Information Choice, Uncertainty, and Expected Returns*

Charles Cao†
David Gempesaw‡
Timothy T. Simin§

July 2019

*We are grateful to Kai Du, Jeremiah Green, Stephen Lenkey, Mehmet Canayaz, Jess Cornaggia, Kimberly Cornaggia, Matthew Gustafson, David Haushalter, Burton Hollifield, Peter Iliev, William Kracaw, Anh Le, Giang Nguyen, Brian Routledge, Joel Vanden, Mitch Warachka (discussant), and Alexey Zhdanov, as well as seminar participants at Miami University, Pennsylvania State University, 2018 Financial Management Association Annual Meeting, and 2019 CMU-PITT-PSU Finance Conference for helpful comments.

†The Pennsylvania State University, Smeal College of Business, 338 Business Bldg, University Park, PA, 16802, phone: (814) 865-7891, email: qxc2@psu.edu

‡Miami University, 2010 Farmer School of Business, 800 E. High Street, Oxford, OH 45056, phone: (513)-529-1560, email: dgempesaw@miamioh.edu

§The Pennsylvania State University, Smeal College of Business, 345 Business Bldg, University Park, PA, 16802, phone: (814) 865-3457, email: tsimin@psu.edu

Information Choice, Uncertainty, and Expected Returns

Abstract

We investigate how information choices impact equity returns and risk. Building upon the theory of Van Nieuwerburgh and Veldkamp (2010), we estimate a learning index that reflects the expected benefits of learning about an asset. High learning index stocks have 6.2% lower returns per year and an order of magnitude lower abnormal volatilities compared to low learning index stocks. Long run patterns in returns and volatilities, other measures of information flow, and the information environment surrounding earnings announcements confirm our interpretation of the learning index. Our findings support the model's predictions and illustrate a novel empirical measure of investor learning.

1 Introduction

Price discovery in the U.S. equity market is driven by the trading decisions of active investors. For every \$1m in passive index strategy trades, active stock selectors trade approximately \$22m.[1] The information acquisition decisions of active investors are crucial for price discovery, but difficult to measure. Recent empirical studies utilize observable outcomes such as the holdings and investment returns of portfolios to analyze the impact of information acquisition on investment performance. In contrast, less is known about the impact of information choices on outcomes for the underlying assets.

In this paper, we investigate the role of investors' learning decisions in determining the cross-section of expected return and risk. Rather than utilizing ex-post portfolio performance measures, we create an empirical proxy for a measure called the learning index (LI) that represents the expected benefits of learning about a particular asset. The learning index is proposed by the rational expectations general equilibrium model of Van Nieuwerburgh and Veldkamp (2010). In this model, investors are able to reduce uncertainty about the future payoffs of particular risky assets before making investment decisions. The model predicts that learning about an asset results in both lower uncertainty and expected return, since an increase in information corresponds to more precise conditional expectations of future payoffs and risk-averse investors prefer to hold assets they know more about. Using our empirical estimate of the learning index, we test how learning impacts risk and return in the context of the equity market.

There are several advantages of using the learning index. Standard asset pricing models do not account for the ability of investors to reduce the risk of particular assets by learning. This omission leads to patterns in pricing errors that can be predicted by the learning index. Estimating the learning index only requires historical return data, so our methodology can be applied to any market or set of assets. Furthermore, the fact that the learning index is derived from theory facilitates interpretation of the measure. The empirical learning index identifies

[1] "Viewpoint: Index investing supports vibrant capital markets," BlackRock, 2017.

assets with the highest expected value of information acquisition and therefore represents a prediction of cross-sectional variation in information flow. Since the learning index utilizes historical returns to predict information choices, which in turn, predict cross-sectional patterns in realized returns and risk, the theory of Van Nieuwerburgh and Veldkamp (2010) and the information choices predicted by the learning index are testable without relying on assumptions about investors' unobservable information sets.[2]

Our novel methodology provides estimates of the learning index for individual stocks at the end of each month from 1964 to 2016. Sorting stocks into LI ranked portfolios reveals a negative cross-sectional relation between LI and stock returns over the following month. For value-weighted portfolios, the average return spread between the highest and lowest quintile portfolios sorted on LI is -0.50% per month or -6.2% per year. Risk-adjusting portfolio returns for exposure to market, size, value, profitability, investment, and momentum factors produces a difference in abnormal returns between the extreme quintiles of -0.52% per month or -6.4% per year. Coefficient estimates from two-stage cross-sectional regressions controlling for stock characteristics typically used as predictors of future stock returns indicate that the difference between the stock with the highest and lowest learning index in an average cross-section is approximately -0.41% per month or -5.0% per year (all else equal). These results support the model's prediction that an increase in information about an asset corresponds to a lower expected return.

We find information choices also impact risk. For each stock-month we construct a measure of abnormal return volatility in the next month relative to the average level of volatility over the prior 12 months. Using value-weighted LI sorted portfolios, we find that abnormal return volatility for high LI stocks is more than an order of magnitude below the abnormal volatility of low LI stocks. Decomposing abnormal return volatility into systematic and idiosyncratic components, we find that LI predicts cross-sectional differences in both components of risk, indicating that learning about an asset reduces both firm-specific

[2]Van Nieuwerburgh and Veldkamp (2009) and Veldkamp (2011) propose this property as an advantage of the theoretical analysis of information choice.

uncertainty and return co-movement with systematic risk factors. These results are robust to using alternative sorting approaches and multivariate cross-sectional regressions of next month systematic, idiosyncratic, or total volatility on LI and a set of control variables. Overall, these results suggest that the observed negative cross-sectional relation between LI and expected return derives from investors' decisions to reduce risk through learning.

We provide a number of analyses to evaluate the interpretation of the learning index as a proxy for learning decisions and information flow. If investors learn information and trade based on that information, prices should move toward their intrinsic values and not revert in the future. Empirically, we find that the difference in risk-adjusted monthly returns between extreme value-weighted LI quintiles is negative and significant for up to eight months following portfolio formation. These differences do not reverse during the subsequent two years, suggesting that the return predictive power of LI is due to investor learning rather than temporary price pressure or mispricing. LI also predicts significant differences in average abnormal return volatility between extreme value-weighted quintiles for seven months after portfolio formation. Using a bivariate portfolio sorting approach to control for firm visibility or informational transparency as measured by firm size, we also observe that stocks with higher values of LI have greater abnormal trading activity, analyst coverage, forecast revisions, improvements in forecast accuracy, EDGAR filing downloads, and news reading activity on Bloomberg terminals.

To demonstrate other possible applications of our estimate of the Van Nieuwerburgh and Veldkamp (2010) learning index, we examine the relationship between LI and the information environment surrounding quarterly earnings announcements. A higher degree of investor learning prior to an earnings announcement should reduce the impact of new information revealed in the announcement and result in a smaller market reaction on average. We find that stocks with a higher learning index in the week before the announcement tend to have smaller market reactions to earnings announcements and attenuated post-earnings announcement drift. High LI stocks also experience a greater level of abnormal trading activity in the week prior to an earnings announcement. These findings are consistent with more information being acquired for high LI stocks and incorporated into prices before earnings announcements.

We perform a number of robustness checks and additional analyses to provide further support for our conclusions. The results of these tests are presented in the Internet Appendix. First, we show that large increases in the learning index are contemporaneously associated with increases in price and volatility. These results demonstrate price appreciation that corresponds to lower expected returns and heightened return volatility that corresponds to the incorporation of information into prices. In addition, we explore the role of cross-sectional variation in information processing costs as reflected by firm complexity. We find that the explanatory power of the learning index for expected returns is stronger among firms with less complicated organizational structures. Finally, we show that our conclusions regarding the relationship between the learning index and the cross-section of risk and return are robust to the use of alternative asset pricing models, alternative measures of risk based on option-implied volatility and CAPM beta, and sample subperiods.

This paper contributes to a line of research featuring empirical applications of noisy rational expectations equilibrium models focused on the information content of prices.[3] The theoretical models underlying the aforementioned literature rely on the assumption that information asymmetry is exogenously determined (e.g., all investors receive a private information signal, or a certain fraction of investors are assumed to be informed). In contrast, our empirical analysis is based on the model of Van Nieuwerburgh and Veldkamp (2010) which treats investors' information sets as endogenous. Kacperczyk, Van Nieuwerburgh, and Veldkamp (2016) construct and test a closely related model of mutual fund managers'

[3]Biais, Bossaerts, and Spatt (2010) argue that prices contain information that is value-relevant to an uninformed investor and document that a price-contingent portfolio based on ex-ante information outperforms a passive index. Banerjee (2011) presents a model that nests the rational expectations and differences of opinion approaches, each of which delivers contrasting predictions regarding how investors use prices. The author finds empirical evidence suggesting that investors exhibit rational expectations and condition their beliefs on prices. Burlacu, Fontaine, Jiminez-Garcés, and Seasholes (2012) develop a measure of information precision and supply uncertainty based on the work of Admati (1985) and investigate its relationship with expected returns.

attention allocation and portfolio choices. While there are a number of common themes between that model and the model underlying our paper, the empirical focus of the two papers is different. These authors concentrate on identifying patterns in mutual fund investment and performance that vary with the business cycle, whereas we are interested in directly estimating the learning index at the individual asset level and using it in cross-sectional analyses.

Our paper also relates to the literature investigating the empirical relationship between information flow, expected returns, and risk.[4] These studies commonly focus on information flows that are exogenous from the investor's perspective. We provide complementary evidence to this literature by demonstrating a cross-sectional link between information, returns, and risk using a measure intended to reflect investors' endogenous learning decisions.

Prior studies have also introduced a number of empirical proxies for investor attention or information acquisition, such as news coverage, abnormal trading volume, extreme one-day returns, and search activity on various platforms.[5] We add to this literature by introducing a theoretically-motivated prediction of investors' learning behavior. This measure has a number of limitations and advantages relative to the other existing proxies. For instance, the empirical learning index only serves as a prediction of investor learning and does not depend on direct observation of information acquisition decisions. Furthermore, the learning index can only be used to make inferences about information flow for the average investor; it cannot be used to differentiate between the information activities of retail investors and

[4]Botosan (1997) finds that greater voluntary disclosure by firms is associated with a lower cost of equity capital. Using firm age as a proxy for uncertainty about future profitability, Pastor and Veronesi (2003) show that firms with lower uncertainty have lower market-to-book ratios and lower volatilities. Pan, Wang, and Weisbach (2015) find that volatility is decreasing in CEO tenure, arguing that uncertainty is reduced over time as investors learn about CEO ability. Using SEC Form 8-K filing frequency as a measure of information intensity, Zhao (2017) demonstrates that information intensity reduces expected uncertainty and expected return.

[5]Barber and Odean (2007) use news coverage, abnormal trading volume, and extreme one-day returns as indirect measures of retail investor attention. Da, Engelberg, and Gao (2011) construct a measure of retail investor attention based on Google search frequency. Recently, researchers have proposed more direct measures of information acquisition, such as download activity of SEC filings (e.g., Drake, Roulstone, and Thornock (2015) and Li and Sun (2017)) or news reading activity on Bloomberg terminals (Ben-Rephael, Da, and Israelsen (2017)).

institutional investors. On the other hand, use of the learning index is advantageous because it is less restricted by data availability. Since estimation of the learning index only requires historical return data, it can be used during earlier time periods as well as for other sets of assets.

The rest of the paper is organized as follows. Section 2 discusses the model of Van Nieuwerburgh and Veldkamp (2010), describes the learning index, and outlines the model's relevant predictions. Section 3 describes the procedure to empirically estimate the learning index and summarizes the data. In Sections 4 and 5, we present empirical results on the cross-sectional explanatory power of the learning index for returns and volatilities. Section 6 provides supporting evidence for the interpretation of the learning index. Section 7 provides concluding remarks.

2 Hypothesis development

Our empirical analysis is based on the rational expectations general equilibrium model of information choice and investment choice developed by Van Nieuwerburgh and Veldkamp (2010). The authors explore the impact of different assumptions regarding learning technologies and investor preferences on the optimal information acquisition strategy. We focus on the version of their model with mean-variance preferences and entropy-based learning because these assumptions lead to a prediction of specialized learning (as opposed to generalized learning), which is consistent with the actual behavior of informed investors.[6]

The model contains multiple risky assets and multiple investors. Prior to investing,

[6] Van Nieuwerburgh and Veldkamp (2010) argue that the entropy-based learning technology is preferable to an additive technology for two reasons. First, it is scale neutral, which means that learning costs are unaffected by the definition of one share of an asset. Second, it leads to a prediction of specialized learning (learning about one asset or risk factor) rather than generalized learning (learning about multiple assets). Specialized learning is more consistent with the empirical observation that concentrated portfolios outperform diversified ones, implying that investors with informational advantages choose to specialize in their information and portfolio choices (e.g., Kacperczyk, Sialm, and Zheng (2005) or Ivković, Sialm, and Weisbenner (2008)). When combined with the entropy technology, the assumption that investors exhibit constant absolute risk aversion (CARA) preferences leads to indifference between any allocation of learning capacity. On the other hand, an investor with mean-variance preferences chooses specialization in learning.

investors have the ability to acquire information about unknown asset payoffs f, which are assumed to be normally distributed with mean vector μ and covariance matrix Σ. The learning decision involves choosing which assets to learn about and how much to learn about them, subject to a learning capacity constraint. The model assumes independent asset payoffs and independent information signals about these payoffs. If assets are correlated, an eigen decomposition can be used to form independent linear combinations of the correlated assets. These synthetic assets can be interpreted as principal components (PC), risk factors, or Arrow-Debreu securities. Specifically, a non-diagonal covariance matrix Σ can be decomposed into an eigenvector matrix Γ and a diagonal eigenvalue matrix Λ: $\Sigma = \Gamma\Lambda\Gamma'$. The eigenvalue matrix contains the variances of the principal components, while the eigenvector matrix contains the loadings of the correlated assets on the principal components. With these assumptions, the investor's information choice is equivalent to choosing the posterior variance of each principal component.

The model takes place over three periods: information choices are made in period 1, investment choices are made in period 2, and payoffs and utility are realized in period 3. The model is solved using backward induction. The optimal investment choice in period 2 is a diversified portfolio that conditions on an investor's prior beliefs, information signal realizations, and prices: $q^* = \frac{1}{\rho}\hat{\Sigma}^{-1}\left(\hat{\mu} - pr\right)$, where ρ is the coefficient of risk aversion, p is a vector of prices, r is the risk-free rate, and $\hat{\mu}$ and $\hat{\Sigma}$ are the posterior mean vector and covariance matrix of payoffs. Similar to Admati (1985), equilibrium prices are a linear function of payoffs and supply shocks x: $pr = A + Bf + Cx$. The coefficient matrices A, B, and C are functions of the posterior beliefs of the average investor, the level of risk aversion, and the asset supply. The average investor can be viewed as a representative investor whose posterior mean $\hat{\mu}_a$ is the average of all investors' posterior means and whose posterior covariance matrix $\hat{\Sigma}_a$ is the harmonic average of all investors' posterior covariances.

In period 1, the optimal information choice is to allocate all learning capacity toward the principal component with the highest value of the learning index. The learning index for PC

i is

$$LI_i = \left(\left((I-B)\mu - A\right)'\Gamma_i\right)^2 \Lambda_i^{-1} + (1-\Lambda_{Bi})^2 + \Lambda_i^{-1}\Lambda_{Ci}^2\sigma_x^2. \tag{1}$$

The first term is equivalent to the prior squared Sharpe ratio of PC i: $\frac{\left(E[\Gamma_i'(f-pr)]\right)^2}{Var[\Gamma_i' f]}$. Alternatively, this term can be viewed as the product of two terms: $E\left[\Gamma_i'(f-pr)\right]$ and $\frac{E[\Gamma_i'(f-pr)]}{Var[\Gamma_i' f]}$, which is equivalent to ρ times the expected investment in PC i. These two terms indicate that the value of learning is greater for an asset with a high expected excess return and a high expected portfolio share. Consequently, there are increasing returns to learning — expecting to hold more of an asset makes it more valuable to learn about that asset, while learning more about an asset makes the asset less risky and more attractive to hold.

The second term reflects expected pricing errors related to the informativeness of prices about payoffs. Λ_{Bi} is the i^{th} eigenvalue of B and captures the relationship between payoffs and prices. When Λ_{Bi} is lower, prices covary less with payoffs, making information about payoffs more valuable to learn. The third term reflects expected pricing errors related to the sensitivity of prices to supply shocks. Λ_i and Λ_{Ci} are the i^{th} eigenvalues of the prior covariance matrix Σ and C, respectively. σ_x^2 is the variance of supply shocks, which is assumed to be the same for all PCs. Holding prior uncertainty constant, higher values of Λ_{Ci} indicate that supply shocks have a greater impact on prices, creating pricing errors that can be exploited by an informed investor.

In equilibrium, ex-ante identical investors specialize by learning about a single factor, but each investor chooses to learn about different factors due to strategic substitutability — investors prefer to learn information that other investors do not know. As more investors learn about a given factor, the expected return on that factor is reduced, which reduces the value of learning about that factor. The model has a unique equilibrium in which the aggregate learning capacity of all investors determines the number of risk factors learned about. However, each individual employs a mixed strategy and randomizes over which of these factors to learn about.

The model generates predictions for the relationships between information choices, risk, and expected returns: an increase in information about an asset leads to a reduction in uncertainty and a lower expected return. The model also provides predictions about the impact of learning on systematic risk exposure and prediction errors from a typical asset pricing model such as the CAPM. Van Nieuwerburgh and Veldkamp (2010) derive a conditional CAPM relation in which risk and expected return are measured conditional on information that the average investor knows. In contrast, the standard unconditional CAPM beta is based only on past return information. Predictions of expected returns from the unconditional CAPM do not account for investors' ability to reduce risk through learning. Learning more information about an individual asset reduces the asset's total risk without changing the asset's correlation with the market risk factor. If investors learn more about an asset, the conditional CAPM beta (i.e., the beta conditional on the information learned by investors) will be lower than the unconditional CAPM beta, and the conditional expected return will be lower than the unconditional expected return. Therefore, the model predicts that learning reduces co-movement with systematic risk factors. The discrepancy between the empirically estimated unconditional risk exposure and the unobserved conditional risk exposure leads to cross-sectional variation in factor model pricing errors that is related to investors' learning decisions.

We apply the model's predictions to the cross-section of domestic equities by estimating the learning index for individual stocks and testing the following hypotheses. The first two hypotheses pertain to the relationship between learning and the cross-section of risk and return.

Hypothesis 1: *Stocks with a higher learning index have lower future returns.*

Hypothesis 2: *Stocks with a higher learning index have lower future volatility.*

We also test a number of hypotheses that are not explicitly derived from the model but relate to the notion that the learning index captures investor learning.

Hypothesis 3a: *The changes in price and volatility predicted by the learning index are long-lasting and do not reverse in the long run.*

Hypothesis 3b: *The learning index is correlated with other measures of information flow, such as analyst coverage.*

Hypothesis 3c: *Stocks with a higher learning index prior to an earnings announcement have smaller market reactions (in absolute value) and greater abnormal trading activity.*

3 Methodology and data

3.1 Estimating the learning index

Our objective is to measure the learning index at the end of each month for each stock in the sample. The estimation procedure follows the approach described in Van Nieuwerburgh and Veldkamp (2009) and Veldkamp (2011). We use a two-year rolling window of weekly returns to construct prices, payoffs, and an estimate of the payoff covariance matrix. We use weekly returns instead of monthly or daily returns in order to increase the number of observations within the window while avoiding the effects of non-synchronous trading. Following convention in the literature, weekly returns are measured from Wednesday to Wednesday. The following steps are performed at each month-end.

Step 1: Construct price (p) and payoff (f) time-series for each stock. The price of each stock is set equal to one in the first week. Stock prices then evolve according to the respective weekly return series. Because prices are assumed to be log-normally distributed, we use log prices to be consistent with the model's assumptions. The stock price in the following week is used as a proxy for the stock's payoff, and excess returns are calculated as $f - pr$, where r is the natural log of one plus the risk-free interest rate. To avoid look-ahead bias in the empirical tests, estimation is only based on information available at the end of the current month. Therefore, the final payoff observation in each window is the price at the end of the last full week in the current month.

Step 2: Convert the cross-section of correlated stocks to a set of uncorrelated assets. Estimate the prior covariance matrix Σ of payoffs from Step 1.[7] Decompose Σ into a diagonal eigenvalue matrix Λ and an eigenvector matrix Γ: $\Sigma = \Gamma\Lambda\Gamma'$. Construct principal component prices $\left(\Gamma' p\right)$, payoffs $\left(\Gamma' f\right)$, and excess returns $\left(\Gamma'(f - pr)\right)$.

Step 3: Estimate the learning index for principal components. The first term of the learning index is estimated by dividing squared average return by the variance of payoffs. The second and third term require estimation of the equilibrium price equation at the principal component level: $\Gamma' pr = \Gamma' A + \Gamma' Bf + \Gamma' Cx$. Since principal components are uncorrelated, this is equivalent to estimating a separate regression for each principal component of its price on a constant and its payoff.[8] The payoff coefficient Λ_B and the regression R^2 are used to compute the second and third term.[9]

Step 4: Estimate the learning index for stocks. Pre-multiply the principal component learning index vector by the eigenvector matrix: $\Gamma\left(LI^{PC}\right)$. The learning index for a given stock is a weighted sum of PC learning indexes where the weights are based on the contribution of the stock to each PC.[10]

[7]To account for heteroskedasticity across individual assets, payoffs are standardized to have zero mean and unit variance prior to computing the covariance matrix and performing the eigen decomposition.

[8]This step involves a time-series regression of two non-stationary variables. The underlying theory suggests that in equilibrium, there exists a linear combination of these variables that is stationary. As such, these variables are said to be cointegrated, and the cointegrating vector can be consistently estimated using OLS. In untabulated analysis, we verify the stationarity of the residuals from this regression.

[9]Estimating $(1-\Lambda_{Bi})^2$: If prices follow the pricing equation $pr = A + Bf + Cx$, then OLS can be used to directly estimate B. The OLS estimate is $\Sigma^{-1}\Sigma B = B$. Since assets are assumed to be independent, B is a diagonal covariance matrix and the eigenvalues of B are the diagonal elements of the matrix. For PC i, the OLS coefficient is a direct estimate of Λ_{Bi}.

Estimating $\Lambda_i^{-1}\Lambda_{Ci}^2\sigma_x^2$: First, compute the unconditional variance of prices: $Var(p) = Var(A + Bf + Cx) = B\Sigma B' + CC'\sigma_x^2$. This expression gives us the total sum of squares of prices. Because the asset supply shocks are assumed to be the regression residual, $CC'\sigma_x^2$ is the unexplained sum of squares and $B\Sigma B'$ is the explained sum of squares. Then $\frac{1-R^2}{R^2}$ corresponds to $\left(B\Sigma B'\right)^{-1}CC'\sigma_x^2$. That is, for asset i, $\Lambda_i^{-1}\Lambda_{Ci}^2\sigma_x^2 = \frac{1-R^2}{R^2}\Lambda_{Bi}^2$.

[10]A well-known practical issue involved in eigen decomposition is that the sign of an eigenvector is arbitrary. While this does not make a difference theoretically, it poses an empirical problem. To resolve this issue, we use the square of the normalized eigenvector elements as weights in calculating the stock learning index. This excludes the possibility of a stock having a negative learning index, which has no theoretical interpretation. Because the eigenvectors are standardized to unit length (i.e., the sum of squares for every eigenvector is one), an eigenvector element squared represents the contribution of the stock to the corresponding principal

Because the number of stocks in each monthly cross-section varies over time, we rank-transform the stock learning index and its components to the interval [0,1] to facilitate interpretation and comparability across cross-sections.

3.2 Data sources and variable definitions

We obtain daily and monthly data for US common stocks listed on the NYSE, AMEX, and NASDAQ from the Center for Research in Security Prices (CRSP) during the period from July 1962 to December 2016. Stock returns are adjusted for delisting following Beaver, McNichols, and Price (2007). To reduce the impact of microstructure issues and the influence of microcaps on the results, we require stocks to have a price greater than \$5 and market capitalization above the 20^{th} NYSE percentile in order to be included in the sample at each month-end. Data for market, size, value, profitability, investment, and momentum risk factors are obtained from Kenneth French's website.[11] Additional data sources include Compustat, Institutional Brokers' Estimate System (I/B/E/S), SEC Electronic Data Gathering, Analysis, and Retrieval (EDGAR) Log Files, Bloomberg, and OptionMetrics. The learning index is estimated over the period July 1964 to December 2016, but certain analyses are limited to a subset of this period based on data availability.

For each stock-month, we construct the following characteristics which have been identified in prior studies as important cross-sectional return predictors. Market beta $\left(\beta^{MKT}\right)$ is calculated from a regression of excess stock returns on excess market returns using daily data from the past year. To account for biases due to infrequent trading, we follow Dimson (1979) by including lagged and lead market returns in this regression. The market beta is the sum of the coefficient estimates for the lagged, current, and lead market return. *SIZE* is the natural logarithm of market value of equity. Book-to-market ratio (BM) is the book value of equity in the latest fiscal year ending in the prior calendar year divided by the market value

component. Therefore, a stock's learning index can be interpreted as a weighted sum of principal component learning indexes, where the weights are proportional to the stock's contribution to each principal component.

[11] mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

of equity at the end of December of the prior calendar year. Profitability ($PROF$) is annual revenues minus cost of goods sold, interest expense, and selling, general, and administrative expenses divided by book equity for the latest fiscal year ending in the prior calendar year. Investment (INV) is the annual percentage change in total assets. Momentum (MOM) is the cumulative return from month $t-11$ to month $t-1$.

Illiquidity ($ILLIQ$) is the absolute value of the monthly stock return divided by the respective monthly trading volume in dollars, scaled by 10^5. Short-term reversal (STR) is the monthly return of the stock over the past month. Long-term reversal (LTR) is the cumulative return from month $t-59$ to month $t-12$. Idiosyncratic volatility ($IVOL$) is the standard deviation of daily residuals within a month from estimation of the Fama and French (2018) six-factor model, which includes market, size, value, profitability, investment, and momentum risk factors.[12] We also compute total return volatility ($RVOL$) as the standard deviation of daily excess returns within a month, and the systematic component of volatility ($SVOL$) as the square root of the difference between $RVOL^2$ and $IVOL^2$, although these two variables are not used as cross-sectional return predictors.

In addition, we construct the following predictors for the cross-section of stock volatility. Return on equity (ROE) is earnings before extraordinary items as of the most recent fiscal quarter end divided by common shareholders' equity as of the end of the previous quarter and multiplied by 100. Volatility of return on equity ($ROEVOL$) is the standard deviation of return on equity over the prior 12 fiscal quarters. Firm age (AGE) is the number of years the firm has existed on CRSP. $DIVD$ is a dividend dummy equal to 1 if the firm paid dividends during the most recent fiscal quarter, and 0 otherwise. Leverage (LEV) is total liabilities scaled by the market value of equity as of the most recent fiscal quarter end. $INVPRC$ is the inverse of the stock price, scaled by 100. R is the monthly stock return in percent. All variable definitions are listed in Table A1.

[12] Results are robust to the use of alternative factor models to estimate systematic and idiosyncratic volatility.

3.3 Descriptive statistics

Table 1 presents time-series averages of monthly cross-sectional summary statistics for the aforementioned stock characteristics. Summary statistics are not presented for LI as it is uniformly distributed between 0 and 1 within each cross-section. In the average month, the average stock in the sample has a market beta of 1.06, market capitalization of $3.62 billion (untabulated), and book-to-market ratio of 0.71. The last row in the table reports time-series summary statistics for the number of stocks in the sample per month. The average (median) number of stocks in the sample in a given month is 1,627 (1,654). Key variable average cross-sectional correlations are included in Table 2. Stocks with high LI have a lower market beta, lower market capitalization, higher book-to-market ratio, lower profitability, lower investment, higher illiquidity, lower past returns over short, intermediate, and long horizons, and higher idiosyncratic volatility. These correlations are generally small.

At first glance, variation in the empirical learning index across assets may seem contradictory to the theoretical equilibrium outcome. In theory, the learning index is equal across assets in equilibrium and investors are indifferent between learning about all assets. In reality, learning is not an instantaneous or frictionless process. Time is required for investors to acquire and process information, update their expectations, and incorporate the information into prices. There may also be limits on the amount an investor can reduce uncertainty about a particular asset or limits on the aggregate learning capacity of investors. These factors would lead to cross-sectional dispersion in the empirical learning index at a single point in time or over short horizons. Importantly, this dispersion diminishes over longer horizons as illustrated by the patterns in Table A2 in the Data Appendix which reports transition probabilities for LI-sorted quintile portfolios over 1-month, 6-month, 12-month, 24-month, and 36-month periods. Consistent with the theoretical expectation, the value of learning about a particular asset declines as more investors learn about it.

4 Learning index and the cross-section of returns

An implication of the model by Van Nieuwerburgh and Veldkamp (2010) is that stocks with a higher learning index have lower future returns. In this section, we investigate the ability of the learning index to predict future stock returns using portfolio sorting analyses and two-stage cross-sectional regressions.

4.1 Portfolio sorting

At the end of each month, stocks are sorted into quintiles based on LI. For each quintile-month, we calculate value-weighted and equal-weighted average portfolio returns in excess of the risk-free rate $\left(R_{p,t} - R_{f,t}\right)$ in the following month as well as the difference in average returns between the extreme quintiles $(5-1)$. Next, we calculate the time-series average return for each of the portfolios. We also measure risk-adjusted excess returns for each portfolio as the alpha (α) from a time-series regression of portfolio excess returns on nested versions of the six-factor model proposed by Fama and French (2018). The six-factor model includes the market $\left(R_{M,t} - R_{f,t}\right)$, size ($SMB$), and value ($HML$) factors of Fama and French (1993), profitability (RMW) and investment (CMA) factors of Fama and French (2015), and a momentum (UMD) factor. Specifically, we estimate time-series regressions for each portfolio p using the six-factor model as well as the nested three-factor and five-factor specifications:

$$\begin{aligned} R_{p,t} - R_{f,t} = \alpha_p + \beta_{1,p}\left(R_{M,t} - R_{f,t}\right) + \beta_{2,p} SMB_t + \beta_{3,p} HML_t \\ + \beta_{4,p} RMW_t + \beta_{5,p} CMA_t + \beta_{6,p} UMD_t + \varepsilon_{p,t}. \end{aligned} \tag{2}$$

Table 3 presents average excess returns and risk-adjusted excess returns for value-weighted (Panel A) and equal-weighted (Panel B) portfolios. We report Newey and West (1987) t-statistics with a maximum lag order of 12 months to account for potential autocorrelation and heteroskedasticity. In Panel A, the highest LI quintile has an average excess return of 0.64% in the month following portfolio formation, while the lowest LI quintile has an average excess monthly return of 1.14%. The difference in excess returns between these quintiles is −0.50% per month (−6.2% per year) and is significant at the 1% level. These results indicate that expected returns are lower on average for high LI stocks compared to low LI stocks.

The next three columns report risk-adjusted returns estimated using various factor models. After controlling for exposure to market, size, and value risk factors, the risk-adjusted return of the $5-1$ spread portfolio remains economically and statistically significant: the monthly three-factor alpha spread is −0.50% with a t-statistic of −4.29. We find qualitatively similar results after adding the profitability, investment, and momentum factors. The five-factor (six-factor) alpha difference between the extreme LI quintiles is −0.67% (−0.52%) per month or −8.4% (−6.4%) per year. Each of these estimates is significant at the 1% level.

Table 3, Panel B reports results using the returns of equal-weighted portfolios. Quintile 5 has an average excess return of 0.83% and quintile 1 has an average excess return of 1.40% per month. The average monthly return of the $5-1$ portfolio is −0.57%. The average differences in three-factor, five-factor, and six-factor alphas between the extreme quintiles are −0.51%, −0.62%, and −0.55% per month (−6.3%, −7.7%, and −6.8% per year), respectively. Thus, we find that the results based on value-weighted portfolios and those based on equal-weighted portfolios are qualitatively and quantitatively similar.

Overall, the results of portfolio sorting indicate that high LI stocks tend to have lower future returns relative to low LI stocks. These results support the prediction that learning is associated with lower expected returns and risk-adjusted returns. The $5-1$ spreads in equal-weighted and value-weighted returns are economically and statistically significant, even after controlling for exposure to several sources of systematic risk. The return differences are not driven solely by stocks in any particular quintile. Rather, average returns and alphas tend to decrease monotonically as LI increases across quintiles. Throughout the remainder of the paper, we use the Fama and French (2018) six-factor model for risk adjustment and volatility decomposition, although conclusions based on alternative factor model specifications are qualitatively similar.

It is useful to distinguish between the $5-1$ portfolio formed based on values of LI in Table 3 and the hypothetical portfolio of an investor that chooses to learn information. The objective of the analysis in Table 4 is to identify whether there is a difference in expected returns between high LI and low LI stocks, not to evaluate the expected portfolio return of a learning investor. Suppose that an investor learns the most about high LI stocks and the least about low LI stocks. This information choice does not imply that she will take a long position in high LI stocks and a short position in LI stocks. Rather, her investment choice for each asset will depend on her posterior information set. The investor uses her information to buy the assets that she expects to have high payoffs and sell the assets that she expects to have low payoffs. Since learning more about an asset makes these expectations more accurate, the investor's expected portfolio return is increasing in her learning capacity. Therefore, while high LI assets have lower equilibrium expected returns compared to low LI assets, an individual investor who learns about these assets has a higher expected portfolio return compared to an uninformed investor.

4.2 Cross-sectional regressions

In this section, we use two-stage cross-sectional regressions to examine the relation between the learning index and expected returns while controlling for other determinants of returns. This approach is appropriate for cross-sectional analysis as it accounts for a time effect in the data (i.e., residuals in a given month are correlated across firms). In the first stage, we estimate monthly cross-sectional regressions of excess stock returns in month $t+1$ on values of LI and a set of ten control variables measured in month t. Of the ten stock characteristics used as controls, the first six are associated with exposure to one of the factors used for risk adjustment in the portfolio sorting analysis. Following the prior literature, we also control for the effects of illiquidity, short-term and long-term return reversals, and idiosyncratic volatility. The full cross-sectional model estimated at the end of each month is

$$\begin{aligned} R_{i,t+1} - R_{f,t+1} = \lambda_0 &+ \lambda_1 LI_{i,t} + \lambda_2 \beta^{MKT}_{i,t} + \lambda_3 SIZE_{i,t} + \lambda_4 BM_{i,t} \\ &+ \lambda_5 PROF_{i,t} + \lambda_6 INV_{i,t} + \lambda_7 MOM_{i,t} + \lambda_8 ILLIQ_{i,t} \\ &+ \lambda_9 STR_{i,t} + \lambda_{10} LTR_{i,t} + \lambda_{11} IVOL_{i,t} + \varepsilon_{i,t+1}. \end{aligned} \tag{3}$$

In the second stage, we calculate the time-series averages of the cross-sectional regression coefficient estimates. As an alternative approach to deal with potential errors-in-variables bias, we also compute Litzenberger and Ramaswamy (1979) precision-weighted time-series coefficient averages following Burlacu et al. (2012) and others, where the weights are inversely proportional to the standard error of the estimates from the first stage.

Table 4 reports equal-weighted average (Panel A) and precision-weighted average (Panel B) slope coefficients, Newey and West (1987) t-statistics (in parentheses), and the average adjusted R^2 for each specification. We begin with a univariate regression of excess return on LI in Column 1. The average slope coefficient is −0.69 with a t-statistic of −4.38. Since values of LI range from zero to one, the reported univariate coefficient estimate for LI can be interpreted as the average return difference between the stock with the highest and lowest value of LI in an average month. As a benchmark, we then estimate a regression of excess return on only the control variables in Column 2. Column 3 presents results from the full regression specification. After controlling for several stock characteristics, the coefficient on LI is −0.42 and remains economically and statistically significant (*t-stat* = −4.60).

Panel B presents precision-weighted average slope coefficients from a similar set of three regressions. In this setting, we continue to find a negative and significant relation between LI and subsequent returns. In the univariate regression, the precision-weighted average coefficient on LI is −0.63 and is significant at the 1% level. Using the full multivariate specification in Column 6, the coefficient of interest is −0.41 with a t-statistic of −4.91.[13]

[13]In untabulated analyses, we find that the results are robust to the inclusion of additional cross-sectional return predictors as controls, including return volatility, skewness, co-skewness, kurtosis, maximum daily return in the past month, share turnover, institutional ownership, number of institutional owners, number

These results reinforce the conclusion that learning is associated with a decrease in expected return, even after controlling for traditional return predictors and assigning more weight to more precise cross-sectional coefficient estimates.

With respect to the control variables, the signs of the coefficient estimates are generally in accordance with the findings of past studies. The significant precision-weighted average coefficient estimates in Panel B indicate that stocks with smaller size, higher book-to-market ratios, higher profitability, lower investment, higher momentum, lower past short-term and long-term returns, and lower idiosyncratic volatility are all associated with higher expected returns. The coefficient estimates for market beta are insignificant in both panels. The precision-weighted average coefficient indicates a negative and significant relation between illiquidity and expected returns. While theory suggests a positive relation between these two variables, Bali, Engle, and Murray (2016) show that the empirical relation between illiquidity and future stock returns becomes negative within stock samples that exclude extremely small or illiquid stocks.

Coefficient estimates reported in Table 4 can be combined with the cross-sectional summary statistics in Table 1 to get a sense of the relative economic importance of each of the explanatory variables. Based on the precision-weighted average coefficient estimates, current monthly returns (STR) carry the strongest explanatory power for next month returns. An increase of one standard deviation in STR results in a decrease in next month return of $0.33\%(= 10.00 \times 0.033)$ on average, all else equal. The explanatory power of the learning index for next month returns is comparable to that of momentum, investment, firm size, and idiosyncratic volatility. Increases of one standard deviation in MOM, INV, $SIZE$, $IVOL$, and LI are associated with average cross-sectional differences in expected monthly return of 0.14%, −0.14%, −0.13%, −0.13%, and −0.12% respectively, holding all other variables constant.[14]

of analyst forecasts, the call-put option implied volatility spread, and the Stambaugh, Yu, and Yuan (2015) mispricing measure.

[14] In recent years, a number of papers have expressed concerns about data mining or p-hacking in the empirical asset pricing literature. Using a replicated set of 447 anomaly variables, Hou, Xue, and Zhang (2017)

5 Learning index and the cross-section of volatility

Another testable hypothesis of the model by Van Nieuwerburgh and Veldkamp (2010) is that learning about an asset leads to a reduction in the posterior variance of the asset's payoff. In this section, we investigate the cross-sectional relationship between the learning index and return volatility.

5.1 Portfolio sorting

We measure abnormal return volatility ($ARVOL$) as the difference between next month return volatility and average return volatility in the prior 12 months, scaled by average return volatility in the prior 12 months and multiplied by 100. As the model predicts that learning also reduces the systematic component of risk, we also construct measures of abnormal systematic volatility and abnormal idiosyncratic volatility. $ASVOL$ is monthly systematic volatility divided by average monthly systematic volatility over the previous 12 months, minus one and multiplied by 100. Similarly, $AIVOL$ is monthly idiosyncratic volatility divided by average monthly idiosyncratic volatility over the previous 12 months, minus one and multiplied by 100.

We sort stocks based on LI into quintiles each month and examine the pattern in time-series means of portfolio average abnormal volatility across quintiles. Table 5 presents value-weighted (Panel A) and equal-weighted (Panel B) portfolio average abnormal volatility. In Panel A, abnormal return volatility is 3.98% lower on average for high LI stocks relative to low LI stocks. This difference is significant at the 1% level. The results in the next two columns suggest that the information choices of investors predict cross-sectional differences in both systematic and idiosyncratic volatility. On average, abnormal systematic (idiosyncratic)

find that the explanatory power of 286 (64%) of anomaly variables become insignificant at the 5% level after controlling for microcaps (stocks below the 20$^{\text{th}}$ NYSE percentile) and using value-weighted portfolios instead of equal-weighted portfolios. In order to deal with the bias introduced by multiple testing, Harvey, Liu, and Zhu (2016) suggest that researchers use a t-statistic of 3.0 as a hurdle for assessing the significance of a new anomaly. With these concerns in mind, we note that the estimates of interest from value-weighted portfolio sorting in Table 3 and multivariate regressions in Table 4 are based on a sample that excludes microcap stocks and have t-statistics exceeding the higher cutoff of 3.0.

volatility in the month following portfolio formation are 4.64% (2.77%) lower for high LI stocks compared to low LI stocks, with a t-statistic of −5.79 (−5.13). We arrive at similar conclusions if abnormal volatility is weighted equally within each portfolio. On average, the differences in $ARVOL$, $ASVOL$, and $AIVOL$ between extreme equal-weighted portfolios is −3.60%, −4.12%, and −3.02%, respectively. Each of these estimates is significant at the 1% level.

Altogether, the results from these sorting analyses indicate that learning is associated with a cross-sectional reduction in both the firm-specific and systematic components of risk. The findings based on the systematic component of volatility do not necessarily imply that the choice to learn about a stock involves the discovery of market-wide or macroeconomic information. Rather, the results support the idea that learning news about a firm can reduce not only firm-specific uncertainty, but also uncertainty arising from co-movement with the market or other common risk factors.[15]

5.2 Cross-sectional regressions

Next, we use two-stage cross-sectional regressions to examine the cross-sectional relationships between the learning index and total return volatility, systematic volatility, and idiosyncratic volatility in a multivariate setting. In the first stage, we estimate monthly cross-sectional regressions of a measure of volatility in month $t+1$ on values of LI and a set of control variables. In the second stage, we calculate the time-series averages and Litzenberger and Ramaswamy (1979) precision-weighted time-series averages of the cross-sectional regression coefficient estimates. The full cross-sectional model estimated at the end of each

[15]For robustness, we consider defining abnormal volatility as the next month volatility relative to current month volatility or relative to average volatility over the prior 3 or 6 months. In addition, we consider using the standard deviation of monthly volatility as the denominator as well as calculating absolute differences instead of relative differences (i.e., using just the numerator) in volatility compared to the prior 1, 3, 6, or 12 months. We also estimate systematic and idiosyncratic volatility using alternative factor model specifications. Finally, instead of examining abnormal volatility, we use a bivariate portfolio sorting approach to examine the relationship between LI and the level of volatility in the following month while controlling for the past 12-month average volatility level. Our conclusions are qualitatively similar under each of these robustness checks. The results are available upon request.

month is

$$
\begin{aligned}
VOL_{i,t+1} = {} & \lambda_0 + \lambda_1 LI_{i,t} + \lambda_2 ROE_{i,t} + \lambda_3 ROEVOL_{i,t} + \lambda_4 AGE_{i,t} \\
& + \lambda_5 DIVD_{i,t} + \lambda_6 LEV_{i,t} + \lambda_7 INVPRC_{i,t} + \lambda_8 SIZE_{i,t} \\
& + \lambda_9 BM_{i,t} + \lambda_{10} MOM_{i,t} + \lambda_{11} STR_{i,t} + \lambda_{12} R_{i,t+1} \\
& + \sum_{j=0}^{11} \gamma_{j,t} VOL_{i,t-j} + \varepsilon_{i,t+1}
\end{aligned}
\tag{4}
$$

where VOL is one of total return volatility ($RVOL$), systematic volatility ($SVOL$), or idiosyncratic volatility ($IVOL$). Pastor and Veronesi (2003) find that stock return volatility is higher for less profitable firms, firms with more volatile profitability, younger firms, and firms that do not pay dividends. Based on this, we include return on equity (ROE), the volatility of return on equity ($ROEVOL$), firm age (AGE), and a dividend dummy ($DIVD$) as controls. Christie (1982) shows that stock return volatility increases after stock prices fall due to a leverage effect, while Duffee (1995) documents a contemporaneous relation between return and volatility. As such, we include financial leverage (LEV), the inverse of stock price ($INVPRC$), and the stock return in the next month (R) as control variables. We also include $SIZE$, BM, MOM, and STR to account for the impact of well-known sources of risk. Finally, we control for 12 lagged monthly values of the respective volatility measure in all specifications since volatility is highly persistent over time. The coefficient estimates on lagged volatilities and the intercept term are not reported in the tables for brevity. Based on the availability of data for the explanatory variables, the sample period for this analysis is December 1974 to December 2016.

Table 6 reports the regression results for total return volatility. Equal-weighted coefficient averages are presented in Panel A. In the first column, we estimate monthly cross-sectional regressions of return volatility in the next month on LI while controlling for lagged monthly volatilities over the past year. With this specification, we find a negative relation between LI and volatility. The coefficient on LI is -1.37 and is significant at the 1% level. This finding supports the prediction that learning is associated with lower uncertainty in the cross-section. In Column 2, we estimate a regression of next month return volatility on only the control variables (including lagged volatility) as a benchmark. Consistent with Pastor and Veronesi (2003), firms with lower return on equity, firms with higher volatility of return on equity, younger firms, and non-dividend-paying firms are all associated with higher stock return volatility. In addition, we find a positive and significant contemporaneous relation between return and volatility as well as between the inverse price level and volatility.

Column 3 of Table 6 presents results from the full regression specification. After controlling for a number of characteristics known to have cross-sectional explanatory power for volatility, we continue to find a negative and significant relation between the learning index and volatility. The coefficient on LI in the full specification is -1.36 with a t-statistic of -7.22. This result suggests that, in the average month, the next month return volatility of the stock with the highest value of LI is 1.36 percentage points lower on average than the stock with the lowest value of LI, holding all other variables constant. Panel B of Table 6 presents precision-weighted averages of the cross-sectional coefficient estimates from three similar specifications. The coefficient estimates for LI in Panel B are comparable in magnitude and significance to those in Panel A. In Column 6, the coefficient on LI is -1.19 and is significant at the 1% level. To obtain an approximation of the relative impact of learning on return volatility, we compare these coefficient estimates to the sample average return volatility reported in Table 1. For the average stock in an average cross-section, an increase in the value of LI from zero to one (all else being equal) is associated with a cross-sectional difference in return volatility of $-\frac{1.35}{34.20} \approx -3.96\%$ based on the equal-weighted average LI coefficient, or $-\frac{1.19}{34.20} \approx -3.48\%$ based on the precision-weighted average LI coefficient.

Based on precision-weighted average coefficient estimates in Column 6, variation in the next month return, current month return, and stock price have the largest impact on return volatility in the following month. All else equal, increases of one cross-sectional standard deviation in R, STR, and $INVPRC$ are associated with average cross-sectional differences of 0.87, −0.86, and 0.68 percentage points in next month return volatility. The explanatory power of the learning index for future monthly return volatility is comparable to that of the book-to-market ratio, dividend dummy, and return on equity. Increases of one standard deviation in LI, BM, $DIVD$, and ROE correspond to cross-sectional decreases in return volatility in the following month of 0.35, 0.33, 0.30, and 0.21 percentage points on average, holding all other variables constant.

In the next two tables, we focus on explaining the systematic and idiosyncratic components of return volatility using a similar cross-sectional multivariate analysis. Table 7 presents equal-weighted averages (Panel A) and precision-weighted averages (Panel B) of coefficient estimates from the systematic volatility regressions. In the first column, we regress $SVOL$ in the next month on LI while controlling for lagged monthly values of $SVOL$ over the past year. The results indicate a negative and significant cross-sectional relation between learning and systematic risk in the following month. The coefficient on LI is −1.05 and is significant at the 1% level. Column 2 reports estimates from a benchmark specification that includes lagged values of $SVOL$ and all explanatory variables besides LI. The coefficient estimates in these columns are consistent with those in Table 6 with respect to sign and significance, with a few exceptions. Firm size and momentum are not significantly related to total return volatility but are positively related to the systematic component of volatility in this specification.

The third column of Table 7 reports results from the regression of next month $SVOL$ on the full set of explanatory variables. After controlling for various stock characteristics associated with volatility, we find that the learning index continues to carry negative and significant explanatory power for cross-sectional variation in systematic volatility during the following month. In Column 3, the average coefficient estimate on LI is −0.79 (t-statistic = −5.54). To evaluate the impact of LI on a relative basis, we compare the LI coefficient estimates from the full specifications to the sample average value of $SVOL$ reported in Table 1. For the average stock in an average cross-section, an increase in the value of LI from zero to one holding all other variables constant is associated with a difference in $SVOL$ of $-\frac{0.79}{22.74} \approx -3.48\%$. In terms of economic significance, the explanatory power of LI for the systematic component of next month volatility is comparable to that of the dividend dummy and book-to-market ratio. We arrive at similar conclusions based on the precision-weighted coefficient averages reported in Panel B.

In Table 8, we repeat the cross-sectional regression analyses using idiosyncratic volatility as the dependent variable and lagged values of idiosyncratic volatility as controls. In Column 1, we regress $IVOL$ in the following month on LI in the current month and 12 lagged monthly values of $IVOL$. The equal-weighted average coefficient estimate on LI is -0.70 (t-statistic = -3.83). Column 2 reports equal-weighted average coefficient estimates from the benchmark specification. Consistent with the findings in the previous two tables, the control variables exhibit significant explanatory power for cross-sectional variation in next month idiosyncratic volatility. We find that firm size and momentum are negatively related to $IVOL$. Thus, it appears that combining the negative effects of these variables on $IVOL$ with their positive effects on $SVOL$ results in the insignificant relations with total volatility reported in Table 6.

After controlling for a number of other stock characteristics, we find that the explanatory power of LI for cross-sectional variation in $IVOL$ becomes stronger. The coefficient estimate from Column 3 indicates that, in the average month, the next month idiosyncratic volatility of the stock with the highest value of LI is 0.93 percentage points lower on average than the stock with the lowest value of LI, all else equal (t-statistic = -6.92). For the average stock in an average cross-section, this estimate corresponds to a cross-sectional decrease in idiosyncratic volatility of $\frac{0.93}{24.46} \approx 3.82\%$. In terms of economic significance, the cross-sectional explanatory power of LI for next month $IVOL$ is comparable to that of the dividend dummy and book-to-market ratio. The results based on precision-weighted coefficient averages in Panel B are qualitatively similar.

In summary, the analyses in this section support the hypothesis that investor learning leads to a cross-sectional reduction in volatility. When combined with the findings in Section 4, the results suggest that this cross-sectional reduction in risk corresponds to a cross-sectional reduction in risk premium or expected return.

6 Support for interpretation of the learning index

6.1 Long-term predictability

In this section, we examine the cross-sectional explanatory power of the learning index for subsequent months up to three years. To the extent that the learning index reflects investors learning fundamental information and incorporating this information into prices, we expect that prices move toward fundamental value and do not reverse in the long run. Alternatively, if the explanatory power of the learning index derives from temporary price movements away from intrinsic value, we expect this mispricing to be eventually corrected over time.

At the end of each month t, we sort stocks into quintiles based on LI and track the difference in value-weighted average returns between the highest LI quintile and the lowest LI quintile $(5-1)$ in each of the 36 months after portfolio formation. Figure 1 presents average monthly returns for the spread portfolio. The average return of the value-weighted spread portfolio $(5-1)$ is most negative in the month immediately following portfolio formation and subsequently moves toward zero. By month $t+5$, the negative average return spread is no longer significant at the 10% level. On a risk-adjusted basis, the value-weighted return spread between the highest and lowest LI quintiles is negative and significant until month $t+8$ and is not significantly different from zero in any of the subsequent months. The results indicate that the explanatory power of LI for returns continues in a declining manner over a period of several months. Furthermore, after adjusting for co-movement with systematic risk factors, the monthly return differences between extreme LI quintiles are not reversed over the subsequent three year period. This finding supports the notion that the cross-sectional explanatory power of LI for returns reflects the effects of prices moving closer to (rather than further from) their fundamental values as investors learn and trade upon new information.

Next, we repeat the portfolio sorting analysis and track the difference in value-weighted average abnormal volatility between the extreme LI quintiles over the subsequent 36 months. In Figure 2, the average spread in $ARVOL$ is negative and significant for seven months after portfolio formation. Beyond this point, all values of $ARVOL$ are not statistically different from zero. This result suggests that the cross-sectional relation between the learning index and risk is not attributable to temporary decreases in volatility. When we decompose abnormal volatility into systematic and idiosyncratic components, we find that the two volatility components exhibit different patterns over the long run. Figure 2 shows that the differences in abnormal systematic volatility predicted by LI tend to reverse to a certain extent over the long run. The spread in value-weighted average $ASVOL$ is negative and significant until month $t+5$, but turns positive and significant in some of the subsequent months. On the other hand, average values of $AIVOL$ for the $5-1$ portfolio presented in Figure 2 are negative and significant until month $t+13$ and are not statistically different from zero for the next two years.

The findings based on long-term volatility predictability suggest that the effects of learning (as measured by the learning index) are more permanent for the idiosyncratic component of risk than for the systematic component of risk. While learning appears to reduce return co-movement with systematic risk factors over the short run, this effect is partially reversed over time. Combined with the results on long-term return predictability, the patterns in long-term volatility predictability suggest that the observed reversal in raw returns in Figure 1 is associated with the reversal in systematic risk. We find no evidence of a reversal in risk-adjusted returns and idiosyncratic risk. In untabulated analyses, we arrive at similar conclusions using alternative factor model specifications for risk adjustment as well as equal-weighted portfolios. In aggregate, the results in this section support the interpretation of the learning index by demonstrating that the cross-sectional differences in risk and risk-adjusted returns predicted by this measure are generally long-lasting.

6.2 Relationship with measures of information flow

In this section, we examine the cross-sectional relation between LI and a number of proxies for investor attention or information demand. We consider measures related to trading activity, analyst coverage, forecast revision and accuracy, SEC filing download activity on EDGAR, and Bloomberg news reading activity. In practice, information acquisition efforts are likely to be constrained by the fact that smaller firms may be less visible to investors, less informationally transparent, or may have less information available for acquisition. As such, for this analysis we use a bivariate dependent portfolio sorting approach based on size and LI. This approach enables investigation of the outcomes of differences in information choices across firms while controlling for the impact of firm visibility, informational transparency, or the amount of acquirable information (as captured by firm size).[16]

At the end of each month, we sort stocks into quintiles based on firm size. Then, within each size quintile, we sort stocks based on LI. Each LI subquintile is combined across size quintiles into a single quintile. This procedure creates portfolios of stocks with differences in LI but similar distributions of size. To verify that our prior conclusions regarding the explanatory power of LI for risk and return are robust to this bivariate sorting approach, we use the same approach to examine patterns in returns and abnormal volatility across LI quintiles while controlling for firm size. The results from these untabulated analyses are qualitatively similar to those presented in Table 3 and Table 5.

Table 9 reports portfolio average values of six different proxies of information flow as well as the respective sample period over which each analysis is performed. The first proxy is abnormal trading activity. According to Barber and Odean (2007), trading activity is likely to increase as investors learn new information about a firm. We first measure trading activity as monthly share turnover, or the total number of shares traded within a month divided by shares outstanding. We then measure abnormal turnover ($ATURN$) as turnover during the

[16] Hong, Lim, and Stein (2000) document that firm size is by far the most significant determinant of analyst coverage. Li and Sun (2017) use firm size as a proxy for visibility and find that size alone explains 40% of cross-sectional variation in EDGAR filing download activity.

current month divided by average monthly turnover over the previous 12 months, minus one and multiplied by 100. Data for this variable are available from CRSP for the full sample period (July 1964 to December 2016). On average, high (low) LI stocks experience a 8.55% (3.68%) increase in monthly share turnover relative to average monthly turnover during the past year. The difference in abnormal turnover between the extreme quintiles is 4.87% with a t-statistic of 5.78. This difference suggests a greater level of abnormal trading activity among stocks expected to be subject to a greater degree of investor learning.

The next three proxies relate to analyst coverage, forecast revisions, and forecast accuracy. We expect greater analyst coverage to be associated with an increase in the information available about a firm. Consistent with this idea, Hong et al. (2000) use analyst coverage as a measure of the rate of information flow. The arrival of new information about a firm should also correspond to a revision of analysts' expectations and more accurate forecasts. Zhang (2008) finds that timely analyst forecast revisions improve market efficiency. Harford, Jiang, Wang, and Xie (2018) show that greater effort by analysts in acquiring information is associated with more frequent forecast revisions and more accurate forecasts. Beginning in July 1984, we measure analyst coverage ($nFCST$) each month as the number of analyst forecasts of earnings per share (EPS) recorded by I/B/E/S for the nearest fiscal quarter. We also measure the number of analyst forecast revisions since the last month ($nREV$). In addition, we construct a measure of the change in forecast accuracy (ΔFA) from one month to the next. First, we calculate the error in the mean forecast for the nearest fiscal quarter as the absolute value of the difference between the mean EPS forecast and the actual EPS as a percentage of the actual EPS. We subtract this error from one to measure forecast accuracy. Finally, we compute the monthly change in forecast accuracy (ΔFA) as forecast accuracy in the current month minus forecast accuracy in the prior month, multiplied by 100. Larger values of (ΔFA) represent increases in forecast accuracy. This variable is computed by firm within a given forecast period so that forecast errors are not compared across different forecast periods.

The evidence indicates a positive association between the learning index and analysts' decisions to follow firms and update forecasts. After controlling for the effects of size, stocks with the highest (lowest) values of LI are covered by an average of 8.62 (7.48) analysts. The difference in coverage is approximately one analyst with a t-statistic of 6.76. On average, 2.51 (2.08) analysts covering a high (low) LI stock revise their forecasts from the prior month. The average difference in the number of forecast revisions is 0.43 and is significant at the 1% level. In an average monthly cross-section in the sample, the average (median) firm has 8.05 (6.70) analyst forecasts and 2.31 (1.30) forecast revisions. As such, the estimates of cross-sectional spreads in analyst coverage and forecast revisions between extreme LI quintiles are also economically significant.

In Column 4 of Table 9, we investigate the relationship between the learning index and changes in forecast accuracy. For all LI quintiles, the average monthly percentage change in forecast accuracy is positive. This pattern implies that on average, the mean forecast estimate become more accurate (relative to the actual realized value) as the fiscal quarter end approaches. Stocks in the highest (lowest) LI quintile have an average improvement in forecast accuracy of 4.32% (2.29%). The difference in ΔFA between the extreme LI quintiles is 2.03% on average (t-statistic = 4.63). Therefore, while the EPS forecasts for all stocks in the sample tend to move closer on average to the actual realized EPS, the monthly increase in accuracy is greater for stocks with higher values of the learning index. In untabulated analyses, we find qualitatively similar results when we compute forecast errors using the median rather than the mean analyst forecast.

The fifth proxy is based on downloads of company filings from the SEC EDGAR database. Although the data are available beginning in January 2003, there are significant known issues with the data due to lost or corrupted log files prior to March 2003 and between September 24, 2005 and May 10, 2006. As such, we begin this analysis in March 2003 and drop months with partial coverage from the sample. Drake et al. (2015) study the determinants and consequences of investor information acquisition using EDGAR download activity as a proxy. Following the methodology of Ryans (2017) to screen out algorithmic download activity, we measure $EDGAR$ as the number of human downloads of a company's SEC filings during the month.[17] After controlling for the size of the firm, we find that the filings of firms with the highest (lowest) values of LI are downloaded approximately 872 (742) times within a month on average. The difference in average EDGAR downloads between these quintiles is approximately 130 with a t-statistic of 3.54. This result supports the notion that investors are more likely to gather information for stocks with higher values of the learning index.

The sixth proxy is based on a measure of Bloomberg news reading activity proposed by Ben-Rephael et al. (2017). Bloomberg provides a variable called "News Heat - Daily Max Readership" that measures readership interest in a company relative to the past 30 days. The variable ranges from 0 to 4, with 0 indicating relatively low interest and 4 indicating unusually high interest. The data for this variable are available beginning February 17, 2010, although historical data are missing for periods between December 2010 and January 2011 as well as between August 2011 and November 2011. As such, we begin this analysis in March 2010 and drop any months with partial coverage from the sample. Following Ben-Rephael et al. (2017), we measure abnormal attention at the daily frequency using a dummy variable that is equal to 1 if the Bloomberg daily maximum is a 3 or 4, and 0 otherwise. We then aggregate this measure to the monthly frequency by computing the total number of days with abnormal attention within a month (BBG). After controlling for size, high (low) LI stocks receive abnormal investor attention during 3.17 (2.81) days within a month on average. The difference in abnormal attention days between high and low LI stocks is 0.36 with a t-statistic of 5.63. Overall, the patterns documented in Table 9 serve as supporting evidence of a relationship between the learning index and information flow.[18]

[17]We obtain summarized EDGAR log file data from James Ryans' website: http://www.jamesryans.com/.

[18]In untabulated analyses, we use multivariate cross-sectional regressions to evaluate the explanatory power of the learning index for analyst coverage as well as EDGAR download activity while controlling for other determinants of these two dependent variables. Our conclusions from these tests are qualitatively similar to those in Table 9.

6.3 Learning prior to earnings announcements

In this section, we examine the relationship between the learning index and the information environment prior to quarterly earnings announcements. If a stock has a high learning index value prior to an earnings announcement, we expect that investors are learning more about a firm and incorporating information into prices before the announcement. If this is true, then the average market reactions to earnings announcements of stocks with high values of LI should be smaller in magnitude than those of low LI stocks. We also expect to observe a higher level of abnormal trading activity prior to the earnings announcement as investors trade upon their information.

Following the accounting literature, we measure the market reaction to earnings announcements using the magnitude of cumulative abnormal returns. Absolute returns can also be interpreted as a simple measure of volatility. Daily abnormal returns are calculated as the difference between the daily stock return and the daily return on a portfolio of firms matched on size and book-to-market ratio. We measure the absolute value of the cumulative abnormal return on the day of the announcement $\left(|CAR|_d\right)$ as well as during a two-day period and five-day period beginning on the announcement date $\left(|CAR|_{d,d+1}\right.$ and $\left.|CAR|_{d,d+4}\right)$. We also examine the magnitude of the post-earnings announcement drift, measured as the absolute value of the cumulative abnormal return during the period beginning two days after the announcement through one day after the following quarterly announcement $\left(|CAR|_{q+1}\right)$. In addition to these measures of market reaction, we construct a measure of abnormal trading activity ($ATURN_{w-1}$), defined as share turnover in the week prior to the earnings announcement announcement relative to average weekly turnover over the past 2 months.

As in the previous section, we use a bivariate dependent sorting approach to control for the effects of firm size.[19] Because earnings announcements occur at various times throughout a month, we estimate the learning index for all firms as of the week prior to

[19] Atiase (1985) and Freeman (1987) show that the amount of information incorporated into stock prices prior to earnings announcements is an increasing function of firm size.

each announcement date in our sample. This approach allows us to measure the expected benefits of learning closer to the announcement date (instead of using values of LI as of the most recent month-end), and it also ensures that the time between measurement of LI and the earnings announcement is relatively uniform across announcements.[20] At the end of each month, all stocks with a quarterly earnings announcement during the month are sorted into quintiles based on market capitalization in the week prior to the earnings announcement, and then based on values of LI from the prior week within each size quintile. Each LI subquintile is then combined across the size quintiles. Due to data availability, the sample period for this analysis is October 1971 to December 2016.

Panel A of Table 10 reports average values and associated t-statistics of the market reaction and trading activity for each quintile. After controlling for firm size, stocks with high values of LI in the prior week tend to have smaller market reactions to quarterly earnings announcements. On average, the $|CAR|$ on the event date is 0.14% smaller for high LI stocks compared to low LI stocks (t-statistic = −2.60). Over a two-day (five-day) period beginning on the announcement date, the difference in market reaction between the extreme LI quintiles is −0.17% (−0.31%). These estimates are significant at the 5% and 1% level, respectively. We also find that the absolute magnitude of the drift in abnormal returns over the quarter following the earnings announcement is smaller for stocks with high values of LI. The average spread in $|CAR|_{q+1}$ between the high and low LI quintiles is −0.42% (t-statistic = −1.98).

The last column in Table 10 indicates a higher degree of abnormal trading activity during the week prior to an earnings announcement for high LI stocks compared to low LI stocks. On average, high (low) LI stocks experience a 5.41% (0.95%) change in share turnover during the week prior to a quarterly earnings announcement relative to average turnover over the past 2 months. The difference in abnormal turnover between the extreme quintiles is 4.46% with a t-statistic of 6.73.

[20]Results are qualitatively similar if we use the learning index as of the end of the month prior to the earnings announcement.

In the lower half of Table 10, we examine how the results from the prior analyses vary with the level of earnings announcement activity during the month. If there are fewer earnings announcements during the month, investors' learning efforts may be more concentrated among these firms. Conversely, during months when many firms are reporting earnings, the learning efforts of investors may be spread out over many firms, and the predictions of the learning index may be less powerful. In order to test this hypothesis, we count the number of earnings announcements in each month and split the sample period into two groups: months with high earnings announcement activity (above the yearly median) and months with low earnings announcement activity (below the yearly median). The rest of the testing procedure remains the same.

For each of the two sample period groups, Table 10 reports time-series means of quintile averages of market reactions and abnormal turnover. Results for months with high activity are presented in Panel B, and results for months with low activity are presented in Panel C. In Panel B, the differences in the four measures of market reaction between extreme LI quintiles are all negative but insignificant. In Panel C, each of these coefficients are negative and significant. During months with relatively low earnings announcement activity, the differences in $|CAR|_d$, $|CAR|_{d,d+1}$, $|CAR|_{d,d+4}$, and $|CAR|_{q+1}$ between the highest and lowest LI quintiles are -0.21%, -0.27%, -0.45%, and -0.77%, with t-statistics of -2.65, -2.46, -3.85, and -2.66, respectively. The fifth column of Table 10 presents differences in abnormal turnover in the week prior to an earnings announcement. The difference in $ATURN_{w-1}$ between extreme LI quintiles is 3.53% during high activity months and 5.40% during low activity months.

In sum, the evidence in this section indicates that stocks with higher values of LI in the week prior to an earnings announcement tend to have smaller abnormal market reactions to the announcement. These stocks also exhibit a greater level of abnormal trading activity during the week before the earnings announcement. Furthermore, we find that the ability of the learning index to predict information flow prior to earnings announcements is stronger during months when learning is concentrated among fewer firms. The findings support the idea that the learning index reflects learning decisions and the flow of information prior to earnings announcements.

7 Conclusion

In this paper, we examine the importance of information choice in determining the cross-section of risk and expected return. Much of the asset pricing literature treats an investor's information set as fixed or exogenously determined. In reality, investors have the choice to learn about assets prior to investing. The model of Van Nieuwerburgh and Veldkamp (2010) accounts for this choice, generating predictions for the optimal learning decisions of a rational investor and the resulting impact on risk and risk premiums across assets. In order to test these predictions, we develop an estimation methodology for the learning index from the model for individual stocks. This measure reflects the expected benefits of learning about a particular asset and serves as a prediction of information flow. Consistent with the model's predictions, we find that the empirical learning index is negatively related to both future return and future volatility in the cross-section.

We provide a number of analyses to support the interpretation of the learning index. First, we show that the differences in risk and risk-adjusted return predicted by the learning index are persistent and do not reverse in the long run, indicating that the explanatory power of this measure is not caused by temporary price pressure or mispricing. In addition, we provide evidence of a contemporaneous relationship between the learning index and abnormal trading activity, analyst coverage, forecast revisions, improvements in forecast accuracy, EDGAR download activity, and Bloomberg news reading activity. We also use the learning index to study the information environment around quarterly earnings announcements. Higher values of the learning index prior to an earnings announcement are associated with greater abnormal trading activity before the announcement and smaller market reactions to the announcement, suggesting that more information has been incorporated into prices for these stocks.

In aggregate, our findings support the theoretical predictions of Van Nieuwerburgh and Veldkamp (2010) and demonstrate a connection between investors' learning decisions and cross-sectional differences in risk and return. This paper illustrates a new empirical approach for predicting information flow that can be used in a variety of other settings to further investigate the role of information choice in asset pricing.

References

Admati, A. (1985). A noisy rational expectations equilibrium for multi-asset securities markets. *Econometrica 53*, 629–658.

Atiase, R. K. (1985). Predisclosure information, firm capitalization, and security price behavior around earnings announcements. *Journal of Accounting Research 23*, 21–36.

Bali, T., R. Engle, and S. Murray (2016). *Empirical Asset Pricing: The Cross Section of Stock Returns*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Banerjee, S. (2011). Learning from prices and the dispersion in beliefs. *Review of Financial Studies 24*, 3025–3068.

Barber, B. and T. Odean (2007). 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.

Beaver, W., M. McNichols, and R. Price (2007). Delisting returns and their effect on accounting-based market anomalies. *Journal of Accounting and Economics 2007*, 341–368.

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

Biais, B., P. Bossaerts, and C. Spatt (2010). Equilibrium asset pricing and portfolio choice under asymmetric information. *Review of Financial Studies 23*, 1503–1543.

Botosan, C. (1997). Disclosure level and the cost of equity capital. *The Accounting Review 72*, 323–349.

Burlacu, R., P. Fontaine, S. Jiminez-Garcés, and M. Seasholes (2012). Risk and the cross section of stock returns. *Journal of Financial Economics 2012*, 511–522.

Christie, A. (1982). The stochastic behavior of common stock variances. *Journal of Financial Economics 10*, 407–432.

Da, Z., J. Engelberg, and P. Gao (2011). In search of attention. *Journal of Finance 66*, 1461–1499.

Dimson, E. (1979). Risk measurement when shares are subject to infrequent trading. *Journal of Financial Economics 7*, 197–226.

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

Duffee, G. (1995). Stock returns and volatility: A firm-level analysis. *Journal of Financial Economics 37*, 399–420.

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

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

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

Freeman, R. N. (1987). The association between accounting earnings and security returns for large and small firms. *Journal of Accounting and Economics 9*, 195–228.

Harford, J., F. Jiang, R. Wang, and F. Xie (2018). Analyst career concerns, effort allocation, and firms' information environment. *Working paper, University of Washington, University at Buffalo, Singapore Management University, and University of Delaware*.

Harvey, C., Y. Liu, and H. Zhu (2016). ... and the cross-section of expected returns. *Review of Financial Studies 29*, 5–68.

Hong, H., T. Lim, and J. Stein (2000). Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *Journal of Finance 55*, 265–295.

Hou, K., C. Xue, and L. Zhang (2017). Replicating anomalies. *Working paper, Ohio State University and University of Cincinnati*.

Ivković, Z., C. Sialm, and S. Weisbenner (2008). Portfolio concentration and the performance of individual investors. *Journal of Financial and Quantitative Analysis 43*, 613–655.

Kacperczyk, M., C. Sialm, and L. Zheng (2005). On the industry concentration of actively managed equity mutual funds. *Journal of Finance 60*, 1983–2011.

Kacperczyk, M., S. Van Nieuwerburgh, and L. Veldkamp (2016). A rational theory of mutual funds' attention allocation. *Econometrica 84*, 571–626.

Li, F. W. and C. Sun (2017). Information acquisition and expected returns: Evidence from EDGAR search traffic. *Working paper, Singapore Management University and Hong Kong University of Science and Technology*.

Litzenberger, R. and K. Ramaswamy (1979). The effect of personal taxes and dividends on capital asset prices: Theory and empirical evidence. *Journal of Financial Economics 7*, 163–195.

Newey, W. and K. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*, 703–708.

Pan, Y., T. Wang, and M. Weisbach (2015). Learning about CEO ability and stock return volatility. *Review of Financial Studies 28*, 1623–1666.

Pastor, L. and P. Veronesi (2003). Stock valuation and learning about profitability. *Journal of Finance 58*, 1749–1789.

Ryans, J. (2017). Using the EDGAR log file data set. *Working paper, London Business School*.

Stambaugh, R., J. Yu, and Y. Yuan (2015). Arbitrage asymmetry and the idiosyncratic volatility puzzle. *Journal of Finance 70*, 1903–1948.

Van Nieuwerburgh, S. and L. Veldkamp (2009). Information immobility and the home bias puzzle. *Journal of Finance 64*, 1187–1215.

Van Nieuwerburgh, S. and L. Veldkamp (2010). Information acquisition and under-diversification. *Review of Economic Studies 77*, 779–805.

Veldkamp, L. (2011). *Information Choice in Macroeconomics and Finance*. Princeton, New Jersey: Princeton University Press.

Zhang, Y. (2008). Analyst responsiveness and the post-earnings-announcement drift. *Journal of Accounting and Economics 46*, 201–215.

Zhao, X. (2017). Does information intensity matter for stock returns? Evidence from Form 8-K filings. *Management Science 63*, 1382–1404.

Table 1: Cross-sectional summary statistics of stock characteristics

	Mean	SD	Percentiles 25[th]	50[th]	75[th]
β^{MKT}	1.06	0.57	0.66	0.99	1.39
$SIZE$	6.50	1.24	5.50	6.24	7.27
BM	0.71	0.48	0.38	0.63	0.92
$PROF$	0.82	0.85	0.39	0.67	1.06
INV	0.17	0.30	0.03	0.10	0.20
MOM	20.70	47.05	−4.28	12.79	34.42
$ILLIQ$	0.22	1.07	0.02	0.06	0.18
STR	1.76	10.00	−3.85	1.07	6.46
LTR	1.11	2.07	0.17	0.64	1.36
$RVOL$	34.20	18.00	22.40	30.47	41.89
$IVOL$	24.46	14.13	15.34	21.38	30.02
$SVOL$	22.74	12.71	14.15	20.15	28.42
ROE	3.38	5.17	1.88	3.37	4.97
$ROEVOL$	4.33	14.11	0.90	1.64	3.24
AGE	23.62	17.84	10.21	17.88	32.89
$DIVD$	0.71	0.42	0.42	1.00	1.00
LEV	2.19	4.03	0.32	0.75	1.67
$INVPRC$	4.11	2.48	2.44	3.48	5.06
R	1.10	9.57	−4.25	0.73	5.98
# of stocks per month	1,627	346	1,600	1,654	1,757

This table reports time-series averages of monthly cross-sectional means, standard deviations, and quartiles of key variables in the paper. The sample includes all NYSE, AMEX, and NASDAQ domestic common stocks with stock price greater than $5 and market capitalization greater than the 20[th] percentile of NYSE stocks at the end of each month. The table summarizes the following characteristics: market beta (β^{MKT}), log of firm size ($SIZE$), book-to-market ratio (BM), profitability ($PROF$), investment (INV), momentum (MOM), illiquidity ($ILLIQ$), short-term reversal (STR), long-term reversal (LTR), return volatility ($RVOL$), idiosyncratic volatility ($IVOL$), systematic volatility ($SVOL$), return on equity (ROE), volatility of return on equity ($ROEVOL$), firm age (AGE), dividend dummy ($DIVD$), leverage (LEV), inverse stock price ($INVPRC$), and monthly return (R). See Table A1 in the Data Appendix for complete variable definitions. The last row in the table reports time-series summary statistics for the number of stocks in the sample per month. The sample period is July 1964 through December 2016.

Table 2: Key variable cross-sectional correlations

	LI	β^{MKT}	$SIZE$	BM	$PROF$	INV	MOM	$ILLIQ$	STR	LTR	$IVOL$
LI	1.00										
β^{MKT}	−0.12	1.00									
$SIZE$	−0.09	−0.02	1.00								
BM	0.07	−0.13	−0.13	1.00							
$PROF$	−0.07	0.11	−0.02	−0.29	1.00						
INV	−0.06	0.20	−0.05	−0.20	0.18	1.00					
MOM	−0.28	0.08	−0.02	−0.09	0.05	0.01	1.00				
$ILLIQ$	0.02	−0.09	−0.19	0.03	−0.01	−0.01	−0.02	1.00			
STR	−0.02	0.01	−0.01	0.02	0.01	−0.01	0.02	0.03	1.00		
LTR	−0.11	0.16	0.02	−0.28	0.17	0.32	−0.03	−0.01	−0.02	1.00	
$IVOL$	0.06	0.35	−0.28	−0.07	0.07	0.15	0.06	0.05	0.18	0.09	1.00

This table reports time-series averages of monthly cross-sectional correlations between variables used as return predictors: Learning index (LI), market beta (β^{MKT}), log of firm size ($SIZE$), book-to-market ratio (BM), profitability ($PROF$), investment (INV), momentum (MOM), illiquidity ($ILLIQ$), short-term reversal (STR), long-term reversal (LTR), and idiosyncratic volatility ($IVOL$). See Table A1 in the Data Appendix for complete variable definitions. The sample period is July 1964 through December 2016.

Table 3: Returns of stocks sorted by learning index

Panel A: Value-weighted portfolios				
Quintile	Excess return	FF3 α	FF5 α	FF6 α
1 (Low LI)	1.138	0.241	0.315	0.246
2	0.976	0.060	0.033	0.025
3	0.861	−0.051	−0.118	−0.044
4	0.758	−0.154	−0.221	−0.116
5 (High LI)	0.638	−0.254	−0.357	−0.271
5−1	−0.500***	−0.495***	−0.672***	−0.517***
(t-stat)	(−3.50)	(−4.29)	(−5.75)	(−3.50)

Panel B: Equal-weighted portfolios				
Quintile	Excess return	FF3 α	FF5 α	FF6 α
1 (Low LI)	1.396	0.284	0.314	0.297
2	1.249	0.135	0.123	0.151
3	1.050	−0.045	−0.078	−0.026
4	0.991	−0.098	−0.162	−0.090
5 (High LI)	0.825	−0.227	−0.305	−0.255
5−1	−0.571***	−0.511***	−0.618***	−0.551***
(t-stat)	(−4.75)	(−5.08)	(−6.40)	(−4.45)

At the end of each month, stocks are sorted into quintiles based on values of the learning index (LI). The table reports the next month value-weighted (Panel A) and equal-weighted (Panel B) quintile average monthly excess return and risk-adjusted excess return (alpha or α). FF3 α is computed with respect to the Fama and French (1993) three-factor model which includes market, size, and value factors. FF5 α is computed with respect to the Fama and French (2015) five-factor model which adds profitability and investment factors to the three aforementioned factors. FF6 α is computed with respect to the Fama and French (2018) six-factor model which adds a momentum factor to the five aforementioned factors. The row labeled "5 − 1" presents the difference in monthly return and alpha between the highest and lowest quintile portfolios. Newey and West (1987) t-statistics and 10%(*), 5%(**), and 1%(***) significance levels for two-sided tests are given for the 5 − 1 portfolio. The sample period is July 1964 to December 2016.

Table 4: Cross-sectional regressions of returns on learning index

	Panel A: Equal-weighted coefficient average			Panel B: Precision-weighted coefficient average		
	(1)	(2)	(3)	(4)	(5)	(6)
LI	-0.691^{***}		-0.416^{***}	-0.633^{***}		-0.405^{***}
	(-4.38)		(-4.60)	(-4.71)		(-4.91)
β^{MKT}		0.061	0.042		-0.063	-0.077
		(0.44)	(0.31)		(-0.51)	(-0.63)
$SIZE$		-0.121^{***}	-0.129^{***}		-0.099^{***}	-0.108^{***}
		(-3.40)	(-3.67)		(-2.99)	(-3.30)
BM		0.085	0.085		0.165^{**}	0.164^{**}
		(0.90)	(0.90)		(2.23)	(2.20)
$PROF$		0.096^{*}	0.093^{*}		0.095^{***}	0.094^{***}
		(1.93)	(1.91)		(2.71)	(2.71)
INV		-0.501^{***}	-0.503^{***}		-0.470^{***}	-0.473^{***}
		(-5.55)	(-5.66)		(-6.03)	(-6.13)
MOM		0.005^{**}	0.004^{**}		0.004^{***}	0.003^{***}
		(2.58)	(2.37)		(3.37)	(2.83)
$ILLIQ$		0.591	0.556		-0.056^{**}	-0.057^{**}
		(0.59)	(0.57)		(-2.34)	(-2.36)
STR		-0.033^{***}	-0.034^{***}		-0.032^{***}	-0.033^{***}
		(-6.47)	(-6.62)		(-6.64)	(-6.79)
LTR		-0.053^{***}	-0.056^{***}		-0.034^{***}	-0.040^{***}
		(-2.82)	(-3.20)		(-2.84)	(-3.41)
$IVOL$		-0.011^{***}	-0.009^{***}		-0.011^{***}	-0.009^{***}
		(-4.24)	(-3.67)		(-4.42)	(-3.85)
Adj R^2	0.010	0.090	0.092	0.010	0.090	0.092

This table presents results from two-stage cross-sectional regressions. At the end of each month, we estimate a cross-sectional regression of next month excess stock return on a set of explanatory variables. Each column reports results for a different regression specification. Explanatory variables include an intercept term, the learning index (LI), market beta (β^{MKT}), firm size ($SIZE$), book-to-market ratio (BM), profitability ($PROF$), investment (INV), momentum (MOM), illiquidity ($ILLIQ$), short-term reversal (STR), long-term reversal (LTR), and idiosyncratic volatility ($IVOL$). See Table A1 in the Data Appendix for complete variable definitions. Panel A reports equal-weighted average slope coefficients and Panel B reports Litzenberger and Ramaswamy (1979) precision-weighted average slope coefficients. The average adjusted R^2 is reported in the last row. The intercept term is not reported for brevity. Newey and West (1987) t-statistics are given in parentheses. 10%(*), 5%(**), and 1%(***) significance levels for two-sided tests are denoted. This regression analysis is based on 814,089 stock-month observations from July 1966 to December 2016 with no missing values for all variables.

Table 5: Volatility of stocks sorted by learning index

	Panel A: Value-weighted portfolios			Panel B: Equal-weighted portfolios		
Quintile	$ARVOL$	$ASVOL$	$AIVOL$	$ARVOL$	$ASVOL$	$AIVOL$
1 (Low LI)	3.502	4.954	2.321	3.106	4.635	2.286
2	1.618	2.703	1.125	1.666	3.054	1.056
3	0.855	2.050	0.297	1.244	2.497	0.691
4	0.934	1.835	0.544	0.583	1.697	0.172
5 (High LI)	−0.472	0.315	−0.452	−0.488	0.512	−0.730
5−1	-3.975^{***}	-4.639^{***}	-2.773^{***}	-3.595^{***}	-4.123^{***}	-3.015^{***}
(t-stat)	(−6.01)	(−5.79)	(−5.13)	(−6.12)	(−6.16)	(−5.82)

At the end of each month, stocks are sorted into quintiles based on values of the learning index (LI). The table reports next month value-weighted (Panel A) and equal-weighted (Panel B) quintile average abnormal return volatility ($ARVOL$), systematic volatility ($ASVOL$), and idiosyncratic volatility ($AIVOL$) relative to the respective average volatility in the prior 12 months. See Table A1 in the Data Appendix for complete variable definitions. The row labeled "$5-1$" presents the difference in abnormal volatility between the highest and lowest quintile portfolios. Newey and West (1987) t-statistics and 10%(*), 5%(**), and 1%(***) significance levels for two-sided tests are given for the $5-1$ portfolio. The sample period is July 1964 to December 2016.

Table 6: Cross-sectional regressions of return volatility on learning index

	Panel A: Equal-weighted coefficient average			Panel B: Precision-weighted coefficient average		
	(1)	(2)	(3)	(4)	(5)	(6)
LI	−1.369***		−1.355***	−1.112***		−1.190***
	(−5.23)		(−7.22)	(−5.90)		(−7.68)
ROE		−0.042***	−0.045***		−0.039***	−0.041***
		(−6.59)	(−6.85)		(−7.98)	(−8.45)
$ROEVOL$		0.045***	0.045***		0.005***	0.005***
		(3.27)	(3.24)		(2.71)	(2.64)
AGE		−0.008***	−0.009***		−0.008***	−0.008***
		(−5.48)	(−5.64)		(−5.96)	(−6.09)
$DIVD$		−0.777***	−0.769***		−0.712***	−0.708***
		(−7.81)	(−8.00)		(−8.40)	(−8.64)
LEV		0.011	0.000		−0.010	−0.016
		(0.32)	(−0.01)		(−0.59)	(−0.96)
$INVPRC$		0.284***	0.287***		0.272***	0.275***
		(10.10)	(10.22)		(10.21)	(10.47)
R		0.088***	0.089***		0.095***	0.096***
		(3.96)	(3.96)		(4.25)	(4.25)
$SIZE$		−0.001	−0.031		−0.043	−0.069
		(−0.01)	(−0.59)		(−0.85)	(−1.37)
BM		−0.880***	−0.842***		−0.790***	−0.757***
		(−6.99)	(−6.80)		(−7.16)	(−6.99)
MOM		0.000	−0.003		0.004	0.001
		(0.10)	(−0.94)		(1.59)	(0.33)
STR		−0.106***	−0.102***		−0.099***	−0.096***
		(−10.61)	(−10.49)		(−11.35)	(−11.26)
Lagged volatilities	Yes	Yes	Yes	Yes	Yes	Yes
Adj R^2	0.493	0.532	0.533	0.493	0.532	0.533

This table presents results from two-stage cross-sectional regressions. At the end of each month, we estimate a cross-sectional regression of next month return volatility ($RVOL$) on a set of explanatory variables. Each column reports results for a different regression specification. Explanatory variables include an intercept term, the learning index (LI), return on equity (ROE), volatility of return on equity ($ROEVOL$), firm age (AGE), a dividend dummy ($DIVD$), leverage (LEV), inverse of stock price ($INVPRC$), firm size ($SIZE$), book-to-market ratio (BM), momentum (MOM), short-term reversal (STR), next month return (R), and 12 lagged values of volatility. See Table A1 in the Data Appendix for complete variable definitions. Panel A reports equal-weighted average slope coefficients and Panel B reports Litzenberger and Ramaswamy (1979) precision-weighted average slope coefficients. The average adjusted R^2 is reported in the last row. The intercept term and coefficient estimates for lagged volatilities are not reported for brevity. Newey and West (1987) t-statistics are given in parentheses. 10%(*), 5%(**), and 1%(***) significance levels for two-sided tests are denoted. This regression analysis is based on 686,245 stock-month observations from December 1974 to December 2016 with no missing values for all variables.

Table 7: Cross-sectional regressions of systematic volatility on learning index

	Panel A: Equal-weighted coefficient average			Panel B: Precision-weighted coefficient average		
	(1)	(2)	(3)	(4)	(5)	(6)
LI	-1.052^{***}		-0.793^{***}	-0.808^{***}		-0.652^{***}
	(−5.27)		(−5.54)	(−5.88)		(−5.68)
ROE		-0.029^{***}	-0.031^{***}		-0.028^{***}	-0.029^{***}
		(−6.46)	(−6.58)		(−7.67)	(−8.02)
$ROEVOL$		0.034^{***}	0.034^{***}		0.003^{**}	0.003^{**}
		(3.17)	(3.15)		(2.26)	(2.22)
AGE		-0.006^{***}	-0.006^{***}		-0.005^{***}	-0.006^{***}
		(−4.18)	(−4.37)		(−4.99)	(−5.14)
$DIVD$		-0.592^{***}	-0.593^{***}		-0.529^{***}	-0.532^{***}
		(−7.29)	(−7.47)		(−7.57)	(−7.82)
LEV		0.032	0.025		0.006	0.002
		(1.20)	(0.94)		(0.46)	(0.17)
$INVPRC$		0.203^{***}	0.205^{***}		0.192^{***}	0.193^{***}
		(9.87)	(9.90)		(10.31)	(10.50)
R		0.054^{***}	0.054^{***}		0.060^{***}	0.060^{***}
		(3.85)	(3.85)		(4.39)	(4.39)
$SIZE$		0.127^{**}	0.108^{*}		0.079	0.063
		(2.31)	(1.96)		(1.51)	(1.20)
BM		-0.633^{***}	-0.621^{***}		-0.556^{***}	-0.547^{***}
		(−6.33)	(−6.29)		(−6.60)	(−6.54)
MOM		0.005^{*}	0.003		0.008^{***}	0.006^{***}
		(1.70)	(1.14)		(3.62)	(2.81)
STR		-0.061^{***}	-0.059^{***}		-0.054^{***}	-0.053^{***}
		(−7.69)	(−7.64)		(−8.30)	(−8.27)
Lagged volatilities	Yes	Yes	Yes	Yes	Yes	Yes
Adj R^2	0.439	0.478	0.479	0.439	0.478	0.479

This table presents results from two-stage cross-sectional regressions. At the end of each month, we estimate a cross-sectional regression of next month systematic volatility ($SVOL$) on a set of explanatory variables. Each column reports results for a different regression specification. Explanatory variables include an intercept term, the learning index (LI), return on equity (ROE), volatility of return on equity ($ROEVOL$), firm age (AGE), a dividend dummy ($DIVD$), leverage (LEV), inverse of stock price ($INVPRC$), firm size ($SIZE$), book-to-market ratio (BM), momentum (MOM), short-term reversal (STR), next month return (R), and 12 lagged values of $SVOL$. See Table A1 in the Data Appendix for complete variable definitions. Panel A reports equal-weighted average slope coefficients and Panel B reports Litzenberger and Ramaswamy (1979) precision-weighted average slope coefficients. The average adjusted R^2 is reported in the last row. The intercept term and coefficient estimates for lagged volatilities are not reported for brevity. Newey and West (1987) t-statistics are given in parentheses. 10%(*), 5%(**), and 1%(***) significance levels for two-sided tests are denoted. This regression analysis is based on 686,245 stock-month observations from December 1974 to December 2016 with no missing values for all variables.

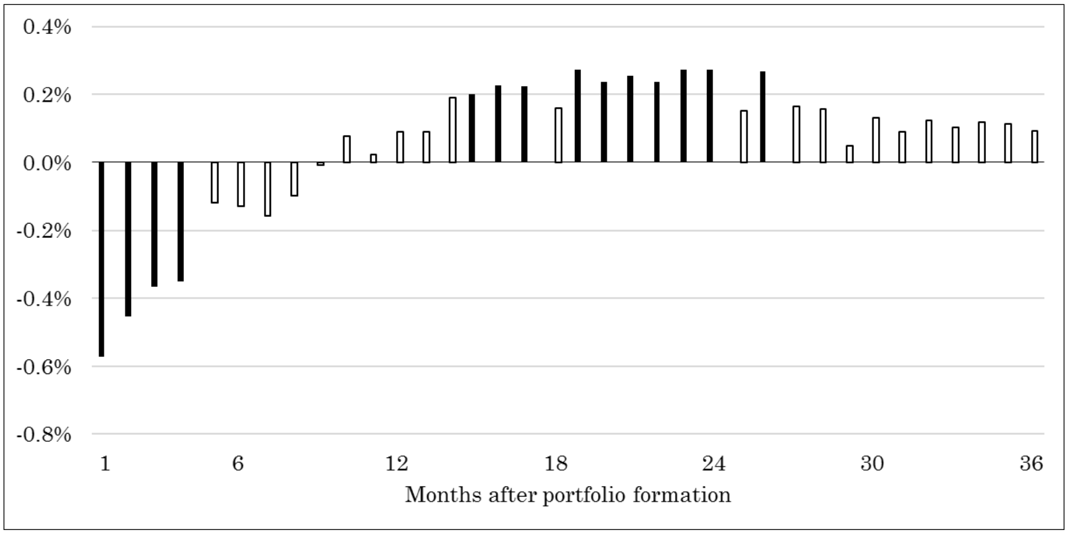
Table 8: Cross-sectional regressions of idiosyncratic volatility on learning index

	Panel A: Equal-weighted coefficient average			Panel B: Precision-weighted coefficient average		
	(1)	(2)	(3)	(4)	(5)	(6)
LI	-0.702^{***}		-0.934^{***}	-0.578^{***}		-0.849^{***}
	(-3.83)		(-6.92)	(-4.11)		(-7.19)
ROE		-0.039^{***}	-0.041^{***}		-0.034^{***}	-0.035^{***}
		(-6.49)	(-6.56)		(-8.07)	(-8.36)
$ROEVOL$		0.040^{***}	0.040^{***}		0.006^{***}	0.006^{***}
		(3.44)	(3.39)		(3.41)	(3.31)
AGE		-0.008^{***}	-0.008^{***}		-0.008^{***}	-0.008^{***}
		(-6.99)	(-7.08)		(-7.13)	(-7.21)
$DIVD$		-0.744^{***}	-0.730^{***}		-0.700^{***}	-0.689^{***}
		(-9.04)	(-9.16)		(-10.48)	(-10.66)
LEV		-0.031	-0.039^{*}		-0.029^{**}	-0.033^{***}
		(-1.33)	(-1.66)		(-2.43)	(-2.77)
$INVPRC$		0.257^{***}	0.258^{***}		0.245^{***}	0.247^{***}
		(11.19)	(11.32)		(10.83)	(11.05)
R		0.071^{***}	0.071^{***}		0.074^{***}	0.074^{***}
		(4.07)	(4.08)		(4.21)	(4.21)
$SIZE$		-0.140^{***}	-0.161^{***}		-0.149^{***}	-0.167^{***}
		(-5.16)	(-5.94)		(-5.78)	(-6.45)
BM		-0.776^{***}	-0.745^{***}		-0.701^{***}	-0.672^{***}
		(-7.87)	(-7.64)		(-7.86)	(-7.65)
MOM		-0.003^{*}	-0.006^{***}		-0.001	-0.003^{*}
		(-1.80)	(-3.00)		(-0.51)	(-1.87)
STR		-0.077^{***}	-0.075^{***}		-0.075^{***}	-0.073^{***}
		(-11.01)	(-10.91)		(-11.47)	(-11.37)
Lagged volatilities	Yes	Yes	Yes	Yes	Yes	Yes
Adj R^2	0.420	0.454	0.455	0.420	0.454	0.455

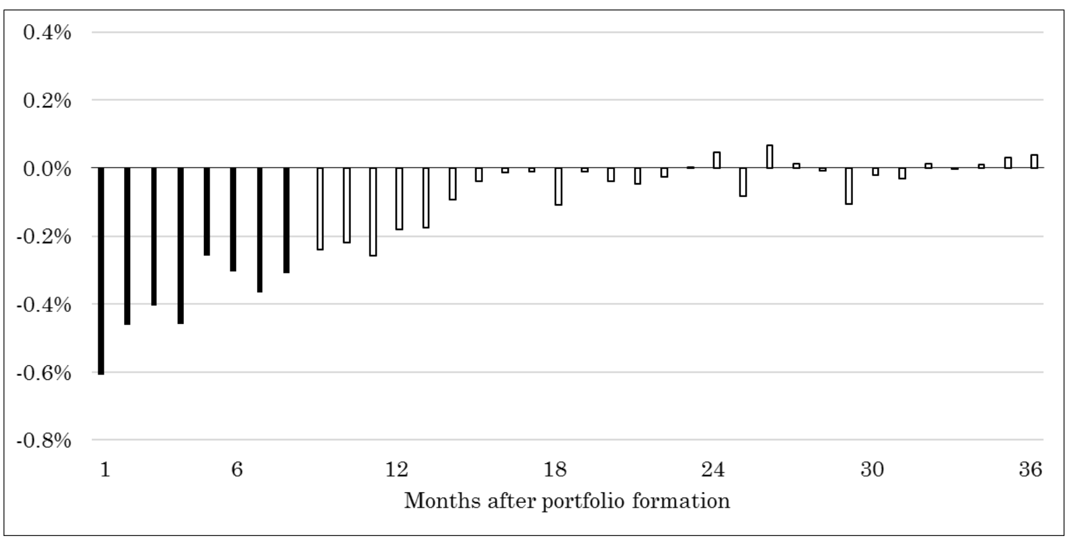
This table presents results from two-stage cross-sectional regressions. At the end of each month, we estimate a cross-sectional regression of next month idiosyncratic volatility ($IVOL$) on a set of explanatory variables. Each column reports results for a different regression specification. Explanatory variables include an intercept term, the learning index (LI), return on equity (ROE), volatility of return on equity ($ROEVOL$), firm age (AGE), a dividend dummy ($DIVD$), leverage (LEV), inverse of stock price ($INVPRC$), firm size ($SIZE$), book-to-market ratio (BM), momentum (MOM), short-term reversal (STR), next month return (R), and 12 lagged values of $IVOL$. See Table A1 in the Data Appendix for complete variable definitions. Panel A reports equal-weighted average slope coefficients and Panel B reports Litzenberger and Ramaswamy (1979) precision-weighted average slope coefficients. The average adjusted R^2 is reported in the last row. The intercept term and coefficient estimates for lagged volatilities are not reported for brevity. Newey and West (1987) t-statistics are given in parentheses. 10%(*), 5%(**), and 1%(***) significance levels for two-sided tests are denoted. This regression analysis is based on 686,245 stock-month observations from December 1974 to December 2016 with no missing values for all variables.

Figure 1: Long-term return predictability:
$LI5 - LI1$ portfolio average return

Excess return



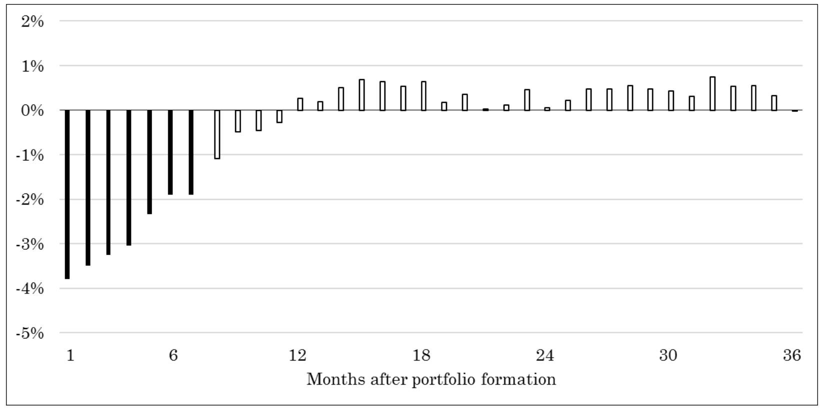
Alpha from Fama and French (2018) six-factor model



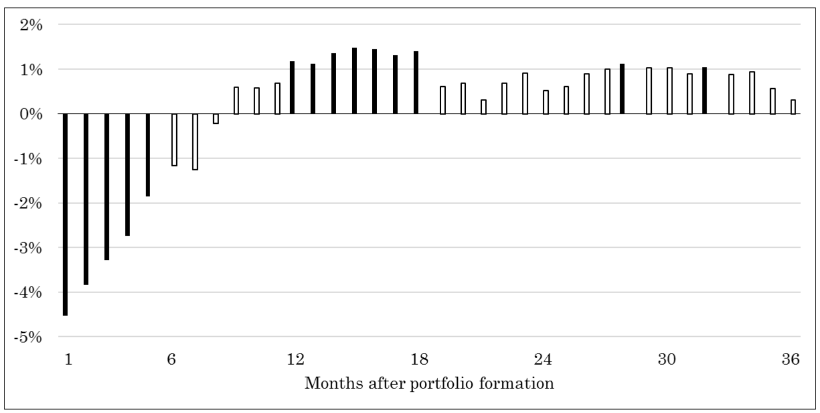
At the end of each month, we sort stocks into quintiles based on values of the learning index (LI) and track the difference in returns between the highest LI quintile and the lowest LI quintile (5 – 1 portfolio) in each of the 36 months following portfolio formation. Stocks are required to have non-missing return observations for all 36 months to be included in the sample. The figure presents value-weighted monthly excess returns and risk-adjusted returns for the 5 – 1 portfolio. Black bars indicate statistical significance at the 10% level. The sample period is July 1964 to December 2016.

Figure 2: Long-term volatility predictability: $LI5 - LI1$ portfolio average abnormal return volatility

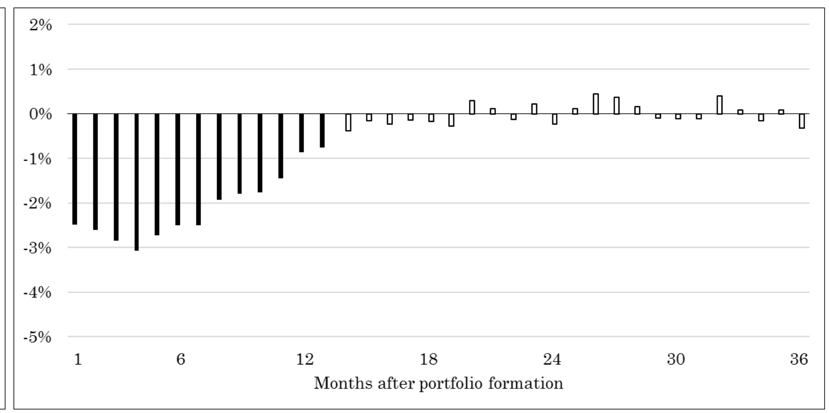
Abnormal return volatility



Abnormal systematic volatility



Abnormal idiosyncratic volatility



At the end of each month, we sort stocks into quintiles based on values of the learning index (LI) and track the difference in abnormal return volatility ($ARVOL$), abnormal systematic volatility ($ASVOL$) , and abnormal idiosyncratic volatility ($AIVOL$) between the highest LI quintile and the lowest LI quintile ($5-1$ portfolio) in each of the 36 months after portfolio formation. Stocks are required to have non-missing volatility observations for all 36 months to be included in the sample. The figure presents value-weighted average abnormal volatilities for the $5-1$ portfolio. Black bars indicate statistical significance at the 10% level. Systematic and idiosyncratic components of volatility are measured using the Fama and French (2018) six-factor model. See Table A1 in the Data Appendix for complete variable definitions. The sample period is July 1964 to December 2016.

Table 9: Relationship with measures of information flow:
Portfolios of stocks sorted by learning index controlling for firm size

Sample period begins:	Jul 1964	Jul 1984	Jul 1984	Jul 1984	Mar 2003	Mar 2010
Quintile	$ATURN$	$nFCST$	$nREV$	ΔFA	$EDGAR$	BBG
1 (Low LI)	3.676	7.483	2.076	2.294	742	2.805
2	6.705	7.767	2.206	2.843	775	2.933
3	8.189	8.047	2.299	3.428	815	3.010
4	9.166	8.361	2.456	4.102	839	3.047
5 (High LI)	8.547	8.621	2.505	4.319	872	3.166
5−1	4.871***	1.139***	0.428***	2.025***	130***	0.361***
(t-stat)	(5.78)	(6.76)	(5.42)	(4.63)	(3.54)	(5.63)

At the end of each month, stocks are sorted into quintiles based on market capitalization. Within each size quintile, stocks are sorted based on values of the learning index (LI). Each LI subquintile is combined across size quintiles into a single quintile. This approach creates portfolios of stocks with differences in LI but similar distributions of size. The table reports the time-series means of quintile averages of six proxies of investor attention or information demand: abnormal monthly share turnover ($ATURN$), number of analyst forecasts ($nFCST$), number of analyst forecast revisions ($nREV$), change in forecast accuracy (ΔFA), number of SEC filing downloads from EDGAR ($EDGAR$), and number of days with abnormal news reading activity on Bloomberg (BBG). See Table A1 in the Data Appendix for complete variable definitions. The row labeled "5 − 1" presents the difference in the respective dependent variable between the highest and lowest quintile portfolios. Newey and West (1987) t-statistics and 10%(*), 5%(**), and 1%(***) significance levels for two-sided tests are given for the 5 − 1 portfolio. The first row of the table header indicates the first month that data are available for the respective dependent variable. All sample periods end in December 2016.

Table 10: Learning prior to earnings announcements: Portfolios of stocks sorted by learning index controlling for firm size

Quintile	$\|CAR\|_d$	$\|CAR\|_{d,d+1}$	$\|CAR\|_{d,d+4}$	$\|CAR\|_{q+1}$	$ATURN_{w-1}$
			Panel A: Full sample period		
1 (Low LI)	2.539	3.939	5.076	12.148	0.951
2	2.463	3.869	4.932	11.894	1.776
3	2.444	3.838	4.911	11.748	2.794
4	2.446	3.839	4.856	11.771	4.112
5 (High LI)	2.403	3.770	4.769	11.726	5.411
5−1	-0.136^{***}	-0.169^{**}	-0.308^{***}	-0.421^{**}	4.460^{***}
(t-stat)	(−2.60)	(−2.14)	(−3.39)	(−1.98)	(6.73)
			Panel B: Months with number of earnings announcements above yearly median		
5−1	−0.066	−0.068	−0.164	−0.074	3.527^{***}
(t-stat)	(−1.03)	(−0.67)	(−1.33)	(−0.27)	(6.24)
			Panel C: Months with number of earnings announcements below yearly median		
5−1	-0.206^{***}	-0.271^{**}	-0.452^{***}	-0.769^{***}	5.395^{***}
(t-stat)	(−2.65)	(−2.46)	(−3.85)	(−2.66)	(4.87)

At the end of each month, all stocks with a quarterly earnings announcement during the month are sorted into quintiles based on market capitalization in the week prior to the earnings announcement. Within each size quintile, stocks are sorted based on values of the learning index (LI) in the week prior to the earnings announcement. Each LI subquintile is combined across size quintiles into a single quintile. This approach creates portfolios of stocks with differences in LI but similar distributions of size. The table reports time-series means of quintile averages of three measures of market reaction and a measure of abnormal trading activity. Market reaction proxies include the absolute value of the cumulative abnormal return on the earnings announcement date d $\left(|CAR|_d\right)$, during a two-day period and five-day period beginning on the announcement date $\left(|CAR|_{d,d+1}\right.$ and $\left.|CAR|_{d,d+4}\right)$, and during the period beginning two days after the earnings announcement date through one day after the firm's next quarterly earnings announcement date $\left(|CAR|_{q+1}\right)$. The proxy for abnormal trading activity is abnormal share turnover in the week prior to the earnings announcement ($ATURN_{w-1}$). See Table A1 in the Data Appendix for complete variable definitions. The row labeled "5 − 1" presents the difference in the respective dependent variable between the highest and lowest quintile portfolios. Newey and West (1987) t-statistics and 10%(*), 5%(**), and 1%(***) significance levels for two-sided tests are given for the 5 − 1 portfolio. The full sample period in Panel A is October 1971 to December 2016. In Panel B and Panel C, the sample period is split into two groups: months with high earnings announcement activity (above the yearly median) and months with low earnings announcement activity (below the yearly median).

Data Appendix

Table A1: Variable definitions

Variables are listed in the order that they are introduced within the body of the paper.

Variable	Definition
LI	Learning index, based on the rational expectations general equilibrium model of information choice and investment choice developed by Van Nieuwerburgh and Veldkamp (2010). The learning index reflects the expected benefits of learning about an asset for a rational average investor. Higher values of the empirical learning index correspond to a greater expected degree of learning and information flow. See Section 3.1 in the text for complete description of variable measurement.
β^{MKT}	Market beta, estimated from a regression of excess stock returns on lagged, current, and lead excess market returns using daily data from the past year.
$SIZE$	Natural logarithm of market value of equity in millions of dollars.
BM	Book-to-market ratio, defined as book value of equity in the latest fiscal year ending in the prior calendar year divided by the market value of equity at the end of December of the prior calendar year.
$PROF$	Profitability, defined as annual revenues minus cost of goods sold, interest expense, and selling, general, and administrative expenses divided by book equity for the latest fiscal year ending in the prior calendar year.
INV	Investment, defined as the annual percentage change in total assets as a decimal.
MOM	Momentum, defined as the cumulative return in percent from month $t-11$ to month $t-1$.
$ILLIQ$	Illiquidity, defined as the absolute monthly return divided by the respective monthly trading volume in dollars, scaled by 10^5.
STR	Short-term reversal, defined as the monthly return in percent over the past month.
LTR	Long-term reversal, defined as the cumulative return as a decimal from month $t-59$ to month $t-12$.
$RVOL$	Return volatility, defined as the standard deviation of daily excess returns within a month.
$IVOL$	Idiosyncratic component of volatility, defined as the standard deviation of daily residuals within a month estimated from a regression of excess stock returns on the six-factor model of Fama and French (2018).
$SVOL$	Systematic component of volatility, defined as the square root of the difference between return variance ($RVOL^2$) and idiosyncratic variance ($IVOL^2$).
α	Risk-adjusted excess return, defined as the intercept from a regression of excess returns on a set of risk factors.
$ARVOL$	Abnormal return volatility, defined as $RVOL$ divided by average $RVOL$ over the previous 12 months, minus one and multiplied by 100.
$AIVOL$	Abnormal idiosyncratic volatility, defined as $IVOL$ divided by average $IVOL$ over the previous 12 months, minus one and multiplied by 100.

Continued on next page

Variable	Definition
$ASVOL$	Abnormal systematic volatility, defined as $SVOL$ divided by average $SVOL$ over the previous 12 months, minus one and multiplied by 100.
ROE	Return on equity, defined as earnings before extraordinary items as of the most recent fiscal quarter end divided by common shareholders' equity as of the end of the previous quarter and multiplied by 100.
$ROEVOL$	Volatility of return on equity, defined as the standard deviation of return on equity over the prior 12 fiscal quarters.
AGE	Firm age, defined as the number of years the firm has existed on CRSP.
$DIVD$	Dummy variable equal to 1 if the firm paid dividends during the most recent fiscal quarter, and 0 otherwise.
LEV	Leverage, defined as total liabilities scaled by the market value of equity as of the most recent fiscal quarter end.
$INVPRC$	Inverse of the stock price, scaled by 100.
R	Monthly return in percent.
$MAXDRET$	Maximum daily return during a 1-month or 3-month horizon (specified in tables).
$MAXWRET$	Maximum weekly return during a 1-month or 3-month horizon (specified in tables).
$MAX\|DRET\|$	Maximum absolute daily return during a 1-month or 3-month horizon (specified in tables).
$MAX\|WRET\|$	Maximum absolute weekly return during a 1-month or 3-month horizon (specified in tables).
$ATURN$	Abnormal monthly share turnover, defined as monthly turnover (total number of shares traded within a month divided by shares outstanding) divided by average monthly turnover over the prior 12 months, minus one and multiplied by 100. $ATURN_{w-1}$ is abnormal weekly share turnover, defined as turnover during the week prior to an earnings announcement date d divided by average weekly turnover over the past 2 months, minus one and multiplied by 100.
$nFCST$	Number of analyst forecasts for the nearest fiscal quarter.
$nREV$	Number of analyst forecast revisions since the last month.
ΔFA	Change in forecast accuracy. The error in the mean forecast is defined for the nearest fiscal quarter as the absolute value of the difference between the mean EPS forecast and the actual EPS as a percentage of the actual EPS. Forecast accuracy is defined as one minus forecast error. The monthly change (in percent) in forecast accuracy is defined as the current month forecast accuracy minus the prior month forecast accuracy, multiplied by 100. This measure is computed by firm and forecast period.
$EDGAR$	Number of human downloads (according to the methodology of Ryans (2017)) of a company's SEC filings from EDGAR during the month.
BBG	Number of days within the month when Bloomberg's "News Heat - Daily Maximum Readership" measure is equal to 3 or 4 out of 4.

Continued on next page

Variable	Definition
$\|CAR\|$	Absolute value of the cumulative abnormal return around a quarterly earnings announcement in percent. Abnormal returns are computed relative to the daily returns of a portfolio matched on size and book-to-market ratio. $\|CAR\|_d$ is computed on the earnings announcement date d. $\|CAR\|_{d-1,d+1}$ is computed over the three-day period around the announcement date. $\|CAR\|_{q+1}$ is computed over the period starting two days after the earnings announcement date through one day after the firm's next quarterly earnings announcement date.
$COMPLEX$	Firm complexity, defined as $1-\sum_{j=1}^{J} s_j^2$, where s_j is the fraction of the firm's total sales generated by industry segment j.
$CVOL$	Call-implied volatility of an at-the-money call option with 30 days to maturity.
$PVOL$	Put-implied volatility of an at-the-money put option with 30 days to maturity.
$ACVOL$	Abnormal call-implied volatility, defined as $CVOL$ divided by average $CVOL$ over the previous 12 months, minus one and multiplied by 100.
$APVOL$	Abnormal put-implied volatility, defined as $PVOL$ divided by average $PVOL$ over the previous 12 months, minus one and multiplied by 100.
$A\beta^{MKT}$	Abnormal market beta, defined as β^{MKT} divided by average monthly β^{MKT} over the previous 12 months, minus one and multiplied by 100.

Table A2: Transition probabilities for portfolios sorted by learning index

	Panel A: 1-month transition matrix				
	$LI1_{t+1}$	$LI2_{t+1}$	$LI3_{t+1}$	$LI4_{t+1}$	$LI5_{t+1}$
$LI1_t$	71.3	23.4	4.6	0.7	0.1
$LI2_t$	23.6	45.5	24.4	5.8	0.8
$LI3_t$	4.5	24.7	41.8	24.1	4.8
$LI4_t$	0.6	5.7	24.4	45.3	24.0
$LI5_t$	0.1	0.7	4.8	24.1	70.3
	Panel B: 6-month transition matrix				
	$LI1_{t+6}$	$LI2_{t+6}$	$LI3_{t+6}$	$LI4_{t+6}$	$LI5_{t+6}$
$LI1_t$	49.5	25.7	13.8	7.6	3.4
$LI2_t$	25.3	27.4	22.0	15.9	9.5
$LI3_t$	13.4	21.8	24.5	22.7	17.6
$LI4_t$	7.0	15.1	22.2	27.0	28.6
$LI5_t$	3.6	9.4	17.3	27.3	42.3
	Panel C: 12-month transition matrix				
	$LI1_{t+12}$	$LI2_{t+12}$	$LI3_{t+12}$	$LI4_{t+12}$	$LI5_{t+12}$
$LI1_t$	32.9	22.9	18.1	14.7	11.3
$LI2_t$	23.0	22.0	20.3	18.4	16.4
$LI3_t$	17.7	20.2	21.0	20.8	20.3
$LI4_t$	13.5	18.2	20.9	23.0	24.4
$LI5_t$	9.3	15.2	20.1	24.8	30.6
	Panel D: 24-month transition matrix				
	$LI1_{t+24}$	$LI2_{t+24}$	$LI3_{t+24}$	$LI4_{t+24}$	$LI5_{t+24}$
$LI1_t$	21.3	20.6	20.0	19.6	18.6
$LI2_t$	19.2	19.8	20.2	20.5	20.3
$LI3_t$	17.9	19.3	20.3	21.0	21.4
$LI4_t$	17.1	18.9	20.4	21.3	22.2
$LI5_t$	16.1	18.6	20.4	21.8	23.2
	Panel E: 36-month transition matrix				
	$LI1_{t+36}$	$LI2_{t+36}$	$LI3_{t+36}$	$LI4_{t+36}$	$LI5_{t+36}$
$LI1_t$	20.2	19.9	19.8	20.1	20.0
$LI2_t$	19.3	19.7	20.2	20.5	20.2
$LI3_t$	18.5	19.7	20.6	20.8	20.4
$LI4_t$	17.9	19.8	20.5	20.8	21.0
$LI5_t$	17.2	19.2	20.2	21.2	22.2

At the end of each month, stocks are sorted into quintiles based on values of the learning index (LI). For each LI quintile in month t, the table reports the time-series average of the percentage of stocks that fall in each LI quintile in month $t+1$ (Panel A), $t+6$ (Panel B), $t+12$ (Panel C), $t+24$ (Panel D), and $t+36$ (Panel E). Within each panel, percentages are calculated using only the stocks that exist in both the initial month and the final month.