregation allows us to calculate the relative frequency with which different historical months are mentioned across all events taking place in a given month. As a concrete example, assume that 1,000 firms each had one event (e.g., one earnings call) in January 2015. Further assume that in 200 of these calls, July 2014 was mentioned at least once. In this case, we would assign July 2014 a 20% probability of being mentioned.[16]

Panel A of Figure 5 shows a heatmap of these probabilities (aggregated to the quarterly level). The x-axis represents the quarter in which corporate events take place, while the y-axis represents the relative frequency with which past quarters are mentioned during these events. A darker shade of red indicates a higher probability of being mentioned. The figure displays a very strong recency effect: the past 8-10 quarters are mentioned with a very high probability. There are also high weights on past crises during extreme episodes. For example, during the 2020 Covid-19 pandemic, the 2008 financial crisis receives a very high weight. Similarly, during the 2008 crisis, the tech bubble burst receives a high weight. To ease the comparison of these patterns with our measure, we show our measure for the same sample period in Panel B of Figure 5.

Visually, the strong overlap between patterns generated by our measure and the patterns extracted from corporate events is apparent. We can also quantify it: the correlation between the two is 0.87 (rank correlation

[15] In Appendix B, we provide further evidence, showing that our measure generates patterns typically observed in subjective beliefs data (Da et al. (2021)).

[16] If a firm has multiple events in a month, we consider a historical month as being mentioned if it was mentioned during at least one of the events.

is 0.79). This strong correlation shows that our measure does a good job at capturing the historical periods that market participants talk about, at least for the aggregate market.

In a further step, we show that our measure also captures such patterns at the individual stock-level. Our tests are simple. For a corporate event of stock l in month t, we ask if our measure captures whether a past month $t-h$ was mentioned during the event. We construct a dummy that is equal to one if, during a corporate event of stock l in month t, the past month $t-h$ was mentioned, where $h \in \{1, 2, 3, \ldots, 180\}$. We regress this dummy on $\pi_{l,t,t-h}$, which is the weight assigned to stock l's past month $t-h$ by our measure if the cue occurs in month t. Thus, our regression is at the *stock* × *current year-month* × *past year-month* level. Accordingly, we cluster standard errors by stock, current year-month, and past year-month.

We present summary statistics for the sample in Table 1 and results from estimating the above regression in Table 2. We find that our measure explains which past periods are mentioned. In columns (1) and (6), we do not include any fixed effects or control variables. In all remaining columns, we include *stock* × *past year-month* fixed effects, which capture the possibility that some past stock-months are more/less likely to be mentioned. Such effects could be due to stock-specific shocks that occurred in the past stock-month and which make it more/less likely that this past stock-month is mentioned. We also include quarter fixed effects to account for seasonality.

The coefficient on $\pi_{l,t,t-h}$ in column (1) implies that a one percentage point (pp) increase in the stock-specific weight on month $t-h$ translates into a 95 pp increase in the probability that month $t-h$ is mentioned during a corporate event in month t. In column (2), we include control variables, constructed following Chen and Zimmermann (2022), as well as the aforementioned fixed effects. With these controls, the magnitude of the effect remains very similar at about 93 pp. At first glance, effect sizes of 95 and 93 pp may seem unreasonably large. However, it is important to note that our measure captures a probability distribution for which the sum of all probabilities must be equal to one. In contrast, there is no restriction on how many past months can be mentioned during a corporate event. Thus, these tests should not be interpreted as regressing one probability on another. Rather, these tests highlight the ability of our measure to distinguish between months that are mentioned versus those that are not. Consistent with this intuition, the deep red lower-left diagonal in Panel A of Figure 5 shows that recent months have a very high probability of being mentioned.

Given the strong degree of recency shown in Panel A of Figure 5, some readers may wonder whether a simple model of extrapolation performs just as well as our measure. Therefore, in column (3), we replace our measure with the weights generated by the exponentially-decaying weighting approach of Greenwood and Shleifer (2014), using their estimated average decay parameter $\lambda = 0.56$. While the coefficient on these exponentially-decaying weights is large and positive, it is not significant.

In column (4), we include both our measure as well as the exponentially-decaying weights as explanatory variables and find that the coefficient on our measure remains very similar, both in magnitude and significance. In column (5), we further augment the regression with *stock* × *current year-month* fixed effects, which capture stock-specific shocks in the current stock-month that affect the likelihood of mentioning any past month. Including these fixed effects shrink the effect size to about 68 pp, but the effect remains both economically and statistically highly significant.

One potential concern with the results in columns (1) through (5) is that they may pick up a spurious correlation. Specifically, communicating recent firm performance is an important part of corporate events like earnings calls, and this tendency just happens to coincide with the strong degree of recency in our measure. The fact that an exponentially-decaying model performs worse than our measure should already alleviate some of these worries. Nevertheless, in the following tests, we show that the overlap is not only due to the recency effect. Specifically, we focus on historical months that are at least five years in the past. Visually, once can think of this as slicing off the deep red lower-left diagonal in Panel A of Figure 5.[17]

In columns (6) through (10), we re-estimate the specifications from the first five columns for these distant months. The coefficient estimates on $\pi_{l,t,t-h}$ are much lower, but still highly significant. In terms of magnitude, these estimates imply that a one percentage point (pp) increase in $\pi_{l,t,t-h}$ translates to a 4-7 pp increase in the probability that the distant month is mentioned during a corporate event in month t. These

[17] As an alternative approach to addressing the concern that managers are prone to mentioning recent firm performance, we extract which periods are mentioned for different types of participants in the corporate event. The transcript of each event identifies the speaker of each sentence, such as an analyst or a manager, during each call. This distinction allows us to analyze these patterns separately for different types of speakers. In Appendix Table IA.3, we find that we can explain which periods are mentioned by analysts as well as managers.

lower magnitudes are due to the fact that managers are unconditionally much less likely to discuss the distant past. The results in columns (8) through (10) also show that the coefficients on the exponentially-decaying weights are precise zeroes, indicating that a model based purely on exponential decay does a poor job at explaining which distant periods are mentioned. Overall, the takeaways from Table 2 are that recency is an important reason for the strong overlap between our measure and the patterns extracted from transcripts, that our measure outperforms a simple model of exponential decay, and that our measure also explains which periods from the distant past are mentioned.

3.4 The Relative Importance of Different Features

Our baseline measure comprises a total of 72 different quantitative features. Here, we analyze how different types of features contribute to explaining the patterns extracted from the transcripts of corporate events. To streamline the analysis, we first classify our features into seven mutually exclusive groups.[18] Each group contains features that are conceptually related. For example, the group "Profitability" contains various measures of profitability, such as return on assets, return on equity, and the gross profit margin. We show the composition of each group in Appendix Table IA.2. For each group, we then reconstruct our measure in two ways: (1) using only the features from the focal group, and (2) using features from all other groups except the focal group. In total, we construct 14 variations of our measure. For each variation, we re-estimate column (1) of Table 2. We report the coefficient on $\pi_{l,t,t-h}$ as well as the adjusted R^2 for each of these regressions in Table 3.

Beginning with the regressions in which we only use one group at a time to construct our measure, we find that there is large variation in how much R^2 each group provides on its own. Features in the "Activity and Operational Efficiency" group provide the most explanatory power, followed by "Leverage and Capital Structure" features. Not surprisingly, "Macroeconomic Variables", which do not vary across firms, do not provide much explanatory power in explaining firm-level patterns.

[18] We constructed these groups in the following way: we retained the macroeconomic variables (Part 1 in Table IA.1) and the stock-level variables (Part 2 in Table IA.1) as standalone groups. We then asked ChatGPT 3.5 to categorize the firm-level variables (Part 3 in Table IA.1) into 5 distinct groups. Table IA.2 shows the exact classification.

If we instead exclude one group at a time, we find that the biggest drop in R^2 results from excluding the "Activity and Operational Efficiency" and "Leverage and Capital Structure" groups – the same groups that also provide the highest standalone R^2. This suggests that these groups contain unique information for explaining which past periods are mentioned, information that is not captured by any other group. In contrast, dropping the "Liquidity and Solvency" group does not reduce R^2 by much, and dropping the "Profitability" group actually increases R^2 slightly, suggesting that other groups are good substitutes for these groups and/or that these groups add more noise than signal. Similarly, the "Stock-Level Variables" and "Market Valuation and Returns" groups each add a decent amount of R^2 on their own, but removing one at a time does not reduce R^2 by much. This is intuitive as these groups are likely good substitutes for each other. Overall, the results in Table 3 suggest that some features in our baseline measure are complements, while others are substitutes for explaining which past periods are mentioned.

3.5 Robustness of Feature Selection

As discussed in Section 3.1, we chose a broad set of features to capture the high-dimensional nature of events. Further, we decided to tie our hands by selecting the majority of our features following Van Binsbergen et al. (2023). It is unclear, however, whether this large number of features is necessary, or how sensitive our results are to the inclusion or exclusion of one or several of these features.

To directly answer these questions, we randomly draw subsets of features and reconstruct our measure using only the randomly drawn features. We begin by drawing 5 features without replacement and reconstruct our measure using only the 5 drawn features. We repeat this process 20 times. We then draw 10, 15, ... , 70 features 20 times, each time reconstructing our measure using only the drawn features. We show how the measures resulting from these random draws compare to our baseline measure in Figure 6.

In the left panel, we plot the distribution of the correlation coefficient between the randomly drawn measures and our baseline measure (which is constructed using all 72 features). A darker blue shading indicates a larger number of drawn features. We find that 95% of the randomly drawn measures have a correlation of at least 0.75 with our baseline measure. As expected, the correlation mechanically increases with the number of drawn features. However, the figure also shows that a relatively small draw of 10-15 features suffices to achieve a correlation of 0.75 or higher.

Next, we re-estimate the regressions in columns (1) and (6) of Table 2 using our randomly drawn measures. We plot the distribution of the adjusted R^2 from these regressions in the middle and right panels of Figure 6. The solid red vertical line in these panels indicates the R^2 for our baseline measure. In both panels, most of the mass is close to the solid red line.

Overall, these results show that there isn't one 'magic' feature that is driving our results. Rather, a random draw of the features – often as small as 10 features – suffices to proxy for the high-dimensional nature of events. This highlights an attractive characteristic of our measure: ex-ante knowledge of which features matter is not necessary. One possible reason for this characteristic is that our measure captures the 'collective' patterns of investors. While some investors may focus on certain features, others prioritize different ones. From an aggregate perspective, even if we rely on only a subset of these features to construct our measure, it is still a relevant proxy for overall investor behavior. Another possible reason is that many of the features in our baseline measure are good substitutes for each other, which is also consistent with our results from the previous section. Going forward, we focus on our baseline measure, as it is the most comprehensive measure, but we find very similar results when we use alternative measures based on fewer features.

3.6 A Measure Based on Qualitative Features

We now discuss an alternative approach for capturing events using qualitative features. Specifically, we use narratives and topical themes extracted from business news, provided by Bybee et al. (2024) in the "Monthly Topic Attention" dataset.[19] This dataset provides a time series for each of 180 distinct topics, indicating how much weight ("attention") was allocated to each topic in the news coverage of the Wall Street Journal, for each month from January 1984 to June 2017.

For our purposes, we can think of each month as being characterized by a 180-dimensional vector, where each dimension represents the weight (or attention) allocated to a given topic. Thus, for each month, the topic vector reflects the landscape of narratives that investors are paying attention to in that month. We

[19] Available here.

use these topic vectors to construct a qualitative variant of our measure. We follow the approach described in Section 3.1, except that we use the 180-dimensional topic vectors when calculating similarity.[20]

In Panel A of Figure 7, we show the distribution of the weights implied by this qualitative measure.[21] For comparison, we display the weights generated by our baseline measure for the same sample period (January 1999 to June 2017) in Panel B of Figure 7. Clearly, the distributions implied by both measures are very similar. Indeed, the correlation between the two is 0.93 (rank correlation is 0.90). This suggests that, despite their conceptual differences, both approaches generate similar weights.[22] The consistency of both approaches is intuitive: the business press covers what is going on in the market and the economy. It is therefore not surprising that these qualitative narratives contain information that overlaps with quantitative data.[23]

At the same time, there are also differences between the two measures. The qualitative measure displays visually striking horizontal lines following extreme events, like the tech bubble burst and the 2008 financial crisis. This indicates that the qualitative aspects of extreme events can be very persistent and slow to fade out.[24] In fact, we document similar patterns in the patterns extracted from transcripts of corporate events (Panel A of Figure 5). More broadly, weights constructed from qualitative features decay more slowly than those constructed only from quantitative features. This slower decay is not only reflected in the horizontal lines in Panel A of Figure 7, but also in the larger shaded red regions throughout the heatmap.[25]

[20] Since the monthly topic vectors are at the market-level, we standardize each of the 180 topic dimensions over time and compute similarity at the market-level.

[21] The figure begins in January 1999, as we need an initial 15 years of data for our rolling window of 15 years.

[22] In Appendix Table IA.5, we also analyze the relative importance of different qualitative features in explaining which past periods managers and analysts mention. We find that the topics "Financial Markets," "Trans/Defense/Local," and "Asset Managers/I-Banks" are the three most useful topics.

[23] In Section 4.2, we use the probability distributions from both approaches to generate similarity-based return expectations. We then use these similarity-based return expectations to explain survey-based return expectations, and find that both approaches yield very similar results.

[24] These findings are consistent with the experimental results of Graeber et al. (2024), who show that stories fade from memory much more slowly than statistics.

[25] For a more formal test, we follow the methodology described in Appendix Section B and calculate the decay speed for our baseline

Throughout most of the paper, we focus on the measure based on quantitative features for the following reasons. First, it can be easily constructed at the firm-level, as all public firms are required to regularly file financial statements with the Securities and Exchange Commission, allowing us to generate firm-specific probability distributions for a large cross section of firms. By contrast, the qualitative measure relies on market-wide news coverage and is therefore available only at the market-level. Indeed, it would be quite challenging to construct a measure based on qualitative features at the firm-level, as it would require a long time series of consecutive news coverage for each firm. While the largest firms are covered regularly in the news, most small and medium firms are covered only occasionally. Further, the measure based on quantitative measures allows us to construct probability distributions not only for a large cross section of firms, but also for a long time series for each firm, as we can draw on the long sample periods available for the CRSP and Compustat databases.

4 Similarity-Based Beliefs

In this section, we use our measure to construct similarity-based return expectations for the aggregate stock market. We then link these similarity-based return expectations with survey-based return expectations. We also construct similarity-based volatility and link it with return volatility implied by stock options.

4.1 Constructing Similarity-Based Return Expectations

We begin by describing how we use the measure introduced in Section 3 to construct similarity-based return expectations for the aggregate stock market. We draw on the market-level probability distributions from Equation (4) and weight historical realized returns of the S&P 500 index with these aggregate weights. This approach yields the similarity-based return expectation for the aggregate market:

$$\tilde{\mathbb{E}}_t[\text{ret}_{Mkt,t+1}] = \sum_{h=1}^{180} \bar{\pi}_{t,t-h}\, \text{ret}_{Mkt,t-h+1} \tag{5}$$

quantitative measure as well as for the qualitative measure. We find that the decay speed for the quantitative measure is higher at 1.247 compared to 1.161 for the qualitative measure. This difference is highly significant ($t = 14.168$).

where $\bar{\pi}_{t,t-h}$ is the weight that our measure assigns to month $t-h$ in month t, and $\text{ret}_{Mkt,t-h+1}$ is the realized return of the S&P 500 index in month $t-h+1$. We choose this weighting approach, as the investor in month t is trying to construct the return expectation for the *next* month, $t+1$.

4.2 Linking Similarity-Based and Survey-Based Return Expectations

We next test whether the similarity-based return expectations constructed using our measure can explain actual investor expectations from surveys. The survey expectations in our tests are from the Gallup investor survey, as it provides a large sample size and follows a consistent methodology (Greenwood and Shleifer (2014)). The Gallup survey does not ask investors for the percentage return they expect to earn in the market, but rather asks investors whether they are bullish or bearish about the market over the next 12 months. We follow Greenwood and Shleifer (2014) and define survey expectations as the percentage point difference between the share of bullish and bearish investors. We source data from Greenwood and Shleifer (2014) for the period from October 1996 to December 2011, and extend the sample period to May 2020 with data sourced directly from Gallup.[26]

Figure 8 gives a visual impression of the relationship between similarity-based and survey-based return expectations. The dashed blue line shows survey-based expectations in month $t+1$.[27] The solid black line shows annualized similarity-based return expectations as of the end of month t, constructed using our measure. Clearly, the two time series comove heavily, especially from 2002 onward. The correlation between the two time series is 0.70 (the rank correlation is 0.69). Overall, the figure suggests that similarity-based and survey-based return expectations are tightly linked. To establish this relationship more rigorously, we estimate the following regression:

$$\text{Survey Expectations}_{t+1} = a + b\tilde{\mathbb{E}}_t^{Ann}[\text{ret}_{Mkt,t+1}] + cX_t + e_t, \tag{6}$$

[26] The Gallup survey elicited survey expectations every three months from 1996-1998, then switched to monthly elicitations, and reverted back to elicitations every three months from 2011 onward. As in Greenwood and Shleifer (2014), some months are missing in the period from 1998-2011.

[27] We use survey expectations from month $t+1$, since survey responses are typically collected throughout a month. This approach ensures that all the information used to construct similarity-based expectations at the end of month t is available to survey respondents throughout the month $t+1$.

where $\tilde{\mathbb{E}}_t^{Ann}[\text{ret}_{Mkt,t+1}]$ is the annualized similarity-based return expectation as of the end of month t, and Survey Expectations$_{t+1}$ are survey-based expectations in month $t+1$ from the Gallup investor survey. X_t is a set of control variables that captures factors which potentially influence investors' belief formation, including the price level (proxied for by Log(P/D)), the risk-free rate, earnings growth, and the unemployment rate. This set of controls follows Greenwood and Shleifer (2014). We show summary statistics of the sample in Panel A of Table 4, and present regression results in Panel A of Table 5.

In column (1) of Panel A of Table 5, we estimate the above regression and find that similarity-based return expectations explain survey-based return expectations. In terms of magnitude, the coefficient implies that a one standard deviation increase in similarity-based return expectations (corresponding to a 3.8 percentage point higher annualized return) is associated with a 15 percentage point increase in the net bullishness of investors about the stock market over the next 12 months. It is also worth noting that our measure drives out the explanatory power of the price level Log(P/D) documented by Greenwood and Shleifer (2014). This result is intuitive, as the price level in log terms is just the sum of historical returns. Since our measure captures the returns from the most similar periods over the past 15 years, the remaining returns that are captured by Log(P/D) are not useful in explaining survey expectations.

One potential concern with this result is that our similarity-based measure might just be a proxy for recent returns, which are known to explain survey expectations (Greenwood and Shleifer (2014)). To address this concern, in column (2), we decompose the similarity-based return expectation into two components, which capture expectations from recent periods (defined as the most recent 12 months) and expectations from distant periods (defined as more than 12 months in the past), respectively. The sum of these two components equals our baseline similarity-based expectation from column (1). We find that both components independently explain survey expectations. In terms of magnitude, a one standard deviation increase in expectations from recent (distant) periods is associated with a 14.4 (7.6) percentage point increase in the net bullishness of investors.

In column (3), we further augment this regression with the cumulative return over the past 12 months, since Greenwood and Shleifer (2014) show that recent returns are an important determinant of investors' return expectations. This control variable is heavily correlated with the component capturing only expectations from recent periods. Despite this, it does not wash out the coefficient on expectations from recent periods. Further, the coefficient on expectations from distant periods remains virtually unchanged.

In column (4), we directly benchmark our measure against a simple model of extrapolation by controlling for the exponentially-weighted average of returns over the past 5 years. To construct this exponentially-weighted return, we first calculate quarterly returns by compounding 3-month returns on a rolling monthly basis. We then use the weighting approach of Greenwood and Shleifer (2014) and their estimated quarterly λ of 0.77 to calculate an exponentially-weighted average return over the past five years. The λ of 0.77 from Greenwood and Shleifer (2014) is estimated using the Gallup investor survey, which is the same survey data that we use in our tests. Thus, if the concern is that our measure is just a fancy way to generate extrapolation, controlling for this exponentially-weighted average return is arguably a high bar to clear for our measure. As the results in column (4) show, expectations from both recent and distant periods continue to explain investors' expectations, highlighting that our measure captures more than a simple model of extrapolation. In columns (5) to (7), we replicate the regressions from columns (2) to (4), using a five-year cutoff to distinguish between recent and distant periods. Our results remain very similar.

Finally, in column (8), we ask whether expectations from distant periods are particularly useful in explaining survey expectations during bad times. These tests are motivated by the results presented in Section 3.2, which show that our measure generates high weights on distant crises during turbulent economic times. We implement this test by interacting expectations from recent and distant periods with a dummy variable that is equal to one during NBER recessions. Indeed, expectations from distant periods have more explanatory power, and expectations from recent periods have less, during recessions. This is intuitive: the 2020 Covid-19 pandemic, for example, may be perceived as similar to the 2008 financial crisis. Similarly, during the 2008 crisis, investors may have drawn on the tech bubble burst.

In Panel B of Table 5, we repeat the analysis from Panel A but use the measure based on qualitative features, which we presented in Section 3.6, to construct similarity-based return expectations. Across all columns, we find very similar results to Panel A, except that the exponentially-weighted past return sometimes washes out the coefficient on expectations from recent periods. However, expectations from distant periods continue to explain survey expectations. Further, in Appendix Table IA.6, we repeat the analysis using a measure that blends quantitative and qualitative features. Here, too, we find very similar results.

Overall, the results in Tables 5 and IA.6 show that similarity-based expectations, constructed using either quantitative or qualitative features, can explain survey expectations. Importantly, the results show that similarity-based expectations capture more than vanilla extrapolation of past returns. The principle of similarity – the underlying driver of our measure – allows it to draw on distant events precisely when the current economic environment resembles these distant events.

4.3 Similarity-Based Volatility

In this section, we test whether similarity can serve as a unifying framework for explaining not only first-moment beliefs (i.e., return expectations), but also higher-moment beliefs in financial markets. These tests also allow us to leverage a key strength of our measure: its rich cross-sectional variation. We use this variation to construct similarity-based volatility for each stock in each month, and link it with option-implied volatility.

We define similarity-based volatility as the standard deviation of monthly returns over the past 180 months, where each historical return is weighted with its associated weight. We begin by constructing stock-specific similarity-based return expectations:

$$\tilde{\mathbb{E}}_t[\text{ret}_{l,t+1}] = \sum_{h=1}^{180} \pi_{l,t,t-h}\, \text{ret}_{l,t-h+1} \tag{7}$$

where $\text{ret}_{l,t-h+1}$ is the realized return of stock l in month $t-h+1$ and $\pi_{l,t,t-h}$ is constructed using Equation (3). Using these stock-specific return expectations, we construct stock-specific similarity-based volatility as follows:

$$\tilde{\sigma}_{l,t} = \sqrt{\sum_{h=1}^{180} \pi_{l,t,t-h} \left(\text{ret}_{l,t-h+1} - \tilde{\mathbb{E}}_t[\text{ret}_{l,t+1}]\right)^2} \tag{8}$$

As our measure of actual perceived volatility, we construct option-implied volatility for each stock in each month, from January 1996 to December 2021. We follow An et al. (2014) and use data from OptionMetrics to calculate the average of put and call implied volatility. In our tests, we regress option-implied volatility in month $t+1$ on annualized similarity-based volatility in month t:

$$\text{Implied Volatility}_{l,t+1} = a + b\tilde{\sigma}^{Ann}_{l,t} + \text{Controls} + \text{FEs} + e_{l,t}. \tag{9}$$

We show summary statistics of the sample in in Panel B of Table 4, and present regression results in Table 6. The estimates in column (1) of Table 6 imply that a one unit increase in similarity-based volatility is associated with an increase in option-implied volatility of 0.644 units. In column (2), we control for perceived volatility from an exponentially-decaying model using the decay parameter $\lambda = 0.56$ from Greenwood and Shleifer (2014). We construct exponential-decay-based volatility as the standard deviation of monthly returns over past months, where each historical return is weighted with exponentially-decaying weights. While the coefficient on similarity-based volatility in column (2) shrinks by about 40% compared to column (1), it remains statistically significant and economically meaningful, emphasizing that our measure captures more than mere extrapolation of recent volatility.

In column (3), we augment the regression with stock and year-month fixed effects. These fixed effects leave the effect of similarity-based volatility largely unchanged. In column (4) we further control for lagged implied volatility, as implied volatility is known to be persistent. The coefficient on similarity-based volatility shrinks substantially once we control for lagged implied volatility, highlighting the importance of this control, but similarity-based volatility continues to explain option-implied volatility.[28] Finally, in column (5), we show that the coefficient on similarity-based volatility remains robust even when we add a host of control variables that capture firm fundamentals that are know to predict returns in the cross-section, namely (i) a firm's size, which we measure as the logarithm of a firm's market capitalization (in million $), (ii) idiosyncratic volatility in %, which we construct following Ang et al. (2006), (iii) asset growth, which we construct as the book asset growth rate (= current book assets – lagged book assets)/lagged book assets), (iv) operating profit following Fama and French (2006), and (v) the logarithm of the book-to-market ratio, where the book-to-market ratio is constructed following Fama and French (1992). In terms of magnitude, the estimate in column (5) implies that a one unit increase in similarity-based volatility is associated with an increase in option-implied volatility of 0.077 units.[29]

[28] In Table IA.7 in the Appendix, we use an alternative approach to handle the persistence in option-implied volatility. We estimate Fama-MacBeth regressions, which only exploit cross-sectional variation, and find similar results.

[29] We would like to highlight one caveat: option-implied volatility consists of subjective beliefs about volatility plus a risk premium, so it is possible that the estimates in Table 6 are colored by the risk premium component. However, our controls are designed to

In summary, the results in this section show that similarity is a potential unifying principle for understanding investor beliefs about various moments in financial markets.[30] From a methodological perspective, the tests in this section also highlight how our measure can be used to easily construct higher-moment beliefs for a range of financial and economic variables.

5 Similarity-Driven Repurchasing Decisions

In our third application, we test whether similarity explains not only beliefs but also investment actions. We focus on the repurchasing decisions of individual investors, revisiting the "repurchase effect" documented by Strahilevitz et al. (2011), who find that investors are more likely to repurchase a stock if they previously sold it for a gain. We hypothesize that similarity is an important mechanism driving this effect. Specifically, if the investor is reminded of a past gain – either by passively recalling it or by actively looking it up – she is more likely to repurchase the associated stock today. Crucially, this implies that identical past realized returns differentially affect the probability of repurchase, depending on their similarity-based weight in the decision process. This is a key prediction that we take to the data.

To test this hypothesis, we use the Barber and Odean (2000) discount brokerage dataset. This setting offers the unique advantage that we can construct *investor-specific* similarity weights. Because we observe each investor's history of holdings and trades, we can construct a distinct database of events, E_i, for each investor. Using this database, we calculate how similar the current period is to past periods in which the investor realized gains or losses. In Section D of the Internet Appendix, we describe the construction of the sample in detail and present the full regression results. Summarizing the results, we find strong evidence for our hypothesis: the interaction between past realized returns and their similarity-based weights explains repurchasing behavior. These results show that when deciding whether to repurchase a stock, investors draw

alleviate these concerns. If the risk premium is time-varying but the same for all stocks, we pick it up with the month fixed effect. If the risk premium is unique to each stock, we pick up the stock-specific constant part with stock fixed effects. Lagged implied volatility captures persistence in the remaining time-varying component.

[30] In Appendix C, we construct similarity-based volatility for the aggregate market and show that it can explain variation in the Volatility Index (VIX) over time.

on their past historical experiences with that stock.

The trading setting allows for another key test. Specifically, since we can construct similarity weights that are unique to each investor as well as similarity weights of the representative investor, we can test which weights better explain trading behavior. If the investor-specific weights dominate, this implies that personally experienced events carry more weight in explaining trading behavior. In contrast, if the representative-investor weights dominate, it implies that investors also draw on historical events that they did not personally experience.[31] In our tests, we find that investor-specific weights dominate. This suggests that investors do not optimally use the entire historical data, but instead draw more heavily on their own, personal experiences.

6 Alternative Explanations

Here, we discuss two alternative explanations for our results. The first possibility is that our measure is merely a relabeling of investors' well-documented tendency to extrapolate recent returns (Barberis (2018)). Our tests are designed to rule out this possibility. First, in our tests linking similarity-based and survey-based return expectations in Table 5, we control for various measures of extrapolation of recent returns. The results from these tests also show that expectations from distant periods (more than five years in the past) are important for explaining investors' expectations, especially during recessions.[32]

The second possibility is that our results are driven by a spurious correlation between current and past periods. For instance, it might be that investors are pessimistic during recessions, and this generates a spurious correlation with past crises in our tests. In other words, this explanation posits that investors completely ignore the past, and only react to the present – but the present is spuriously correlated with the

[31] For both weighting schemes, the outcome being weighted is always a personally realized return. The difference between investor-specific and representative-investor weights lies in the set of months included in the normalization (the interference term), which either includes only the investor's own holding history versus the full 180-month history of the stock.

[32] Further, in the trading tests in Section D of the Internet Appendix, we control for *stock* × *year-month* fixed effects. If investors merely repurchased stocks that recently performed well, these fixed effects would soak up this tendency. A simple extrapolation model cannot explain these results.

past. While this explanation might plausibly explain the results linking similarity-based and survey-based return expectations in Table 5, it cannot explain the results linking similarity-based and option-implied volatility in Table 6. The reason is that these tests are conducted in the cross-section of stocks, allowing us to include year-month fixed effects. These fixed effects account for differences in optimism/pessimism across time periods, ruling out that such differences are driving the results.[33]

7 Conclusion

In this paper, we propose a framework in which investors draw on past periods that are similar to the current environment when forming their beliefs or making decisions. Motivated by our framework, we construct a similarity-based measure that specifies which historical periods investors draw on at a given point in time. We validate this measure by showing that it matches the historical periods that managers and analysts discuss during corporate events. We also show that this similarity-based weighting of history explains return expectations from investor surveys, higher moment beliefs implied by option prices, and the trading decisions of individual investors.

On the methodological front, we provide a straightforward, theory-based approach for constructing similarity-based beliefs for a wide range of economic and financial variables. For instance, using similarity-based probability distributions generated by our measure, it is possible to construct similarity-based expectations for macro-variables like GDP growth or the inflation rate, as well as for firm-level variables such as earnings or cash flow growth. Being able to construct similarity-based expectations for different variables

[33] In the trading tests in Section D of the Internet Appendix, we further address this concern in a number of ways, as these tests exploit variation both in the cross-section of stocks as well as in the cross-section of investors. By including *stock* × *year-month* fixed effects, we can show that investors behave differently when looking at the same stock in the same month. The variation in this setting also allows us to include *account* × *year-month* fixed effects, which control for differences in optimism/pessimism across investors for the same time period. In our tightest tests, we show that for the same investor in the same month, the current environment differentially cues experiences of stocks that were previously liquidated in the same past month. Even with these stringent controls, we continue to exploit variation across investors who face different interference levels due to their unique trading histories. Spurious correlations between the current and past periods cannot explain these results.

may be particularly valuable when survey data are not available. A further advantage of our approach is that it is very flexible. By expanding or reducing the set of features used to construct the measure, future studies can tailor it to different applications. Thus, our approach may help researchers study the role of similarity in other domains of the economy.

Our findings also open a path to better understand the foundations – cognitive or otherwise – that drive the effects we document. Our empirical results are consistent with a broad class of mechanisms in which agents try to identify the current state of the world and look back to see what happened in similar states. For instance, such similarity-based weighting maps onto nearest-neighbor estimation, where investors efficiently use data from the most relevant precedents, as well as models of reinforcement learning, where agents in complex environments generalize from past events by mapping current states to similar past states. Alternatively, our results are also consistent with theories of associative memory, in which recall is selectively skewed towards more similar past experiences, or with explicit cognitive reasoning, in which investors actively seek out informative historical analogies. Regardless of whether the underlying driver is a subconscious process or a conscious calculation, our results suggest that similarity is a fundamental principle governing how financial market participants process historical information to form their beliefs and decide on their actions.

References

Adam, K. and Nagel, S. (2023). Expectations data in asset pricing. In *Handbook of Economic Expectations*. Elsevier.

An, B.-J., Ang, A., Bali, T. G., and Cakici, N. (2014). The joint cross section of stocks and options. *Journal of Finance*, 69(5):2279–2337.

Ang, A., Hodrick, R. J., Xing, Y., and Zhang, X. (2006). The cross-section of volatility and expected returns. *Journal of Finance*, 61(1):259–299.

Barber, B. M. and Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance*, 55(2):773–806.

Barberis, N. (2018). Psychology-based models of asset prices and trading volume. In *Handbook of behavioral economics: applications and foundations 1*, volume 1, pages 79–175. Elsevier.

Barberis, N., Greenwood, R., Jin, L., and Shleifer, A. (2015). X-capm: An extrapolative capital asset pricing model. *Journal of Financial Economics*, 115(1):1–24.

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1):173–196.

Ben-David, I. and Hirshleifer, D. (2012). Are investors really reluctant to realize their losses? trading responses to past returns and the disposition effect. *Review of Financial Studies*, 25(8):2485–2532.

Benartzi, S. and Thaler, R. H. (1995). Myopic loss aversion and the equity premium puzzle. *Quarterly Journal of Economics*, 110(1):73–92.

Bodoh-Creed, A. L. (2020). Mood, memory, and the evaluation of asset prices. *Review of Finance*, 24(1):227–262.

Bordalo, P., Burro, G., Coffman, K., Gennaioli, N., and Shleifer, A. (2025). Imagining the future: memory, simulation, and beliefs. *Review of Economic Studies*, 92(3):1532–1563.

Bordalo, P., Conlon, J. J., Gennaioli, N., Kwon, S. Y., and Shleifer, A. (2023). Memory and probability. *Quarterly Journal of Economics*, 138(1):265–311.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2020). Memory, attention, and choice. *Quarterly Journal of Economics*, 135(3):1399–1442.

Brunnermeier, M., Farhi, E., Koijen, R. S., Krishnamurthy, A., Ludvigson, S. C., Lustig, H., Nagel, S., and Piazzesi, M. (2021). Perspectives on the future of asset pricing. *Review of Financial Studies*, 34(4):2126–2160.

Bybee, L., Kelly, B., Manela, A., and Xiu, D. (2024). Business news and business cycles. *Journal of Finance*, 79(5):3105–3147.

Cassella, S. and Gulen, H. (2018). Extrapolation bias and the predictability of stock returns by price-scaled variables. *Review of Financial Studies*, 31(11):4345–4397.

Charles, C. (2022). Memory and trading. *Working Paper*.

Charles, C. (2025). Memory moves markets. *Review of Financial Studies*, 38(6):1641–1686.

Chen, A., Hoberg, G., and Zhang, M. B. (2025). Haven't we seen this before? return predictions from 200 years of news. *Return Predictions from*, 200.

Chen, A. Y. and Zimmermann, T. (2022). Open source cross sectional asset pricing. *Critical Finance Review*, 11(2):207–264.

Chen, Z. and Huang, J. (2025). Memory and beliefs in financial markets: A machine learning approach. *Working Paper*.

Choi, J. J., Laibson, D., Madrian, B. C., and Metrick, A. (2009). Reinforcement learning and savings behavior. *Journal of Finance*, 64(6):2515–2534.

Colonnelli, E., Gormsen, N. J., and McQuade, T. (2024). Selfish corporations. *Review of Economic Studies*, 91(3):1498–1536.

Croushore, D. and Stark, T. (2001). A real-time data set for macroeconomists. *Journal of Econometrics*, 105(1):111–130.

Da, Z., Huang, X., and Jin, L. J. (2021). Extrapolative beliefs in the cross-section: What can we learn from the crowds? *Journal of Financial Economics*, 140(1):175–196.

Enke, B., Schwerter, F., and Zimmermann, F. (2024). Associative memory, beliefs and market interactions. *Journal of Financial Economics*, forthcoming.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465.

Fama, E. F. and French, K. R. (2006). Profitability, investment and average returns. *Journal of Financial Economics*, 82(3):491–518.

Frydman, C., Barberis, N., Camerer, C., Bossaerts, P., and Rangel, A. (2014). Using neural data to test a theory of investor behavior: An application to realization utility. *Journal of Finance*, 69(2):907–946.

Gennaioli, N., Leva, M., Schoenle, R., and Shleifer, A. (2024). How inflation expectations de-anchor: The role of selective memory cues. Technical report, National Bureau of Economic Research.

Gennaioli, N. and Shleifer, A. (2010). What comes to mind. *Quarterly Journal of Economics*, 125(4):1399–1433.

Giglio, S., Maggiori, M., Stroebel, J., and Utkus, S. (2021). Five facts about beliefs and portfolios. *American Economic Review*, 111(5):1481–1522.

Gilboa, I. and Schmeidler, D. (1995). Case-based decision theory. *Quarterly Journal of Economics*, 110(3):605–639.

Gödker, K., Jiao, P., and Smeets, P. (2025). Investor memory. *The Review of Financial Studies*, 38(6):1595–1640.

Goetzmann, W. N., Watanabe, A., and Watanabe, M. (2022). Evidence on retrieved context: How history matters. *Working Paper*.

Graeber, T., Roth, C., and Zimmermann, F. (2024). Stories, statistics, and memory. *Quarterly Journal of Economics*, 139(4):2181–2225.

Greenwood, R. and Shleifer, A. (2014). Expectations of returns and expected returns. *Review of Financial Studies*, 27(3):714–746.

Hartzmark, S. M. (2015). The worst, the best, ignoring all the rest: The rank effect and trading behavior. *Review of Financial Studies*, 28(4):1024–1059.

Jiang, Z., Liu, H., Peng, C., and Yan, H. (2025). Investor memory and biased beliefs: Evidence from the field. *Quarterly Journal of Economics*, forthcoming.

Jin, L. J. and Sui, P. (2022). Asset pricing with return extrapolation. *Journal of Financial Economics*, 145(2):273–295.

Kahana, M. J. (2012). *Foundations of human memory*. OUP USA.

Kaustia, M. and Knüpfer, S. (2008). Do investors overweight personal experience? evidence from ipo subscriptions. *Journal of Finance*, 63(6):2679–2702.

Li, K., Mai, F., Shen, R., and Yan, X. (2021). Measuring corporate culture using machine learning. *Review of Financial Studies*, 34(7):3265–3315.

Malmendier, U. (2021a). Experience effects in finance: Foundations, applications, and future directions. *Review of Finance*, 25(5):1339–1363.

Malmendier, U. (2021b). Exposure, experience, and expertise: Why personal histories matter in economics. *Journal of the European Economic Association*, 19(6):2857–2894.

Malmendier, U. and Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking? *Quarterly Journal of Economics*, 126(1):373–416.

Malmendier, U. and Nagel, S. (2016). Learning from inflation experiences. *Quarterly Journal of Economics*, 131(1):53–87.

Malmendier, U. and Wachter, J. A. (2024). Memory of past experiences and economic decisions. In Kahana, M. and Wagner, A., editors, *The Oxford Handbook of Human Memory*, volume 2 of *Applications*. Oxford University Press.

Mullainathan, S. (2002). A memory-based model of bounded rationality. *Quarterly Journal of Economics*, 117(3):735–774.

Nagel, S. and Xu, Z. (2022). Asset pricing with fading memory. *Review of Financial Studies*, 35(5):2190–2245.

Odean, T. (1998). Are investors reluctant to realize their losses? *Journal of Finance*, 53(5):1775–1798.

Shefrin, H. and Statman, M. (1985). The disposition to sell winners too early and ride losers too long: Theory and evidence. *Journal of Finance*, 40(3):777–790.

Strahilevitz, M. A., Odean, T., and Barber, B. M. (2011). Once burned, twice shy: How naive learning, counterfactuals, and regret affect the repurchase of stocks previously sold. *Journal of Marketing Research*, 48(SPL):S102–S120.

Van Binsbergen, J. H., Han, X., and Lopez-Lira, A. (2023). Man versus machine learning: The term structure of earnings expectations and conditional biases. *Review of Financial Studies*, 36(6):2361–2396.

Voigt, M. (2023). Investor beliefs and asset prices under selective memory. *Working Paper*.

Wachter, J. A. and Kahana, M. J. (2024). A retrieved-context theory of financial decisions. *Quarterly Journal of Economics*, 139(2):1095–1147.

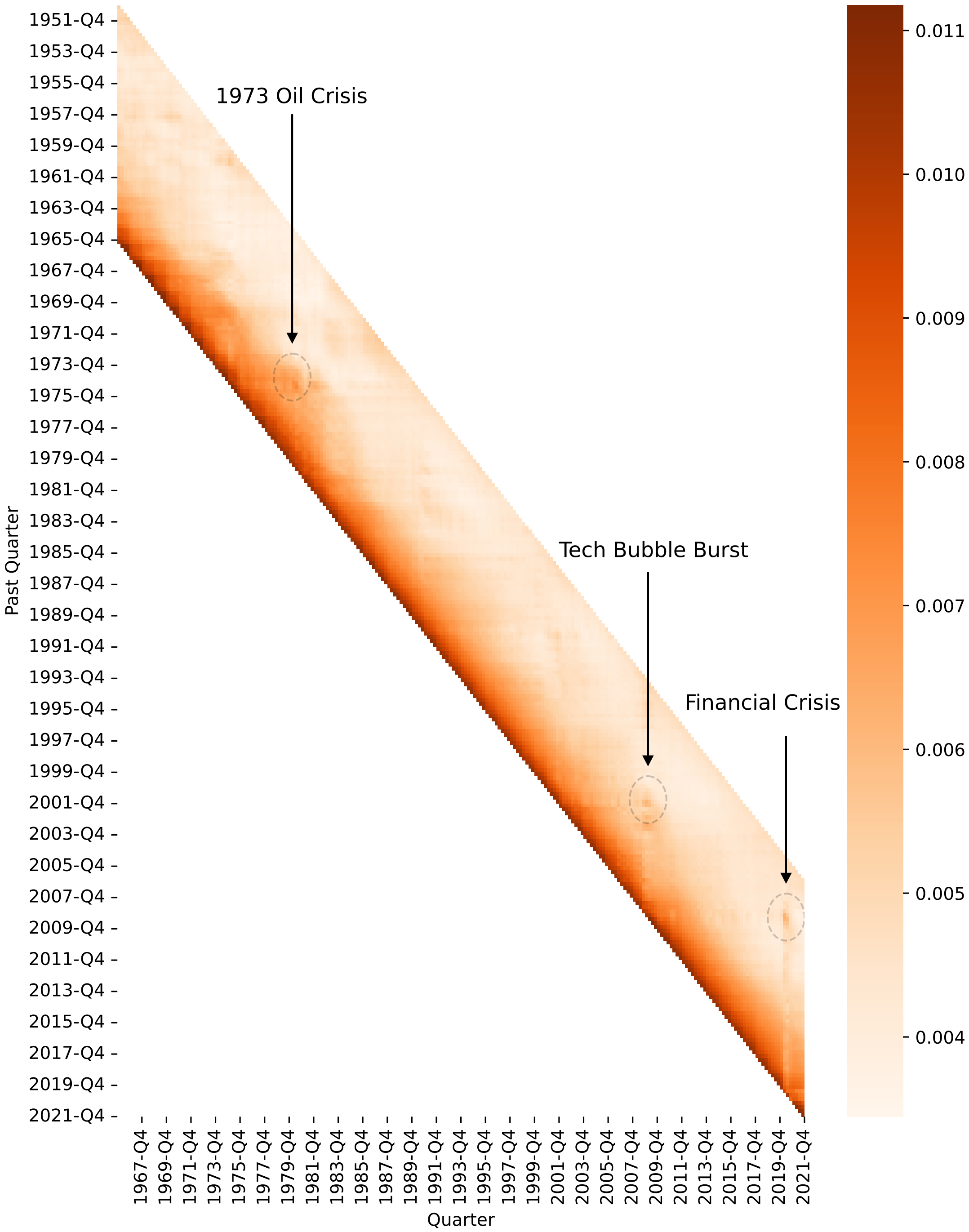


Figure 1: Similarity-Based Weights from our Measure

This figure displays the weights that our measure assigns to the past 60 quarters (= 15 years) for each quarter from 1966 to 2021. The x-axis indicates the quarter in which the hypothetical cue occurs, while the y-axis indicates the weights assigned to past quarters. A darker shade of red indicates a higher weight. This figure is also repeated at the beginning of the manuscript.

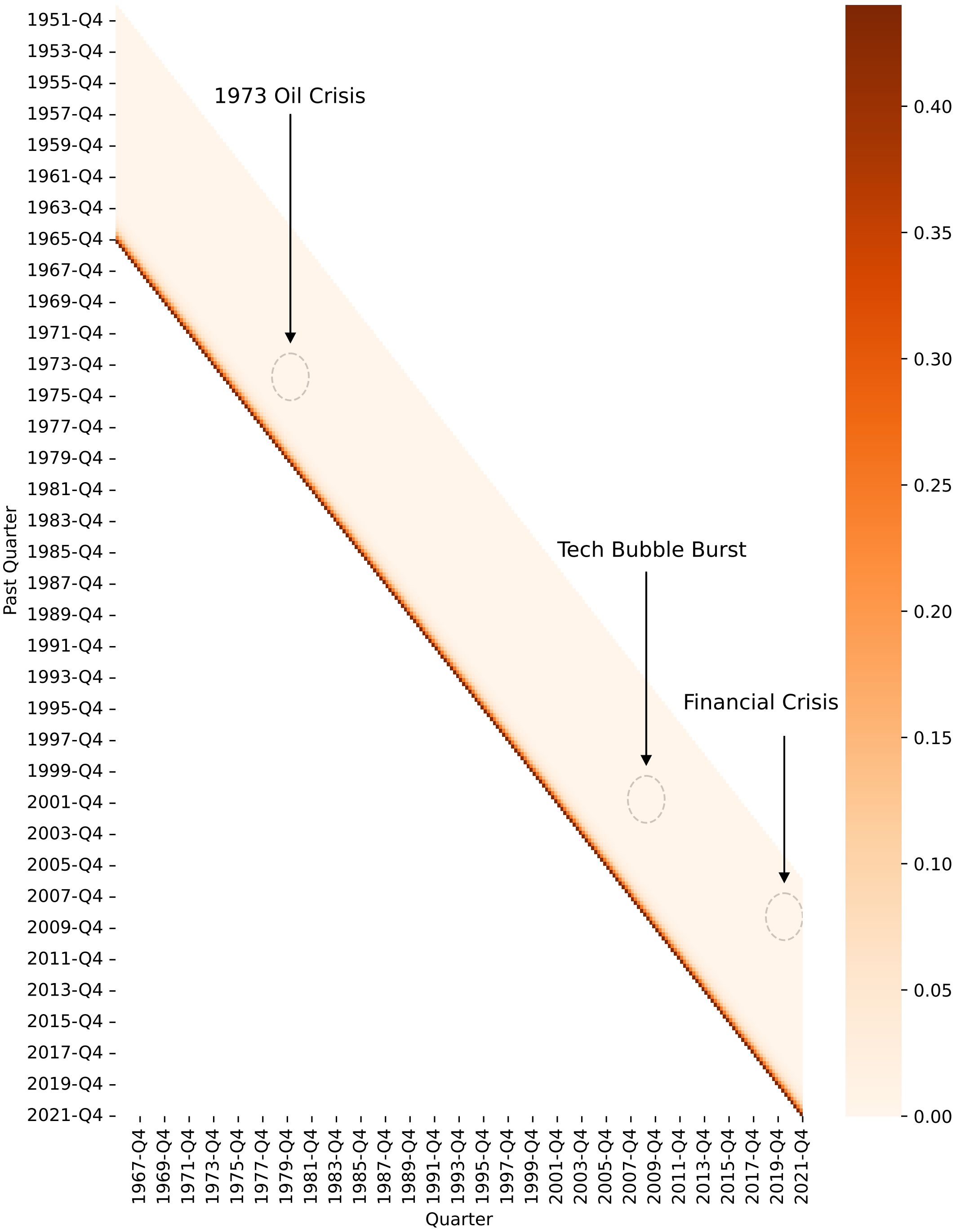


Figure 2: Weights from a Simple Model of Exponential Decay

This figure displays the weights that a simple model of exponential decay assigns to the past 60 quarters (= 15 years) for each quarter from 1966 to 2021. The weight assigned to the j-lagged quarter is given by:

$$\omega_j = \frac{\lambda^j}{\Sigma_{k=0}^{59} \lambda^k},$$

where $\lambda = 0.56$ following Greenwood and Shleifer (2014).

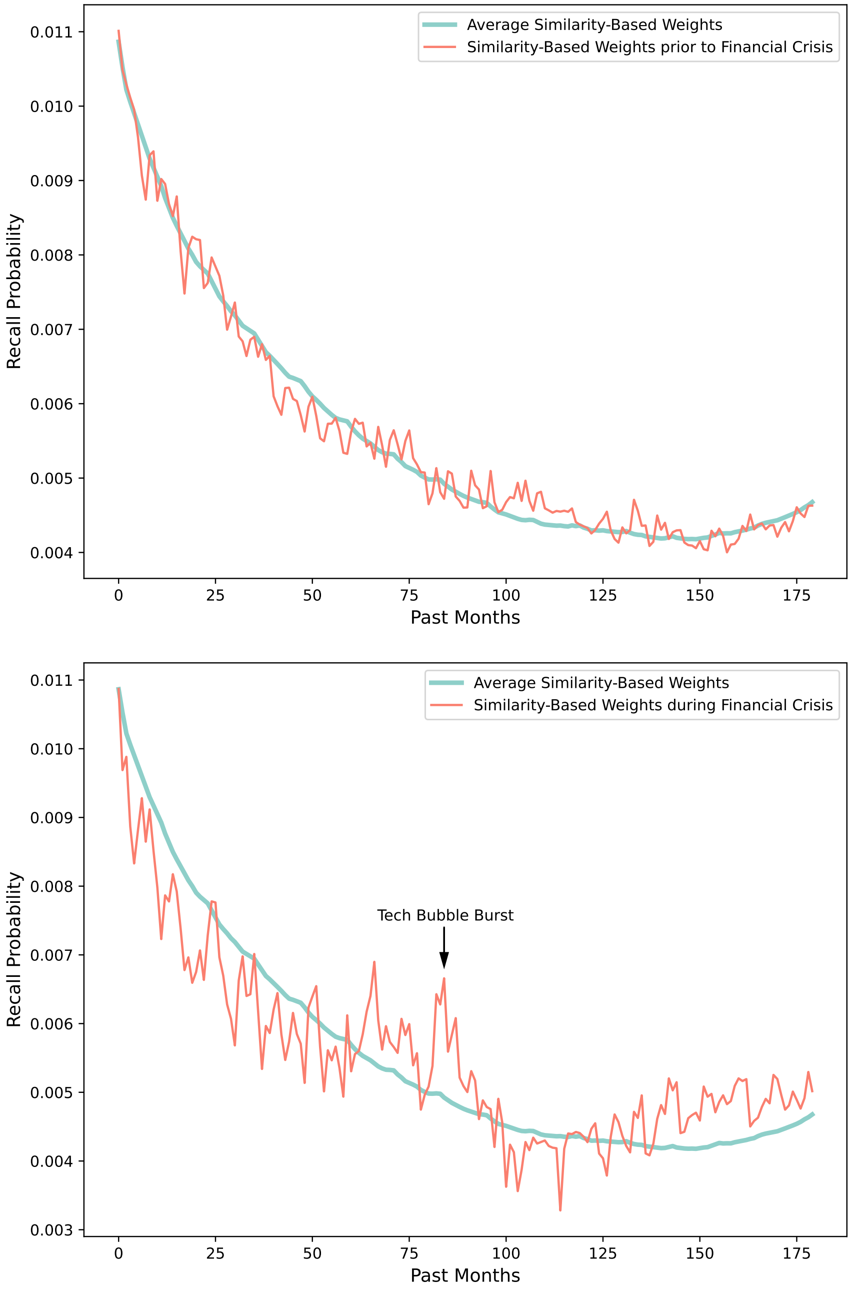


Figure 3: Similarity-Based Weights from our Measure: Before and During the 2008 Financial Crisis

This figure displays the weights constructed using our measure immediately before and during the 2008 Financial Crisis. The x-axis shows the number of months between the cueing month and the historical target month, while the y-axis shows the weights assigned to historical months by our measure. In the upper panel, we present weights constructed in June 2007 (in red), which is a date that is immediately before the outbreak of the 2008 Financial Crisis. In the lower panel, we present weights constructed in December 2008 (in red), which is a date that is during the 2008 Financial Crisis. In both panels, we also display the average weights for our entire sample in green.

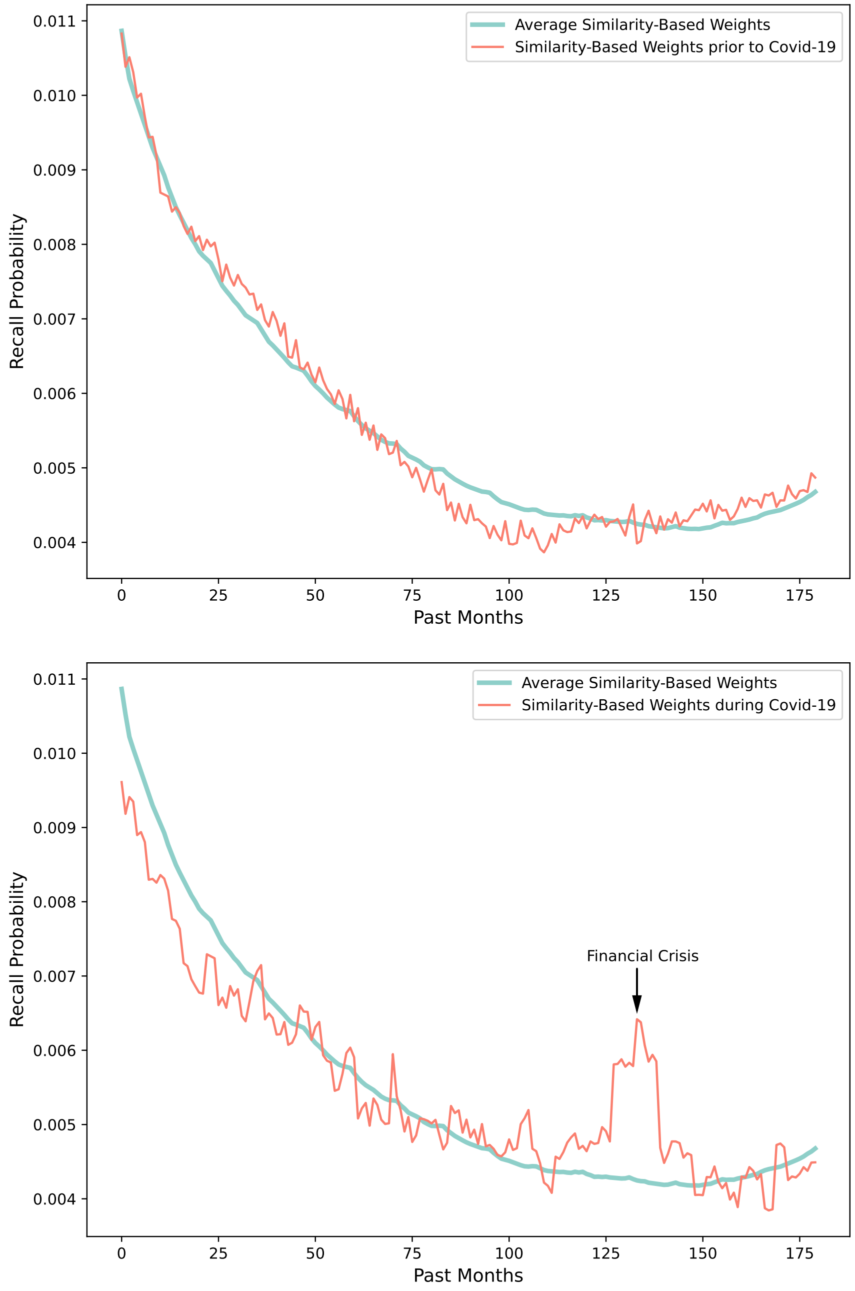
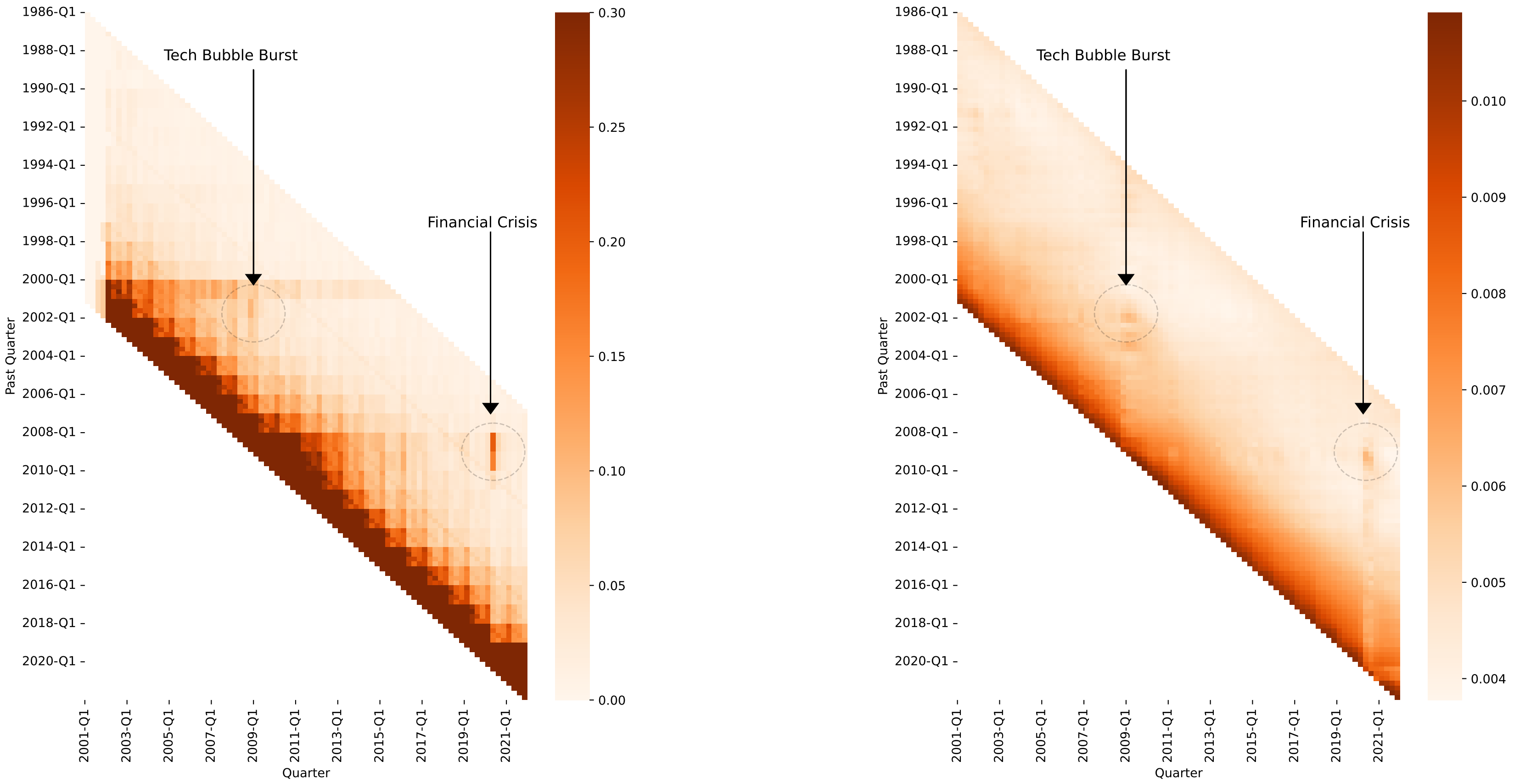


Figure 4: Similarity-Based Weights from our Measure: Before and During the Covid-19 Pandemic

This figure displays the weights constructed using our measure immediately before and during the Covid-19 Pandemic. The x-axis shows the number of months between the cueing month and the historical target month, while the y-axis shows the weights assigned to historical months by our measure. In the upper panel, we present weights constructed in December 2019 (in red), which is a date that is immediately before the outbreak of the 2020 Covid-19 Pandemic. In the lower panel, we present weights constructed in June 2020 (in red), which is a date that is during the 2020 Covid-19 Pandemic. In both panels, we also display the average weights for our entire sample in green.



(a) Panel A: Historical Dates Discussed During Corporate Events

(b) Panel B: Similarity-Based Weights from our Measure

Figure 5: Validation of our Measure

This figure compares the dates that are discussed during corporate events with the similarity-based weights constructed using our measure. Panel A displays the probability that a previous quarter is mentioned during a corporate event. The sample includes corporate events taking place between 2001 and 2021. The x-axis represents the quarter in which the corporate event occurs, while the y-axis represents the probability with which each of the past 60 quarters (= 15 years) is mentioned during the event. A darker shade of red indicates a higher probability, and we require the darkest color to be capped at a probability of 0.3. Panel B replicates Figure 1 for the same sample period (January 2001 to December 2021), showing the similarity-based weights constructed using our measure.

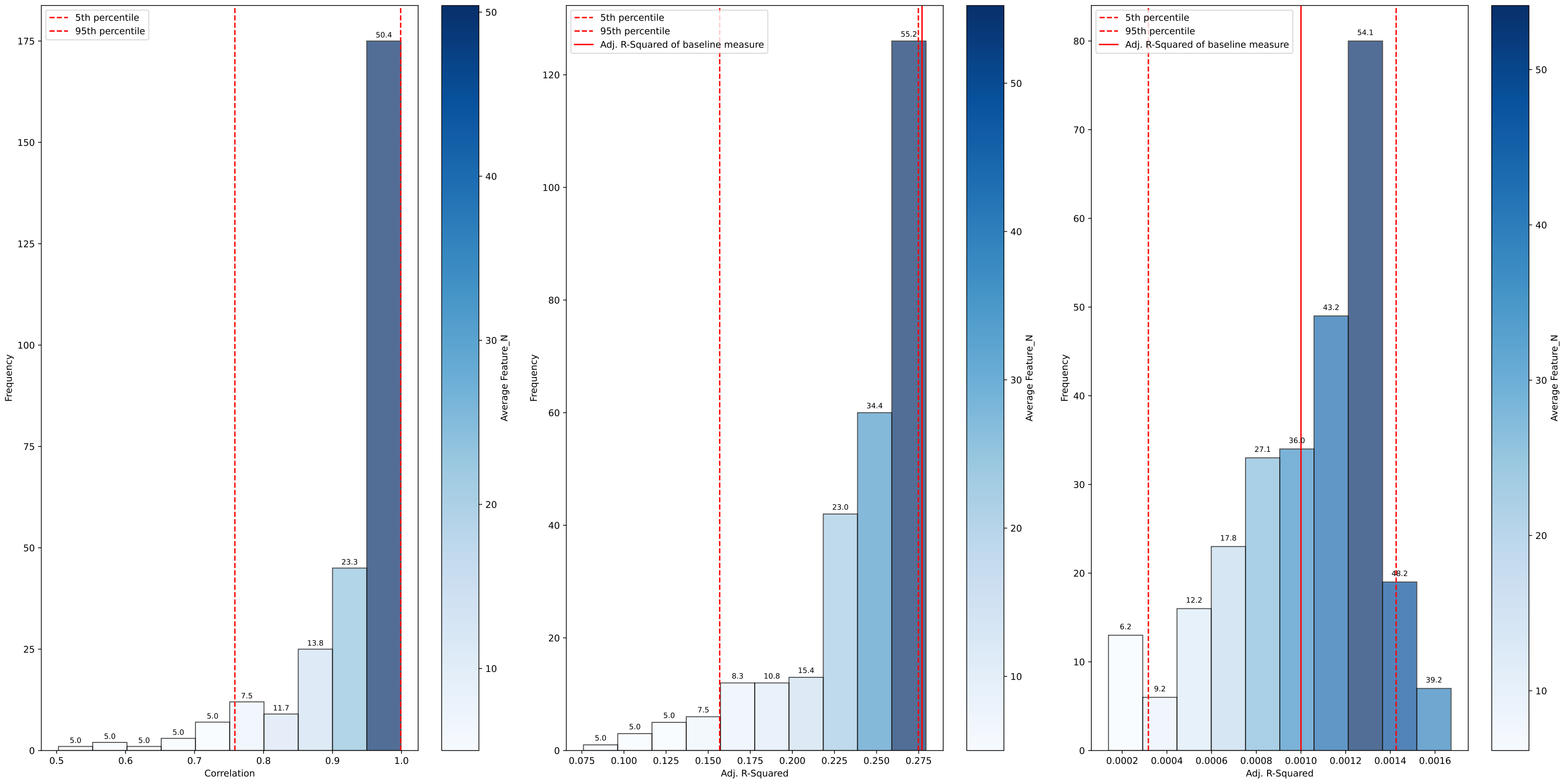
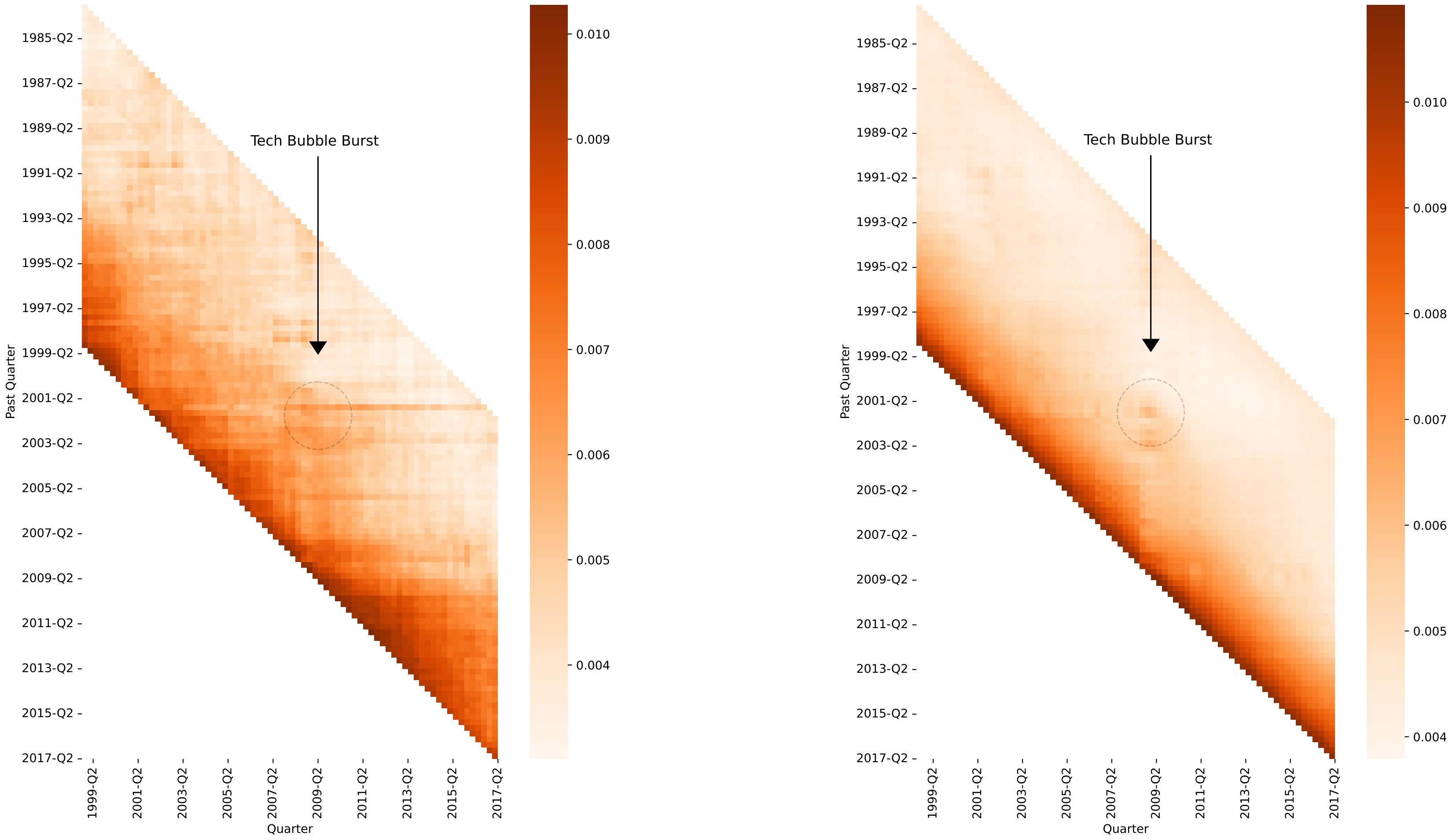


Figure 6: Measures based on Randomly Drawn Features

This figure presents statistics comparing the measures based on randomly drawn features to our baseline measure (which is based on all 72 features). The left panel shows the distribution of correlation coefficients between the randomly drawn measures and our baseline measure. The middle and right panels report the distributions of the adjusted R^2 when we re-estimate the regressions from columns (1) and column (6), respectively, of Table 2 using the randomly drawn measures. The dashed lines in all panels indicate the 5th and 95th percentile. The solid red lines in the middle and right panels indicate the adjusted R^2 of our baseline measure. The number on top of each bar indicates the average number of randomly drawn features, and a darker blue shading indicates a higher number of randomly drawn features.



(a) Panel A: Similarity-Based Weights from the Qualitative Measure

(b) Panel B: Similarity-Based Weights from the Baseline Measure

Figure 7: Similarity-Based Weights from our Measure: Qualitative vs. Quantitative Features

This figure shows similarity-based weights generated by our measure when using only qualitative features (Panel A) and only quantitative features (Panel B). Panel A is constructed using a measure based on the Monthly Topic Attention data released by Bybee et al. (2024). The x-axis indicates the quarter in which hypothetical cue occurs, while the y-axis indicates the weight that the measure assigns to past quarters. A darker shade of red indicates a higher weight. Panel B replicates Figure 1 for the same sample period (January 1999 to June 2017) as Panel A, showing the weights generated by our baseline measure.

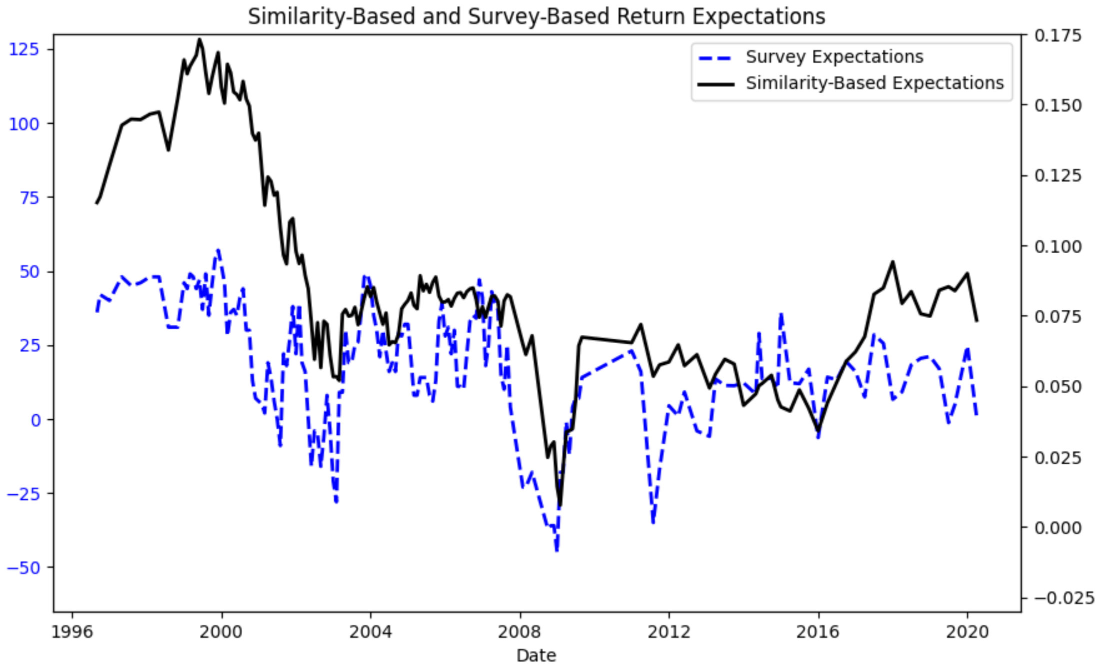


Figure 8: Similarity-Based and Survey-Based Return Expectations

This figure plots similarity-based return expectations (as of the end of month t) and survey-based return expectations (elicited over the course of month $t + 1$). Similarity-based return expectations are annualized return expectations of the S&P 500 index derived from our measure. Survey-based return expectations are the percentage point difference in bullish and bearish investors from the Gallup investor survey. We source survey data from Greenwood and Shleifer (2014) for the period from October 1996 to December 2011, and extend the sample period to May 2020 with data sourced directly from Gallup.

Table 1: Summary Statistics for Corporate Event Data

This table presents summary statistics for the sample used in the tests displayed in Table 2. The sample covers corporate events taking place between January 2001 and December 2021. Month Mentioned is an indicator that equals one if, during a firm l's corporate event in month t, any participant mentions the historical month $t-h$ at least once, and zero otherwise. We focus on the previous 180 months, i.e., $h \in \{1, 2, 3, \dots, 180\}$. The similarity-based weight $\pi_{l,t,t-h}$ captures the stock-specific weight that our measure assigns to month $t-h$ in month t. Exponential Weight is the weight on month $t-h$ derived from a simple model of exponential decay following Greenwood and Shleifer (2014) using $\lambda = 0.56$. Size is the logarithm of a firm's market capitalization (in million $). Turnover is the monthly dollar trading volume over market capitalization at the end of the month. The firm's book-to-market ratio (BM) is constructed following Fama and French (1992), using book equity from (at least) six months ago and market capitalization from the most recent December. Ivol is idiosyncratic volatility (in %) from CAPM regressions and is constructed following Ang et al. (2006), using daily data from the past month. Price is the firm's stock price (in $) at the end of the month.

	N	Mean	Median	Std.Dev	P25	P75	Min	Max
Month Mentioned	15,239,350	0.168	0.000	0.374	0.000	0.000	0.000	1.000
$\pi_{l,t,t-h}$	15,239,350	0.006	0.005	0.002	0.004	0.007	0.001	0.014
Exponential Weight	15,239,350	0.030	0.000	0.124	0.000	0.000	0.000	1.000
Size	15,239,350	7.570	7.581	1.985	6.278	8.906	0.591	14.659
Turnover	15,239,350	0.176	0.133	0.148	0.084	0.216	0.001	0.836
BM	15,239,350	0.601	0.505	0.693	0.293	0.791	-26.782	27.284
Ivol	15,239,350	0.018	0.014	0.015	0.009	0.021	0.001	0.655
Price	15,239,350	44.919	29.600	77.189	14.230	52.960	0.054	3440.160

Table 2: Validation of our Measure using Transcripts of Corporate Events

This table shows that our measure explains which historical dates are discussed during corporate events. The dependent variable is an indicator that equals one if, during a firm l's corporate event in month t, any participant mentions the historical month $t-h$ at least once, and zero otherwise. The similarity-based weight $\pi_{l,t,t-h}$ is the stock-specific weight that our measure assigns to month $t-h$ in month t. Exponential Weight is the weight on month $t-h$ derived from a simple model of exponential decay following Greenwood and Shleifer (2014) using λ = 0.56. The first five columns focus on the previous 180 months, i.e., $h \in \{1, 2, 3, \ldots, 180\}$. The last five columns focus on months that are at least 5 years in the past, i.e., $h \in \{61, 62, 63, \ldots, 180\}$. Columns (2) to (4) and (7) to (9) include *stock* × *past year-month* fixed effects as well as control variables. The control variables are as of the end of month $t-1$. Columns (5) and (10) further include *stock* × *current year-month* fixed effects. These fixed effects soak up the control variables. In all columns except (1) and (6), we also include quarter fixed effects. Standard errors are clustered by stock, current year-month, and past year-month, and the t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

Dependent Variable:	Month Mentioned									
Sample:	Past 15 Years					At Least 5 Years in the Past				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\pi_{l,t,t-h}$	95.400***	92.563***		90.818***	67.818***	5.266***	7.127***		6.912***	3.956***
	(61.19)	(47.61)		(43.90)	(37.08)	(13.94)	(14.93)		(14.05)	(9.26)
Exponential Weight			580.990	262.915	275.649			0.000	0.000	0.000
			(1.62)	(1.49)	(1.55)			(1.58)	(1.55)	(1.59)
Size		-0.021***	-0.075***	-0.021***			-0.005***	-0.006***	-0.005***	
		(-5.75)	(-8.11)	(-5.79)			(-3.57)	(-4.04)	(-3.49)	
Turnover		0.017	0.014	0.018*			0.014***	0.015***	0.014***	
		(1.54)	(0.56)	(1.66)			(2.64)	(2.77)	(2.73)	
BM		-0.002	-0.008*	-0.002			-0.002**	-0.002**	-0.002**	
		(-1.02)	(-1.83)	(-1.02)			(-2.15)	(-2.18)	(-2.14)	
Ivol		-0.681***	-1.592***	-0.692***			-0.051	-0.048	-0.050	
		(-6.05)	(-6.36)	(-6.15)			(-0.78)	(-0.71)	(-0.76)	
Price		-0.000***	-0.001***	-0.000***			-0.000***	-0.000***	-0.000***	
		(-2.81)	(-3.19)	(-2.83)			(-3.28)	(-3.33)	(-3.27)	
Adjusted R-Squared	0.277	0.391	0.260	0.393	0.464	0.001	0.076	0.075	0.076	0.212
N	15,239,350	15,239,350	15,239,350	15,239,350	15,239,350	10,153,585	10,153,585	10,153,585	10,153,585	10,153,585
Stock x Current Year-Month FE	NO	NO	NO	NO	YES	NO	NO	NO	NO	YES
Stock x Past Year-Month FE	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES
Quarter FE	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES

Table 3: The Relative Importance of Different Quantitative Features

This table presents results from tests that analyze the relative importance of different quantitative features. We classify quantitative features into seven mutually exclusive groups: "Activity and Operational Efficiency," "Leverage and Capital Structure," "Liquidity and Solvency," "Market Valuation and Returns," "Stock-Level Variables," "Profitability," and "Macroeconomic Variables". Table IA.2 shows which features are classified into which group. For each group, we construct our measure in two ways: (1) using only the features from the focal group, and (2) using features from all other groups except the focal group. This approach yields a total of 14 variations of our measure, each based on a different set of features. For each variation of our measure, we re-estimate column (1) of Table 2 and report the coefficient on $\pi_{l,t,t-h}$ as well as the adjusted R^2.

	Included		Excluded	
Features:	Coef	Adjusted R^2	Coef	Adjusted R^2
Activity and Operational Efficiency	61.523	0.219	92.119	0.261
Leverage and Capital Structure	50.142	0.178	93.943	0.261
Liquidity and Solvency	52.004	0.163	93.371	0.271
Market Valuation and Returns	53.040	0.137	88.984	0.270
Stock-Level Variables	37.311	0.133	92.335	0.268
Profitability	42.942	0.130	98.002	0.286
Macroeconomic Variables	25.974	0.034	89.412	0.275

Table 4: Summary Statistics for Tests of Similarity-Based Beliefs

This table presents summary statistics for the samples used in the tests displayed in Tables 5 and 6. Panel A displays summary statistics for the sample used in Table 5. The sample period for these tests ranges from October 1996 to May 2020 for our baseline measure and from October 1996 to June 2017 for the measure based on qualitative features. We source Gallup Survey Expectations from Greenwood and Shleifer (2014) until December 2011, and extend the sample until May 2020 with data sourced directly from Gallup. Gallup Survey Expectations are defined as the percentage point difference in bullish and bearish investors from the Gallup investor survey. Similarity-Based Expectation ($\tilde{\mathbb{E}}^{Ann}_{Mkt,t}[\text{ret}]$) is the annualized return expectation for the S&P 500 index derived from our baseline measure, which is based on quantitative features. We also construct a Similarity-Based Expectation using a measure based on qualitative features. We decompose the baseline similarity-based expectation into expectations from recent and distant periods, where recent refers either to the most recent 12 months or the most recent five years, and distant refers either to more than 12 months or more than five years in the past. The sum of recent and distant expectations equals the baseline similarity-based expectation. We also construct this decomposition for the similarity-based expectations based on qualitative features, but for brevity we do not show the decomposition in this table. The cumulative return over the previous 12 months, the exponentially-weighted past return, the logarithm of the price-dividend ratio (Log(P/D)), the risk-free rate, the earnings growth rate, and the unemployment rate are constructed following Greenwood and Shleifer (2014). US Rec is a dummy variable that is equal to one during NBER recessions. Panel B displays summary statistics for the sample used in Table 6. The sample period for these tests ranges from January 1996 to December 2021. Option-implied volatility is constructed following An et al. (2014). Similarity-based volatility ($\tilde{\sigma}^{Ann}_{l,t}$) is the standard deviation of monthly returns over the past 180 months, where each historical return is weighted with its associated similarity-based weight. Exponential-decay-based volatility is the standard deviation of monthly returns over past months, where each historical return is weighted with exponentially-decaying weights, using the decay parameter $\lambda = 0.56$ from Greenwood and Shleifer (2014). Size is the logarithm of a firm's market capitalization (in million \$), idiosyncratic volatility (in %) is constructed following Ang et al. (2006), asset growth is the book asset growth rate (= current book assets – lagged book assets)/lagged book assets), operating profit is constructed following Fama and French (2006), and the book-to-market ratio is constructed following Fama and French (1992).

Panel A: Return Expectations

	N	Mean	Median	Std.Dev	P25	P75	Min	Max
Gallup Survey Expectations	171	18.079	18.303	20.636	7.000	34.000	-45.000	57.000
Similarity-Based Expectation	171	0.091	0.085	0.038	0.065	0.102	0.006	0.172
Similarity-Based Expectation (Qual.)	149	0.089	0.084	0.037	0.062	0.102	0.008	0.170
Exp from Recent Periods (<=12 mths)	171	0.005	0.011	0.021	-0.006	0.018	-0.064	0.038
Exp from Distant Periods (>12 mths)	171	0.086	0.075	0.033	0.065	0.114	0.031	0.150
Exp from Recent Periods (<=5 yrs)	171	0.033	0.031	0.038	0.001	0.059	-0.053	0.110
Exp from Distant Periods (>5 yrs)	171	0.058	0.063	0.026	0.045	0.073	-0.021	0.093
Cumulative Return (past 12 mths)	171	0.067	0.118	0.183	-0.047	0.186	-0.441	0.438
Exponentially-Weighted Past Return	171	0.019	0.028	0.034	0.001	0.040	-0.098	0.071
US Rec	171	0.123	0.000	0.329	0.000	0.000	0.000	1.000
Log(P/D)	171	4.059	4.041	0.235	3.941	4.202	3.324	4.502
Risk-free Rate	171	1.005	1.007	0.019	0.994	1.016	0.950	1.101
Earnings Growth	171	0.042	0.134	0.330	-0.055	0.195	-0.886	0.767
Unemployment	171	5.455	5.000	1.576	4.400	5.800	3.500	14.700

Panel B: Volatility Perceptions

	N	Mean	Median	Std.Dev	P25	P75	Min	Max
Option-Implied Volatility	205,906	0.392	0.349	0.187	0.265	0.467	0.100	1.473
Similarity-Based Volatility	205,906	0.405	0.372	0.155	0.290	0.490	0.160	1.031
Exponential-Decay-Based Volatility	205,906	0.299	0.247	0.204	0.168	0.365	0.022	1.609
Size	205,906	8.033	7.845	1.620	6.835	9.093	3.721	14.813
Idiosyncratic Volatility (3F)	205,906	0.017	0.014	0.011	0.010	0.020	0.000	0.339
Asset Growth	205,906	0.120	0.057	0.461	-0.011	0.151	-0.929	47.725
Operating Profit	205,906	0.367	0.279	3.619	0.180	0.399	-187.667	204.250
log(Book-to-Market Ratio)	205,906	-0.927	-0.862	0.793	-1.337	-0.419	-9.872	3.063

Table 5: Similarity-Based and Survey-Based Return Expectations of the S&P 500 Index

This table shows that similarity-based return expectations explain survey-based return expectations. In all columns, the dependent variable is the difference in the percentage of bullish and bearish investors from the Gallup investor survey (elicited over the course of month $t + 1$). In column (1), the main independent variable is the similarity-based return expectation for the S&P 500 index (as of the end of month t). In column (2), we decompose this similarity-based expectation into expectations from recent periods (most recent 12 months) and from distant periods (more than 12 months in the past). The sum of these two components equals the similarity-based expectation from column (1). In column (3), we add the cumulative return over the past 12 month as an additional control, following Greenwood and Shleifer (2014). In column (4), we control for the exponentially-weighted average return over the past five years. To construct this exponentially-weighted return, we first calculate quarterly returns by compounding 3-month returns on a rolling monthly basis. We then use the weighting approach of Greenwood and Shleifer (2014) and their estimated quarterly λ of 0.77 to calculate an exponentially-weighted average return over the past five years. Columns (5) - (7) mirror columns (2) - (4), except that the cutoff between recent and distant periods is five years instead of 12 months. In column (8), we interact expectations from recent and distant periods with a dummy variable equal to one during NBER recessions. All columns include the same control variables as Panel B of Table 3 in Greenwood and Shleifer (2014). Panel A uses our baseline measure (sample period: October 1996 to May 2020). Panel B uses the measure based on qualitative features (sample period: October 1996 to June 2017). t-statistics, in parentheses, are Newey-West adjusted with twelve lags. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

Panel A: Baseline Measure

	Gallup Survey Expectations							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Similarity-Based Expectation	393.708***							
	(5.91)							
Exp from Recent Periods (<=12 mths)		685.113***	569.203**	328.357**				
		(10.52)	(2.47)	(2.08)				
Exp from Distant Periods (>12 mths)		229.203***	228.124***	196.305***				
		(4.49)	(4.48)	(3.58)				
Exp from Recent Periods (<=5 yrs)					406.455***	202.179***	121.518*	148.440*
					(6.28)	(3.70)	(1.73)	(1.95)
Exp from Distant Periods (>5 yrs)					336.970***	271.280***	237.670***	236.250***
					(4.46)	(4.06)	(3.40)	(3.96)
Exp from Recent Periods (<=5 yrs) X US Rec								-151.748*
								(-1.89)
Exp from Distant Periods (>5 yrs) X US Rec								1243.513***
								(5.00)
US Rec								-71.300***
								(-4.47)
Cumulative Return (past 12 mths)			14.676			61.741***		20.032
			(0.53)			(5.39)		(1.38)
Exponentially-Weighted Past Return				209.842**			351.544***	265.144***
				(2.30)			(6.49)	(3.49)
Log(P/D)	-10.587	4.540	5.187	6.809	-9.413	6.997	9.253	8.102
	(-0.73)	(0.43)	(0.49)	(0.60)	(-0.65)	(0.72)	(0.87)	(0.75)
Risk-free Rate	-159.299***	-113.846**	-112.010**	-101.615**	-158.009***	-104.694**	-86.874*	-77.445*
	(-3.19)	(-2.39)	(-2.37)	(-2.08)	(-3.18)	(-2.29)	(-1.79)	(-1.68)
Earnings Growth	19.761***	0.152	-1.067	1.622	20.515***	-6.895	-2.372	-4.010
	(2.93)	(0.02)	(-0.14)	(0.22)	(2.81)	(-0.85)	(-0.34)	(-0.47)
Unemployment	0.297	-0.337	-0.296	-0.349	0.507	-0.411	-0.829	-0.325
	(0.26)	(-0.29)	(-0.26)	(-0.29)	(0.41)	(-0.38)	(-0.67)	(-0.20)
Adjusted R-Squared	0.580	0.657	0.655	0.665	0.584	0.655	0.676	0.720
N	171	171	171	171	171	171	171	171

Panel B: Measure Based on Qualitative Features

	Gallup Survey Expectations							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Similarity-Based Expectation (Qual.)	360.142***							
	(3.55)							
Exp from Recent Periods (<=12 mths)		799.303***	532.065**	224.772				
		(7.20)	(2.41)	(0.84)				
Exp from Distant Periods (>12 mths)		256.447***	246.268***	188.839**				
		(3.53)	(3.77)	(2.57)				
Exp from Recent Periods (<=5 yrs)					373.870***	214.248***	125.121	30.940
					(4.27)	(3.41)	(1.53)	(0.36)
Exp from Distant Periods (>5 yrs)					290.980***	278.868***	237.726**	138.377
					(2.99)	(2.90)	(2.44)	(1.46)
Exp from Recent Periods (<=5 yrs) X US Rec								-320.062***
								(-3.59)
Exp from Distant Periods (>5 yrs) X US Rec								1114.367***
								(7.20)
US Rec								-61.875***
								(-5.74)
Cumulative Return (past 12 mths)			29.513			67.303***		12.938
			(1.22)			(5.89)		(0.78)
Exponentially-Weighted Past Return				280.913**			369.628***	370.104***
				(2.23)			(6.36)	(4.49)
Log(P/D)	-4.388	-0.213	3.032	6.400	-3.518	6.843	8.261	19.651
	(-0.24)	(-0.02)	(0.25)	(0.46)	(-0.19)	(0.58)	(0.63)	(1.32)
Risk-free Rate	-206.673***	-145.090***	-143.424***	-137.413***	-205.597***	-137.209***	-122.365**	-113.761**
	(-3.64)	(-2.83)	(-2.81)	(-2.63)	(-3.76)	(-2.66)	(-2.33)	(-2.38)
Earnings Growth	22.096***	0.513	-2.008	2.899	23.135***	-7.771	-2.695	-7.767
	(3.46)	(0.07)	(-0.27)	(0.35)	(3.23)	(-1.00)	(-0.36)	(-0.89)
Unemployment	1.264	0.124	0.231	-0.240	1.426	0.151	-0.836	-2.110
	(0.60)	(0.06)	(0.12)	(-0.12)	(0.66)	(0.08)	(-0.43)	(-1.24)
Adjusted R-Squared	0.549	0.649	0.649	0.666	0.558	0.651	0.680	0.739
N	149	149	149	149	149	149	149	149

Table 6: Similarity-Based Volatility and Option-Implied Volatility

This table shows that similarity-based volatility explains option-implied volatility in the cross-section of stocks. In all columns, the dependent variable is option-implied volatility in month $t + 1$, constructed following An et al. (2014). All independent variables are as of the end of month t. Similarity-based volatility is the standard deviation of monthly returns over the past 180 months, where each historical return is weighted with its associated similarity-based weight. Exponential-decay-based volatility is the standard deviation of monthly returns over past months, where each historical return is weighted with exponentially-decaying weights, using the decay parameter λ = 0.56 from Greenwood and Shleifer (2014). In column (1), we only include similarity-based volatility as an independent variable. In column (2), we control for exponential-decay-based volatility. In column (3), we add stock and year-month fixed effects. In column (4) we control for lagged option-implied volatility, and in column (5) we control for size, idiosyncratic volatility (following Ang et al. (2006)), asset growth, operating profit (following Fama and French (2006)), and the logarithm of the book-to-market ratio (following Fama and French (1992)). We multiply all coefficients by 100. Standard errors are clustered by stock and month, and the t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

	Option-Implied Volatility				
	(1)	(2)	(3)	(4)	(5)
Similarity-Based Volatility	64.418***	38.500***	36.981***	12.963***	7.689***
	(34.47)	(22.69)	(13.45)	(9.96)	(6.79)
Exponential-Decay-Based Volatility		36.722***	18.328***	6.270***	5.258***
		(22.77)	(23.51)	(15.90)	(14.26)
Option-Implied Volatility (lagged)				61.511***	55.787***
				(52.02)	(47.45)
Size					-2.259***
					(-13.01)
Idiosyncratic Volatility (3F)					129.236***
					(19.65)
Asset Growth					0.276**
					(2.01)
Operating Profit					0.005
					(0.72)
log(Book-to-Market Ratio)					-0.075
					(-0.82)
Adjusted R-Squared	0.285	0.399	0.695	0.804	0.810
N	205,906	205,906	205,906	205,906	205,906
Month FE	NO	NO	YES	YES	YES
Stock FE	NO	NO	YES	YES	YES

INTERNET APPENDIX

A Construction of the Corporate Event Sample

In this section, we describe how we construct the sample that we use in our tests in Section 3.3 of the paper. We collect transcripts of corporate events from Refinitiv StreetEvents for January 2001 to December 2021. The set of corporate events covered by these transcripts includes Earnings Calls, M&A Calls, Sales Calls, Analyst Meetings as well as Corporate Conference Presentations. Q&A sessions of these calls are also included in the transcripts. Each transcript of an event provides a verbatim representation of what was spoken and by whom during the event, along with metadata that allows us to match the company to the Compustat database. This metadata includes the ticker symbol header, company name, event title, and date of the event. We match company names in the transcripts to the corresponding 'GVKEY' (the unique Compustat identifier) following Li et al. (2021). We also match the sample to CRSP and retain only stocks with a share code of 10 or 11 and an exchange code of 1, 2, or 3.

We parse the text of each transcript and segment it into sentences. For each sentence, we identify the dates mentioned in it using *sutime*, a Python wrapper for Stanford CoreNLP's SUTime Java library.[34] *sutime* allows us to identify date formats in both absolute and relative forms. Absolute date formats are phrases that directly represent a specific date, such as "2011-03-31" or "2011-Q3". When dealing with absolute dates, *sutime* can interpret date strings in various formats; it recognizes that strings like "2011-03-31" and "March 31, 2011" are equivalent. In contrast, relative date formats involve phrases that indirectly convey a date, such as "two weeks ago". For instance, if a conference call occurs on December 31, 2011, and the phrase "two weeks ago" is mentioned, *sutime* codes the corresponding date as December 17, 2011. *sutime* labels both absolute and relative date formats as belonging to the category *date*.

There are other forms of date-related terms that *sutime* can detect. For instance, in a sentence containing the phrase "century long commitment", *sutime* would classify the term "century long" as belonging to the category *duration*. We disregard sentences with terms falling into this category. In our sample, we focus only on dates from sentences with phrases that are grouped into the *date* category.

[34] Link: https://pypi.org/project/sutime/

The identified dates mainly fall into three different frequencies: days (e.g., 2011-03-31), quarters (e.g., 2011-Q3), and years (e.g., 2011). To ensure consistency with the monthly frequency of our measure, we aggregate dates to the monthly level. Specifically, we assume that when referring to a date of lower frequency, all corresponding months within that date are included. For instance, if a conference participant mentions the year 2011, we treat it as if every month in 2011 was mentioned. Moreover, since our focus is on similarity with past periods, we only focus on dates that occurred before the date of the corporate event, and ignore references to the future.

B Variation in the Decay Speed

Our goal in this section is to show that our measure can generate patterns that are typically observed in subjective beliefs data. Specifically, our measure generates weights that are largely consistent with empirical patterns on return extrapolation documented in Da et al. (2021). In this study, the authors show that investors extrapolate from past returns, placing higher weights on recent returns compared to distant returns. Moreover, these extrapolative weights decay faster for small firms and value stocks.

Here, we show that our measure generates variation in the decay speed in the cross-section of stocks that is consistent with the findings in Da et al. (2021). We construct our measure of decay speed for stock l in month t as follows:

$$\text{Decay Speed}_{l,t} = \frac{\pi_{l,t,t}}{\pi_{l,t,t-12}} \tag{IA.1}$$

We show how decay speed varies along four firm-level characteristics, which are each measured as of the end of each month: the logarithm of a firm's market capitalization (in million \$), the book-to-market ratio following Fama and French (1992), the idiosyncratic volatility (in %) from CAPM regressions following Ang et al. (2006), and the stock price (in \$). At the end of each month, we sort stocks into deciles based on these characteristics and calculate both the equal-weighted and value-weighted average of decay speed.

Table IA.4 presents our findings. The first row of Panel A shows that, when weighting stocks equally in each portfolio, the decay speed of the smallest firms is 10.7% (= 1.34/1.21 - 1) faster than that of the largest firms. We find differences of a similar magnitude when we compare stocks in the extreme deciles of idiosyncratic volatility and stock price, as shown in the third and fourth rows, with weights decaying faster for volatile and low-priced stocks. We also find that decay speed is lower for growth stocks, but the magnitude of this difference is smaller. Panel B shows similar results when we value-weight stocks in each portfolio.

Overall, the consistency between our findings and those of Da et al. (2021) suggests a link between similarity-based weighting and return extrapolation. We view the results of these tests as suggestive evidence that our measure is linked with subjective beliefs data from the field. In Section 4.2 of the paper, we empirically establish the connection between similarity-based and survey-based return expectations.

C Linking Similarity-Based Volatility and the VIX

In this section, we use our measure to construct similarity-based volatility for the aggregate market and link it to the actual perceived volatility captured by the Volatility Index (VIX). We construct similarity-based volatility for the aggregate market as the standard deviation of monthly S&P 500 index returns over the past 180 months, where each historical return is weighted with its associated similarity-based weight. Specifically, we construct similarity-based volatility for the aggregate market as follows:

$$\tilde{\sigma}_{Mkt,t} = \sqrt{\sum_{h=1}^{180} \bar{\pi}_{t,t-h} \left(\text{ret}_{Mkt,t-h+1} - \tilde{\mathbb{E}}_{Mkt,t}[\text{ret}_{t+1}]\right)^2} \tag{IA.2}$$

where $\bar{\pi}_{t,t-h}$ is constructed using Equation (4) and $\tilde{\mathbb{E}}_{Mkt,t}[\text{ret}_{t+1}]$ is constructed using Equation (5).

We proxy for actual perceived volatility at the market-level using the VIX. This is a natural choice, as the VIX is designed to capture the market's annualized expectation of volatility over the next 30 days. We download the VIX from the website of the Federal Reserve Bank of St. Louis, calculate the average of daily VIX over the course of each month, and normalize it by dividing it by 100. We provide summary statistics of the VIX and similarity-based volatility in Panel B of Table IA.8.

We begin by visualizing the relationship between similarity-based volatility and the VIX in Figure IA.2. The sample period ranges from January 1991 to December 2021. The dashed blue line represents the average daily VIX over the course of month $t+1$, while the solid black line represents similarity-based volatility as of the end of month t. The two time series track each other well, with the VIX being more volatile over the sample period. The correlation between the two time series is 0.26 (the rank correlation is 0.35).

Next, we regress the VIX (from month $t + 1$) on similarity-based volatility (as of the end of month t) and present the results in Table IA.9. Column (1) shows that similarity-based volatility explains the VIX. In terms of magnitude, a one-standard-deviation increase in similarity-based volatility (0.011) is associated with a 0.020-unit increase in the VIX, which corresponds to about 26% of its standard deviation. In column (2), we control for the same control variables as in column (1) of Table 5 as well as the realized volatility over the past 180 months. With these controls, the coefficient on similarity-based volatility more than triples. The reason for this sharp increase is that similarity-based and realized volatility are strongly correlated. Since the coefficient on realized volatility is negative and significant, economically, these results imply that the VIX is high when similarity-based volatility exceeds what was actually realized.

D Similarity-Driven Repurchasing Decisions

In this section, we use our measure to revisit previously documented evidence that past trading experiences affect investors' future repurchasing decisions (Kaustia and Knüpfer (2008); Choi et al. (2009); Strahilevitz et al. (2011)). We show that the similarity-based weight assigned to a past trading experience significantly influences an investor's likelihood of repurchasing a stock, highlighting the role of similarity not only for beliefs, but also for actions. Further, as we discuss in more detail below, the trading setting allows us to construct similarity-based probability distributions for each investor individually, thus allowing us to test the role of similarity not only in the cross-section of stocks, but also in the cross-section of investors.

D.1 Empirical Design

Consider an investor who debates the question *"Should I repurchase stock l?"*. Strahilevitz et al. (2011) find that the answer to this question depends on how well stock l previously performed for the investor. In particular, if the stock was previously sold for a gain, the investor is more likely to repurchase the stock. We propose that similarity is the governing principle driving this effect. Specifically, we hypothesize that when an investor debates the above question, it brings to mind current stock-level and contextual features, which lead the investor to draw on past experiences. Suppose that the investor previously sold stock l for a gain. We hypothesize that if the investor is reminded of this gain – either by passively recalling it or by actively looking it up – she is more likely to repurchase stock l today. Crucially, this implies that identical past realized returns differentially affect the probability of repurchase, depending on their similarity-based weight in the decision process. This is a key prediction that we will take to the data.

We chose this setting because it is a particularly clean setting to test for similarity-driven effects in trading. First, by analyzing repurchasing decisions, the stocks we focus on are not part of the investor's current portfolio. This helps attenuate concerns that the effects we are documenting are driven by passive attention spillover effects, as the investor does not see any of these stocks in her current portfolio statement. Second, this setting allows us to sidestep several well-known trading patterns, such as the disposition effect (Shefrin and Statman (1985); Odean (1998)) and the rank effect (Hartzmark (2015)). These patterns apply to stocks currently held by the investor and primarily concern selling decisions, not repurchasing behavior.

We test our hypothesis using data from a large US discount brokerage. These are the same data as in Barber and Odean (2000). The data include the holdings and trades of approximately 78,000 households between January 1991 and November 1996. We retain only common stocks, drop trades with negative commissions, and match the data to CRSP for information on stock prices. Since our measure is constructed monthly, we aggregate each investor's holdings and trades to the monthly level. This means that we pool all trades (buys, sells, liquidations, and repurchases) that the investor executes in a given month. Since we are interested in analyzing investors' repurchasing decisions, our sample consists of all stocks that an investor can repurchase in a given month. These are stocks that an investor once held and subsequently liquidated, but that the investor has not (yet) repurchased since the latest liquidation. Once an investor repurchases a stock, we exclude that stock from the sample going forward, until the investor potentially liquidates the position again. For each position of stock l in account i, we use the past trading history to calculate the weighted average purchase price (WAPP). When an investor fully liquidates a position in a month, we use the WAPP and the average transacted price in the liquidation month to calculate the realized return $Ret_{i,l,t-h}$ (winsorized at the 1st and 99th percentiles). We only consider positions for which we know all historical purchase prices, so that we can accurately calculate the WAPP. To this end, we exclude positions that investors held in the first month of the position files, since we do not know at which price the investor acquired these positions.

In our tests, we estimate the following regression:

$$\text{Buy}_{i,l,t} = a\,\text{Ret}_{i,l,t-h} + b\,\pi_{i,l,t,t-h} + c\,\text{Ret}_{i,l,t-h} \times \pi_{i,l,t,t-h} + \text{Controls} + \text{FEs} + \epsilon_{i,l,t} \qquad \text{(IA.3)}$$

The dependent variable in this regression is a dummy variable, $\text{Buy}_{i,l,t}$, which is equal to one if investor i repurchases stock l in month t. The variable $Ret_{i,l,t-h}$ is the return that investor i realized when liquidating stock l in month $t-h$. Finally, the variable $\pi_{i,l,t,t-h}$ is the similarity-based weight assigned to the stock-month in which the past return of stock l was realized. This weight is unique to investor i.[35] Standard errors in this regression are clustered by account, liquidation month, and current month.

To construct similarity weights at the investor-level, we assume that a stock enters an investor's database, E_i, when the investor i first acquires the stock (the "entering month"). We also retain the assumption of narrow framing at the stock-level. We use Equation (2) to calculate investor-specific similarity, $S_i(e_{l,t-h}, \kappa_{l,t})$, which is the cosine similarity between the vector of features from the current stock-month and the vector of features from the stock-month in which the investor realized the return $Ret_{i,l,t-h}$. We also construct investor-specific interference (or normalization) as $I_i(\kappa_{l,t}) = \sum_{e'_l \in E_i} S_i(e'_l, \kappa_{l,t})$, that is, the sum of all $S_i(e'_l, \kappa_{l,t})$ across months in which stock l is in investor i's database E_i, starting with the entering month and ending with month $t-1$. We then apply Equation (3) at the investor-level to construct $\pi_{i,l,t,t-h}$:

$$\pi_{i,l,t,t-h} \equiv \frac{S_i(e_{l,t-h}, \kappa_{l,t})}{I_i(\kappa_{l,t})} \qquad \text{(IA.4)}$$

Intuitively, $\pi_{i,l,t,t-h}$ is the probability that investor i draws on the stock-month in which $Ret_{i,l,t-h}$ was realized when cued in month t. This investor-specific probability can be interpreted analogously to the weight constructed for a representative investor in previous tests. The key difference is that the investor-specific weight is based on the database E_i, which is unique to investor i.

The coefficients in the above regression have intuitive interpretations. Coefficient a captures the effect of past realized returns on the probability of repurchase if the similarity weight implied by our measure is zero. Coefficient b captures the effect of the similarity weight on the likelihood of repurchase if the past realized return is zero. Finally, coefficient c captures the additional effect of the similarity weight on the likelihood of repurchase for a given (non-zero) past realized return.

[35] The granular nature of the data allows us to construct investor-specific similarity weights for these tests. However, we find similar results when we use the coarser weights of a representative investor, which we have used in our tests so far.

Across specifications, we include different sets of fixed effects, which we discuss in more detail below. We also control for the return of stock l between the liquidation month and the end of month $t-1$, since Strahilevitz et al. (2011) show that this return, which the investor could have hypothetically realized had she not sold the stock, affects the probability of repurchase. In other words, this allows us to control for the role of regret. Moreover, following Ben-David and Hirshleifer (2012), we control for the logarithm of the initial purchase price (WAPP), the square root of the number of days between initial purchase and liquidation, and the stock volatility calculated using daily returns over the 250 days preceding the initial purchase.

D.2 Results

Before showing the results from estimating Equation (IA.3), we briefly discuss the summary statistics of our sample, presented in Table IA.10. On average, there are about 6 stocks that an investor can repurchase in a given month.[36] These are stocks that the investor previously held and subsequently liquidated. Investors realized an average gain of 6.60% when liquidating previously-held positions. The unconditional probability of repurchasing a previously-held stock is low at only 0.5%, but as we show next, the probability of drawing on previous gains and losses strongly affects the likelihood of repurchase.

Table IA.11 presents our regression results. We first replicate the finding of Strahilevitz et al. (2011) using a simple linear probability model.[37] We regress the repurchase dummy $\text{Buy}_{i,l,t}$ on the past realized return, $Ret_{i,l,t-h}$, and the above control variables. We also include *account* × *year-month* and *stock* × *year-month* fixed effects. The first set of fixed effects controls for characteristics of investor i in month t that might be driving the propensity to repurchase, such as investor wealth, sophistication, or the investor's optimism/pessimism about the current market environment. The second set of fixed effects controls for stock-level information revealed in month t that might be driving repurchasing behavior of all investors.

In column (1), the coefficient on past realized returns is positive and significant. In terms of magnitude, this coefficient implies that a realized return of 6.60% (the average in our sample) increases the likelihood of repurchase by 0.025 percentage points (pp). This effect, while statistically significant, is small – even

[36] For comparison, investors hold about 2 stocks on average in their portfolio in a given month.

[37] Strahilevitz et al. (2011) use both ratio analysis and a hazard model to document their finding.

compared to the relatively low unconditional probability of repurchase in our sample of 0.5%. However, as we show in the remaining columns of Table IA.11, this average effect masks strong heterogeneity with respect to the similarity-based weight assigned to the stock-month in which the past return of stock l was realized.

In column (2), we augment the regression with the similarity-based weight, $\pi_{i,l,t,t-h}$, and its interaction with the previously realized return, $Ret_{i,l,t-h}$. The coefficients on the similarity-based weight as well as on the interaction term are positive, economically meaningful, and highly significant. Consider first the coefficient on the similarity-based weight. This coefficient implies that a 10% weight on the month in which the investor liquidated the stock is associated with a 0.355 pp increase in the likelihood of repurchasing the stock (for a realized return of zero). This effect size is roughly 71% of the unconditional probability of repurchase. Further, the coefficient on the interaction term implies that for the same weight of 10%, if the investor previously realized a return of 6.60%, the repurchase probability increases by an additional 0.084 pp. Compared to the unconditional probability of repurchase (0.5%), this additional effect represents about a 17% increase.[38]

In column (3), we add $S_i(e_{l,t-h}, \kappa_{l,t})$ and $I_i(\kappa_{l,t})$ as separate regressors, and interact each with the previously realized return, $Ret_{i,l,t-h}$. Including similarity and interference separately as independent variables allows for a targeted test of the mechanics of our framework. Specifically, since these two forces have opposing effects on the final weight – similarity increases the weight, while interference (the sum of all similarities) reduces it – the interaction of $Ret_{i,l,t-h}$ with similarity should have a positive coefficient and the interaction of $Ret_{i,l,t-h}$ with interference should have a negative coefficient. This is precisely what we find. In terms of magnitude, for a previously realized return of 6.60%, a one standard deviation increase in similarity (one std. dev. = 0.195) implies a 0.01 pp increase in the probability of repurchase. Conversely, a one standard deviation increase in interference (one std. dev. = 10.597) implies a 0.018 pp decrease in the

[38] One potential concern with this specification is that the set of features that we use to construct the similarity-based weights include the stock's monthly return as well as other variables that are driven by monthly returns, like the book-to-market ratio. Using these features to construct the weights might lead to mechanical effects. To address this concern, in Table IA.14 we exclude all return-related variables when constructing similarity-based weights and replicate columns (2) through (5) of Table IA.11. We find that the results are virtually identical.

probability of repurchase.

Finally, in columns (4) and (5), we further tighten up these regressions by including *account* × *current year-month* × *liquidation year-month* fixed effects. Intuitively, including these fixed effects allows us to compare an investor's repurchasing decisions for stocks that were previously liquidated in the same month. While these stocks have been out of the investor's portfolio for the same amount of time, they generally differ in their previously realized returns $Ret_{i,l,t-h}$, as well as their similarity, $S_i(e_{l,t-h}, \kappa_{l,t})$, and interference, $I_i(\kappa_{l,t})$, resulting in different weights, $\pi_{i,l,t,t-h}$.[39] Including these very tight fixed effects does not change the coefficients on the interaction terms.[40]

In Table IA.12, we interact the previously realized return $Ret_{i,l,t-h}$ not only with the investor-specific similarity weight, but also with the similarity weight of a representative investor. This allows us to test whether the investor-specific weights, which are based on an investor's unique past experiences, dominate the generic weights of a representative investor. In column (1), we show that the interaction term with the representative-investor weights is positive and significant. However, once we add the interaction with the investor-specific weights in column (2), these investor-specific weights dominate and wash out the coefficient on the interaction term with the representative-investor weights. This also holds true when we include *account* × *current year-month* × *liquidation year-month* fixed effects in columns (3) and (4). These results suggest that investors do not optimally use the entire historical data. Rather, they rely more strongly on similarity to personal experiences (captured by investor-specific weights) than on similarity to the entire history (captured by the representative-investor weights).

In the tests in Table IA.11, we focus on the weighting of the realization month $t - h$, i.e., the month in which a previous gain or loss was realized. We focus on the realization month because previous work on investor behavior strongly suggests that the realization of a gain or loss has a particularly strong grip on investors' minds (Odean (1998); Frydman et al. (2014)). However, evaluating a stock for a potential repurchase, a sophisticated investor might focus on how similar today's environment is to the month in which

[39] These fixed effects also rule out that we are merely picking up a horizon effect, where investors simply repurchase recently-liquidated stocks.

[40] As yet another variation of the tests in this section, we use realized returns per month to capture the speed with which a previous gain or loss was realized. We find very similar results and present them in Table IA.16.

she previously *opened* a position that turned out to be a gain or loss. Therefore, in Table IA.13, we repeat the analysis using the weight assigned to the month $t - o$ in which a previous position was opened, $\pi_{i,l,t,t-o}$, and interact it with the previously realized return, $Ret_{i,l,t-h}$. We find that the interaction of the previously realized return with the weight on the opening month is positive and significant. In columns (2) and (4), we run a horse race between the weights on the opening and the realization month. While both interaction terms are positive, only the effect for the realization month remains significant in our preferred specification in column (4).

D.2.1 The Role of Encoding Strength for Similarity-Driven Repurchases

In this section, we examine whether the encoding strength of an experience matters for the effects documented in Table IA.11. We hypothesize that, conditional on all else being equal, an investor is more likely to draw on a strongly encoded experience.

To capture the encoding strength of a past trading experience, we use the attention that the investor devoted to her portfolio in the month of the experience. We proxy for this attention using the total number of transactions that the investor executed in the month of the experience. Specifically, we construct a dummy variable that is equal to one if the investor executed at least six transactions in the month of the experience, and zero otherwise.[41]

In our tests, we augment Equation (IA.3) with a triple interaction term between the attention dummy, the past realized return, $Ret_{i,l,t-h}$, and the investor-level weight, $\pi_{i,l,t,t-h}$. If the intensity of encoding matters, the coefficient on this triple interaction term should be positive. Intuitively, a positive coefficient on the triple interaction implies that investors draw more heavily on experiences that were encoded in months with high attention, and therefore have a stronger effect on the likelihood of repurchase. Notably, the direction of this effect depends on the sign of the previously realized return: a positive realized return increases the likelihood of repurchase, while a negative realized return decreases it.

Table IA.15 presents the results. In both columns, we find that the coefficient on the triple interaction term is positive and significant. Consider the estimates in column (1), which includes *account* × *year-month*

[41] The 75th percentile of transactions per month is six.

and *stock* × *year-month* fixed effects. The coefficient on the triple interaction term implies that for a realized return of 6.60% and a similarity-based weight of 10%, the likelihood of repurchase is about 0.032 pp higher if the past return was realized in a month in which the investor paid a lot of attention to her portfolio. In column (2), in which we include extremely tight *account* × *current year-month* × *liquidation year-month* fixed effects, the effect size is very similar. Overall, the results in this table show that even for past experiences that have identical realized returns and that have equal similarity-based weights, those that are encoded more strongly have a stronger effect on the likelihood of repurchase.

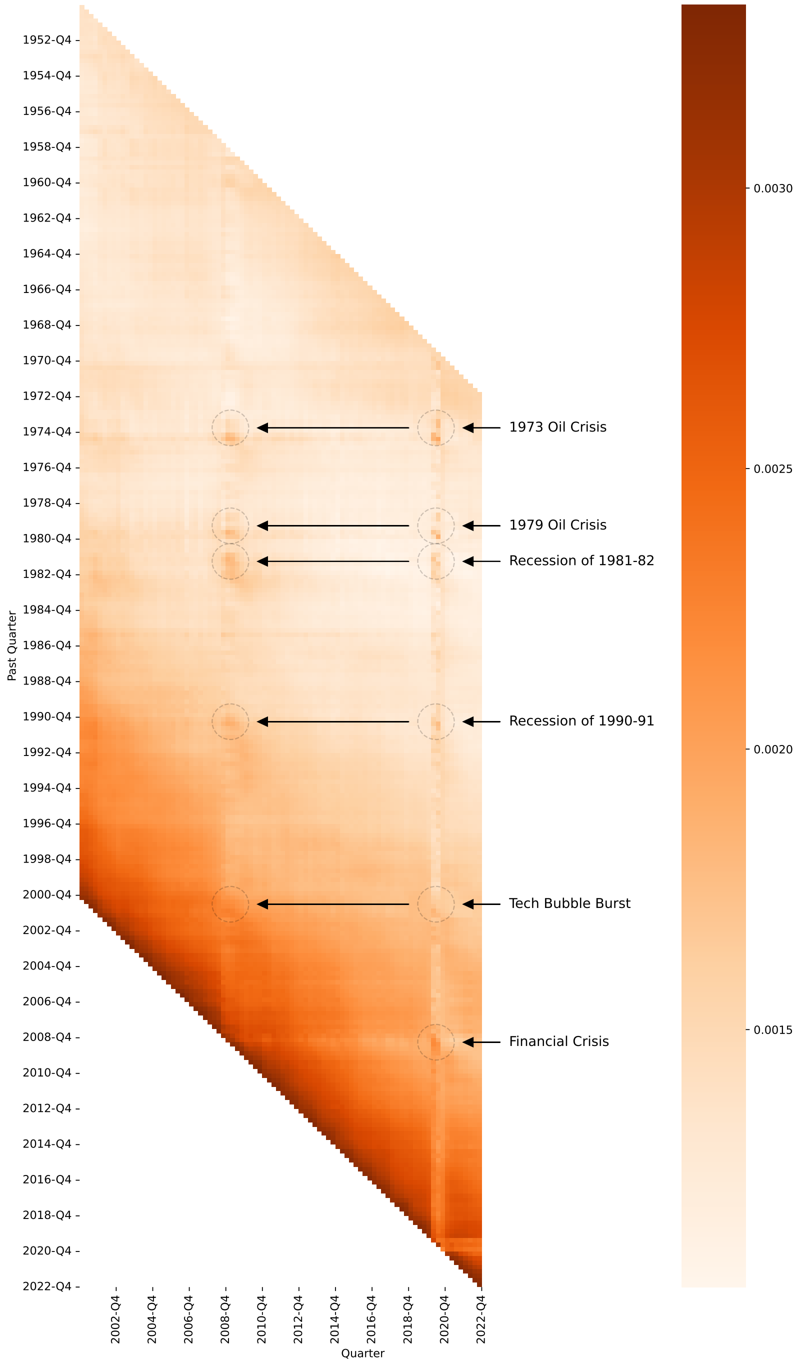


Figure IA.1: Similarity-Based Weights from our Measure: Going Back 50 Years

This figure displays the weights that our measure assigns to each of the past 200 quarters (= 50 years) for each quarter from 2001 to 2021. The x-axis indicates the quarter in which the hypothetical cue occurs, while the y-axis indicates the weight that our measure assigns to past quarters. A darker shade of red indicates a higher weight.

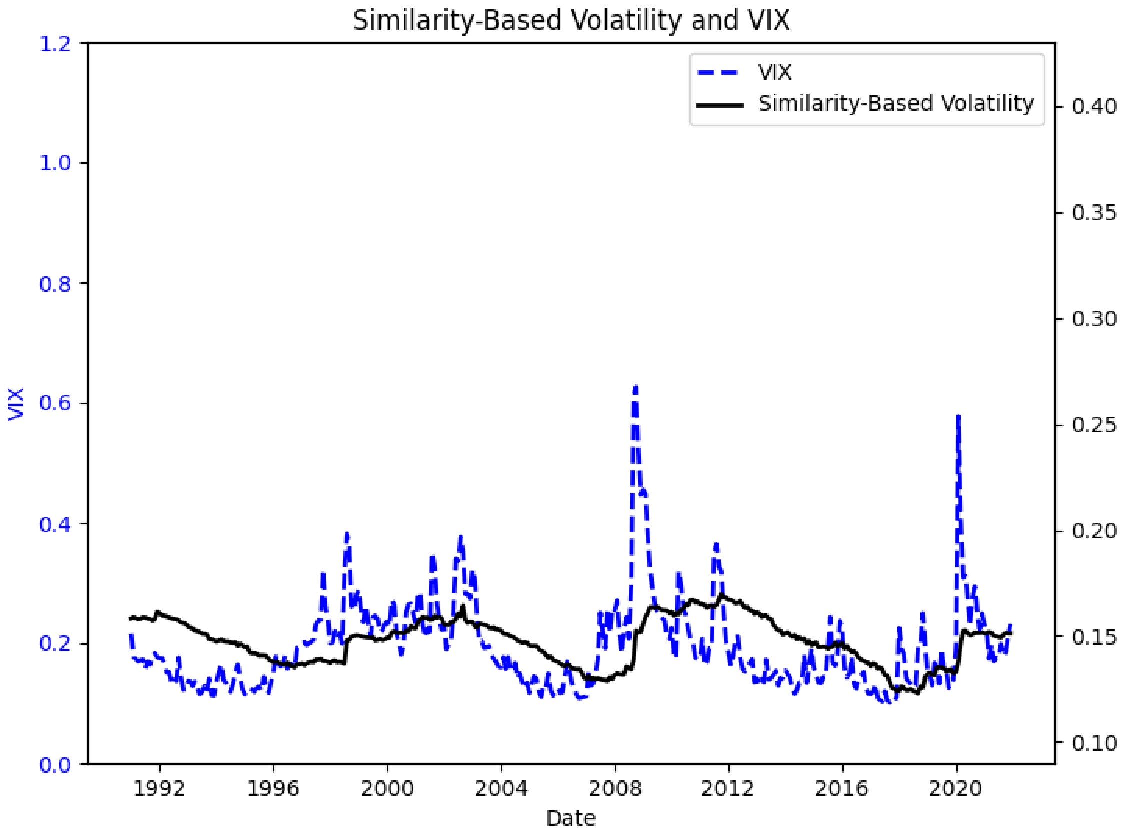


Figure IA.2: Similarity-Based Volatility and the VIX

This figure plots similarity-based volatility (as of the end of month t) and the average daily VIX over the course of month $t + 1$. VIX is divided by 100. Similarity-Based Volatility is the annualized volatility of the S&P 500 index derived from our measure.

Table IA.1: Set of Features

Part 1. Macroeconomic Variables			
con_g	Log Difference of Consumption in Goods and Services	IPT_g	Log Difference of Industrial Production Index
GDP_g	Log Difference of Real GDP	unemployment	Unemployment Rate
INFLATION	Inflation Rate		
Part 2. Stock-Level Variables			
return	Monthly Stock Return	PRICE	Stock Price
DOLLARVOL	Dollar Volume	VOL	Trading Volume
Part 3. Firm-Level Variables			
Accrual	Accruals/Average Assets	adv_sale	Advertising Expenses/Sales
aftret eq	After-tax Return on Average Common Equity	aftret equity	After-tax Return on Total Stockholders Equity
aftret invcapx	After-tax Return on Invested Capital	at_turn	Asset turnover
bm	Book/Market	capei	Shillers Cyclically Adjusted P/E ratio
capital_ratio	Capitalization Ratio	cash_debt	Cash Flow/Total Debt
cash_lt	Cash Balance/Total Liabilities	cash_ratio	Cash Ratio
cfm	Cash Flow Margin	curr_debt	Current Liabilities/Total Liabilities
curr_ratio	Current Ratio	debt_asset	Total Debt/Total Assets
debt_at	Total Debt/Total Assets	debt_capital	Total Debt/Capital
debt_ebitda	Total Debt/EBITDA	debt_invcap	Long-term Debt/Invested Capital
divyield	Dividend Yield	dltt_be	Long-term Debt/Book Equity
dpr	Dividend Payout Ratio	efftax	Effective Tax Rate
equity invcap	Common Equity/Invested Capital	evm	Enterprise Value Multiple
fcf_ocf	Free Cash Flow/Operating Cash Flow	gpm	Gross Profit Margin
GProf	Gross Profit/Total Assets	int_debt	Interest/Average Long-term Debt
int_totdebt	Interest/Average Total Debt	intcov	After-tax Interest Coverage
intcov_ratio	Interest Coverage Ratio	inv_turn	Inventory Turnover
invt_act	Inventory/Current Assets	lt_ppent	Total Liabilities/Total Tangible Assets
npm	Net Profit Margin	ocf_lct	Operating CF/Current Liabilities
opmad	Operating Profit Margin After Depreciation	opmbd	Operating Profit Margin Before Depreciation
pay_turn	Payables Turnover	pcf	Price/Cash flow
pe_exi	P/E (Diluted, Excl. EI)	pe_inc	P/E (Diluted, Incl. EI)
PEG_trailing	Trailing P/E to Growth ratio	pretret_earnat	Pre-tax Return on Total Earning Assets
pretret_noa	Pre-tax return on Net Operating Assets	profit_lct	Profit Before Depreciation/Current Liabilities
ps	Price/Sales	ptb	Price/Book
ptpm	Pre-tax Profit Margin	quick_ratio	Quick Ratio (Acid Test)
rd_sale	Research and Development/Sales	rect_act	Receivables/Current Assets
rect_turn	Receivables Turnover	roa	Return on Assets
roce	Return on Capital Employed	roe	Return on Equity
sale_equity	Sales/Stockholders Equity sale	invcap	Sales/Invested Capital
sale_nwc	Sales/Working Capital	short_debt	Short-Term Debt/Total Debt
totdebt_invcap	Total Debt/Invested Capital		

Table IA.2: Groups of Variables for Tests in Section 3.4

Group 1: Macroeconomic Variables			
con_g	Log Difference of Consumption in Goods and Services	IPT_g	Log Difference of Industrial Production Index
GDP_g	Log Difference of Real GDP	unemployment	Unemployment Rate
INFLATION	Inflation Rate		
Group 2: Stock-Level Variables			
return	Monthly Stock Return	PRICE	Stock Price
DOLLARVOL	Dollar Volume	VOL	Trading Volume
Group 3: Profitability			
aftret eq	After-tax Return on Average Common Equity	aftret equity	After-tax Return on Total Stockholders Equity
aftret invcapx	After-tax Return on Invested Capital	gpm	Gross Profit Margin
GProf	Gross Profit/Total Assets	npm	Net Profit Margin
opmad	Operating Profit Margin After Depreciation	opmbd	Operating Profit Margin Before Depreciation
pretret_earnat	Pre-tax Return on Total Earning Assets	pretret_noa	Pre-tax return on Net Operating Assets
ptpm	Pre-tax Profit Margin	roa	Return on Assets
roce	Return on Capital Employed	roe	Return on Equity
accrual	Accruals/Average Assets	efftax	Effective Tax Rate
Group 4: Liquidity and Solvency			
cash_ratio	Cash Ratio	cfm	Cash Flow Margin
curr_debt	Current Liabilities/Total Liabilities	curr_ratio	Current Ratio
ocf_lct	Operating CF/Current Liabilities	quick_ratio	Quick Ratio (Acid Test)
invt_act	Inventory/Current Assets	cash_lt	Cash Balance/Total Liabilities
fcf_ocf	Free Cash Flow/Operating Cash Flow		
Group 5: Leverage and Capital Structure			
capital_ratio	Capitalization Ratio	cash_debt	Cash Flow/Total Debt
debt_asset	Total Debt/Total Assets	debt_at	Total Debt/Total Assets
debt_capital	Total Debt/Capital	debt_ebitda	Total Debt/EBITDA
debt_invcap	Long-term Debt/Invested Capital	dltt_be	Long-term Debt/Book Equity
equity invcap	Common Equity/Invested Capital	int_totdebt	Interest/Average Total Debt
int_debt	Interest/Average Long-term Debt	lt_ppent	Total Liabilities/Total Tangible Assets
short_debt	Short-Term Debt/Total Debt	totdebt_invcap	Total Debt/Invested Capital
intcov_ratio	Interest Coverage Ratio	intcov	After-tax Interest Coverage
Group 6: Market Valuation and Returns			
bm	Book/Market	capei	Shillers Cyclically Adjusted P/E ratio
evm	Enterprise Value Multiple	pcf	Price/Cash flow
pe_exi	P/E (Diluted, Excl. EI)	pe_inc	P/E (Diluted, Incl. EI)
PEG_trailing	Trailing P/E to Growth ratio	ps	Price/Sales
ptb	Price/Book	divyield	Dividend Yield
dpr	Dividend Payout Ratio		
Group 7: Activity and Operational Efficiency			
at_turn	Asset turnover	inv_turn	Inventory Turnover
pay_turn	Payables Turnover	rect_turn	Receivables Turnover
sale_equity	Sales/Stockholders Equity	sale_invcap	Sales/Invested Capital
sale_nwc	Sales/Working Capital	adv_sale	Advertising Expenses/Sales
rect_act	Receivables/Current Assets	profit_lct	Profit Before Depreciation/Current Liabilities
rd_sale	Research and Development/Sales		

Table IA.3: Analysts' and Managers' Mention of Historical Dates

This replicates Table 2 using patterns extracted from sentences spoken by analysts (Panel A) and managers (Panel B) during corporate events. The dependent variable is an indicator that equals one if, during a firm's corporate event in month t, an analyst (Panel A) or manager (Panel B) mentions the historical month $t - h$ at least once, and zero otherwise. The similarity-based weight $\pi_{l,t,t-h}$ captures the stock-specific weight that our measure assigns to month $t - h$ in month t. Exponential Weight is the weight on month $t - h$ derived from a simple model of exponential decay following Greenwood and Shleifer (2014) using λ = 0.56. The first five columns focus on the previous 180 months, i.e., $h \in \{1, 2, 3, \ldots, 180\}$. The last five columns focus on months that are at least 5 years in the past, i.e., $h \in \{61, 62, 63, \ldots, 180\}$. Columns (2) to (4) and (7) to (9) include *stock* × *past year-month* fixed effects as well as control variables. The control variables are as of the end of month $t - 1$. Columns (5) and (10) further include *stock* × *current year-month* fixed effects. These fixed effects soak up the control variables. In all columns except (1) and (6), we also include quarter fixed effects. Standard errors are clustered by stock, current year-month, and past year-month, and the t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

Panel A: Analysts

Dependent Variable:	Month Mentioned									
Sample:	Past 15 Years					At Least 5 Years in the Past				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\pi_{l,t,t-h}$	52.393***	55.668***		53.215***	47.289***	0.977***	1.389***		1.363***	0.904***
	(41.18)	(34.19)		(26.38)	(24.23)	(8.96)	(8.15)		(7.97)	(5.42)
Exponential Weight			555.844	369.465	391.393			0.000	0.000	0.000
			(1.63)	(1.56)	(1.59)			(1.61)	(1.57)	(1.58)
Size		-0.003**	-0.035***	-0.003**			-0.001*	-0.001**	-0.001*	
		(-2.01)	(-7.19)	(-2.15)			(-1.92)	(-2.52)	(-1.87)	
Turnover		0.024***	0.023*	0.026***			0.007***	0.008***	0.007***	
		(3.49)	(1.69)	(3.78)			(4.45)	(4.48)	(4.48)	
BM		0.000	-0.003	0.000			-0.001***	-0.001***	-0.001***	
		(0.27)	(-1.58)	(0.29)			(-3.18)	(-3.23)	(-3.18)	
Ivol		-0.344***	-0.887***	-0.360***			0.021	0.021	0.021	
		(-6.08)	(-6.88)	(-6.38)			(0.70)	(0.71)	(0.71)	
Price		-0.000***	-0.000***	-0.000***			-0.000***	-0.000***	-0.000***	
		(-3.42)	(-3.38)	(-3.47)			(-3.38)	(-3.50)	(-3.37)	
Adjusted R-Squared	0.194	0.286	0.189	0.295	0.344	0.000	0.009	0.008	0.009	0.147
N	15,239,350	15,239,350	15,239,350	15,239,350	15,239,350	10,153,586	10,153,586	10,153,586	10,153,586	10,153,586
Stock x Current Year-Month FE	NO	NO	NO	NO	YES	NO	NO	NO	NO	YES
Stock x Past Year-Month FE	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES
Quarter FE	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES

Panel B: Managers										
Dependent Variable:	Month Mentioned									
Sample:	Past 15 Years					At Least 5 Years in the Past				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\pi_{l,t,t-h}$	91.471***	89.174***		87.446***	66.115***	4.595***	6.119***		5.933***	3.451***
	(57.78)	(44.51)		(41.56)	(35.90)	(13.42)	(14.36)		(13.58)	(8.98)
Exponential Weight			566.534	260.266	274.665			0.000	0.000	0.000
			(1.62)	(1.49)	(1.55)			(1.57)	(1.54)	(1.57)
Size		-0.018***	-0.070***	-0.018***			-0.004***	-0.005***	-0.004***	
		(-5.06)	(-7.80)	(-5.10)			(-3.03)	(-3.51)	(-2.96)	
Turnover		0.024**	0.021	0.025**			0.010**	0.012**	0.011**	
		(2.15)	(0.88)	(2.27)			(2.27)	(2.42)	(2.36)	
BM		-0.002	-0.008*	-0.002			-0.002**	-0.002**	-0.002**	
		(-0.94)	(-1.82)	(-0.94)			(-2.29)	(-2.33)	(-2.29)	
Ivol		-0.722***	-1.599***	-0.733***			-0.048	-0.045	-0.047	
		(-6.50)	(-6.63)	(-6.59)			(-0.80)	(-0.74)	(-0.79)	
Price		-0.000***	-0.001***	-0.000***			-0.000***	-0.000***	-0.000***	
		(-2.82)	(-3.18)	(-2.84)			(-3.30)	(-3.35)	(-3.29)	
Adjusted R-Squared	0.269	0.385	0.256	0.387	0.458	0.001	0.074	0.074	0.075	0.208
N	15,239,350	15,239,350	15,239,350	15,239,350	15,239,350	10,153,586	10,153,586	10,153,586	10,153,586	10,153,586
Stock x Current Year-Month FE	NO	NO	NO	NO	YES	NO	NO	NO	NO	YES
Stock x Past Year-Month FE	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES
Quarter FE	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES

Table IA.4: Cross-Sectional Variation in the Decay Speed

This table presents the relation between stock-level characteristics and the decay speed. Decay speed for stock l in month t is defined as follows:

$$\text{Decay Speed}_{l,t} = \frac{\pi_{l,t,t}}{\pi_{l,t,t-12}}$$

A higher value indicates a faster decay speed. The table shows how decay speed varies along four firm-level characteristics, which are each measured as of the end of the month: the logarithm of a firm's market capitalization (in million $), the book-to-market ratio following Fama and French (1992), the idiosyncratic volatility (in %) from CAPM regressions following Ang et al. (2006), and the stock price (in $). The sample runs from January 1966 to December 2021. The t-statistics for tests that the difference between the highest and lowest decile is equal to zero are displayed in parentheses.

Panel A: Equal-Weighted Memory Decay Speed

Model	Low	P2	P3	P4	P5	P6	P7	P8	P9	High	High-Low
Size	1.34	1.31	1.30	1.28	1.27	1.25	1.24	1.24	1.23	1.21	–0.13
											(–29.33)
BM	1.26	1.25	1.25	1.26	1.27	1.26	1.27	1.27	1.28	1.30	0.04
											(8.56)
Ivol	1.22	1.23	1.23	1.24	1.25	1.27	1.28	1.30	1.32	1.34	0.12
											(29.37)
Price	1.34	1.32	1.29	1.27	1.26	1.25	1.24	1.24	1.23	1.22	–0.12
											(–21.17)

Panel B: Value-Weighted Memory Decay Speed

Model	Low	P2	P3	P4	P5	P6	P7	P8	P9	High	High-Low
Size	1.33	1.31	1.30	1.28	1.26	1.25	1.24	1.24	1.23	1.19	–0.14
											(–32.72)
BM	1.19	1.20	1.22	1.22	1.22	1.22	1.23	1.23	1.23	1.27	0.09
											(9.55)
Ivol	1.18	1.20	1.21	1.22	1.22	1.23	1.25	1.26	1.28	1.29	0.11
											(20.33)
Price	1.32	1.29	1.27	1.25	1.23	1.23	1.22	1.22	1.21	1.19	–0.13
											(–18.20)

Table IA.5: The Relative Importance of Different Qualitative Features

This table presents results from tests that analyze the relative importance of different qualitative features. The qualitative features are classified into 23 mutually exclusive groups ("topics") following Bybee et al. (2024). For each topic, we construct our measure in two ways: (1) using only the features from the focal topic, and (2) using features from all other topics except the focal topic. This approach yields a total of 46 variations of our measure, each based on a different set of topics. For each variation of our measure, we re-estimate column (1) of Table 2, aggregated to the market level, and report the coefficient on $\bar{\pi}_{t,t-h}$ as well as the adjusted R^2. Topics are ranked in descending order of R^2 values for included variable sets. In case of a tie, topics are further ranked in ascending order of R^2 values for excluded variable sets.

	Included		Excluded	
Features:	Coef	Adjusted R^2	Coef	Adjusted R^2
Financial Markets	56.077	0.384	137.360	0.575
Trans/Defense/Local	75.110	0.377	128.231	0.560
Asset Managers/I-Banks	55.120	0.367	130.798	0.567
Technology	50.840	0.350	133.616	0.574
Industry	75.869	0.287	128.068	0.572
Negotiations	50.278	0.259	131.022	0.573
Science/Language	57.138	0.255	128.353	0.566
Social/Cultural	56.886	0.254	127.477	0.562
Challenges	47.817	0.246	131.537	0.575
Activism/Language	52.314	0.245	129.478	0.569
Management	38.529	0.244	131.866	0.574
International Affairs	63.818	0.239	126.293	0.563
Buyouts & Bankruptcy	53.479	0.226	129.722	0.571
Political Leaders	48.306	0.215	128.115	0.566
Economic Growth	51.682	0.185	126.601	0.562
Entertainment	41.010	0.173	129.224	0.570
Banks	43.185	0.173	127.769	0.565
Government	49.253	0.166	126.868	0.566
Terrorism/Mideast	31.218	0.152	128.793	0.570
Oil & Mining	41.491	0.145	127.268	0.564
Courts	32.239	0.125	128.488	0.568
Corporate Earnings	30.970	0.096	129.136	0.572
Labor/Income	38.827	0.089	126.238	0.566

Table IA.6: Similarity-Based and Survey-Based Return Expectations of the S&P 500 Index: Blending Quantitative and Qualitative Features

This table repeats the analysis from Table 5 using a measure that blends both quantitative and qualitative features. The weight on each month is the simple average of the weights generated by our baseline measure (which is based on quantitative features) and the measure based on qualitative features. In all columns, the dependent variable is the difference in the percentage of bullish and bearish investors from the Gallup investor survey (elicited over the course of month $t+1$). In column (1), the main independent variable is the similarity-based return expectation of the S&P 500 index derived from the blended measure (as of the end of month t). In column (2), we decompose this similarity-based expectation into expectations from recent periods (most recent 12 months) and from distant periods (more than 12 months in the past). The sum of these two components equals the similarity-based expectation from column (1). In column (3), we add the cumulative return over the past 12 months as an additional control, following Greenwood and Shleifer (2014). In column (4), we control for the exponentially-weighted average return over the past five years. To construct this exponentially-weighted return, we first calculate quarterly returns by compounding 3-month returns on a rolling monthly basis. We then use the weighting approach of Greenwood and Shleifer (2014) and their estimated quarterly λ of 0.77 to calculate an exponentially-weighted average return over the past five years. Columns (5) - (7) mirror columns (2) - (4), except that the cutoff between recent and distant periods is five years instead of 12 months. In column (8), we interact expectations from recent and distant periods with a dummy variable equal to one during NBER recessions. All columns include the same control variables as Panel B of Table 3 in Greenwood and Shleifer (2014). The sample period ranges from October 1996 to June 2017. t-statistics, in parentheses, are Newey-West adjusted with twelve lags. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

	Gallup Survey Expectations							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Similarity-Based Expectation (Blend)	295.988***							
	(2.84)							
Exp from Recent Periods (<=12 mths)		538.241***	-164.468	-205.972				
		(5.00)	(-1.33)	(-1.55)				
Exp from Distant Periods (>12 mths)		181.992**	174.218***	101.420				
		(2.51)	(2.92)	(1.26)				
Exp from Recent Periods (<=5 yrs)					301.817***	93.189	-8.338	-120.469
					(3.21)	(1.15)	(-0.08)	(-1.07)
Exp from Distant Periods (>5 yrs)					247.300**	225.252**	177.861*	51.188
					(2.39)	(2.53)	(1.92)	(0.54)
Exp from Recent Periods (<=5 yrs) X US Rec								-420.290***
								(-4.29)
Exp from Distant Periods (>5 yrs) X US Rec								844.534***
								(5.57)
US Rec								-38.964***
								(-3.77)
Cumulative Return (past 12 mths)			90.675***			74.338***		11.009
			(6.20)			(5.44)		(0.72)
Exponentially-Weighted Past Return				416.578***			404.472***	402.346***
				(5.72)			(6.37)	(4.69)
Log(P/D)	3.074	12.439	16.970	17.234	4.219	18.991	20.757	39.183**
	(0.15)	(0.84)	(1.26)	(1.04)	(0.22)	(1.59)	(1.57)	(2.28)
Risk-free Rate	-230.160***	-204.285***	-150.076***	-144.605***	-230.518***	-143.473***	-124.193**	-81.919
	(-3.74)	(-3.22)	(-3.05)	(-2.96)	(-3.80)	(-2.74)	(-2.44)	(-1.66)
Earnings Growth	19.644***	4.653	-1.572	8.934	20.378***	-11.213	-5.822	-8.279
	(3.18)	(0.65)	(-0.23)	(1.05)	(3.01)	(-1.39)	(-0.79)	(-1.00)
Unemployment	0.953	0.381	0.017	-0.732	1.062	-0.524	-1.533	-2.927*
	(0.47)	(0.19)	(0.01)	(-0.37)	(0.52)	(-0.28)	(-0.78)	(-1.73)
Adjusted R-Squared	0.538	0.582	0.637	0.654	0.539	0.644	0.675	0.725
N	149	149	149	149	149	149	149	149

Table IA.7: Similarity-Based Volatility and Option-Implied Volatility: Fama-MacBeth Regressions

This table shows that similarity-based volatility explains option-implied volatility in the cross-section of stocks. In all columns, the dependent variable is option-implied volatility in month $t + 1$, constructed following An et al. (2014). All independent variables are as of month t. Similarity-based volatility is the standard deviation of monthly returns over the past 180 months, where each historical return is weighted with its associated similarity-based weight. Exponential-decay-based volatility is the standard deviation of monthly returns over past months, where each historical return is weighted with exponentially-decaying weights, using the decay parameter λ = 0.56 from Greenwood and Shleifer (2014). In all columns, we estimate Fama-MacBeth regressions. In column (1), we only include similarity-based volatility as an independent variable. In column (2), we control for exponential-decay-based volatility. In column (3), we control for lagged option-implied volatility, and in column (4) we control for size, idiosyncratic volatility (following Ang et al. (2006)), asset growth, operating profit (following Fama and French (2006)), and the logarithm of the book-to-market ratio (following Fama and French (1992)). We multiply all coefficients by 100. Standard errors are Newey-West adjusted with twelve lags, and the t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

	Option-Implied Volatility			
	(1)	(2)	(3)	(4)
Similarity-Based Volatility	69.936***	52.146***	14.184***	10.965***
	(18.62)	(14.49)	(13.52)	(12.55)
Exponential-Decay-Based Volatility		26.870***	7.031***	5.312***
		(16.08)	(9.06)	(9.56)
Option-Implied Volatility (lagged)			72.472***	66.752***
			(47.76)	(37.43)
Size				-0.690***
				(-7.81)
Idiosyncratic Volatility (3F)				122.886***
				(10.70)
Asset Growth				0.049
				(0.57)
Operating Profit				-0.108**
				(-2.37)
log(Book-to-Market Ratio)				0.226***
				(4.22)
Adjusted R-Squared	0.439	0.500	0.745	0.758
N	205,906	205,906	205,906	205,906

Table IA.8: Summary Statistics for Table IA.9

This table presents summary statistics for the sample used in Table IA.9. The sample period for these tests ranges from January 1990 to December 2021. VIX is the average daily Volatility Index over the course of a month. We normalize VIX by dividing it by 100. Similarity-Based Volatility is the annualized volatility of the S&P 500 index, constructed using our measure. The logarithm of the price-dividend ratio (Log(P/D)), the risk-free rate, the earnings growth rate, and the unemployment rate are constructed following Greenwood and Shleifer (2014). Realized Volatility is the the annualized realized volatility of monthly S&P 500 index returns over the past 180 months.

	N	Mean	Median	Std.Dev	P25	P75	Min	Max
VIX Index	372	0.194	0.175	0.077	0.139	0.228	0.101	0.627
Similarity-Based Volatility	372	0.148	0.149	0.011	0.138	0.157	0.123	0.170
Log(P/D)	372	3.946	3.947	0.254	3.841	4.071	3.292	4.502
Risk-free Rate	372	1.005	1.006	0.018	0.993	1.016	0.950	1.101
Earnings Growth	372	0.230	0.125	1.043	-0.052	0.202	-0.886	7.935
Unemployment	372	5.902	5.500	1.742	4.600	6.800	3.500	14.700
Realized Volatility (Past 180 Months)	372	0.153	0.152	0.009	0.145	0.160	0.138	0.174

Table IA.9: Similarity-Based Volatility of the S&P 500 Index

This table shows that similarity-based volatility predicts the VIX. In both columns, the dependent variable is the average daily VIX over the course of month $t + 1$. The main independent variable is annualized similarity-based volatility of the S&P 500 index, constructed using our measure (as of the end of month t). Column (2) adds control variables. The sample period ranges from January 1990 to December 2021. t-statistics, in parentheses, are Newey-West adjusted with twelve lags. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

	VIX	
	(1)	(2)
Similarity-Based Volatility	1.807***	6.137***
	(2.81)	(3.51)
Realized Volatility (Past 180 Months)		-6.947***
		(-3.18)
Log(P/D)		0.038
		(0.77)
Risk-free Rate		0.929***
		(3.19)
Earnings Growth		-0.008
		(-1.02)
Unemployment		0.011***
		(2.96)
Adjusted R-Squared	0.065	0.254
N	372	372

Table IA.10: Summary Statistics for Tests of Similarity-Driven Repurchasing Decisions

This table presents summary statistics for the samples used in the tests displayed in Tables IA.11, IA.13, and IA.15. The data in these tests are the same as in Barber and Odean (2000) and the sample period ranges from January 1991 to November 1996. Number of Stocks for Repurchase is the number of distinct stocks that the investor once owned but does not own in month t. Number of Stocks in Portfolio is the number of distinct stocks that the investor owns in month t. These two variables are at the account-month level. Repurchase Dummy (Buy) is a dummy variable that is equal to one if investor i repurchases a previously-held stock l in month t. Ret is the return that investor i realized when liquidating a previously-held position l in month $t-h$. The similarity-based weight $\pi_{i,l,t,t-h}$ is the investor-specific weight that our measure assigns to the month in which the investor liquidated a previously-held position in stock l. It is constructed for each investor individually using the investor's historical holdings. Return between Sell and Repurchase is the return that a stock realized between the previous liquidation and a (potential) repurchase. The logarithm of the initial purchase price (ln(WAPP)), the square root of the number of days between initial purchase and liquidation ($\sqrt{\text{Time Owned}}$), and the volatility calculated using daily returns over the 250 days preceding the initial purchase (Return Volatility) are constructed following Ben-David and Hirshleifer (2012).

	N	Mean	Median	Std.Dev	P25	P75	Min	Max
Number of Stocks for Repurchase	1,121,190	5.737	3.000	7.730	2.000	6.000	1.000	438.000
Number of Stocks in Portfolio	1,121,190	2.460	0.000	5.581	0.000	3.000	0.000	645.000
Repurchase Dummy (Buy)	6,452,764	0.005	0.000	0.072	0.000	0.000	0.000	1.000
Ret	6,452,764	0.066	0.044	0.308	-0.089	0.182	-0.693	1.400
$\pi_{i,l,t,t-h}$	6,452,764	0.057	0.032	0.084	0.020	0.060	0.003	1.000
Return between Sell and Repurchase	6,452,764	0.327	0.085	1.140	-0.143	0.477	-0.999	138.342
$\ln(\text{WAPP})$	6,452,764	2.349	2.431	1.229	1.727	3.008	-2.773	15.032
$\sqrt{\text{Time Owned}}$	6,452,764	2.502	2.236	1.258	1.414	3.162	1.000	8.367
Return Volatility	6,452,764	0.032	0.029	0.018	0.021	0.039	0.005	0.409

Table IA.11: Similarity-Driven Repurchasing Decisions

This table shows that the similarity-based weight on a past trading experience explains the likelihood that an investor repurchases a stock. The dependent variable in all columns is a dummy that is equal to one if the investor repurchases a previously-held stock in month t. The main independent variable in column (1) is the return that the investor realized when liquidating a previously-held position. The main independent variables in columns (2) and (4) are (i) the return that the investor realized when liquidating a previously-held position, (ii) the investor-specific similarity weight $\pi_{i,l,t,t-h}$ on the month in which the previously-held position was liquidated, and (iii) the interaction of (i) and (ii). Columns (3) and (5) break out the investor-specific similarity weight separately by investor-specific similarity and interference. All columns include a set of control variables as well as *stock* × *current year-month* fixed effects. Columns (2) and (3) further include account × current year-month fixed effects, and columns (4) and (5) include *account* × *liquidation year-month* × *current year-month* fixed effects instead. All coefficients are multiplied by 100. Standard errors are clustered by account, liquidation month, and current month, and the t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

	Repurchase Dummy (Buy)				
	(1)	(2)	(3)	(4)	(5)
Ret	0.384***	-0.174***	0.447***	-0.132**	0.538***
	(8.72)	(-4.16)	(8.68)	(-2.33)	(5.66)
$\pi_{i,l,t,t-h}$		3.553***		-1.591***	
		(21.69)		(-5.08)	
$\pi_{i,l,t,t-h}$ x Ret		12.734***		11.551***	
		(17.08)		(11.91)	
Similarity			1.621***		-0.164**
			(18.07)		(-2.41)
Interference			0.003		0.047***
			(1.56)		(13.61)
Similarity x Ret			0.828***		0.994***
			(10.69)		(9.35)
Interference x Ret			-0.026***		-0.036***
			(-12.50)		(-10.67)
Return between Sell and Repurchase	-0.131***	-0.105***	-0.076***	-0.104***	-0.094***
	(-6.40)	(-6.48)	(-5.30)	(-5.13)	(-4.65)
$\ln(\text{WAPP})$	0.350***	0.225***	0.266***	0.195***	0.218***
	(7.91)	(6.62)	(7.09)	(4.37)	(4.89)
$\sqrt{\text{Time Owned}}$	-0.010	0.027***	-0.070***	-0.059***	-0.186***
	(-1.46)	(5.76)	(-7.79)	(-4.54)	(-10.09)
Return Volatility	1.167	1.771**	2.932***	3.484***	3.311***
	(1.21)	(2.15)	(3.18)	(3.39)	(3.37)
Adjusted R-Squared	0.056	0.058	0.057	0.051	0.051
N	6,452,764	6,452,764	6,452,764	2,862,961	2,862,961
Stock x Current Month FE	YES	YES	YES	YES	YES
Account x Current Month FE	YES	YES	YES	NO	NO
Account x Past Month x Current Month FE	NO	NO	NO	YES	YES

Table IA.12: Similarity-Driven Repurchasing Decisions: Investor-Specific Weights vs. Representative Investor Weights

This table shows the repurchase effect using both the investor-specific similarity weight, $\pi_{i,l,t,t-h}$, as well as the similarity weight of a representative investor, $\pi_{rep,l,t,t-h}$. The dependent variable in all columns is a dummy that is equal to one if the investor repurchases a previously-held stock in month t. All columns include a set of control variables as well as *stock* × *current year-month* fixed effects. Columns (1) and (2) further include *account* × *current year-month* fixed effects, and columns (3) and (4) include *account* × *liquidation year-month* × *current year-month* fixed effects instead. All coefficients are multiplied by 100. Standard errors are clustered by account, liquidation month, and current month, and the t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

	Repurchase Dummy (Buy)			
	(1)	(2)	(3)	(4)
Ret	0.075	-0.170***	0.071	-0.165**
	(1.46)	(-3.69)	(1.15)	(-2.65)
$\pi_{rep,l,t,t-h}$	34.231***	16.353***	-3.250	-1.556
	(18.66)	(8.78)	(-1.24)	(-0.53)
$\pi_{rep,l,t,t-h}$ x Ret	25.648***	2.824	23.628***	2.753
	(11.15)	(1.10)	(7.14)	(0.78)
$\pi_{i,l,t,t-h}$		3.147***		-1.571***
		(18.76)		(-4.90)
$\pi_{i,l,t,t-h}$ x Ret		12.540***		11.317***
		(14.96)		(10.68)
Return between Sell and Repurchase	-0.101***	-0.093***	-0.103***	-0.105***
	(-5.72)	(-6.02)	(-5.04)	(-5.09)
$\ln(\text{WAPP})$	0.340***	0.236***	0.218***	0.193***
	(8.42)	(7.02)	(4.81)	(4.24)
$\sqrt{\text{Time Owned}}$	-0.032***	0.013**	-0.035***	-0.058***
	(-5.44)	(2.63)	(-3.12)	(-4.46)
Return Volatility	1.857*	2.037**	2.976***	3.485***
	(1.99)	(2.45)	(3.09)	(3.39)
Adjusted R-Squared	0.057	0.058	0.052	0.051
N	6,452,764	6,452,764	2,883,462	2,862,961
Stock x Current Month FE	YES	YES	YES	YES
Account x Current Month FE	YES	YES	NO	NO
Account x Past Month x Current Month FE	NO	NO	YES	YES

Table IA.13: Similarity-Driven Repurchasing Decisions: Opening vs. Realization Month

This table shows the repurchase effect for similarity-based weights on the month in which the previous return was realized (the "realization month" $t - h$), $\pi_{i,l,t,t-h}$, as well as on the month in which the position was initially opened (the "opening month" $t - o$), $\pi_{i,l,t,t-o}$. Columns (1) and (3) replicate columns (2) and (4) of Table IA.11, using the weight on the opening month. Columns (2) and (4) include both weights. The dependent variable in all columns is a dummy that is equal to one if the investor repurchases a previously-held stock in month t. All columns include a set of control variables as well as *stock* × *current year-month* fixed effects. Columns (1) and (2) further include *account* × *current year-month* fixed effects, and columns (3) and (4) include *account* × *liquidation year-month* × *current year-month* fixed effects instead. All coefficients are multiplied by 100. Standard errors are clustered by account, liquidation month, and current month, and the t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

	Repurchase Dummy (Buy)			
	(1)	(2)	(3)	(4)
Ret	-0.102**	-0.145***	-0.042	-0.119**
	(-2.40)	(-3.58)	(-0.74)	(-2.32)
$\pi_{i,l,t,t-o}$	3.528***	-0.753	-1.463***	3.632***
	(21.23)	(-1.19)	(-4.88)	(3.48)
$\pi_{i,l,t,t-o}$ x Ret	13.557***	6.813***	11.932***	3.501
	(15.62)	(3.40)	(11.06)	(1.26)
$\pi_{i,l,t,t-h}$		4.274***		-5.223***
		(6.89)		(-4.39)
$\pi_{i,l,t,t-h}$ x Ret		6.544***		8.364***
		(4.16)		(3.57)
Return between Sell and Repurchase	-0.107***	-0.105***	-0.103***	-0.105***
	(-6.48)	(-6.45)	(-5.07)	(-5.16)
$\ln(\text{WAPP})$	0.231***	0.229***	0.203***	0.191***
	(6.73)	(6.76)	(4.54)	(4.30)
$\sqrt{\text{Time Owned}}$	0.038***	0.026***	-0.059***	-0.051***
	(7.68)	(5.81)	(-4.53)	(-4.08)
Return Volatility	1.674**	1.806**	3.513***	3.434***
	(2.03)	(2.19)	(3.40)	(3.37)
Adjusted R-Squared	0.058	0.058	0.051	0.051
N	6,452,764	6,452,764	2,863,256	2,862,961
Stock x Current Month FE	YES	YES	YES	YES
Account x Current Month FE	YES	YES	NO	NO
Account x Past Month x Current Month FE	NO	NO	YES	YES

Table IA.14: Similarity-Driven Repurchasing Decisions: Excluding Return-Related Variables When Constructing Similarity-Based Weights

This table repeats the analysis from Table IA.11, but excludes all return-related variables when constructing similarity-based weights. The dependent variable in all columns is a dummy that is equal to one if the investor repurchases a previously-held stock in month t. The main independent variables in columns (1) and (3) are (i) the return that the investor realized when liquidating a previously-held position, (ii) the investor-specific similarity weight on the month in which the previously-held position was liquidated, and (iii) the interaction of (i) and (ii). Columns (2) and (4) break out the investor-specific weight separately by investor-specific similarity and interference. All columns include a set of control variables as well as *stock* × *current year-month* fixed effects. Columns (1) and (2) further include *account* × *current year-month* fixed effects, and columns (3) and (4) include *account* × *liquidation year-month* × *current year-month* fixed effects instead. All coefficients are multiplied by 100. Standard errors are clustered by account, liquidation month, and current month, and the t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

	Repurchase Dummy (Buy)			
	(1)	(2)	(3)	(4)
Ret	-0.172***	0.482***	-0.128**	0.614***
	(-4.12)	(10.03)	(-2.27)	(6.46)
$\pi_{i,l,t,t-h}$	3.534***		-1.608***	
	(21.65)		(-5.18)	
$\pi_{i,l,t,t-h}$ x Ret	12.710***		11.507***	
	(17.12)		(11.96)	
Similarity		1.389***		-0.273***
		(17.41)		(-4.04)
Interference		0.001		0.045***
		(0.34)		(13.50)
Similarity x Ret		0.748***		0.880***
		(10.45)		(9.48)
Interference x Ret		-0.026***		-0.036***
		(-12.31)		(-10.74)
Return between Sell and Repurchase	-0.106***	-0.085***	-0.104***	-0.095***
	(-6.50)	(-5.70)	(-5.13)	(-4.68)
$\ln(\text{WAPP})$	0.224***	0.262***	0.196***	0.220***
	(6.58)	(6.83)	(4.39)	(4.93)
$\sqrt{\text{Time Owned}}$	0.027***	-0.061***	-0.059***	-0.181***
	(5.69)	(-6.92)	(-4.56)	(-9.88)
Return Volatility	1.779**	2.999***	3.495***	3.321***
	(2.16)	(3.21)	(3.40)	(3.38)
Adjusted R-Squared	0.058	0.057	0.051	0.051
N	6,452,764	6,452,764	2,862,961	2,862,961
Stock x Current Month FE	YES	YES	YES	YES
Account x Current Month FE	YES	YES	NO	NO
Account x Past Month x Current Month FE	NO	NO	YES	YES

Table IA.15: Similarity-Driven Repurchasing Decisions: The Role of Encoding Strength

This table shows that the strength with which a past trading experience is encoded affects the strength of the similarity-driven repurchase effect. Columns (1) and (2) augment columns (2) and (4) of Table IA.11, respectively, with an additional interaction with the dummy variable Attention. This dummy is equal to one if the investor executed at least six transactions in the month of the experience, and zero otherwise. The 75th percentile of transactions executed per month is six, so this dummy captures months in which the number of transactions is above the 75th percentile. All coefficients are multiplied by 100. Standard errors are clustered by account, liquidation month, and current month, and the t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

	Repurchase Dummy (Buy)	
	(1)	(2)
Attention x $\pi_{i,l,t,t-h}$ x Ret	4.804***	4.534**
	(4.10)	(2.44)
$\pi_{i,l,t,t-h}$ x Ret	10.775***	8.431***
	(13.89)	(6.29)
Attention x $\pi_{i,l,t,t-h}$	-0.368	-1.733***
	(-1.34)	(-2.76)
Attention x Ret	-0.084	-0.038
	(-1.62)	(-0.39)
Attention	0.019	
	(1.39)	
Ret	-0.123***	-0.078
	(-3.16)	(-1.24)
$\pi_{i,l,t,t-h}$	3.767***	-0.239
	(19.35)	(-0.43)
Return between Sell and Repurchase	-0.104***	-0.103***
	(-6.46)	(-5.09)
$\ln(\text{WAPP})$	0.228***	0.199***
	(6.70)	(4.44)
$\sqrt{\text{Time Owned}}$	0.028***	-0.054***
	(5.74)	(-4.19)
Return Volatility	1.774**	3.473***
	(2.15)	(3.37)
Adjusted R-Squared	0.058	0.051
N	6,452,764	2,862,961
Stock x Current Month FE	YES	YES
Account x Current Month FE	YES	NO
Account x Past Month x Current Month FE	NO	YES

Table IA.16: Similarity-Driven Repurchasing Decisions: Per-Unit-Time Realized Returns

This table repeats the analysis from Table IA.11, but uses the realized return per month to construct the variable Ret. The dependent variable in all columns is a dummy that is equal to one if the investor repurchases a previously-held stock in month t. The main independent variable in column (1) is the return per month that the investor realized when liquidating a previously-held position. The main independent variables in columns (2) and (4) are (i) the return per month that the investor realized when liquidating a previously-held position, (ii) the investor-specific similarity weight on the month in which the previously-held position was liquidated, and (iii) the interaction of (i) and (ii). Columns (3) and (5) break out the investor-specific weight separately by investor-specific similarity and interference. All columns include a set of control variables as well as *stock* × *current year-month* fixed effects. Columns (1), (2), and (3) further include *account* × *current year-month* fixed effects, and columns (4) and (5) include *account* × *liquidation year-month* × *current year-month* fixed effects instead. All coefficients are multiplied by 100. Standard errors are clustered by account, liquidation month, and current month, and the t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

	Repurchase Dummy (Buy)				
	(1)	(2)	(3)	(4)	(5)
Ret (per month)	2.047***	0.401**	-0.963**	0.365	-0.556
	(8.76)	(2.40)	(-2.25)	(1.60)	(-1.03)
$\pi_{i,l,t,t-h}$		3.479***		-1.739***	
		(21.62)		(-5.55)	
$\pi_{i,l,t,t-h}$ x Ret (per month)		22.460***		18.741***	
		(13.11)		(9.40)	
Similarity			1.555***		-0.213***
			(18.04)		(-3.15)
Interference			0.002		0.046***
			(1.20)		(13.47)
Similarity x Ret (per month)			8.020***		7.474***
			(9.96)		(8.90)
Interference x Ret (per month)			-0.098***		-0.120***
			(-9.27)		(-7.46)
Return between Sell and Repurchase	-0.148***	-0.131***	-0.092***	-0.133***	-0.114***
	(-7.57)	(-7.67)	(-6.48)	(-6.41)	(-5.78)
$\ln(\text{WAPP})$	0.312***	0.178***	0.213***	0.127***	0.156***
	(8.48)	(6.09)	(6.53)	(3.01)	(3.67)
$\sqrt{\text{Time Owned}}$	0.005	0.035***	-0.055***	-0.057***	-0.173***
	(0.94)	(7.47)	(-6.96)	(-4.30)	(-9.60)
Return Volatility	1.361	1.571*	3.031***	3.244***	3.203***
	(1.36)	(1.88)	(3.20)	(3.05)	(3.11)
Adjusted R-Squared	0.056	0.058	0.057	0.051	0.051
N	6,452,764	6,452,764	6,452,764	2,862,961	2,862,961
Stock x Current Month FE	YES	YES	YES	YES	YES
Account x Current Month FE	YES	YES	YES	NO	NO
Account x Past Month x Current Month FE	NO	NO	NO	YES	YES