Investor Attention and Stock Returns [*]

Jian Chen[†] Guohao Tang[‡] Jiaquan Yao[§] Guofu Zhou[¶]

First draft: April 2018

Current version: August 2020

[*] We are grateful to Hendrik Bessembinder (Managing Editor) and an anonymous referee for very insightful and helpful comments that improve the paper substantially. We also thank Koustav De (discussant), Joey Engelberg, Michael Hasler, Tse-Chun Lin, Lin Peng, Dragon Tang, Qunzi Zhang (discussant), and seminar participants at Beijing University, Emory University, Georgia State University, Indiana University, Renmin University of China, Shanghai University of Finance and Economics, Sichuan University, Southwestern University of Finance and Economics, Washington University in St. Louis, Zhejiang University, and conference participants at the 2018 Conference on Financial Predictability and Big Data, 2019 China Finance Review International Conference, 2019 Chinese Finance Annual Meeting, and 2019 FMA Annual Meeting for insightful comments. Chen and Yao acknowledge financial support from the National Natural Science Foundation of China (Grant No.71671148 and No.71502152), respectively. Part of the work was undertaken while Chen and Tang were visiting Washington University in St. Louis.

[†] Department of Finance, School of Economics, Xiamen University, China, 361005; e-mail: jchenl@xmu.edu.cn.

[‡] College of Finance and Statistics, Hunan University, Changsha, China, 410082; e-mail: ghtang@hnu.edu.cn.

[§] Corresponding author: School of Management, Jinan University, China, 510632; e-mail: jiaquanyao@gmail.com.

[¶] Olin School of Business, Washington University in St. Louis, St. Louis, Missouri, 63130; e-mail: zhou@wustl.edu; phone number: 314-935-6384.

Investor Attention and Stock Returns

Abstract

We propose an investor attention index based on proxies in the literature, and find that it predicts the stock market risk premium significantly, both in-sample and out-of-sample, while every proxy individually has little predictive power. The index is extracted by using the partial least squares, but the results are similar by the scaled principal component analysis. Moreover, the index can deliver sizable economic gains for mean-variance investors in asset allocation. The predictive power of the investor attention index stems primarily from the reversal of temporary price pressure and from the stronger forecasting ability for high-variance stocks.

JEL classifications: C22, C53, G11, G12, G17

Keywords: Investor Attention, Stock Return Predictability, Partial Least Square (PLS), Principal Component Analysis (PCA), Out-of-sample Forecast

I. Introduction

Attention is a scarce cognitive resource (Kahneman (1973)), and a growing body of research investigates its impact on cross-sections of stock prices, including Peng and Xiong (2006), Barber and Odean (2008), Dellavigna and Pollet (2009), Hou, Peng, and Xiong (2009), Da, Engelberg, and Gao (2011), Lou (2014), and Ben-Rephael, Da, and Israelsen (2017). Meanwhile, theoretical models, such as that of Peng and Xiong (2006), suggest that limited attention leads investors to focus on market- and sector-wide information more than on firm-specific information, implying a link between investor attention and market returns. However, there are limited empirical studies on the ability of investor attention to predict the aggregate stock market returns. Li and Yu (2012) and Yuan (2015) seem to be the only such studies; they find only in-sample evidence of predictability. Nonetheless, since Goyal and Welch (2008), researchers now focus on using out-of-sample to test the market predictability, which has not been addressed in the attention literature. In fact, we find that existing individual attention proxies have limited power in predicting the market in- and out-of-sample.

In this paper, we use collectively 12 individual attention proxies instead of individual ones, and show that their common component matters to the stock market and this component is well extracted by using the information aggregating methods of partial least squares (PLS), scaled principal component analysis (sPCA), and principal component analysis (PCA). Our paper makes three major contributions to the literature. First, we show for the first time that investor attention matters at the market level: it can strongly predict the stock market in- and out-of-sample when individual proxies are used collectively via the PLS, sPCA, and PCA approaches, and can yield sizable economic gains to mean-variance investors. Second, we show that investor attention is much more important than previously thought. If investor attention only influences the stock prices in a cross-section, its role is limited in the broad scope of finance. However, if it has an impact on the aggregate market, its role increases immensely. As Cochrane (2008) emphasizes, the market risk premium has a profound impact on asset pricing, corporate finance, and the entire economy and its predictability is one of the central issues in finance. However, existing studies do not provide sufficient evidence on the ability of investor attention in predicting the market. Our study does.

Third, similar to the investor sentiment index of Baker and Wurgler (2006), our study provides an investor attention index.[1] It captures related information in all individual proxies, making it a comprehensive measure of market-level attention. Thus, it can be used to examine the impact of investor attention in many contexts, such as in applications wherever the investor sentiment index is used. Hence, the impact of the aggregate investor attention index goes beyond its predictability on market risk premium.

In aggregating attention information, we select 12 popular individual attention proxies as components based on their real-time availability. They are abnormal trading volume (Barber and Odean (2008)); extreme returns (Barber and Odean (2008)); past returns (Aboody, Lehavy, and Trueman (2010)); nearness to 52-week high and nearness to historical high (Li and Yu (2012)); analyst coverage (Hirshleifer and Teoh (2003), Peng (2005), and Hirshleifer, Hsu, and Li (2013)); changes in advertising expenses (Lou (2014)); mutual fund inflow and outflow; media coverage (Barber and Odean (2008) and Fang and Peress (2009)); search-traffic on the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system (Lee, Ma, and Wang (2015), Drake, Roulstone, and Thornock (2015), and Drake, Jennings, Roulstone, and Thornock (2017)); and Google search volume (Da et al. (2011)).[2] Since most existing attention measures are at the firm level, we aggregate them first into measures at the market level whenever needed, and then aggregate further the individual market-level measures into a single index of aggregate investor attention.

Our primary aggregation method is PLS. As is the case for aggregating investor sentiment proxies in Huang, Jiang, Tu, and Zhou (2015), it is reasonable to assume that true investor attention is unobservable, and each individual measure is simply a proxy of it. Statistically, we need to extract the true attention that is related to stock returns from the proxies by removing all noises of the individual errors irrelevant to stock returns. As shown by Wold (1966), the pioneer of the PLS method, Kelly and Pruitt (2013, 2015), and Light, Maslov, and Rytchkov (2017), among others, PLS is an efficient method to obtain the aggregated attention from all individual proxies. The result (A^{PLS}) is one of our aggregated attention indices, which utilizes all information in individual

[1] See, e.g., Zhou (2018), for a review on investor sentiment. Recently, Jiang, Lee, Martin, and Zhou (2019) and Chen, Tang, Yao, and Zhou (2020) propose manager sentiment and employee sentiment indices, respectively.

[2] Ben-Rephael et al. (2017) use Bloomberg search as a measure of institutional attention. We exclude this measure, because the Bloomberg data sample (available since February 17, 2010) is too short for our purpose here.

proxies as well as that in the market return.

We also use PCA and the recently developed sPCA of Huang, Jiang, Li, Tong, and Zhou (2020). The PCA method extracts an index that explains the maximum variation of the proxies, not necessarily the returns. By design, PCA has a limitation in capturing the maximum information that is related to stock returns Kelly and Pruitt (2015). To better capture predictability, Huang et al. (2020) improve the PCA method by scaling each predictor according to its predictive power for future stock returns. Intuitively, their sPCA puts more weights on predictors that are more important in forecasting future returns. Thus, we have two alternative aggregate attention indices based on the PCA and sPCA approaches, denoted as A^{PCA} and A^{sPCA}, respectively.[3]

Using the PLS attention measure, A^{PLS}, as a single predictor, we find that the in-sample R^2 is 2.15%, with a highly significant slope of -0.64%, in the predictive regression of monthly excess returns of the stock market on A^{PLS} for the period from January 1980 to December 2017. This predictability exists up to two years, but the magnitude of regression slope shrinks with the increase in prediction horizons, indicating that the predictability effect weakens in the long run. We find similar evidence for the alternative two attention indices A^{sPCA} and A^{PCA}. The in-sample R^2 of A^{sPCA} for monthly market returns is 1.26%, with a regression coefficient of -0.49% which is statistically significant. The predictability effect becomes weak at the longer prediction horizons. A^{PCA} also predicts the market returns significantly across the forecasting horizons, except for the one-month horizon. In comparison with the individual attention proxies, our aggregate attention measures show stronger forecasting power for the stock market returns, suggesting that using the proxies collectively outperforms using them individually in terms of the return predictability.

Moreover, we compare the predictive power of aggregate investor attention with common return predictors, the economic variables used by Goyal and Welch (2008) and the investor sentiment index of the Baker and Wurgler (2006). We find that the aggregate investor attention measures maintain strong predictability after controlling for them. The results suggest that the aggregate investor attention contains unique forecasting information for the stock market, which cannot be explained by the economic fundamentals and investor sentiment.

[3] Recently, Da, Hua, Hung, and Peng (2020) propose attention measures by differentiating retail and institutional investors. They also use the PLS and PCA methods to construct the attention measures.

On critical out-of-sample assessment, we employ two evaluation metrics, Campbell and Thompson (2008)'s R^2_{OS} statistic and Clark and West (2007)'s *mean squared forecasting errors (MSFE)-adjusted* statistic, following studies in the predictability literature. The results show that all three aggregate attention measures deliver statistically significant R^2_{OS}'s across prediction horizons in the out-of-sample period from January 1995 to December 2017, except for A^{PCA} at the monthly horizons. Moreover, the magnitude of R^2_{OS}'s is economically sizable. The R^2_{OS}'s are 6.60%, 2.31%, and 2.39% per annum, respectively, for A^{PLS}, A^{sPCA}, and A^{PCA}. By contrast, we find that the individual attention proxies have limited predictability out-of-sample.

Whether significant predictability of investor attention can yield sizable economic gains is an important question. With the superior forecasting performance of aggregate investor attention, we show that they can indeed lead to sizable investment gains for a mean-variance investor from an asset allocation perspective. The annualized certainty equivalent return (CER) gains are 4.55%, 2.78%, and 5.00% at the annual horizons for A^{PLS}, A^{sPCA}, and A^{PCA} respectively, when the investor allocates investments between the market and risk-free rate. Moreover, investment portfolios based on the aggregate investor attention have large annualized Sharpe ratios. For example, A^{PLS} generates a Sharpe ratio of 0.74 at the monthly horizon, larger than that of the market portfolio, 0.50. Our asset allocation results are robust to a proportional transaction cost of 0.50%.

Our empirical findings are important for three reasons. First, they show, for the first time, that investor attention matters to the aggregate stock market both statistically and economically, highlighting its unrecognized significant role in asset pricing. Second, relying on any individual attention measure, the true predictive power of investor attention is likely to be understated. Instead, our aggregate investor attention uses all individual proxies collectively via the efficient aggregating approaches of PLS, sPCA, or PCA. The aggregated indices summarize the most relevant information in individual proxies, and therefore, they outperform the extant individual attention measures. Third, A^{PLS} can be used like the investor sentiment index of Baker and Wurgler (2006) for other applications.

To further understand why the aggregate investor attention predicts future market returns negatively, we explore possible underlying economic mechanisms. We find that the negative predictability primarily stems from the reversal of temporary price pressure. Barber and Odean (2008) and Da et al. (2011) argue that individual investors are net buyers of attention-grabbing stocks. The buying flow pushes up the price temporarily and this attention-driven price pressure reverts to fundamentals subsequently. Our empirical findings are consistent with their interpretations. We find that high investor attention increases the net buying, but this increase slows down at the subsequent month and diminishes in the long run. However, our results cannot rule out the possibility that net selling follows the high investor attention, as suggested by Yuan (2015). Moreover, attention can also be positively related to future returns, as shown by Gervais, Kaniel, and Mingelgrin (2001) empirically, and by Andrei and Hasler (2020) theoretically, over relatively short horizons and with stocks of high attention.

Cross-sectionally, we find that aggregate investor attention negatively predicts excess returns of stock portfolios sorted on market beta and idiosyncratic volatility. Thus, our results demonstrate that the negative return predictability is pervasive in the cross-section, consistent with our findings at the aggregate market level. Moreover, we find that there is large cross-sectional variation in predictability. The regression slope is more negative for high-beta stocks and for those with high idiosyncratic volatility. Han, Hirshleifer, and Walden (2020) document that investors tend to be attracted to high-variance stocks (high-beta stocks and stocks with high idiosyncratic volatility), pushing their prices upward and thereby depressing their expected returns. Our empirical findings are consistent with theirs.

The remainder of the paper is organized as follows. Section II describes the data and methodology for constructing the investor attention. Section III provides the empirical results. Section IV explores the economic source of the return predictability. Section V concludes.

II. Data and Investor Attention Construction

A. Individual Attention Proxies

We use 12 major individual attention proxies: abnormal trading volume (Barber and Odean (2008)); extreme returns (Barber and Odean (2008)); past returns (Aboody et al. (2010)); nearness to the Dow 52-week high and nearness to the Dow historical high (Li and Yu (2012)); analyst coverage

(Hirshleifer and Teoh (2003), Peng (2005), and Hirshleifer et al. (2013)); changes in advertising expenses (Lou (2014)); media coverage (Barber and Odean (2008), Fang and Peress (2009)); mutual fund inflow and outflow; Google search volume (Da et al. (2011)); and the search-traffic on EDGAR (Lee et al. (2015), Drake et al. (2015), and Drake et al. (2017)). We follow the literature in constructing these 12 attention proxies. Except for the nearness to the 52-week high and nearness to the historical high, we first construct the firm-level attention measures and next aggregate them to the market level.

The detailed construction is as follows:

- Abnormal trading volume (A^{AVol}): We first compute the ratio of trading volume at the end of each month to the average over the previous 1 year for each stock (NYSE/AMEX/NASDAQ). Then, we calculate the equal-weighted abnormal trading volume across all stocks as the market-level attention measure. We obtain the cross-sectional equity trading volumes from the Center for Research in Security Prices (CRSP) database for January 1980 to December 2017.

- Extreme returns (A^{ERet}): We first calculate the ratio of returns at the end of each month to the average over the previous 1 year for each stock (NYSE/AMEX/NASDAQ). Then, we calculate the equal-weighted extreme returns across all stocks as the market-level attention measure. We obtain the cross-sectional equity returns from the CRSP database for January 1980 to December 2017.

- Past returns (A^{PRet}): We define the past return as the monthly cumulative return over the prior 12 months for each stock (NYSE/AMEX/NASDAQ). Then, we calculate the equal weighted past return across all stocks as the measure for the aggregate stock market. We obtain the cross-sectional equity returns from the CRSP database for January 1980 to December 2017.

- Nearness to the Dow 52-week high (A^{52wH}) and nearness to the Dow historical high (A^{HisH}): Let p_t denote the monthly level of the Dow stock index. $p_{52w,t}$ and $p_{max,t}$ represent its 52-week (12-month) high and historical high, respectively. We define the monthly nearness to the Dow 52-week high as the ratio of the current Dow index at month t to its 52-week high, $x_{52w,t} = \frac{p_t}{p_{52w,t}}$, and the monthly nearness to the Dow historical high as the ratio of the current

Dow index at month t to its historical high, $x_{max,t} = \frac{p_t}{p_{max,t}}$. We obtain the Dow stock index from Yahoo Finance for January 1980 to December 2017.

- Analyst coverage ($A^{\#AC}$): We first count the total number of analyst 1-year ahead forecasts of earnings per share for each stock (NYSE/AMEX/NASDAQ) within a month. Then, we calculate the equal-weighted number of analyst's earnings forecasts across all stocks as the measure of the aggregate stock market. We obtain the number of analyst's earnings forecasts from the Institutional Brokers Estimate System (I/B/E/S) database for January 1980 to December 2017.

- Changes in advertising expenses (A^{CAD}): We first compute the changes in the log values of advertising expenditure from year $t-1$ to year t for each stock (NYSE/AMEX/NASDAQ).[4] We next calculate the equal-weighted changes in advertising expenses across all stocks as the monthly measure for the aggregate stock market. We obtain the advertising expenditure values from COMPUSTAT for January 1980 to December 2017.

- Mutual fund inflow (A^{Inflow}) and outflow ($A^{Outflow}$): We define the mutual fund inflow as the monthly total net asset value of shares sold, which includes new shares sold and other sales, for each fund. We define the mutual fund outflow as the monthly redemption of each fund. We then compute the equal-weighted mutual fund inflow and outflow, respectively, across all funds as the market-level measure. The mutual fund data are from the CRSP mutual fund database for January 2004 to December 2017.[5]

- Media coverage (A^{Media}): We define media coverage as the total number of news articles published on the Dow Jones Newswires during the month for each stock. We then calculate the average media coverage across all stocks as the measure for the aggregate stock market. We obtain the news data from the RavenPack database from January 2004 to December 2017.

- Google search volume (A^{Google}): We follow Da et al. (2011) and compute the monthly search

[4] We keep the change in advertising expenses the same as the previous year if a firm does not report its annual fundamentals in a given year.

[5] The data of mutual fund inflow and outflow are available since 2000. However, we use a shorter sample in order to reconcile with the attention measures that are available only since 2004, such as, the Google search volume.

frequency from Google Trends based on stock tickers. We then calculate the average Google search volume across all stocks as the market-level search volume. The data sample period runs from January 2004 to December 2017.

- Search-traffic on EDGAR (A^{EDGAR}): For each stock, we first count the number of EDGAR downloads for this firm's statements during a given month. We then calculate the average EDGAR downloads across all stocks as the market-level attention measure. The raw EDGAR file data are available to download at https://www.sec.gov/data/edgar-log-file-data-set.html. We follow Lee et al. (2015) and exclude the search records of all daily IPs that download more than 50 unique firms' filings. The sample period for the search records is from January 2004 to June 2017.

We aggregate the firm-level attention measures to the market level using the equal weighting method, which is also used by Rapach, Ringgenberg, and Zhou (2016) and Jondeau, Zhang, and Zhu (2019). Equal weighting is likely to be more informative than value weighting in aggregating the firm-level attention information, because it treats the attention information across a wide variety of firms equally. In contrast, value weighting places more emphasis on firms with large capitalization. Intuitively, when investors allocate attention to more stocks (large, mid, and small cap stocks), this more likely indicates that the investor attention allocated to the aggregate market increases. Thus, to avoid the domination of large cap stocks, we use equal weighting to aggregate the firm-level measure to the market level.

[Insert Table 1 and Table 2 about here]

Table 1 reports the median, quartile (75% and 25%) distributions, skewness, and first-order autocorrelation coefficient (ρ) of the 12 individual attention proxies. All attention variables are standardized to have mean of 0 and variance of 1. As the table shows, the values of median vary from -0.24 for A^{EDGAR} to 0.38 for A^{HisH}. A^{EDGAR} has the largest 75% quartile and the smallest 25% quartile among all variables. Table 2 provides the pairwise correlations among the attention proxies. We observe that most individual attention proxies are positively correlated, with several exceptions that have negligible negative values. The correlation coefficients range from -0.37 to 0.80, suggesting that these 12 attention proxies capture both common and different aspects of investor attention, and hence, using a specific proxy is unlikely to be complete in terms of the aggregate effect of investor attention on the stock market.

B. Aggregate Investor Attention

In this subsection, we tend to use the individual attention proxies collectively, unlike existing studies, which often examine one of them. We interpret true investor attention as an unobservable variable and any of the 12 attention measures as simply a proxy of the unobservable variable. Then, it is clearly desirable to extract the common component of the true attention by removing noises.

1. Factor Structure Model

We consider a forecasting model based on investor attention,

$$r_{t+1} = \alpha + \beta A_t^* + \varepsilon_{t+1} \ , \tag{1}$$

where r_{t+1} is realized excess stock return at time $t+1$, A_t^* is the true but unobservable investor attention at time t, and ε_{t+1} is a noise term that is unpredictable and unrelated to A_t^*. Model (1) implies that the true investor attention A_t^* is related to the subsequent stock return, which is consistent with the predictions of attention theories, such as those of Peng and Xiong (2006).

Next, we assume a linear factor structure for the attention proxies. Let $A_t = (A_{1,t}, \ldots, A_{N,t})'$ denote an $N \times 1$ vector of individual investor attention proxies at period t, N be the number of proxies, which is 12 in our case, and A_t correspond to the attention variables described in Subsection II.A. The structural model for $A_{i,t}$ $(i = 1, \ldots, N)$ is given by

$$A_{i,t} = \eta_{i,0} + \eta_{i,1} A_t^* + \eta_{i,2} E_t + e_{i,t} \ , \tag{2}$$

where A_t^* is the true but unobservable investor attention in model (1), $\eta_{i,1}$ is the factor loading that summarizes the sensitivity of attention proxy $A_{i,t}$ to the true attention A_t^*, E_t is the common approximation error component of all the proxies that are irrelevant to stock returns, and $e_{i,t}$ is the idiosyncratic noise associated with measure i only.

To determine the unique role of investor attention in the stock market, we tend to efficiently estimate A_t^*, the collective contribution to the true yet unobservable investor attention. The key idea here is to impose the factor structure (2) on the proxies to estimate A_t^*, and at the same time, to eliminate their common approximation error E_t and the idiosyncratic noise $e_{i,t}$ from the estimation process. To do so, we use three approaches: PCA, PLS, and sPCA. Consequently, we have three estimated aggregate investor attention indices, A^{PCA}, A^{PLS}, and A^{sPCA}, corresponding to the three approaches.

2. Principal Component Analysis (PCA)

The PCA is the simplest and most popular approach. It extracts the first principal component of $A_{i,t}$ as the aggregate attention measure that has the maximum representation of the total variations of the 12 individual attention proxies. By its econometric design, the PCA approach can separate A_t^* from $e_{i,t}$ and hence, capture the common attention information in individual attention proxies. This approach has been widely used in the literature on stock return predictability, such as in studies by Baker and Wurgler (2006), Ludvigson and Ng (2007), and Neely, Rapach, Tu, and Zhou (2014), among many others.

However, the major shortcoming of the PCA is that it may fail to eliminate the common measurement or observation errors (E_t) unrelated to the stock returns in individual attention proxies. In fact, it captures only the maximum common variations of predictors, and thus, incorporates the E_t into the estimation process as well. As Kelly and Pruitt (2013, 2015) show, the components that best describe the predictors' variation are not necessarily the most useful factors for forecasting. Thus, it is possible that PCA may fail to generate significant forecasts for future stock returns, even when stock returns are indeed strongly predictable by the true investor attention A_t^*. To overcome this econometric difficulty, we employ the PLS below, which is pioneered by Wold (1966) and further developed by Kelly and Pruitt (2013, 2015) and Light et al. (2017).

3. Partial Least Square (PLS)

The PLS approach extracts A_t^* from the individual attention proxies according to its covariance with future stock returns and chooses a linear combination of the attention proxies that is optimal for forecasting. In doing so, PLS can be implemented in the following two steps of ordinary least squares (OLS) regressions.

The first step is a time-series regression of each individual attention proxy at month t on the future realized excess stock return (as a proxy for expected excess returns), r_{t+1},

$$A_{i,t} = \pi_0 + \pi_i r_{t+1} + u_{i,t} \ , \tag{3}$$

where $A_{i,t}$ is the i-th individual attention proxy. The coefficient of π_i in the first-step time-series regression (3) captures the sensitivity of the attention proxy $A_{i,t}$ to the investor attention A_t^* instrumented by future stock return r_{t+1}. Because the future stock return r_{t+1} is driven by A_t^*, as shown in model (1), attention proxies are related to the predictable component of stock returns and are uncorrelated with the unpredictable errors. Therefore, the coefficient π_i approximately describes how each attention proxy depends on the true investor attention A_t^*.

The second-step regression is a cross-sectional regression for each time period t,

$$A_{i,t} = c_t + A_t^{PLS} \hat{\pi}_i + v_{i,t} \ , \tag{4}$$

where $\hat{\pi}_i$ is the loading estimated in regression (3) and A_t^{PLS}, the regression slope, is the PLS attention measure at time t. In the regression (4), the first-step regression loading becomes the independent variable, and A_t^{PLS} is the regression slope to be estimated.

Intuitively, PLS exploits the factor nature of the joint system, Equations (1) and (2), to infer the relevant attention factor A_t^{PLS}. If the true factor loading π_i were known, we could consistently estimate A_t^{PLS} by simply running cross-sectional regressions of $A_{i,t}$ on π_i period by period. However, because π_i is unknown, the first-stage regression slopes provide an approximate estimation of how $A_{i,t}$ depends on A_t^{PLS}. In other words, PLS uses time $t+1$ stock returns to discipline the dimension reduction to extract A_t^* relevant for forecasting, and discards the common and idiosyncratic components, such as E_t and $e_{i,t}$, which are irrelevant for forecasting.

As Kelly and Pruitt (2015) document, because the proxies may be measured with noise, the first-stage regression takes an errors-in-variables form and the second-stage regression produces an estimate for a unique but unknown rotation of the latent factor (A_t^* in our case). However, since the relevant factor space is spanned by the common component of proxies, a predictive regression of realized returns on the estimated PLS factor delivers consistent forecasts of expected returns driven by the latent factor.

In the empirical implementation, we use the full sample data from January 1980 to December 2017 to estimate the PLS attention index and investigate its in-sample return predictability. Specifically, in the time-series regression (3), we estimate the loadings (π_i) for A^{ERet}, A^{PRet}, A^{52wH}, A^{HisH}, A^{CAD}, A^{AVol}, and $A^{\#AC}$ from January 1980 to December 2017, A^{Inflow}, $A^{Outflow}$, A^{Media}, and A^{Google} from January 2004 to December 2017, and A^{EDGAR} from January 2004 to June 2017. In the second step, we run the cross-sectional regression (4) for each time t from January 1980 to December 2017 and estimate the A_t^{PLS} based on the available loadings π_i for each period. Thus, we obtain monthly PLS-based aggregate investor attention A_t^{PLS} from January 1980 to December 2017.

For the out-of-sample forecasting, the standard approach is to repeat these two steps by truncating the unknown observations at month t. Specifically, in the first step, the latest return that we can use on the right-hand side of regression (3) is r_t, and therefore, the last observation of the individual attention measure on the left-hand side of regression (3) is $A_{i,t-1}$. In the second step, we run the cross-sectional regressions for months 1 through t. In summary, for out-of-sample forecasting, we construct all inputs to the forecast using data observed no later than month t. Moreover, because under the mild assumption that the relationship between investor attention and expected stock returns is stable over time, the slope π_i can be estimated more precisely by using the averaging of π_i over all previous periods instead of only the most recent π_i, as suggested by Light et al. (2017).

4. Scaled Principal Component Analysis (sPCA)

In addition to the PCA and PLS approaches, we also use sPCA, which is recently proposed by Huang et al. (2020). As stated in Subsection II.B.2, while the PCA factor maximally represents the total variations of predictors, it ignores the forecasting target and therefore, is an unsupervised learning technique for dimension reduction. In contrast, sPCA is designed to use the target information to guide dimension reduction.

The sPCA is implemented in two steps. First, a panel of scaled attention predictors, $(\beta_1 A_{1,t}, \ldots, \beta_N A_{N,t})$, is constructed, where the scaled coefficient β_i $(i = 1, \ldots, N)$ is the slope from the predictive regression of the realized stock excess returns (r_{t+1}) on the i-th attention proxy $(A_{i,t})$,

$$r_{t+1} = \alpha_i + \beta_i A_{i,t} + \varepsilon_{t+1} \,. \tag{5}$$

In the second step, the conventional PCA is applied to $(\beta_1 A_{1,t}, \ldots, \beta_N A_{N,t})$, the panel of scaled predictors. Then, the first principal component is the sPCA-based aggregate investor attention, A^{sPCA}. For out-of-sample forecasting, like the implementation of the PLS approach, we recursively estimate the regression (5) using the data observed no later than month t.

Intuitively, the scaled series $\beta_i A_{i,t}$ reflects the i-th attention proxy's predictive power on the future returns. A proxy with strong forecasting power receives a larger weight (i.e., higher absolute value of β_i), whereas a predictor with weak forecasting power receives a smaller weight. In summary, the sPCA performs the PCA on the scaled attention proxies, rather than on the raw proxies.

[Insert Figure 1 about here]

Figure 1 displays the time series of the three attention indices, A^{PCA}, A^{PLS}, and A^{sPCA}, from January 1980 to December 2017. We observe that the aggregate investor attention indices, measured by PLS, PCA, and sPCA, are time varying for our sample from January 1980 to December 2017. In general, they decrease in the economic recessions, in line with the empirical finding from Sicherman, Loewenstein, Seppi, and Utkus (2016) who show that investor attention falls by 9.5%

after the market declines. This phenomenon, called "selective attention" or the *ostrich effect*, is introduced by Karlsson, Loewenstein, and Seppi (2009).

III. Empirical Results

A. Forecasting Stock Market Returns

In this section, we explore the forecasting power of aggregate investor attention for the stock market excess return, which is defined as the difference between the value-weighted aggregate stock return and T-bill rate from the CRSP database. The univariate predictive regression is

$$R_{t+h} = \alpha + \beta A_t + \varepsilon_{t+h} \ , \tag{6}$$

where R_{t+h} is the average stock market excess return over the prediction horizon h, $h = 1$, 3, 6, 12, and 24 months, and A_t is one of the aggregate attention indices, A^{PLS}, A^{sPCA}, and A^{PCA}, constructed by the PLS, sPCA, and PCA approaches, respectively. We test the in-sample predictive ability of A_t by estimating the regression (6) for January 1980 to December 2017. Specifically, we inspect the estimate of β ($\hat{\beta}$) in regression (6). The null hypothesis is that A_t has no predictive ability; that is, $\beta = 0$, and regression (6) reduces to the constant expected return model ($R_{t+1} = \alpha + \varepsilon_{t+1}$). Under the alternative hypothesis, β is different from zero, and A_t contains information useful for predicting R_{t+1}. A time-varying expected stock return model applies. We use the Hodrick (1992) standard error and the Newey and West (1987) standard error, respectively, to compute the t-statistic corresponding to $\hat{\beta}$.[6]

[Insert Table 3 about here]

Panel A of Table 3 reports the forecasting results for A^{PLS}. We observe that A^{PLS} significantly predicts market excess returns and this predictability persists up to one year. More specifically,

[6] In studying the predictability over longer horizons, the Newey-West test can seriously overreject in finite samples due to the interaction between the persistent regressor and serially correlated errors. The Hodrick standard error, which uses the moving-average structure of the aggregated error, performs better (Ang and Bekaert (2007)).

at the monthly horizon, the β estimate is -0.64% with a t-statistic of -2.66 (-2.85) based on the Hodrick (Newey-West) standard error. For longer prediction horizons, although the β estimate is still negative, it shrinks to -0.21% in magnitude at the two-year horizon. Thus, the return predictability becomes weaker in the long run.

Theoretically, the sign of the β coefficient in Eq. (1) is not conclusive. On the one hand, Barber and Odean (2008) argue that individual investors are net buyers of attention-grabbing stocks, and consequently high attention leads to contemporaneous positive price pressure and thus lower future returns, which is consistent with Peng and Xiong (2006). On the other hand, Gervais et al. (2001) find that attention, as captured by trading volume, is positively related to stock's visibility, which can increase the stock value, and their empirical evidence is consistent with Andrei and Hasler (2020) whose model allows for either positive or negative slopes (depending in general on the news relative to its mean). However, their study focuses on daily and weekly data and on high attention stocks only, but we focus on monthly market returns. It appears that prices can move in one direction in the short run and in the opposite direction in the long run, so their results are different from ours. To further strengthen the economic explanation, we, in Section IV, provide additional analysis, which is consistent with Da et al. (2011), to support the argument of Barber and Odean (2008) for the economic driving forces for our results.

Economically, the magnitude of β estimate is sizable. Because we standardize all predictors to have zero mean and unit variance, our result for the monthly horizon implies that a one-standard deviation increase in A^{PLS} leads to a 0.64% decrease in the next month's expected stock market return. If we annualize this size, it equals 7.68%, which is comparable with conventional macroeconomic predictors. For example, a one-standard deviation increase in the dividend–price ratio, the consumption–wealth ratio, and the net payout ratio tend to increase the risk premium by 3.60%, 7.39%, and 10.2% per annum, respectively (see, e.g., Lettau and Ludvigson (2001), Boudoukh, Michaely, Richardson, and Roberts (2007)).

In addition, the regression R^2 provides another metric to evaluate the economic significance of the forecasting ability of A^{PLS}. At the monthly horizon, the in-sample R^2 equals 2.15%, which is economically large. Our result implies that A^{PLS} can explain the time variation of monthly market excess returns by 2.15%. With the increase in prediction horizon, the R^2 peaks at the annual horizon, with a value of 7.65%, and subsequently declines to 5.62% at the two-year horizon.

Panel B reports the forecasting results for A^{sPCA}. We find that like A^{PLS}, A^{sPCA} has negative predictive power for market excess returns. Specifically, the regression slope is -0.49% at the monthly horizon, which is statistically significant with a t-statistic of -2.43 (-2.29) based on the Hodrick (Newey-West) standard error. The coefficient estimate remains significant up to one year, but its absolute value maximizes at the quarter horizon and thereafter decreases with the prediction horizon. Thus, consistent with findings from Panel A, results for A^{sPCA} also suggest that the return predictability weakens in the long run. Moreover, the in-sample R^2 at the monthly horizon, 1.26%, is economically sizable, and it maximizes at the quarter horizon, with a value of 4.86%.

We find similar evidence for A^{PCA} in Panel C. The coefficients of A^{PCA} are negative across prediction horizons and statistically significant except for the estimate at the monthly horizon. The return predictability effect is strong at the quarter horizon, becomes slightly weaker at the semi-annual and annual horizons, and largely diminishes at the two-year horizon, in line with findings from Panels A and B. In addition, we observe that basically, the β estimates of A^{PCA} are smaller in magnitude than those of A^{PLS} and A^{sPCA} from Panels A and B. It is plausible that PCA is a less efficient method than PLS and sPCA in aggregating the information from the individual attention proxies. PCA finds only the best predictor representing the variation of the individual attention proxies, and it cannot effectively remove the common noise from the proxies, as stated in Section II.B.

As a comparison, we examine the forecasting abilities of the 12 individual attention proxies for future market excess returns. Table IA.1 of Internet Appendix presents the results. We observe that only two proxies (A^{HisH} and A^{Inflow}) can predict the future market excess returns negatively and significantly at the monthly horizon. The number of predictors that have significant forecasting power increases to four at the annual horizon. Our results demonstrate that individual attention proxies have limited return predictability. As Section II demonstrates, the noises in measuring the market-wide attention are likely to impair their abilities to predict the future market returns. Thus, relying on a single proxy fails to explore the aggregate effect of investor attention on the stock market. Evidently, it is desirable to extract the common component from the individual attention proxies by largely removing noises. We meet this objective by using the PLS, sPCA, and PCA approaches, and find the evidence of strong predictability in Table 3.

In summary, we show that the aggregate investor attention indices constructed by PLS, sPCA, and PCA (A^{PLS}, A^{sPCA}, and A^{PCA}) exhibit statistically and economically significant in-sample predictive power for the future market excess returns. This predictability persists up to one or two years. Our finding suggests that investor attention indeed plays an important role in the aggregate stock market, which is consistent with the prediction of attention theories and complements extant empirical studies that find an impact of investor attention only on cross-sectional stock returns. Our aggregate attention measure outperforms the individual attention proxies in predicting the stock market, because it captures the most relevant information in true investor attention from the individual proxies by removing noises that may impair the aggregate effect of investor attention on the market.

B. Comparison with Economic Variables

Our compelling evidence shows the strong predictability of aggregate investor attention indices. We further examine whether the forecasting information comes from the business cycle-related fundamentals. To address this issue, we control for a set of economic variables commonly used by the forecasting literature. The predictive regression is,

$$R_{t+h} = \alpha + \beta A_t + \phi \mathbf{X}_t + \varepsilon_{t+h} \, , \tag{7}$$

where R_{t+h} is the average stock market excess return over the prediction horizon h, A_t is one of the attention indices A^{PLS}, A^{sPCA}, and A^{PCA} at time t, and $\mathbf{X}_t$ represents a vector of economic variables from Goyal and Welch (2008). Goyal and Welch (2008) suggest 14 economic variables and the data is available from Amit Goyal's website, http://www.hec.unil.ch/agoyal/. Appendix A shows a description of these 14 variables. Using all variables together in one regression may result in the multicollinearity problem. Thus, in our model specification, we use 8 of them, including DP, DY, EP, BW, SVAR, LTR, TMS, and DFY. We find our results robust to use alternative regression specifications in Table IA.2 of Internet Appendix.

[Insert Table 4 about here]

Panel A of Table 4 reports the estimation results for A^{PLS}. We find that the regression slopes on A^{PLS} remain statistically significant after controlling for the economic variables, suggesting that the impact of investor attention on aggregate stock market cannot be explained by the economic fundamentals. In addition, the coefficient estimates are large in magnitude. For example, at the monthly horizon, the β estimate of A^{PLS} is -0.90%, indicating the economic significance. The magnitude shrinks with the increase in prediction horizon. Also large are the regression R^2's. The adjusted R^2 increases from 5.43% at the monthly horizon to 13.99% at the annual horizon. Then, combining investor attention with economic predictors can generate strong forecasting power for the aggregate stock market.

In Panels B and C, we observe similar results for A^{sPCA} and A^{PCA}. After controlling for economic variables, A^{sPCA} and A^{PCA} still predict future stock market returns significantly, except for A^{PCA} at the monthly horizon. The regression slopes are economically sizable and their absolute values decrease with the horizon. Moreover, incorporating investor attention index A^{sPCA} (or A^{PCA}) into the regressions based on economic variables deliver large in-sample R^2's, which reach 13.63% (13.80%) per annum.

C. Comparison with Investor Sentiment

We next compare the aggregate investor attention with the sentiment-related predictor in term of forecasting ability for market returns. On the one hand, Da et al. (2011) argue that because attention is a necessary condition for generating sentiment, increased investor attention, especially that coming from "noise" traders prone to behavioral bias, likely leads to stronger sentiment. On the other hand, increased attention to genuine news may increase the rate at which information is incorporated into prices and thus, may attenuate sentiment. Our analysis in this subsection is important for understanding the unique role of investor attention in predicting the market.

We employ the investor sentiment index of Baker and Wurgler (2006) (S^{BW}), which has been widely used by the literature, such as Baker and Wurgler (2007), Yu and Yuan (2011), Baker, Wurgler, and Yuan (2012), Stambaugh, Yu, and Yuan (2012), and others. The data is available from

Jeffrey Wurgler's website, http://people.stern.nyu.edu/jwurgler/. To compare investor attention and sentiment, we first compute their correlations. The correlation coefficient between A^{PLS} and S^{BW} is 0.37, that between A^{sPCA} and S^{BW} is 0.04, and that between A^{PCA} and S^{BW} is 0.01. The low correlation coefficients imply that investor attention contains information distinct from that of investor sentiment.

[Insert Table 5 about here]

We next analyze the incremental forecasting power of investor attention after controlling for sentiment, based on the following predictive regression,

$$R_{t+h} = \alpha + \beta A_t + \phi S_t^{BW} + \varepsilon_{t+h} \,, \tag{8}$$

where R_{t+h} is the average stock market excess return over the prediction horizon h, A_t is one of the investor attention indices A^{PLS}, A^{sPCA}, and A^{PCA} at time t, and S_t^{BW} represents the Baker and Wurgler (2006) investor sentiment index at time t. Table 5 reports the estimation results. We observe that after controlling for the investor sentiment S^{BW}, the regression slopes on investor attention indices A^{PLS}, A^{sPCA}, and A^{PCA} remain statistically significant across prediction horizons, except for A^{PCA} at the monthly horizon. This finding implies that investor attention contains unique information in forecasting the stock market, which complements the sentiment predictor. In addition, our results demonstrate that using investor attention and investor sentiment jointly in one regression can generate powerful return predictability. The in-sample R^2 per annum is as large as 9.30% for A^{PLS} and also sizable for A^{sPCA} and A^{PCA}. Thus, understanding the impact of investor attention on the market is meaningful, because it contains information distinct from investor sentiment.

D. Out-of-sample Performance

Even though the in-sample analysis provides parameter estimates that are more efficient and thus, more precise return forecasts by utilizing all available data, Goyal and Welch (2008), among others, argue that out-of-sample tests seem to be more relevant for assessing genuine return predictability in real time. In this subsection, we evaluate the out-of-sample forecasting power of aggregate investor attention for market excess returns.

Basically, we start with an initialization period to estimate the predictive regression (6) based on each attention measure to produce the first out-of-sample forecast. The forecast return is

$$\widehat{R}_{t+h} = \widehat{\alpha}_t + \widehat{\beta}_t \, A_t \ , \tag{9}$$

where $\widehat{\alpha}_t$ and $\widehat{\beta}_t$ are the OLS estimates of regression (6). We recursively estimate regression (6) and repeatedly construct the monthly out-of-sample forecasts according to Equation (9) for the following periods, until we reach the end of the sample period. Moreover, following Campbell and Thompson (2008) and Pettenuzzo, Timmermann, and Valkanov (2014), we impose an economic restriction on forecast returns, that the expected risk premium must be positive to be consistent with theory.

In the empirical implementation, we use the initial period of January 1980 to December 1994 and therefore the out-of-sample forecast evaluation period spans from January 1995 to December 2017. We choose the length of the initial in-sample estimation period so that the observations are enough to estimate the initial parameters precisely and the out-of-sample period is relatively long to evaluate the forecast.[7] Importantly, as stated in Section II, we construct the month-t aggregate investor attention (A_t^{PLS}, A_t^{sPCA}, or A_t^{PCA}) using the available data observed no later than this month to predict the month-$t+1$ return out-of-sample. In addition, when constructing the PLS investor attention out-of-sample, we use the averaging of π_i, the slope in the first-step regression (3), over all previous periods, as suggested by Light et al. (2017). Our results are robust to using the most recent π_i estimate and alternative averaging schemes, like the averaging over past 5 or 10 years.

To evaluate the out-of-sample performance, we employ the common Campbell and Thompson (2008)'s R^2_{OS} and Clark and West (2007)'s *MSFE-adjusted* statistic. The R^2_{OS} measures the proportional reduction in mean squared forecast error (MSFE) for the predictive regression forecast vis-á-vis the benchmark forecast. When $R^2_{OS} > 0$, the predictive regression forecast outperforms the benchmark forecast in terms of MSFE. The prevailing benchmark is the average excess return from

[7] Barbara and Inoue (2012) and Hansen and Timmermann (2012) show that the out-of-sample tests of predictive ability have better size properties when the forecast evaluation period is a relatively large proportion of the available sample, as in our case.

the beginning of the sample through month t. This forecast corresponds to the constant expected excess return model, Equation (6) with $\beta = 0$, and implies that returns are not predictable, as in the canonical random walk with drift model for the log of stock prices. To ascertain whether the predictive regression forecast delivers a statistically significant improvement in MSFE, we use Clark and West (2007)'s *MSFE-adjusted* statistic to test the null hypothesis that the historical average MSFE is less than or equal to that of the predictive regression forecast against the alternative hypothesis that the historical average MSFE is greater than that of the predictive regression forecast, corresponding to H_0: $R^2_{OS} \leq 0$ against H_A: $R^2_{OS} > 0$.

[Insert Table 6 about here]

Table 6 presents the out-of-sample results. We find that all three aggregate attention indices A^{PLS}, A^{sPCA}, and A^{PCA} generate positive R^2_{OS}'s, which are statistically significant according to the *MSFE-adjusted* statistics, except for A^{PCA} at the monthly horizon. Then, our results demonstrate that the MSFE of out-of-sample forecasts generated by aggregate investor attention is significantly lower than that of the historical average. In addition, the magnitude of R^2_{OS}'s is economically sizable. For example, the R^2_{OS} of A^{PLS} equals 2.04% at the monthly horizon, and increases to 6.60% at the annual horizon. Owing to a large unpredictable component inherent in stock returns, the R^2's of stock return forecasts are typically small. Campbell and Thompson (2008) argue that a monthly out-of-sample R^2 of 0.5% can generate significant economic value. Clearly, the R^2_{OS}'s of aggregate attention indices are much larger than 0.5%, suggesting substantial economic significance (Kandel and Stambaugh (1996)). We will analyze this issue at the next subsection.

As a comparison, we examine the out-of-sample performance of individual attention proxies. Due to data constraints, we only show results in Table IA.3 of the Internet Appendix for the 7 proxies (A^{AVol}, A^{ERet}, A^{PRet}, A^{52wH}, A^{HisH}, $A^{\#AC}$, and A^{CAD}) from January 1995 to December 2017. We observe that only three of the 7 proxies, A^{52wH}, A^{HisH}, and A^{CAD}, deliver positive and significant R^2_{OS}'s, 1.12%, 2.49%, and 2.40%, respectively, at the annual horizon. Nonetheless, the values are still smaller than that of A^{PLS} (6.60%). Thus, consistent with our conclusion for the in-sample results, relying on a single proxy tends to underestimate the predictive power of aggregate investor attention for the market.

To summarize, Table 6 reveals that aggregate investor attention, constructed by the PLS, sPCA, or PCA methods, displays strong out-of-sample forecasting power for stock market returns. In contrast, few of the individual attention proxies can significantly predict the out-of-sample market returns. Our empirical findings are of great importance to the attention literature. First, they demonstrate for the first time that investor attention can predict the aggregate stock market out of sample, like its impact on cross-sectional returns, which emphasizes the unrecognized role of investor attention in asset pricing. Second, the predictive power of investor attention for the market would be understated without efficient information aggregation.

E. Asset Allocation Analysis

Given the strong predictive power of aggregate investor attention, its economic value is still unclear. It would be of interest to find out whether aggregate investor attention generates substantial economic value for investors if they utilize the forecasting information of aggregate investor attention rather than totally ignore it. Then, in this subsection, we assess the economic value from an asset allocation perspective.

Following Kandel and Stambaugh (1996), Campbell and Thompson (2008), and Ferreira and Santa-Clara (2011), we consider a mean-variance investor who uses return forecasts to make asset allocation decisions across risky stocks and risk-free bills. The investor rebalances the portfolio at the end of the next month. The weights of equities in the portfolio are determined by

$$w_t = \frac{1}{\gamma} \frac{\widehat{R}_{t+1}}{\widehat{\sigma}^2_{t+1}} \,, \tag{10}$$

where γ is the degree of risk aversion, $\widehat{R}_{t+1}$ is the out-of-sample forecast of stock excess return, and $\widehat{\sigma}^2_{t+1}$ is the forecast of its variance. Similarly to Campbell and Thompson (2008), we assume that the investor uses a 5-year moving window of past returns to estimate the variance of future stock returns. In addition, we restrict w_t to lie between 0 and 1.5 to exclude short sales and have 50% leverage at most.

The investor allocates $1 - w_t$ of the portfolio to risk-free bills, and the realized portfolio return

(R^p_{t+1}) at time $t+1$ is

$$R^p_{t+1} = w_t R_{t+1} + R^f_{t+1} \,, \tag{11}$$

where R^f_{t+1} is the risk-free return. The certainty equivalent return (CER) of the portfolio is

$$CER_p = \widehat{\mu}_p - 0.5\,\gamma \widehat{\sigma}^2_p \,, \tag{12}$$

where $\widehat{\mu}_p$ and $\widehat{\sigma}^2_p$ are the sample mean and variance, respectively, of the investor's portfolio over the forecast evaluation period. We can interpret CER as the risk-free return that an investor is willing to accept instead of holding the risky portfolio. The CER gain is the difference between the CER for the investor who uses a predictive regression forecast of monthly returns generated by Equation (9) and that for an investor who uses the historical average forecast. We multiply this difference by 12 so that we can interpret it as the annual portfolio management fee that an investor would be willing to pay to access the predictive regression forecast. In addition to the CER gain, we also compute the annualized Sharpe ratios of portfolio R^p_t to evaluate the investment performance. Considering the existence of transaction cost in real investment, we check the robustness of our asset allocation results after deducting a proportional transaction cost of 50 basis points. In this way, we measure the direct economic value of return predictability.

To analyze the economic value of return predictability at longer horizons, we follow Rapach et al. (2016) and assume that the investor rebalances at the same frequency as the forecast horizon. For the quarterly horizon, at the end of the quarter, the investor uses a predictive regression or historical average forecast of the excess return over the next three months ($h = 3$) and the allocation rule given by Equation (10) to determine the stock weight for the next three months; at the end of the next quarter, the investor updates the quarterly predictive regression or historical average forecast and determines the new weight (so that the investor uses nonoverlapping return forecasts). The investor follows analogous procedures for semi-annual and annual return forecasts and rebalancing.

[Insert Table 7 about here]

Table 7 reports the asset allocation results for the out-of-sample period from January 1995 to December 2017. We assume a risk aversion coefficient of five and find our results robust to alternative reasonable coefficient values. We observe that the return forecasts of aggregate investor attention generate extremely sizable investment profits across prediction horizons, except for A^{PCA} at the monthly horizon. More specifically, the CER gain of A^{PLS} is 3.99% at the monthly horizon, implying that an investor would be willing to pay an annual fee of up to 399 basis points (bps) to access the predictive regression forecasts of A^{PLS}. This large economic value also exists at the quarter, semi-annual, and annual horizons. Similarly to A^{PLS}, A^{sPCA} and A^{PCA} also generate substantial economic values. The CER gain of A^{sPCA} is 3.11% at the monthly horizon and still sizable at longer horizons, although it slightly decreases to 2.78% at the annual horizon. A^{PCA} generates a maximum CER gain of 5.00% at the annual horizon, suggesting large investment profits. After considering the transaction cost of 0.5%, the economic value remains sizable. The net-of-transaction-cost CER gains of A^{PLS} (A^{sPCA}) range from 3.04% to 4.41% (1.93% to 2.70%) across horizons. Also economically large is that of A^{PCA} at the annual horizon, which is 4.96%.

In addition, the investment portfolio based on aggregate investor attention generates remarkably large Sharpe ratios. As Table 7 shows, the annualized Sharpe ratio is 0.74 for A^{PLS} at the monthly horizon, while the market has a Sharpe ratio of 0.50. After deducting the 50 bps transaction cost, it is 0.67, which is still economically large. Thus, our result indicates that the market timing strategy based on investor attention A^{PLS} outperforms the naive buy-and-hold strategy. In the long run, the Sharpe ratio without transaction cost (with 50 bps transaction cost) decreases to 0.43 (0.42) at the annual horizon. We find similar results for A^{sPCA} and A^{PCA}. Investment portfolios based on these two alternative attention measures also deliver substantial Sharpe ratios, ranging from 0.38 to 0.67 (0.36 to 0.48) for A^{sPCA} (A^{PCA}). These results are robust to the 50 bps transaction cost.

In summary, there are potentially large investment profits in the asset allocation based on aggregate investor attention, suggesting substantial economic values for mean-variance investors. This analysis then emphasizes the important role of investor attention on the aggregate stock market from an asset allocation perspective.

IV. Economic Explanation

Our empirical results show that high investor attention predicts a subsequent low stock return. In this section, we explore the possible economic sources of the negative return predictability.

A. Investor Attention and Aggregate Order Imbalance

Barber and Odean (2008) argue that when investors search stocks to buy, they have to select from thousands of candidates. However, when they select those to sell, they can only sell what they already have. Hence, investors are more likely to buy those attention-grabbing stocks, which results in temporary positive price pressure. Da et al. (2011) find supporting evidence that a positive abnormal google search volume, a proxy of investor attention, predicts higher stock prices in the next two weeks and this impact weakens thereafter. Most importantly, the price pressure tends to revert in the fourth week and is almost completely reversed in one year.

According to Barber and Odean (2008) and Da et al. (2011), the negative predictability of our investor attention indices might come from the reversal of temporary price pressure. High attention may result in net buying flow of individual investors, which pushes up the price. The temporary price pressure reverts to fundamentals in subsequent periods. Consequently, the high investor attention proceeds the future low stock returns.

To exploit this interpretation, we test the relationship between aggregate order flow and investor attention index. We follow Lee and Ready (1991) and Barber and Odean (2008) and define the monthly order flow at firm level as,

$$OF_{i,t} = \frac{TNB_{i,t} - TNS_{i,t}}{TNB_{i,t} + TNS_{i,t}}, \tag{13}$$

where $OF_{i,t}$ is buy-sell imbalance for firm i at month t, $TNB_{i,t}$ is the total number of purchase of stock i within month t, and $TNS_{i,t}$ is the total number of sales of stock i within month t. We compute $OF_{i,t}$ by using the tick-by-tick transaction data from the Trade and Quote database (TAQ) over the time from 1993 to 2017. The aggregate market-level order flow AOF_t is the value-weighted $OF_{i,t}$. We use the following regression to examine the impact of investor attention on aggregate buy-sell imbalance,

$$\Delta AOF_{t+h} = \alpha + \beta_1 A_t + \beta_2 Ret_t + \beta_3 Ret_{t-1} + \beta_4 Ret_{t-2} + \varepsilon_{t+h} \,, \tag{14}$$

where ΔAOF_{t+h} is the change in AOF_t over the period h, and $h = 0$, 1, 6, and 12 months; A_t represents one of the attention indices: A^{PLS}, A^{sPCA}, and A^{PCA}; and Ret_t, Ret_{t-1}, and Ret_{t-2} are stock market returns at time t, $t-1$, and $t-2$, respectively. $h = 0$ refers to a contemporaneous relationship between ΔAOF_t and A_t.

[Insert Table 8 about here]

Panels A, B, and C of Table 8 report the estimation results for A^{PLS}, A^{sPCA}, and A^{PCA}, respectively. In Panel A, we find evidence that A^{PLS} strongly influences the change in aggregate order flow when $h = 0$. The regression slope on A^{PLS} is positive at 0.12% with a t-statistic of 2.71, implying that high investor attention significantly increases the net buying. This is consistent with findings of Barber and Odean (2008) and Da et al. (2011). We find similar evidence for A^{sPCA} and A^{PCA} in Panels B and C. The coefficient of A^{sPCA} (A^{PCA}) from the contemporaneous regression is 0.05% (0.02%) with a t-statistic of 3.75 (1.73), supporting the interpretation that net buying flow of individual investors is a source of the negative predictability of attention indices.

Since the negative return predictability of aggregate attention index may stem from the future price reversal, we then expect that the impact of A_t on changes in order imbalance should become reverse at the subsequent month, indicating that the increase in net buying slows down following high investor attention. Our results from Table 8 support this conjecture. In Panel A, the coefficient estimate of A^{PLS} from the monthly predictive regression is -0.14% with a t-statistic of -2.99, indicating that the net buying stops increasing at subsequent period $t+1$ and hence the temporary price pressure tends to revert. At longer horizons, the impact of A^{PLS} on buy-sell imbalance completely diminishes. Our finding is consistent with the interpretation of Da et al. (2011) that there should be a price reversal in the long run if attention-driven net buying results in a temporary price pressure. In addition, our results cannot rule out the possibility of selling pressure in subsequent periods. Yuan (2015) finds that investors are likely to sell more stocks following attention-grabbing events because of the disposition effect and rebalance needs. Our results for A^{sPCA} and A^{PCA} from

Panels B and C are analogous to those from Panel A. When $h = 1$, the regression slope on A^{sPCA} (A^{PCA}) is negative at -0.13% (-0.07%) with a t-statistic of -3.02 (-1.64), suggesting that the net buying decreases at the subsequent month $t+1$.

In summary, results from Table 8 demonstrate that the aggregate investor attention reflects a strong link to the trading behavior of individual investors. High attention indicates significantly increases in the net buying flow of individual investors in aggregate, resulting in a temporary price pressure. Subsequently, this net buying flow slows down and as a result, the price tends to revert. The above evidence provides a possible economic mechanism through which investor attention predicts future market returns negatively.

An alternative explanation is that high attention may indicate the desire of investors to obtain information about firm fundamentals, especially during news release time such as an earnings announcement. The point is that, empirically, it is very difficult to distinguish between the effects of attention and the efforts on information acquisition. Nevertheless, since more information acquisition is likely to reduce information asymmetry, lowering the risk and hence lowering the risk premium. In this case, more attention is likely to predict low future returns too.

B. Predictability across Characteristic-based Portfolios

Our above evidence shows that the negative return predictability of investor attention primarily stems from the reversal of temporary price pressure in the long run. In this subsection, we explore the cross-sectional variation in aggregate investor attention's effects on stock returns.

Han et al. (2020) argue that high-variance stocks (high-beta stocks and stocks with high idiosyncratic volatility) are more likely to attract investor's attention. We then consider portfolios sorted by these two characteristics. We obtain the value-weighted returns of 10 market beta-sorted portfolios from the website of Kenneth R. French. Additionally, we compute the value-weighted returns of 10 portfolios sorted on idiosyncratic volatility, which is defined as the residual volatility from regressing a stock's daily excess returns on market returns over the prior year. To test the cross-sectional predictability, we estimate the following predictive regression,

$$R^i_{t+h} = \alpha^i + \beta^i A_t + \varepsilon^i_{t+h} \,, \tag{15}$$

where R^i_{t+h} represents the average excess returns of each portfolio over the horizon h, and A_t is one of the aggregate investor attention indices, A^{PLS}, A^{sPCA}, and A^{PCA}.

[Insert Figure 2 about here]

Figure 2 plots the coefficient estimates (β^i) of the above regressions for characteristic decile portfolios over the sample period from January 1980 to December 2017. Results show that most coefficient estimates are negative across the portfolios, suggesting that the negative return predictability of aggregate investor attention (A_t) is pervasive in the cross-section, consistent with our findings at the aggregate market level. More importantly, we detect large cross-sectional variation in the regression slope estimates. The slope is more negative for high-beta stocks and those with high idiosyncratic volatility. Our finding strongly supports the argument of Han et al. (2020) that investors tend to be attracted to high-variance stocks, pushing their prices upward and thereby depressing their expected returns.

Overall, our analysis for the characteristic-based portfolios shows that aggregate investor attention has strong and negative predictive power pervasively across portfolios. This predictability varies cross-sectionally. It is particularly strong for high-variance stocks, which helps us to better understand the negative predictive power of investor attention for the stock market.

V. Conclusion

In this study, we investigate, for the first time, the collective predictive power of investor attention measures for the aggregate stock market; this is in contrast to existing studies, which mainly focus on cross-section predictability and on the use of individual predictors. We aggregate individual investor attention measures by using three approaches: PLS, PCA, and an improved PCA approach, sPCA. We find that the aggregated investor attention indices predict the subsequent monthly stock market returns negatively and significantly. However, this predictability becomes weaker with the increase in prediction horizons. In contrast, individual attention proxies have limited return predictability. The predictive power of our attention indices is greater than that of using common return predictors and is still present after controlling for investor sentiment. Moreover, the strong predictability exists out-of-sample and hence delivers sizable economic value for mean-variance investors in asset allocation.

Our study highlights the important role of investor attention in the stock market and in finance in general. We identify two economic sources of the negative predictability. The predictive power of aggregate investor attention for stock market is likely derived from the reversal of temporary price pressure caused by net buying and from the stronger power for high-variance stocks. Future research may extend our information aggregation approach to other asset markets or other countries, and may also apply aggregate investor attention wherever the investor sentiment index is applied. While the purpose of the current study is to show that investor attention matters for the market risk premium, it would be of interest to investigate how its predictive power could be further improved with machine learning tools (see, e.g., Gu, Kelly, and Xiu (2020) and Rapach and Zhou (2019)).

Appendix A. Detailed Description of Economic Variables

In the robustness check, we control for the following 14 economic variables of Goyal and Welch (2008).

- Dividend–price ratio (log), DP: log of a 12-month moving sum of dividends paid on the S&P 500 index minus the log of stock prices (S&P 500 index).
- Dividend yield (log), DY: log of a 12-month moving sum of dividends minus the log of lagged stock prices.
- Earnings–price ratio (log), EP: log of a 12-month moving sum of earnings on the S&P 500 index minus the log of stock prices.
- Dividend–payout ratio (log), DE: log of a 12-month moving sum of dividends minus the log of a 12-month moving sum of earnings.
- Stock return variance, SVAR: sum of squared daily returns on the S&P 500 index.
- Book-to-market ratio, BM: ratio of book value to market value for the Dow Jones Industrial Average.[8]
- Net equity expansion, NTIS: ratio of a 12-month moving sum of net equity issues by NYSE-listed stocks to the total end-of-year market capitalization of NYSE stocks.
- Treasury bill rate, TBL: interest rate on a 3-month Treasury bill (secondary market).
- Long-term yield, LTY: long-term government bond yield.
- Long-term return, LTR: return on long-term government bonds.
- Term spread, TMS: long-term yield minus the Treasury bill rate.
- Default yield spread, DFY: difference between BAA- and AAA-rated corporate bond yields.
- Default return spread, DFR: long-term corporate bond return minus the long-term government bond return.

[8] We compute the logarithm of the book-to-market ratio in the empirical analysis.

- Inflation, INFL: calculated from the consumer price inflation (CPI) for all urban consumers; we use lagged 2-month inflation in the regression to account for the delay in CPI releases.

References

Aboody, D., Lehavy, R., Trueman, B., 2010. Limited attention and the earnings announcement returns of past stock market winners. *Review of Accounting Studies* 15, 317–344.

Andrei, D., Hasler, M., 2020. Dynamic attention behavior under return predictability. *Management Science* 66, 2906–2928.

Ang, A., Bekaert, G., 2007. Stock return predictability: Is it there? *Review of Financial Studies* 20, 651–707.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. *Journal of Finance* 61, 1645–1680.

Baker, M., Wurgler, J., 2007. Investor sentiment in the stock market. *Journal of Economic Perspectives* 21, 129–152.

Baker, M., Wurgler, J., Yuan, Y., 2012. Global, local, and contagious investor sentiment. *Journal of Financial Economics* 104, 272–287.

Barbara, R., Inoue, A., 2012. Out-of-sample forecast tests robust to the choice of window size. *Journal of Business and Economic Statistics* 30, 432–453.

Barber, B. M., Odean, T., 2008. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies* 21, 785–818.

Ben-Rephael, A., Da, Z., Israelsen, R., 2017. It depends on where you search: Institutional investor attention and underreaction to news. *Review of Financial Studies* 30, 3009–3047.

Boudoukh, J., Michaely, R., Richardson, M., Roberts, M., 2007. On the importance of measuring payout yield: Implications for empirical asset pricing. *Journal of Finance* 62, 877–915.

Campbell, J., Thompson, S., 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies* 21, 1509–1531.

Chen, J., Tang, G., Yao, J., Zhou, G., 2020. Employee sentiment and stock returns, Working Paper, Washington University in St. Louis.

Clark, T. E., West, K. D., 2007. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics* 138, 291–311.

Cochrane, J. H., 2008. The dog that did not bark: A defense of return predictability. *Review of Financial Studies* 21, 1533–1575.

Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. *Journal of Finance* 66, 1461–99.

Da, Z., Hua, J., Hung, C., Peng, L., 2020. Market returns and a tale of two types of attention, Working Paper, City University of New York.

Dellavigna, S., Pollet, J., 2009. Investor inattention and friday earnings announcements. *Journal of Finance* 64, 709–749.

Drake, M., Jennings, J., Roulstone, D., Thornock, J., 2017. The comovement of investor attention. *Management Science* 63, 2847–2867.

Drake, M., Roulstone, D., Thornock, J., 2015. The determinants and consequences of information acquisition via EDGAR. *Contemporary Accounting Research* 32, 1128–1161.

Fang, L., Peress, J., 2009. Media coverage and the cross-section of stock returns. *Journal of Finance* 64, 2023–2052.

Ferreira, M. A., Santa-Clara, P., 2011. Forecasting stock market returns: The sum of the parts is more than the whole. *Journal of Financial Economics* 100, 514–537.

Gervais, S., Kaniel, R., Mingelgrin, D., 2001. The high-volume return premium. *Journal of Finance* 56, 877–919.

Goyal, A., Welch, I., 2008. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21, 1455–1508.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. *Review of Financial Studies* 33, 2223–2273.

Han, B., Hirshleifer, D., Walden, J., 2020. Social transmission bias and investor behavior, Working Paper, NBER.

Hansen, P. R., Timmermann, A., 2012. Choice of sample split in out-of-sample forecast evaluation, Working Paper, University of California at San Diego.

Hirshleifer, D., Hsu, P., Li, D., 2013. Innovative efficiency and stock returns. *Journal of Financial Economics* 107, 632C654.

Hirshleifer, D., Teoh, S., 2003. Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics* 36, 337C386.

Hodrick, R., 1992. Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *Review of Financial Studies* 5, 357–386.

Hou, K., Peng, L., Xiong, W., 2009. A tale of two anomalies: The implications of investor attention for price and earnings momentum, Working Paper, Ohio State University.

Huang, D., Jiang, F., Li, K., Tong, G., Zhou, G., 2020. Scaled PCA: A new approach to dimension reduction, Working Paper, Washington University in St. Louis.

Huang, D., Jiang, F., Tu, J., Zhou, G., 2015. Investor sentiment aligned: A powerful predictor of stock returns. *Review of Financial Studies* 28, 791–837.

Jiang, F., Lee, J., Martin, X., Zhou, G., 2019. Manager sentiment and stock returns. *Journal of Financial Economics* 132, 126–149.

Jondeau, E., Zhang, Q., Zhu, X., 2019. Average skewness matters. *Journal of Financianl Economics* 134, 29–47.

Kahneman, D., 1973. Attention and effort, Englewood Cliffs, NJ: Prentice-Hall.

Kandel, S., Stambaugh, R., 1996. On the predictability of stock returns: An asset allocation perspective. *Journal of Finance* 51, 385–424.

Karlsson, N., Loewenstein, G., Seppi, D., 2009. The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty* 38, 95–115.

Kelly, B., Pruitt, S., 2013. Market expectations in the cross-section of present values. *Journal of Finance* 68, 1721–1756.

Kelly, B., Pruitt, S., 2015. The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics* 186, 294–316.

Lee, C., Ma, P., Wang, C., 2015. Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics* 116, 410–431.

Lee, C., Ready, M., 1991. Inferring trade direction from intraday data. *Journal of Finance* 46, 733–746.

Lettau, M., Ludvigson, S., 2001. Consumption, aggregate wealth, and expected stock returns. *Journal of Finance* 56, 815–849.

Li, J., Yu, J., 2012. Investor attention, psychological anchors, and stock return predictability. *Journal of Financial Economics* 104, 401–419.

Light, N., Maslov, D., Rytchkov, O., 2017. Aggregation of information about the cross section of stock returns: A latent variable approach. *Review of Financial Studies* 30, 1339–1381.

Lou, D., 2014. Attracting investor attention through advertising. *Review of Financial Studies* 27, 1797–829.

Ludvigson, S., Ng, S., 2007. The empirical risk-return relation: A factor analysis approach. *Journal of Financial Economics* 83, 171–222.

Neely, C., Rapach, D., Tu, J., Zhou, G., 2014. Forecasting the equity risk premium: The role of technical indicators. *Management Science* 60, 1772–1791.

Newey, W., West, K., 1987. A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–708.

Peng, L., 2005. Learning with information capacity constraints. *Journal of Financial Quantitative Analysis* 40, 307C329.

Peng, L., Xiong, W., 2006. Investor attention, overconfidence and category learning. *Journal of Financial Economics* 80, 563–602.

Pettenuzzo, D., Timmermann, A., Valkanov, R., 2014. Forecasting stock returns under economic constraints. *Journal of Financial Economics* 114, 517–553.

Rapach, D., Ringgenberg, M., Zhou, G., 2016. Short interest and aggregate stock returns. *Journal of Financianl Economics* 121, 46–65.

Rapach, D., Zhou, G., 2019. Time-series and cross-sectional stock return forecasting: New machine learning methods, Working Paper, Washington University in St. Louis.

Sicherman, N., Loewenstein, G., Seppi, D., Utkus, S., 2016. Financial attention. *Review of Financial Studies* 29, 863–897.

Stambaugh, R., Yu, J., Yuan, Y., 2012. The short of it: Investor sentiment and anomalies. *Journal of Financial Economics* 104, 288–302.

Wold, H., 1966. Estimation of principal components and related models by iterative least squares, in P. R. Krishnaiaah (eds.), *Multivariate Analysis*, 391-420. New York: Academic Press.

Yu, J., Yuan, Y., 2011. Investor sentiment and the mean-variance relation. *Journal of Financial Economics* 100, 367–381.

Yuan, Y., 2015. Market-wide attention, trading, and stock returns. *Journal of Financial Economics* 116, 548–564.

Zhou, G., 2018. Measuring investor sentiment. *Annual Review of Financial Economics* 10, 239–259.

Figure 1. Time-varying Aggregate Investor Attention

This figure plots the time series of aggregate investor attention constructed by the partial least squares (PLS), scaled principal component analysis (sPCA), and principal component analysis (PCA) approaches. Grey shadow bars denote National Bureau of Economic Research (NBER) recessions. The sample period is January 1980 to December 2017. All attention indices are standardized to have mean of 0 and variance of 1.

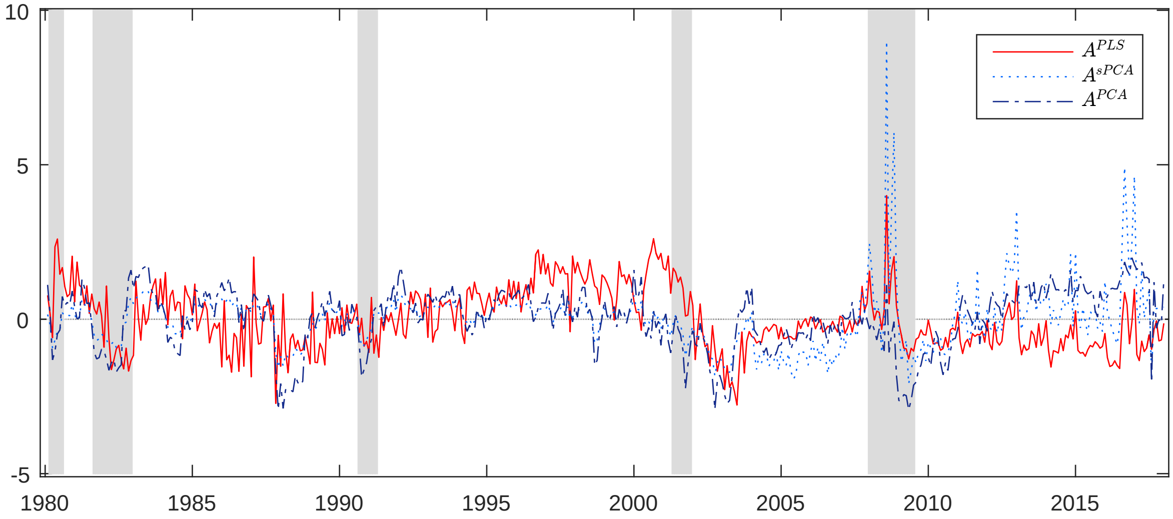

Figure 2. Forecasting Characteristic-sorted Portfolios

This figure plots the regression coefficients (β) from the following univariate predictive regressions,

$$R^i_{t+h} = \alpha^i + \beta^i A_t + \varepsilon^i_{t+h} \ ,$$

where R^i_{t+h} is the average excess return of each one of the portfolios sorted on market beta and idiosyncratic volatility over the prediction horizon h, $h=1$, 6, and 12 months; A_t is one of the aggregate investor attention indices, A^{PLS}, A^{sPCA}, and A^{PCA}. We sort firms into 10 deciles according their characteristics. Decile 1 refers to firms in the lowest decile and decile 10 refers to firms in the highest decile. The sample period is from January 1980 to December 2017.

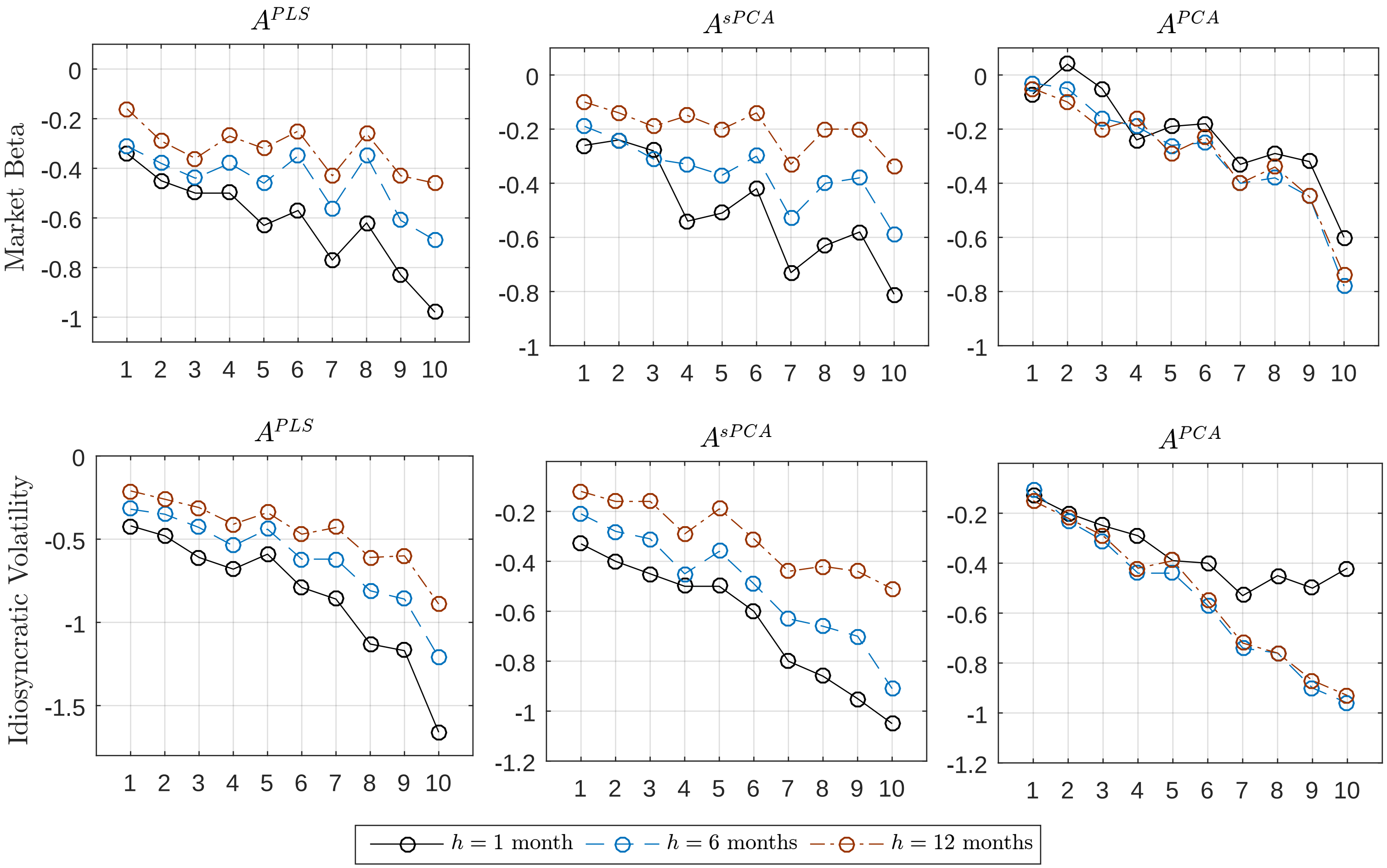

Table 1. Summary Statistics

This table reports the median, 25% and 75% quartiles, skewness, and first-order autocorrelation coefficient ($\rho(1)$) of the 12 individual attention proxies: abnormal trading volume (A^{AVol}), extreme returns (A^{ERet}), past returns (A^{PRet}), nearness to the Dow 52-week high (A^{52wH}), nearness to the Dow historical high (A^{HisH}), analyst coverage ($A^{\#AC}$), change in advertising expenses (A^{CAD}), mutual fund inflow (A^{Inflow}), mutual fund outflow ($A^{Outflow}$), media coverage (A^{Media}), Google search volume (A^{Google}), and search-traffic on Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system (A^{EDGAR}). All attention variables are standardized to have mean of 0 and variance of 1.

	25% Quartile	Median	75% Quartile	Skewness	$\rho(1)$	Sample Period
A^{AVol}	−0.71	−0.07	0.57	0.82	0.46	1980:01–2017:12
A^{ERet}	−0.52	0.05	0.56	−0.36	0.19	1980:01–2017:12
A^{PRet}	−0.66	−0.03	0.46	0.83	0.92	1980:01–2017:12
A^{52wH}	−0.16	0.34	0.65	−2.18	0.89	1980:01–2017:12
A^{HisH}	−0.47	0.38	0.77	−1.41	0.94	1980:01–2017:12
$A^{\#AC}$	−0.75	−0.09	0.69	0.55	0.84	1980:01–2017:12
A^{CAD}	−0.27	0.11	0.58	−1.54	0.97	1980:01–2017:12
A^{Inflow}	−0.55	−0.10	0.31	4.73	0.33	2004:01–2017:12
$A^{Outflow}$	−0.63	−0.19	0.23	2.27	0.57	2004:01–2017:12
A^{Media}	−0.75	0.01	0.63	0.46	−0.06	2004:01–2017:12
A^{Google}	−0.58	−0.04	0.65	0.18	0.75	2004:01–2017:12
A^{EDGAR}	−0.85	−0.24	0.96	0.51	0.96	2004:01–2017:06

Table 2. Correlations of Attention Proxies

This table shows the pairwise correlations of the 12 individual attention proxies: abnormal trading volume (A^{AVol}), extreme returns (A^{ERet}), past returns (A^{PRet}), nearness to the Dow 52-week high (A^{52wH}), nearness to the Dow historical high (A^{HisH}), analyst coverage ($A^{\#AC}$), change in advertising expenses (A^{CAD}), mutual fund inflow (A^{Inflow}), mutual fund outflow ($A^{Outflow}$), media coverage (A^{Media}), Google search volume (A^{Google}), and search-traffic on Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system (A^{EDGAR}). All attention variables are standardized to have mean of 0 and variance of 1. The sample period is from January 2004 through June 2017.

	A^{ERet}	A^{PRet}	A^{52wH}	A^{HisH}	$A^{\#AC}$	A^{CAD}	A^{Inflow}	$A^{Outflow}$	A^{Media}	A^{Google}	A^{EDGAR}
A^{AVol}	−0.10	0.21	0.06	0.11	−0.11	0.11	0.11	0.06	0.02	0.12	−0.04
A^{ERet}		−0.12	0.13	0.16	0.03	0.10	0.23	0.26	0.08	0.08	0.02
A^{PRet}			0.60	0.24	0.14	−0.37	−0.08	−0.21	0.10	−0.28	−0.04
A^{52wH}				0.80	0.39	0.12	−0.03	−0.08	0.14	−0.01	0.20
A^{HisH}					0.42	0.52	0.08	0.12	0.13	0.24	0.35
$A^{\#AC}$						0.03	0.11	0.06	0.22	0.27	0.63
A^{CAD}							−0.01	0.06	0.00	0.23	−0.03
A^{Inflow}								0.79	0.12	0.17	0.17
$A^{Outflow}$									0.02	0.23	0.32
A^{Media}										0.04	−0.02
A^{Google}											0.48

Table 3. In-sample Forecasting Results

This table reports results from following predictive regression,

$$R_{t+h} = \alpha + \beta A_t + \varepsilon_{t+h} ,$$

where R_{t+h} is the average stock market excess return over the prediction horizon h, h = 1, 3, 6, 12, and 24 months, and A_t is one of the attention measures A^{PLS}, A^{sPCA}, and A^{PCA}, constructed by the partial least square, scaled principal component analysis, and principal component analysis approaches, respectively. Panels A, B, and C show the results for A^{PLS}, A^{sPCA}, and A^{PCA}, respectively. In each panel, the estimates of regression slopes (β) and R^2s are reported in percentage form. Brackets below the slope estimates report the t-statistics based on the Hodrick (1992) standard errors (Hodrick-t) and Newey and West (1987) standard errors (NW-t). The sample period is from January 1980 to December 2017. We standardize all predictors to have zero mean and unit variance.

	$h=1$	$h=3$	$h=6$	$h=12$	$h=24$
Panel A: Results for A^{PLS}					
β (%)	−0.64	−0.56	−0.50	−0.37	−0.21
Hodrick-t	[−2.66]	[−2.32]	[−2.39]	[−2.35]	[−1.60]
NW-t	[−2.85]	[−2.58]	[−2.86]	[−2.91]	[−2.07]
R^2 (%)	2.15	4.54	6.97	7.65	5.62
Panel B: Results for A^{sPCA}					
β (%)	−0.49	−0.58	−0.38	−0.23	−0.05
Hodrick-t	[−2.43]	[−2.11]	[−1.96]	[−1.81]	[−0.58]
NW-t	[−2.29]	[−2.42]	[−2.23]	[−2.04]	[−0.69]
R^2 (%)	1.26	4.86	3.91	2.86	0.28
Panel C: Results for A^{PCA}					
β (%)	−0.21	−0.32	−0.26	−0.28	−0.16
Hodrick-t	[−1.04]	[−1.78]	[−1.56]	[−2.10]	[−1.77]
NW-t	[−1.00]	[−1.80]	[−1.70]	[−2.47]	[−2.16]
R^2 (%)	0.22	1.48	1.90	4.24	3.03

Table 4. Comparison with Economic Variables

This table reports results from following predictive regression,

$$R_{t+h} = \alpha + \beta A_t + \phi \mathbf{X}_t + \varepsilon_{t+h} \ ,$$

where R_{t+h} is the average stock market excess return over the prediction horizon h, $h = 1$, 6, and 12 months, A_t is one of the attention measures A^{PLS}, A^{sPCA}, and A^{PCA} at time t, and $\mathbf{X}_t$ represents a vector of economic variables from Goyal and Welch (2008). Panels A, B, and C show results for A^{PLS}, A^{sPCA}, and A^{PCA}, respectively. In each panel, reported are estimates of regression slopes and adjusted R^2s in percentage form. Brackets below the slope estimates report the t-statistics based on the Hodrick (1992) standard errors. The sample period is from January 1980 to December 2017. We standardize all predictors to have zero mean and unit variance.

	Panel A: Results for A^{PLS}			Panel B: Results for A^{sPCA}			Panel C: Results for A^{PCA}		
	$h = 1$	$h = 6$	$h = 12$	$h = 1$	$h = 6$	$h = 12$	$h = 1$	$h = 6$	$h = 12$
A_t	−0.90	−0.52	−0.31	−0.47	−0.39	−0.25	−0.36	−0.28	−0.29
	[−4.06]	[−2.95]	[−2.67]	[−2.42]	[−1.88]	[−1.78]	[−1.58]	[−1.84]	[−2.40]
DP	0.98	0.11	−0.03	0.62	−0.17	−0.20	−0.06	−0.70	−0.83
	[0.45]	[0.11]	[−0.03]	[0.27]	[−0.17]	[−0.20]	[−0.02]	[−0.62]	[−0.82]
DY	0.69	0.56	0.52	0.49	0.56	0.54	0.96	0.91	1.05
	[0.32]	[0.58]	[0.62]	[0.22]	[0.59]	[0.63]	[0.39]	[0.91]	[1.20]
EP	0.72	0.07	0.12	0.43	−0.07	0.06	0.37	−0.12	0.03
	[1.22]	[0.16]	[0.44]	[0.74]	[−0.17]	[0.21]	[0.61]	[−0.27]	[0.12]
BM	−2.48	−0.53	−0.36	−1.46	0.01	−0.07	−1.16	0.26	0.11
	[−3.32]	[−0.72]	[−0.56]	[−1.76]	[0.02]	[−0.11]	[−1.33]	[0.32]	[0.17]
SVAR	−0.80	0.00	0.06	−0.70	0.08	0.11	−0.80	0.00	0.06
	[−3.85]	[−0.02]	[0.51]	[−3.26]	[0.58]	[0.82]	[−3.63]	[−0.01]	[0.54]
LTR	0.44	0.20	0.13	0.43	0.19	0.12	0.45	0.20	0.12
	[2.24]	[2.54]	[2.51]	[2.27]	[2.24]	[2.25]	[2.31]	[2.48]	[2.32]
TMS	0.18	0.07	0.28	0.41	0.20	0.36	0.38	0.18	0.34
	[0.81]	[0.30]	[1.70]	[1.86]	[0.91]	[2.13]	[1.63]	[0.77]	[1.99]
DFY	0.48	−0.01	−0.02	0.34	−0.10	−0.06	0.24	−0.17	−0.15
	[1.17]	[−0.02]	[−0.08]	[0.81]	[−0.27]	[−0.24]	[0.51]	[−0.45]	[−0.56]
Adj. R^2	5.43	8.69	13.99	3.41	7.19	13.63	2.82	4.70	13.80

Table 5. Comparison with Investor Sentiment

This table reports results from following predictive regression,

$$R_{t+h} = \alpha + \beta A_t + \phi S_t^{BW} + \varepsilon_{t+h} \ ,$$

where R_{t+h} is the average stock market excess return over the prediction horizon h, $h = 1$, 6, and 12 months, A_t is one of the attention measures A^{PLS}, A^{sPCA}, and A^{PCA} at time t, and S_t^{BW} represents the investor sentiment index of Baker and Wurgler (2006). Panels A, B, and C show results for A^{PLS}, A^{sPCA}, and A^{PCA}, respectively. In each panel, reported are estimates of regression slopes and adjusted R^2s in percentage form. Brackets below the slope estimates report the t-statistics based on the Hodrick (1992) standard errors. The sample period is from January 1980 to December 2017. We standardize all predictors to have zero mean and unit variance.

	Panel A: Results for A^{PLS}			Panel B: Results for A^{sPCA}			Panel C: Results for A^{PCA}		
	$h=1$	$h=6$	$h=12$	$h=1$	$h=6$	$h=12$	$h=1$	$h=6$	$h=12$
A_t	−0.56	−0.41	−0.30	−0.48	−0.36	−0.21	−0.21	−0.26	−0.27
	[−2.15]	[−1.81]	[−1.80]	[−2.44]	[−1.91]	[−1.85]	[−1.08]	[−1.81]	[−2.35]
S_t^{BW}	−0.22	−0.24	−0.21	−0.41	−0.38	−0.30	−0.43	−0.39	−0.31
	[−1.03]	[−1.30]	[−1.20]	[−2.01]	[−2.08]	[−1.74]	[−2.12]	[−2.17]	[−1.78]
Adj. R^2	1.94	7.99	9.30	1.70	7.52	7.56	0.74	5.77	9.07

Table 6. Out-of-sample Forecasting Results

This table reports the out-of-sample R^2_{OS}'s and *MSFE-adjusted* statistics for predicting the average stock market returns over the prediction horizon h based on one of the attention indices: A^{PLS}, A^{sPCA}, and A^{PCA}. $h = 1$ month, 3 months, 6 months, and 12 months. All of the predictors and regression slopes are estimated recursively using the data available at the forecast formation time t. The out-of-sample period is from January 1995 to December 2017. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Panel A: Results for A^{PLS}		Panel B: Results for A^{sPCA}		Panel C: Results for A^{PCA}	
	R^2_{OS} (%)	*MSFE* *-adjusted*	R^2_{OS} (%)	*MSFE* *-adjusted*	R^2_{OS} (%)	*MSFE* *-adjusted*
$h = 1$	2.04***	3.05	1.20**	1.80	−0.50	−0.39
$h = 3$	3.84***	3.38	3.94***	3.42	1.39**	1.89
$h = 6$	5.05***	3.76	3.44***	3.14	1.52**	1.79
$h = 12$	6.60***	3.90	2.31**	2.53	2.39**	2.34

Table 7. Asset Allocation Performance

This table reports the annualized CER gains (in percentage) and annualized Sharpe ratios for a mean-variance investor with a risk-aversion coefficient of five, who allocates assets between the market and risk-free bills using the out-of-sample forecasts based on A^{PLS}, A^{sPCA}, or A^{PCA} over the prediction horizon h. $h = 1$ month, 3 months, 6 months, and 12 months. We consider two scenarios: zero transaction cost and a proportional transaction cost of 50 basis points per transaction. The out-of-sample period is from January 1995 through December 2017.

	Panel A: Results for A^{PLS}		Panel B: Results for A^{sPCA}		Panel C: Results for A^{PCA}	
	CER Gain	Sharpe Ratio	CER Gain	Sharpe Ratio	CER Gain	Sharpe Ratio
No Transaction Cost						
$h = 1$	3.99	0.74	3.11	0.67	−1.20	0.36
$h = 3$	3.17	0.57	2.90	0.55	1.84	0.48
$h = 6$	3.10	0.51	2.61	0.47	2.98	0.48
$h = 12$	4.55	0.43	2.78	0.38	5.00	0.45
50 bps Transaction Cost						
$h = 1$	3.22	0.67	1.93	0.56	−1.60	0.31
$h = 3$	3.05	0.54	2.53	0.51	1.72	0.46
$h = 6$	3.04	0.49	2.48	0.44	2.93	0.46
$h = 12$	4.41	0.42	2.70	0.37	4.96	0.44

Table 8. Relation with Order Imbalance

This table reports results from the following regression,

$$\Delta AOF_{t+h} = \alpha + \beta_1 A_t + \beta_2 Ret_t + \beta_3 Ret_{t-1} + \beta_4 Ret_{t-2} + \varepsilon_{t+h} \ ,$$

where ΔAOF_{t+h} is the change in aggregate order flow over the period h, and $h = 0$, 1, 6, and 12 months; A_t represents one of the attention indices: A^{PLS}, A^{sPCA}, and A^{PCA}; and Ret_t, Ret_{t-1}, and Ret_{t-2} are stock market return at time t, $t-1$, and $t-2$, respectively. $h = 0$ refers to a contemporaneous relationship between ΔAOF_t and A_t. Panels A, B, and C show results for A^{PLS}, A^{sPCA}, and A^{PCA}, respectively. In each panel, reported are estimates of regression slopes and adjusted R^2s in percentage form. Brackets below the slope estimates report the t-statistics based on the Hodrick (1992) standard errors. The sample period is from January 1993 to December 2017. We standardize all predictors to have zero mean and unit variance.

	Panel A: Results for A^{PLS}				Panel B: Results for A^{sPCA}				Panel C: Results for A^{PCA}			
	$h=0$	$h=1$	$h=6$	$h=12$	$h=0$	$h=1$	$h=6$	$h=12$	$h=0$	$h=1$	$h=6$	$h=12$
A_t	0.12	−0.14	−0.01	0.01	0.05	−0.13	−0.01	0.01	0.02	−0.07	0.00	0.01
	[2.71]	[−2.99]	[−0.38]	[0.97]	[3.75]	[−3.02]	[−0.51]	[0.68]	[1.73]	[−1.64]	[0.16]	[0.76]
Ret_t	0.74	−0.68	−0.10	−0.05	0.74	−0.69	−0.10	−0.05	0.72	−0.67	−0.10	−0.05
	[7.61]	[−4.79]	[−6.16]	[−3.91]	[7.53]	[−5.07]	[−6.52]	[−3.73]	[7.04]	[−4.36]	[−5.46]	[−3.64]
Ret_{t-1}	−0.75	−0.03	0.02	0.01	−0.76	−0.03	0.02	0.01	−0.76	−0.01	0.02	0.00
	[−6.08]	[−0.56]	[1.66]	[0.84]	[−6.06]	[−0.52]	[1.63]	[0.76]	[−6.25]	[−0.13]	[1.42]	[0.51]
Ret_{t-2}	0.03	0.09	0.01	0.01	0.02	0.10	0.01	0.01	0.02	0.11	0.01	0.01
	[0.46]	[1.75]	[0.75]	[1.21]	[0.35]	[2.15]	[0.79]	[1.05]	[0.30]	[2.07]	[0.71]	[0.80]
Adj. R^2	39.25	19.07	10.27	8.53	38.86	19.08	10.28	8.31	39.09	18.40	10.20	8.08

Internet Appendix for

Investor Attention and Stock Returns

July 2020

This Internet Appendix repots the results for supplementary and robustness tests as described below:

Table IA 1. In-sample Forecasting Results for Individual Attention Proxies

This table reports results from following predictive regression,

$$R_{t+h} = \alpha + \beta A_t + \varepsilon_{t+h},$$

where R_{t+h} is the average stock market excess return over the prediction horizon h, $h =$ 1, 6, and 12 months, and A_t denotes one of the 12 attention proxies: abnormal trading volume (A^{AVol}), extreme returns (A^{ERet}), past returns (A^{PRet}), nearness to the Dow 52-week high (A^{52wH}), nearness to the Dow historical high (A^{HisH}), analyst coverage ($A^{\#AC}$), change in advertising expenses (A^{CAD}), mutual fund inflow (A^{Inflow}), mutual fund outflow ($A^{Outflow}$), media coverage (A^{Media}), Google search volume (A^{Google}), and search-traffic on Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system (A^{EDGAR}). Panels A, B, and C show the results for A^{AVol}, A^{ERet}, A^{PRet}, A^{52wH}, A^{HisH}, $A^{\#AC}$, and A^{CAD} from January 1980 to December 2017; for A^{Inflow}, $A^{Outflow}$, A^{Media}, and A^{Google} from January 2004 to December 2017; and for A^{EDGAR} from January 2004 to June 2017, respectively. In each panel, the estimates of regression slopes (β), *t*-statistics (*t*-stat.) based on the Hodrick (1992) standard errors, and R^2s are reported.

	$h = 1$			$h = 6$			$h = 12$		
	β (%)	*t*-stat.	R^2 (%)	β (%)	*t*-stat.	R^2 (%)	β (%)	*t*-stat.	R^2 (%)
Panel A: 1980:01-2017:12									
A^{AVol}	0.15	0.67	0.12	0.12	1.02	0.42	0.04	-0.60	0.08
A^{ERet}	-0.28	-0.91	0.41	-0.14	-0.85	0.56	-0.09	-1.21	0.49
A^{PRet}	-0.22	-1.51	0.26	-0.27	-2.17	2.06	-0.29	-2.00	4.84
A^{52wH}	-0.15	-0.89	0.12	-0.18	-2.02	0.95	-0.22	-2.46	2.82
A^{HisH}	-0.32	-1.95	0.54	-0.36	-4.42	3.61	-0.32	-2.84	5.63
$A^{\#AC}$	0.29	1.48	0.45	0.18	0.91	0.87	0.18	1.04	1.80
A^{CAD}	-0.23	-1.29	0.28	-0.22	-1.38	1.32	-0.26	-2.06	3.81
Panel B: 2004:01-2017:12									
A^{Inflow}	-0.54	-2.61	1.79	-0.58	-1.71	8.49	-0.18	-1.20	1.70
$A^{Outflow}$	-0.76	-1.59	3.45	-0.44	-1.05	4.76	-0.20	-1.03	2.12
A^{Media}	0.06	0.20	0.02	-0.09	-0.70	0.21	-0.06	-0.47	0.19
A^{Google}	-0.09	-0.40	0.05	-0.27	-0.97	1.84	-0.28	-1.15	3.20
Panel C: 2004:01-2017:06									
A^{EDGAR}	0.40	1.31	0.92	0.34	1.08	2.75	0.32	1.00	4.64

Table IA 2. Additional Results of Comparison with Economic Variables

This table reports results from following predictive regression,

$$R_{t+h} = \alpha + \beta A_t + \phi \boldsymbol{X}_t + \varepsilon_{t+h},$$

where R_{t+h} is the average stock market excess return over the prediction horizon h, $h =$ 1, 3, 6, and 12 months, A_t is one of the attention measures A^{PLS}, A^{SPCA}, and A^{PCA} at time *t*, and $\boldsymbol{X}_t$ represents a vector of economic variables from Goyal and Welch (2008), including dividend–payout ratio (DE), net equity expansion (NTIS), treasury bill rate (TBL), long-term yield (LTY), default yield spread (DFY), and inflation (INFL). Panels A, B, and C show results for A^{PLS}, A^{SPCA}, and A^{PCA}, respectively. In each panel, reported are estimates of regression slopes and adjusted R^2s in percentage form. Brackets below the slope estimates report the *t*-statistics based on the Hodrick (1992) standard errors. The sample period is from January 1980 to December 2017.

	Panel A: Results for A^{PLS}			Panel B: Results for A^{SPCA}			Panel C: Results for A^{PCA}		
	$h = 1$	$h = 6$	$h = 12$	$h = 1$	$h = 6$	$h = 12$	$h = 1$	$h = 6$	$h = 12$
A	-0.63	-0.55	-0.34	-0.47	-0.39	-0.24	-0.18	-0.23	-0.24
	[-2.35]	[-2.28]	[-2.13]	[-3.06]	[-2.59]	[-2.08]	[-0.98]	[-1.64]	[-1.93]
DE	0.03	0.17	0.09	0.03	0.16	0.08	0.00	0.12	0.03
	[0.14]	[0.75]	[0.56]	[0.16]	[0.95]	[0.64]	[-0.02]	[0.63]	[0.23]
NTIS	0.13	0.24	0.13	-0.11	0.03	0.00	-0.02	0.11	0.05
	[0.46]	[0.74]	[0.52]	[-0.41]	[0.10]	[0.00]	[-0.05]	[0.31]	[0.20]
TBL	0.04	0.31	-0.40	-0.52	-0.19	-0.72	-0.48	-0.15	-0.68
	[0.07]	[0.44]	[-0.87]	[-0.87]	[-0.32]	[-1.63]	[-0.77]	[-0.23]	[-1.47]
LTY	-0.19	-0.26	0.42	0.23	0.11	0.66	0.18	0.05	0.61
	[-0.26]	[-0.40]	[0.93]	[0.35]	[0.19]	[1.46]	[0.25]	[0.08]	[1.30]
DFR	0.42	0.09	0.02	0.39	0.07	0.00	0.44	0.10	0.03
	[2.70]	[1.70]	[0.42]	[1.99]	[0.82]	[0.07]	[2.50]	[1.22]	[0.50]
INFL	0.14	-0.21	-0.16	0.16	-0.20	-0.15	0.20	-0.17	-0.14
	[0.39]	[-1.02]	[-1.38]	[0.42]	[-1.01]	[-1.36]	[0.52]	[-0.85]	[-1.22]
Adj. R^2	1.74	8.65	11.60	1.12	5.53	9.28	0.19	3.01	9.35

Table IA 3. Additional Out-of-sample Forecasting Results for Individual Attention Proxies

This table reports the out-of-sample R_{OS}^2's and *MSFE-adjusted* statistics for predicting the average stock market returns over the prediction horizon h based on one of the 7 attention proxies: abnormal trading volume (A^{AVol}), extreme returns (A^{ERet}), past returns (A^{PRet}), nearness to the Dow 52-week high (A^{52wH}), nearness to the Dow historical high (A^{HisH}), analyst coverage ($A^{\#AC}$), change in advertising expenses (A^{CAD}). Panels A, B, and C show the results for $h =$ 1, 6, and 12 months, respectively. All of the predictors and regression slopes are estimated recursively using the data available at the forecast formation time *t*. The out-of-sample period is from January 1995 to December 2017. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Panel A: $h =$ 1		Panel B: $h =$ 6		Panel C: $h =$ 12	
	R_{OS}^2 (%)	*MSFE-adjusted*	R_{OS}^2 (%)	*MSFE-adjusted*	R_{OS}^2 (%)	*MSFE-adjusted*
A^{AVol}	-0.77	-0.73	-0.45	-0.02	-0.76	-0.99
A^{ERet}	-0.13	0.26	-0.44	0.19	-1.47	-2.03
A^{PRet}	-0.39	0.18	-3.11	0.04	-4.32	0.85
A^{52wH}	-0.63	-0.65	-4.41	-3.80	1.12**	2.15
A^{HisH}	-0.66	-0.12	-3.23	0.04	2.49***	3.57
$A^{\#AC}$	-0.20	1.04	-0.93	-0.08	-3.55	-1.94
A^{CAD}	-0.51	-0.25	-3.07	-1.82	2.40***	3.85