The Review of Financial Studies

Information Choice, Uncertainty, and Expected Returns

Charles Cao
The Pennsylvania State University

David Gempesaw
Miami University

Timothy T. Simin
The Pennsylvania State University

We investigate how information choices affect equity returns and risk. Building on an existing theoretical model of information and investment choice, we estimate a learning index that reflects the expected benefits of learning about an asset. High learning index stocks have lower future returns and risk compared to low learning index stocks. Analysis of a conditional asset pricing model, long-run patterns in returns and volatilities, other measures of information flow, and the information environment surrounding earnings announcements reinforce our interpretation of the learning index. Our findings support the model's predictions and illustrate a novel empirical measure of investor learning. (*JEL* G12, G14)

Received July 8, 2019; editorial decision October 16, 2020 by Editor Stijn Van Nieuwerburgh.

Price discovery in the U.S. equity market is driven by the trading decisions of active investors. For every $1 million in passive index strategy trades, active investors trade approximately $22 million.[1] The information acquisition decisions of active investors are crucial for price discovery, but difficult to measure. Recent empirical studies have utilized observable outcomes, such as the holdings and investment returns of portfolio managers, to analyze the impact

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, Yanhui Zhao, and Alexey Zhdanov, as well as seminar participants at Miami University, Pennsylvania State University, the 2018 Financial Management Association Annual Meeting, and the 2019 CMU-PITT-PSU Finance Conference for helpful comments. Special thanks go to Stijn Van Nieuwerburgh (the editor) and two anonymous referees, whose comments have significantly improved the quality of the paper. Send correspondence to David Gempesaw, dgempesaw@miamioh.edu.

[1] See Black Rock (2017).

The Review of Financial Studies 34 (2021) 5977–6031

doi:10.1093/rfs/hhaa146 Advance Access publication January 25, 2021

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) (hereafter VNV). 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 lower 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 affects risk and return in the context of the equity market.

Using the learning index in this context comes with several advantages. 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 requires only 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 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 VNV and the information choices predicted by the learning index are testable without relying on assumptions about investors' unobservable information sets. Van Nieuwerburgh and Veldkamp (2009) and Veldkamp (2011) propose this property as an advantage of the theoretical analysis of information choice.

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 -0.49% per month or -6.1% 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 -0.49% per month or -6.0% per year. Coefficient estimates from two-stage cross-sectional multivariate regressions indicate that one cross-sectional standard deviation increase in LI is associated with an average cross-sectional reduction in expected monthly return of 7 basis points, holding all other

variables constant (for comparison, an increase of one standard deviation in firm size is associated with a reduction in the next month's return of 16 basis points). 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 affect risk. Using a bivariate portfolio sorting approach to control for average volatility over the past 12 months, we find that annualized return volatility in the following month is 2 percentage points lower for high LI stocks compared to low LI stocks. Decomposing 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 uncertainty and return comovement with systematic risk factors. These conclusions are robust to the use of multivariate cross-sectional and panel regressions of the next month's 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 returns derives from investors' decisions to reduce risk through learning.

We perform a number of analyses to evaluate the interpretation of the empirical learning index as a proxy for learning decisions and information flow. According to the VNV model, learning more about an asset lowers its market risk. We find that estimates of the impact of learning on capital asset pricing model (CAPM) beta are consistently negative over time, and accounting for that negative impact in a conditional version of the CAPM produces better expectations of returns than a number of well-known asset pricing models. We also 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 2 years, suggesting that the return predictive power of LI is due to investor learning rather than temporary price pressure or mispricing. Using a bivariate portfolio sorting approach to control for firm visibility or informational transparency, 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 VNV 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. The predictive ability of

the learning index is stronger during months when learning is concentrated among fewer firms as well as for earnings announcements containing a large surprise. These findings are consistent with more information being acquired for high LI stocks and incorporated into prices before earnings announcements.

As a final test, we examine whether the explanatory power of the learning index for expected returns is due to the effects on learning intensity of information supply or shocks to uncertainty. Information supply is measured using either stock price asynchronicity (one minus the R^2 from the market model regression) or analyst forecast coverage. Shocks to uncertainty are measured using either quarterly changes in earnings volatility or abnormal dispersion in analysts' forecasts. After controlling for these additional determinants of learning intensity, we continue to find a negative and significant relation between LI and the next month's returns.

We perform a number of robustness checks to provide support for our main conclusions. We verify that the explanatory power of the learning index for expected returns is robust to the use of alternative factor models. To evaluate the robustness of the learning index estimation procedure, we repeat our main portfolio sorting and regression analyses using two alternative versions of the learning index and arrive at qualitatively similar conclusions. The Internet Appendix reports the results of these tests.

This paper contributes to a line of research featuring empirical applications of noisy rational expectations equilibrium models focused on the information content of prices.[2] 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 VNV model, which treats investors' information sets as endogenous. Kacperczyk, Van Nieuwerburgh, and Veldkamp (2016) construct and test a closely related model of mutual fund managers' attention allocation and portfolio choices. While a number of common themes run between that model and the model underlying our paper, the empirical focus of each paper 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.

[2] 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 evidence suggesting that investors exhibit rational expectations and condition their beliefs on prices. Burlacu et al. (2012) develop a measure of information precision and supply uncertainty based on the work of Admati (1985) and investigate its relationship with expected returns.

Our paper also relates to the literature investigating the empirical relationship between information flow, expected returns, and risk.[3] 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 1-day returns, and search activity on various platforms.[4] 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 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 is applicable to different time periods, markets, and types of assets.

1. Overview of the Model

Our empirical analysis is based on the rational expectations model of information choice and investment choice developed by Van Nieuwerburgh and Veldkamp (2010). In the paper, the authors explore the impact of different assumptions regarding learning technologies and investor preferences on the optimal information acquisition strategy in partial equilibrium. We focus on the general equilibrium version of the model with mean variance preferences and an entropy learning technology, which is discussed in the internet appendix to their paper.[5] Combining these particular assumptions leads to a prediction of

[3] 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 decreases with the length of CEO tenure and argue that uncertainty is reduced over time as investors learn about a CEO's 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.

[4] Barber and Odean (2007) use news coverage, abnormal trading volume, and extreme 1-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 (Drake, Roulstone, and Thornock 2015) or news reading activity on Bloomberg terminals (Ben-Rephael, Da, and Israelsen 2017).

[5] The internet appendix to VNV can be found at https://www0.gsb.columbia.edu/faculty/lveldkamp/papers/portfolio_appdx.pdf.

specialized learning as opposed to generalized learning by investors.[6] In this section, we present an overview of their model and the details relevant to our empirical analysis.

The model contains N risky assets, a risk-free asset with return r, and a continuum of atomless investors. The per capita supply of the risky assets is $\overline{x}+x$, where $\overline{x}$ is a positive constant vector and x is a random vector with known mean and variance and zero covariance across assets: $x \sim N(0, \sigma_x^2 I)$.[7] The vector x can be interpreted as shocks due to liquidity, hedging, or life cycle needs of traders, noise trading, errors by investors when inverting prices, uninformed trading, or any other trading activity based on reasons orthogonal to prices and payoffs. The risky asset supply leads to noise in prices which prevents perfectly revealing prices and creates an incentive to learn.

The model unfolds over three periods. In period 1, ex ante identical investors 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 and a no-forgetting constraint. In period 2, investors observe an $N \times 1$ vector of information signals $\eta = f + e_\eta$, where $e_\eta \sim N(0, \Sigma_\eta)$. Each investor's information signal is an independent draw from the distribution they have chosen.[8] Information from signal realizations is combined with information from prior beliefs and asset prices to form posterior beliefs $\hat{\mu}$ and $\hat{\Sigma}$. Conditioning on this information, investors choose portfolio allocations subject to the budget constraint $W = W_0 r + q'(f - pr)$, where p is a vector of asset prices determined by market clearing. In period 3, investors receive asset payoffs and realize utility. Investors have mean variance preferences with an absolute risk-aversion coefficient ρ:[9]

$$U_1 = E_1\left[\rho E_2[W] - \frac{\rho^2}{2} V_2[W]\right]. \tag{1}$$

The model assumes independent asset payoffs and independent information signals about these payoffs. The assumption of independent asset payoffs

[6] VNV argue that entropy-based learning technology is preferable to additive technology since the former is scale neutral, which means that learning costs are unaffected by the definition of one share of an asset, and it leads to a prediction of specialized learning (learning about one asset or risk factor) rather than generalized learning (learning about multiple assets). While the specialization of learning is model specific and not a generic feature of entropy models (see Sims 2006), specialized learning is 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; Ivković, Sialm, and Weisbenner 2008). When combined with 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.

[7] Other related papers using this assumption include Van Nieuwerburgh and Veldkamp (2009), Kacperczyk, Van Nieuwerburgh, and Veldkamp (2016), and Kacperczyk, Nosal, and Stevens (2019).

[8] Each investor observes an unbiased signal of the truth that is combined with independent noise. The noise can be viewed as the error investors make when processing and interpreting information.

[9] See equation (18) of VNV.

is without loss of generality.[10] If asset payoffs 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 nondiagonal 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 risk factors, while the eigenvector matrix contains the loadings of the correlated assets on the risk factors.

The assumption of independent information signals is not without loss of generality, but makes the model tractable by reducing the signal covariance matrix Σ_η from $N(N+1)/2$ distinct elements to N.[11] This assumption implies that the variance of posterior beliefs has the same eigenvectors as the variance of prior beliefs, and the prior and posterior covariance matrices of the risk factors are diagonal. The interpretation of this assumption is that investors take the correlation structure of risk factors as given and choose how much to reduce risk through learning. This assumption does not necessarily prevent investors from learning about multiple risk factors. With independent assets and signals, the investor's information choice becomes equivalent to choosing the posterior variance (eigenvalue) for each asset.

The model is solved using backward induction. In period 2, the optimal investment choice is a diversified portfolio that conditions on posterior beliefs:

$$q^* = \frac{1}{\rho} \hat{\Sigma}^{-1} \left(\hat{\mu} - pr \right). \tag{2}$$

Similar to Admati (1985), equilibrium prices are a linear function of payoffs and supply shocks x:

$$pr = A + Bf + Cx. \tag{3}$$

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 (see proposition 1 of the internet appendix to VNV). Note that asset prices are a function of the posterior mean and variance of the average investor. This result is because the VNV model features a continuum agent economy as in Admati (1985). If prices take this form, then posterior beliefs can be expressed as follows using standard Bayesian updating:[12]

$$\begin{aligned} \hat{\mu} &\equiv E[f | \mu, \eta, p] \\ &= (\Sigma^{-1} + \Sigma_\eta^{-1} + \Sigma_p^{-1})^{-1} (\Sigma^{-1} \mu + \Sigma_\eta^{-1} \eta + \Sigma_p^{-1} B^{-1} (rp - A))^{-1} \end{aligned} \tag{4}$$

[10] See section 1.3 of VNV.

[11] See section A.1 of the internet appendix to VNV. In section B of their internet appendix, the authors show that relaxing this assumption (allowing correlated signals) does not change the learning behavior predicted by the model.

[12] See equations (9) and (10) of the internet appendix to VNV.

$$\hat{\Sigma} \equiv V[f|\mu, \eta, p] = (\Sigma^{-1} + \Sigma_{\eta}^{-1} + \Sigma_{p}^{-1})^{-1} \tag{5}$$

Because prior information and prices are taken as given from the investor's perspective, signal variances and posterior belief variances are uniquely associated. Therefore, the information allocation choice can be modeled as choosing posterior belief variance $\hat{\Sigma}$ instead of Σ_{η} .

Substituting (2) into (1) gives

$$U_1 = E_1\left[\frac{1}{2}(\hat{\mu} - pr)'\hat{\Sigma}^{-1}(\hat{\mu} - pr)\right] + rW_0. \tag{6}$$

At time 1, payoffs and prices are unknown, so $(\hat{\mu} - pr)$ is a random variable, and (6) is a noncentral χ^2-distributed random variable. Taking the expectation of (6) gives

$$U_1 = \frac{1}{2}\text{Trace}\left(\hat{\Sigma}^{-1}V_1[\hat{\mu} - pr]\right) + \frac{1}{2}E_1[\hat{\mu} - pr]'\hat{\Sigma}^{-1}E_1[\hat{\mu} - pr] + rW_0, \tag{7}$$

where $E_1[\hat{\mu} - pr] = (I - B)\mu - A$ and $V_1[\hat{\mu} - pr] = \Sigma - \hat{\Sigma} + B\Sigma B' + CC'\sigma_x^2 - 2B\Sigma$.[13]

The information choice is subject to two constraints. The first constraint is an information capacity constraint. With entropy-based learning, this means that the total amount of learning (i.e., the distance between Σ and $\hat{\Sigma}$) cannot exceed capacity K.[14] The second constraint is a nonnegative learning constraint, which means that investors cannot forget information contained in prior beliefs or prices. After incorporating the assumption of independent information signals and rearranging, the information choice problem in period 1 can be written as[15]

$$\begin{aligned} &\max_{\hat{\Sigma}_1, \ldots, \hat{\Sigma}_N} \frac{1}{2}\left[-N + \sum_{i=1}^{N} LIF_i \frac{\Sigma_i}{\hat{\Sigma}_i}\right] \\ &\text{subject to} \quad \prod_{i=1}^{N} \frac{\Sigma_i}{\hat{\Sigma}_i} = K \quad \text{and} \quad \left(\hat{\Sigma}_i^{-1} \geq \Sigma_i^{-1} + \Lambda_{pi}^{-1},\right) \quad \forall i \end{aligned} \tag{8}$$

where i refers to risk factor i.

---

[13] See section A.3 and equation (11) of the internet appendix to VNV.

[14] Two aspects of information acquisition cost should be considered: how much learning capacity to acquire and how to allocate that learning capacity across assets. VNV concentrate on the second of these two aspects. Since the capacity allocation decision only depends on the level of capacity required, the authors assume that the optimal level of capacity acquired is given exogenously by an unspecified utility cost function. Therefore, the cost of learning in their model is defined by the amount of learning capacity required to learn from a given information source. This cost is reflected in the entropy learning constraint. The constraint implies that investors use up their learning capacity to learn from private signals as well as prices.

[15] See equation (14) in the internet appendix to VNV.

In period 1, the optimal information choice is to allocate all learning capacity toward the factor with the highest value of the learning index. The learning index for factor i is[16]

$$LIF_i = \left(((I-B)\mu - A)'\Gamma_i\right)^2 \Lambda_i^{-1} + (1-\Lambda_{Bi})^2 + \Lambda_i^{-1}\Lambda_{Ci}^2\sigma_x^2. \quad (9)$$

The first term on the right side of Equation (9) is equivalent to the prior squared Sharpe ratio for factor i: $\frac{\left(E\left[\Gamma_i'(f-pr)\right]\right)^2}{Var[\Gamma_i' f]}$. Alternatively, this term can be viewed as the product of two terms: $E\left[\Gamma_i'(f-pr)\right]$ and $\frac{E\left[\Gamma_i'(f-pr)\right]}{Var[\Gamma_i' f]}$, which is equivalent to ρ times the expected investment in factor i. These two terms indicate that the value of learning is greater for a factor with a high expected excess return and a high expected portfolio share. Consequently, expecting to hold more of an asset makes learning about that asset more valuable, while learning more about an asset makes the asset less risky and more attractive to hold.

The second term on the right side of Equation (9) reflects expected pricing errors related to the informativeness of prices about payoffs for factor i. Λ_{Bi} is the ith 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 for factor i. Λ_i and Λ_{Ci} are the ith 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 an informed investor can exploit.

Since learning choices in VNV are over risk factors, we follow the method proposed in Van Nieuwerburgh and Veldkamp (2009) and Veldkamp (2011) to generate a learning index for individual stocks by premultiplying the risk factor learning indexes by the eigenvector matrix containing the loadings of the correlated assets on the risk factors.

$$LI = \Gamma(LIF) \quad (10)$$

The vector LI contains the learning index for each asset in the cross-section. We provide more detail in the methodology section below.

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

[16] See equations (12) and (13) in the internet appendix to VNV.

about. However, each individual employs a mixed strategy and randomizes over which of these factors to learn about.[17] Furthermore, since information signal realizations and posterior beliefs are random variables, the portfolio choices of ex ante identical investors will also differ based on the actual signal realization each investor receives.

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. In proposition 4 of the internet appendix to VNV, the authors 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 on past return information only.

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 comovement 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.

As is the case with any theoretical model, the model of VNV relies on a number of simplifying assumptions. Many of these assumptions have long been used in finance and economics, such as mean variance preferences and entropy-based learning constraints. While we base our empirical analysis on this model, our objective is not to argue that the model is a perfect representation of reality. One of the important and unique characteristics of the VNV model (which the authors themselves recognize) is that it can be tested using observable data, unlike alternative models with stylized assumptions about information endowments or dynamic overlapping generations models that may arguably be more realistic.

For instance, Mondria (2010) extends the VNV model by relaxing the assumption that investors take the structure of risk factors as given. In this model, investors can choose to observe a linear combination of risk factor payoffs as a private signal. Changing this assumption leads to novel predictions

[17] See section A.4 of the internet appendix to VNV. In equilibrium, expected returns, learning incentives, and aggregate information allocation choices all can be expressed as functions of the information set of the average investor.

about price comovement and the transmission of volatility shocks that are not present in the VNV model. Yet, unlike the VNV model which assumes N assets, the model only features two assets, which makes it unsuitable for direct empirical testing in the equity market. Spiegel (1998) proposes an overlapping generations model with multiple securities and homogeneously informed agents in order to explain the discrepancy between dividend and stock price volatilities. Watanabe (2008) extends this work by incorporating heterogeneous information and asymmetric information precision. The dynamic nature of these models produces multiple equilibria and a rich set of predictions about volatility, comovement in asset returns, and trading activity. However, these models assume information sets and signal precisions are exogenously determined, so they cannot be used to investigate predictions about information choices.

Another alternative approach to test information theories would be to use direct measures of information acquisition. Gargano and Rossi (2018) use a novel brokerage account data set that captures web activity on the brokerage website at the individual investor level. This data set allows them to study the relationship between information acquisition and performance from the investor's standpoint.[18] While this approach of studying information certainly grants some novel insights, it is hamstrung by other limitations related to data availability (e.g., the sample period is only 18 months). The learning index of the VNV model can be estimated for any time period and any set of assets with historical price information. While the learning index cannot be used to evaluate individual investor performance, it can be used to investigate outcomes at the asset level.

Thus, despite the potential limitations of the VNV model, we believe that focusing on an empirically testable and widely applicable model of information choice still represents an important step towards understanding the role of information choice in financial markets. In addition, recent empirical research qualitatively supports the prediction of the VNV model that investors specialize in their learning choices, leading to concentrated portfolios. These findings can provide at least some indication that the model is a reasonable approximation of investor learning.

2. Hypotheses

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.

[18] See Mondria, Wu, and Zhang (2010), Mondria and Quintana-Domeque (2012), and Cziraki, Mondria, and Wu (2020) for other examples of work using internet search and news article counts as measures of information acquisition.

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

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

The empirical asset pricing literature has produced a myriad of asset characteristics and risk factor models to explain cross-sectional stock returns. Although these characteristics and factors are not present in the VNV model, we incorporate them into our empirical analyses in order to control for omitted variable biases, reverse causality, and other alternative explanations for the empirical relationships we observe.

The third hypothesis relates to the model's prediction that learning reduces comovement with systematic risk factors, which is not captured by the unconditional CAPM.

Hypothesis 3. Stocks with a higher learning index have lower betas than what the CAPM predicts.

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 4a. The changes in price and volatility predicted by the learning index are long-lasting and do not reverse in the long run.

Hypothesis 4b. The learning index is correlated with other measures of information flow, such as analyst coverage.

Hypothesis 4c. Stocks with a higher learning index prior to an earnings announcement have smaller market reactions and greater abnormal trading activity.

Hypothesis 4d. The negative relation between the learning index and the market reaction to an earnings announcement is stronger during months with fewer announcements and for announcements with a larger unexpected component in announced earnings.

Hypothesis 4e. The explanatory power of the learning index for future returns is not attributable to the effects on learning intensity of information supply or shocks to uncertainty.

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 2-year rolling window of weekly returns to construct prices, payoffs, and an estimate of the payoff covariance matrix. The return measure we use is the holding period total return, which reflects the change in the total value of an investment and includes both capital appreciation and dividends. 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 nonsynchronous trading and other microstructure effects (see, e.g., Lo and Wang 2006). Following the convention in the literature, we measure weekly returns from Wednesday to Wednesday. The general approach is to use an eigenvalue decomposition to transform a set of correlated stocks to principal components (factors), then estimate the principal component learning index, and finally use the eigenvector matrix to transform the principal component learning index back to a learning index for the correlated stocks. In particular, 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 distributed lognormally, we use logarithmic 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 $f - pr$ are calculated as $\ln(\frac{P_{t+1}}{P_t}) - \ln(1+r_f)$, where r_f is the risk-free interest rate. To avoid look-ahead bias in the empirical tests, we base the estimation 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. To account for heteroscedasticity across individual assets, payoffs are standardized to have zero mean and unit variance prior to computing the covariance matrix and performing the eigen decomposition. Standardizing payoffs prior to estimating principal components avoids overweighting stocks with high idiosyncratic volatility when extracting the principal components and leads to extracted factors with evenly distributed explanatory power across individual assets.[19] Decompose Σ into a diagonal eigenvalue matrix Λ and an eigenvector matrix

[19] By definition, principal component analysis maximizes the total variance of each extracted factor. Since return volatilities of individual stocks are largely dominated by idiosyncratic volatility, extracted factors based on unstandardized returns will be significantly influenced by idiosyncratic returns. See Xu (2007) for further information on the approach of applying principal components analysis to the correlation structure of asset returns. Other papers using this approach include Stock and Watson (2002) and Bai and Ng (2002).

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 excess return by the variance of payoffs. The second and third terms 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.[20] The payoff coefficient Λ_B and the regression R^2 are used to compute the second and third terms.[21]

Step 4: Estimate the learning index for stocks. Premultiply the principal component learning index vector by the eigenvector matrix: $\Gamma(LIF)$. 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.[22]

Because these weights are the same across the three components of a given factor learning index for a given stock, the three components of the stock-level learning index are highly correlated. Multicollinearity problems make it impossible to assess the relative importance and independent effect of each component. While it is theoretically possible to decompose the factor-level learning indexes, doing this empirically is challenging. The ordering of principal components (according to the amount of variance explained) can vary from one month to the next, so the factors cannot be uniquely identified over time. In addition, the potential number of principal components varies with the size of the cross-section. We are primarily interested in examining the

---

[20] This step involves a time-series regression of two nonstationary 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 an ordinary least squares (OLS) regression. In untabulated analysis, we verify the stationarity of the residuals from this regression.

[21] When estimating $\left(1 - \Lambda_{Bi}\right)^2$, if prices follow the pricing equation $pr = A + Bf + Cx$, then an OLS regression can be used to directly estimate 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, $\Lambda_i^{-1} \Lambda_{Ci}^2 \sigma_x^2 = \frac{1-R^2}{R^2} \Lambda_{Bi}^2$.

[22] A well-known practical issue involved in eigen decomposition is that the sign of an eigenvector is arbitrary. Because of this, 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 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. In any given cross-section, we use the principal components with eigenvalues greater than or equal to one. Because the number of stocks and the number of principal components vary over time, we scale the learning index in each month by the ratio of the number of stocks divided by the number of principal components in that month. This is equivalent to scaling the learning index for all stocks by a constant within each cross-section and does not affect the cross-sectional explanatory power of the measure.

explanatory power of the learning index for the cross-section of stock returns, not principal component returns. One well-known disadvantage of principal component analysis is limited interpretability. As such, we do not attempt to directly analyze the principal components and only utilize them in order to apply the VNV model to a set of correlated assets following the authors' suggestion.

Variation in the empirical learning index across assets may seem contradictory to the theoretical equilibrium outcome that the learning index is equal across assets. However, ample evidence indicates that market frictions cause cross-sectional dispersion in learning over short time periods. Menzly and Ozbas (2010) claim that investor specialization promotes partial incorporation of information into prices since a subset of investors fails to recover informative signals from observed prices as in Grossman and Stiglitz (1980). This causes market information segmentation resulting in return predictability across economically related firms. Additionally, investors may be limited by how much they can reduce uncertainty about a particular asset or by their aggregate learning capacity. For instance, predictability due to the slow propagation of information flow across economically related firms is attributed to attention constraints of investors (Cohen and Frazzini 2008), investor recognition (Hou and Moskowitz 2005), and other market frictions or institutional constraints (Hou 2007).

Transactional frictions, such as constraints on short sales, can also lead to cross-sectional differences in learning. Stambaugh, Yu, and Yuan (2015) term the differing willingness or ability of traders to short stocks, coupled with cross-sectional differences in the supply of shortable stocks, arbitrage asymmetry and cite a long list of studies addressing the impact of arbitrage asymmetry in the equity market. Another potential cause of cross-sectional frictions to learning may be found in changes to market sentiment. The long-run information content of market sentiment is plausible behaviorally if people's optimism or pessimism develops into a consensus view indicating that the importance of sentiment may build over time, and rationally if arbitrage forces, which are likely to eliminate short-run mispricing, fail at longer horizons. An example of this limit to arbitrage is the noise trader risk described in De Long et al. (1990). Brown and Cliff (2005) find that while sentiment predicts longer-run returns, they find little predictive evidence for near-term returns.

The actions of insiders and specialists also affect the speed of information transmission. In noisy rational expectations models where informed traders can trade in the options market first (Easley, O'Hara, and Srinivas 1998; An et al. 2014), the existence of noise traders allows informed traders to mask the information content of trades, leading to inefficient and hence predictable prices. Brennan and Hughes (1991) describe how brokers have pricing incentives to produce information on low price stocks. Lowered prices due to stock splits direct the attention paid to a firm by investment analysts. Hong and Li (2019) find evidence that the cessation of trading by insiders who routinely trade predicts future returns and fundamentals.

Corwin and Coughenour (2008) present results showing specialists with limited attention shift effort to the most actively trading stocks, causing less price improvement and higher transaction costs for the remaining stocks they cover. Sudden shifts in attention by information producers and changes in insider trading can plausibly cause changes in the cross-sectional learning environment.

These types of frictions 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 appendix, which reports transition probabilities for LI-sorted quintile portfolios over 1-, 6-, 12-, 24-, and 36-month periods. Consistent with the theoretical expectation, the value of learning about a particular asset declines as more investors learn about it.

3.2 Data sources and variable definitions

We obtain daily and monthly data for U.S. 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 20th 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.[23] Additional data sources include Compustat, Institutional Brokers' Estimate System (I/B/E/S), SEC Electronic Data Gathering, Analysis, and Retrieval (EDGAR) Log Files, and Bloomberg. 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 weekly data from the past 2 years. To account for biases due to infrequent trading, we follow Dimson (1979) and include 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 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.

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

Investment (*INV*) is the annual percentage change in total assets. Momentum (*MOM*) is the cumulative return from months $t-11$ to $t-1$.

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 months $t-59$ to $t-12$. 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.[24] 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 one if the firm paid dividends during the most recent fiscal quarter, and zero 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 (expressed as a percentage). Table A1 lists all variable definitions.

3.3 Descriptive statistics

Table 1 presents time-series averages of monthly cross-sectional summary statistics for the aforementioned stock characteristics. In the average month, the average stock in the sample has a market beta of 1.07, market capitalization of \$3.62 billion (the logarithm of market capitalization is tabulated), 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). Table 2 includes average cross-sectional correlations for key variables. 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.

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

Table 1
Cross-sectional summary statistics of stock characteristics

			Percentiles		
	Mean	SD	25th	50th	75th
LI	0.460	0.589	0.111	0.242	0.566
β^{MKT}	1.066	0.531	0.685	0.998	1.371
SIZE	6.503	1.244	5.501	6.244	7.269
BM	0.705	0.483	0.377	0.626	0.921
PROF	0.822	0.847	0.389	0.670	1.063
INV	0.172	0.303	0.030	0.097	0.205
MOM	20.700	47.046	−4.275	12.788	34.418
ILLIQ	0.217	1.071	0.018	0.059	0.183
STR	1.760	10.003	−3.848	1.073	6.458
LTR	1.111	2.072	0.170	0.640	1.361
RVOL	34.203	18.001	22.403	30.474	41.891
IVOL	24.464	14.126	15.339	21.376	30.022
SVOL	22.738	12.714	14.153	20.154	28.420
ROE	3.384	5.165	1.881	3.369	4.971
ROEVOL	4.325	14.109	0.900	1.644	3.244
AGE	23.621	17.840	10.213	17.875	32.890
DIVD	0.709	0.422	0.422	1.000	1.000
LEV	2.188	4.033	0.318	0.747	1.669
INVPRC	4.106	2.483	2.436	3.483	5.055
R	1.102	9.571	−4.246	0.730	5.980
# 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 20th percentile of NYSE stocks at the end of each month. The table summarizes the following characteristics: the learning index (*LI*), market beta (β^{MKT}), logarithm 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 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
Cross-sectional correlations for key variables

	LI	β^{MKT}	*SIZE*	*BM*	*PROF*	*INV*	*MOM*	*ILLIQ*	*STR*	*LTR*	*IVOL*
LI	1.00										
β^{MKT}	−0.08	1.00									
SIZE	−0.08	−0.02	1.00								
BM	0.04	−0.13	−0.13	1.00							
PROF	−0.04	0.11	−0.02	−0.29	1.00						
INV	−0.04	0.20	−0.05	−0.20	0.18	1.00					
MOM	−0.19	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	−0.01	0.01	−0.01	0.02	0.01	−0.01	0.02	0.03	1.00		
LTR	−0.07	0.16	0.02	−0.28	0.17	0.32	−0.03	−0.01	−0.02	1.00	
IVOL	0.03	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}), logarithm 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 appendix for complete variable definitions. The sample period is July 1964 through December 2016.

4. Learning Index and the Cross-Section of Returns

An implication of the model by VNV 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, two-stage cross-sectional regressions, and panel 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:

$$\begin{aligned} R_{p,t}-R_{f,t}=&\alpha_p+\beta_{1,p}\left(R_{M,t}-R_{f,t}\right)+\beta_{2,p}SMB_t+\beta_{3,p}HML_t \\ &+\beta_{4,p}RMW_t+\beta_{5,p}CMA_t+\beta_{6,p}UMD_t+\varepsilon_{p,t}. \end{aligned} \tag{11}$$

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 heteroscedasticity. In panel A, the highest *LI* quintile has an average excess return of 0.24% in the month following portfolio formation, while the lowest *LI* quintile has an average excess monthly return of 0.73%. The difference in excess returns between these quintiles is −0.49% per month (−6.1% 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 −0.48% with a t-statistic of −3.89. 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 −0.66% (−0.49%) per month or −8.1% (−6.0%) per year. Each of these estimates is significant at the 1% level.

Table 3
Returns of stocks sorted by learning index

Quintile	Excess return	FF3 α	FF5 α	FF6 α
		A. Value-weighted portfolios		
1 (Low *LI*)	0.734	0.227	0.296	0.220
2	0.556	0.043	0.020	0.033
3	0.485	−0.037	−0.099	−0.034
4	0.360	−0.163	−0.233	−0.140
5 (High *LI*)	0.242	−0.253	−0.359	−0.265
5−1	−0.491***	−0.480***	−0.655***	−0.485***
t-stat	(−3.49)	(−3.89)	(−5.43)	(−3.35)
		B. Equal-weighted portfolios		
1 (Low *LI*)	1.020	0.303	0.334	0.317
2	0.832	0.115	0.093	0.119
3	0.666	−0.035	−0.064	−0.007
4	0.567	−0.127	−0.182	−0.112
5 (High *LI*)	0.454	−0.206	−0.290	−0.240
5−1	−0.566***	−0.510***	−0.623***	−0.558***
t-stat	(−4.68)	(−5.05)	(−6.68)	(−4.54)

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's value-weighted (panel A) and equal- weighted (panel B) average monthly excess return and risk-adjusted excess return (alpha or α) for each quintile. 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 are given for the 5 − 1 portfolio. The sample period is July 1964 to December 2016. * p <.1; ** p <.05; *** p <.01 (significance levels for two-sided tests are indicated).

Table 3, panel B reports results using the returns of equal-weighted portfolios. Quintile 5 has an average excess return of 0.45% and quintile 1 has an average excess return of 1.02% 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 −0.56% per month (−6.3%, −7.7%, and −6.9% 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 models 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 3 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).

To facilitate interpretation of the coefficient estimates, we standardize all explanatory variables within each month to a mean of zero and standard deviation of one. In the first stage, we estimate monthly cross-sectional regressions of excess stock returns in month $t+1$ on values of the learning index 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.[25] The full cross-sectional model estimated at the end of each month is

$$\begin{aligned} R_{i,t+1} - R_{f,t+1} = {} & \lambda_0 + \lambda_1 LI_{i,t} + \lambda_2 \beta_{i,t}^{MKT} + \lambda_3 SIZE_{i,t} \\ & + \lambda_4 BM_{i,t} + \lambda_5 PROF_{i,t} + \lambda_6 INV_{i,t} \\ & + \lambda_7 MOM_{i,t} + \lambda_8 ILLIQ_{i,t} + \lambda_9 STR_{i,t} \\ & + \lambda_{10} LTR_{i,t} + \lambda_{11} IVOL_{i,t} + \varepsilon_{i,t+1}. \end{aligned} \tag{12}$$

[25] We find that the results are robust to the inclusion of additional cross-sectional return predictors as controls, including return volatility, skewness, coskewness, kurtosis, maximum daily return in the past month, share turnover, institutional ownership, number of institutional owners, number of analysts' forecasts, the call-put option implied volatility spread, and the Stambaugh, Yu, and Yuan (2015) mispricing measure. We also repeat the cross-sectional regression analysis of returns without standardizing the explanatory variables within each month. As expected, the magnitudes of the coefficient estimates differ from the respective estimates in our main tests. Nevertheless, we continue to find that the learning index is negatively and significantly related to future returns.

Table 4
Cross-sectional regressions of returns on learning index

	(1)	(2)
LI	−0.139***	−0.069***
	(−4.12)	(−3.90)
β^{MKT}		0.002
		(0.02)
SIZE		−0.161***
		(−3.75)
BM		0.063
		(1.51)
PROF		0.069**
		(2.13)
INV		−0.140***
		(−5.79)
MOM		0.177***
		(2.81)
ILLIQ		−0.043***
		(−2.73)
STR		−0.342***
		(−6.81)
LTR		−0.083***
		(−3.25)
IVOL		−0.120***
		(−3.70)
Adj. R^2	.005	.091

This table presents results from two-stage cross-sectional regressions. At the end of each month, we estimate a cross-sectional regression of the next month's excess stock return on a set of explanatory variables. All explanatory variables are standardized within each month to a mean of zero and standard deviation of one. Each column reports the average slope coefficients for each different regression specification. The explanatory variable of interest is the learning index (*LI*). Control variables include 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 appendix for complete variable definitions. 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. This regression analysis is based on 814,089 stock-month observations from July 1966 to December 2016 with no missing values for all variables. $^*p < .1$; $^{**}p < .05$; $^{***}p < .01$ (significance levels for two-sided tests are indicated).

In the second stage, we calculate the time-series averages of the cross-sectional regression coefficient estimates.[26]

Table 4 reports average 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 from this regression is negative and significant at the 1% level. The reported univariate coefficient estimate can be interpreted as the average return difference associated with a one standard deviation difference in the learning index in an average month. Column 2 presents results using the full regression specification. After controlling for several stock characteristics, the coefficient for the learning index remains negative and statistically significant. The average slope coefficient for *LI* is −0.069 (*t*-statistic=−3.90).

[26] 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. The precision-weighted results are qualitatively similar to those reported in Table 4.

The signs of the coefficient estimates for the control variables are generally in accordance with the findings of past studies. The significant coefficient estimates indicate that stocks with lower size, 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 and book-to-market ratio are positive but insignificant.[27] The results indicate 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. Our sample does not include stocks with market capitalization below the 20th NYSE percentile threshold or price less than $5.

Because the explanatory variables are standardized, the coefficient estimates reported in Table 4 can be compared to get a sense of the relative economic importance of each of the variables in explaining the next month's returns. Based on the coefficient estimates in column 2, the short-term reversal effect carries the strongest explanatory power for the next month's returns. An increase of one standard deviation in *STR* results in a decrease in the next month's return of 0.34% on average, all else equal. Increases of one standard deviation in *MOM*, *SIZE*, *INV*, *IVOL*, *LTR*, and *LI* are associated with average cross-sectional differences in expected monthly return of 0.18%, −0.16%, −0.14%, −0.12%, −0.08%, and −0.07% respectively, holding all other variables constant. Thus, even after controlling for several well-known return predictors, the explanatory power of the learning index for expected returns is still economically significant.[28]

4.3 Panel regressions

We estimate Equation (12) as a panel regression to further evaluate the relationship between the learning index and expected returns. In these regressions, we use stock fixed effects and/or month fixed effects and two-way clustered standard errors at the stock-month level. Explanatory variables are standardized to facilitate interpretation of the coefficient estimates.

---

[27] Fama and French (2015) discuss the redundancy of HML in the five-factor model.

[28] 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 452 anomaly variables, Hou, Xue, and Zhang (2020) find that the explanatory power of 65% of these variables become insignificant at the 5% level after controlling for microcaps (stocks below the 20th NYSE percentile) and using value-weighted portfolios instead of equal-weighted portfolios. To address 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 Tables 4 and 5 are based on a sample that excludes microcap stocks and have t-statistics exceeding the higher cutoff of 3.0.

Table 5
Panel regressions of returns on learning index

	(1)	(2)	(3)	(4)
LI	−0.126***	−0.139***	−0.126***	−0.137***
	(−4.03)	(−4.53)	(−4.05)	(−4.67)
β^{MKT}	0.008	−0.008	0.008	−0.028
	(0.09)	(−0.12)	(0.09)	(−0.42)
SIZE	−0.154***	−1.716***	−0.154***	−1.971***
	(−3.62)	(−11.02)	(−3.60)	(−15.73)
BM	0.075	0.148**	0.075	0.055
	(1.53)	(1.98)	(1.52)	(0.96)
PROF	0.087**	0.037	0.087**	0.012
	(2.39)	(1.17)	(2.38)	(0.44)
INV	−0.162***	−0.135***	−0.162***	−0.112***
	(−6.35)	(−4.26)	(−6.30)	(−3.77)
MOM	0.160**	0.181**	0.160**	0.204***
	(2.14)	(2.49)	(2.14)	(2.99)
ILLIQ	−0.058***	−0.068***	−0.058***	−0.060***
	(−3.18)	(−3.00)	(−3.17)	(−2.83)
STR	−0.197***	−0.198***	−0.197***	−0.184***
	(−3.08)	(−3.15)	(−3.08)	(−2.99)
LTR	−0.074**	−0.101**	−0.074**	−0.079**
	(−2.26)	(−2.57)	(−2.22)	(−2.11)
IVOL	−0.115**	−0.079**	−0.115**	−0.086**
	(−2.08)	(−2.18)	(−2.08)	(−2.45)
Stock FE	No	Yes	No	Yes
Month FE	No	No	Yes	Yes
R^2	.001	.005	.002	.007

This table presents results from panel regressions of the next month's excess stock return on a set of explanatory variables. Each column reports results for a different regression specification. All explanatory variables are standardized within each month to a mean of zero and standard deviation of one. Explanatory variables include 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 appendix for complete variable definitions. The within R^2 is reported in the last row. Standard errors are clustered at the stock-month level for all specifications. This regression analysis is based on 814,089 stock-month observations from July 1966 to December 2016 with no missing values for all variables. * $p < .1$; ** $p < .05$; *** $p < .01$ (significance levels for two-sided tests are indicated).

Consistent with the results from the cross-sectional return regressions, the results in Table 5 indicate a statistically and economically significant negative relationship between the learning index and the next's month returns. In column 1, the regression is estimated with no fixed effects included. The coefficient for *LI* is −0.126 with a *t*-statistic of −4.03. When only stock fixed effects are added in column 2, the coefficient for *LI* is −0.139 (*t*-statistic = −4.53). When only month fixed effects are added in column 3, the coefficient for *LI* is −0.126 (*t*-statistic = −4.05). When both stock and month fixed effects are included in column 4, the coefficient for *LI* is −0.137 (*t*-statistic = −4.67).

Based on the magnitudes of the coefficient estimates in column 4, the explanatory power of the learning index for future returns is comparable to that of investment, momentum, and short-term reversal. The signs of the significant coefficients are consistent with those in the cross-sectional regressions in Table 4. After controlling for stock and month fixed effects, the coefficients for book-to-market and profitability are insignificant.

5. Learning Index and the Cross-Section of Volatility

In this section, we investigate the cross-sectional relationship between the learning index and return volatility. The model predicts that learning reduces uncertainty relative to investors' prior beliefs. As such, in all of our main volatility tests, we examine the relationship between the learning index and future volatility, while controlling for prior volatility.

5.1 Portfolio sorting

In this section, we use a bivariate sorting procedure to examine the relationship between *LI* and future volatility, while controlling for past volatility (average return volatility over the prior 12 months). As the model predicts that learning also reduces the systematic component of risk, we also use systematic and idiosyncratic volatilities as dependent variables (and past systematic volatility or past idiosyncratic volatility as the respective control variable).[29] At the end of each month, stocks are sorted into quintiles based on average volatility over the past 12 months. Within each volatility quintile, stocks are sorted based on values of *LI*, and then each *LI* subquintile is combined across volatility quintiles into a single quintile.

Table 6 presents value-weighted average (panel A), equal-weighted average (panel B), and median (panel C) values of return volatility (*RVOL*), systematic volatility (*SVOL*), and idiosyncratic volatility (*IVOL*) expressed as a percentage for each *LI* quintile. Based on value-weighted averages in panel A, the next month's return volatility is 2.09% lower on average for high *LI* stocks relative to low *LI* stocks after controlling for past volatility. This difference is significant at the 1% level. Results are qualitatively similar when we examine equal-weighted average volatility or portfolio median volatility. The results in the next column suggest that the information choices of investors predict cross-sectional differences in systematic volatility. In panel A, *SVOL* in the month following portfolio formation is 2.27% lower for high *LI* stocks compared to low *LI* stocks after controlling for past *SVOL*, with a t-statistic of -5.90. Again, we arrive at similar conclusions using equal-weighted average or median systematic volatility in panels B and C. In the last column, we examine the relationship between the learning index and idiosyncratic volatility. After controlling for past *IVOL*, the difference in the next month's *IVOL* between extreme *LI* quintiles is negative but insignificant in panel A. However, this difference is significant using the equal-weighted portfolio average or portfolio median: -0.58% in panel B and -0.79% in panel C (t-statistics of -3.28 and -4.85, respectively).

At first glance, the spreads in volatility predicted by the learning index may not seem economically significant in relation to the overall level of

[29] For robustness, we also use a bivariate independent sorting approach and alternative factor model specifications to measure systematic and idiosyncratic volatility. Our conclusions are qualitatively similar under each of these robustness checks.

Table 6
Volatility of stocks sorted by learning index Bivariate dependent sorting

Quintile	*RVOL*	*SVOL*	*IVOL*
	A. Value-weighted portfolios		
1 (Low *LI*)	28.71	21.20	18.29
2	27.56	20.03	17.88
3	27.37	19.71	17.99
4	27.13	19.39	18.01
5 (High *LI*)	26.62	18.93	17.91
5−1	−2.09***	−2.27***	−0.39
t-stat	(−4.33)	(−5.90)	(−1.22)
	B. Equal-weighted portfolios		
1 (Low *LI*)	34.25	23.07	24.21
2	33.81	22.65	24.02
3	33.54	22.42	23.90
4	33.35	22.25	23.82
5 (High *LI*)	32.91	21.80	23.63
5−1	−1.34***	−1.27***	−0.58***
t-stat	(−4.40)	(−5.31)	(−3.28)
	C. Portfolio median		
1 (Low *LI*)	31.09	20.67	21.64
2	30.46	20.20	21.30
3	30.15	19.98	21.12
4	29.96	19.83	20.98
5 (High *LI*)	29.59	19.50	20.85
5−1	−1.50***	−1.18***	−0.79***
t-stat	(−5.62)	(−5.76)	(−4.85)

The table presents results for the cross-sectional relationship between the learning index (*LI*) and volatility in the following month, controlling for past volatility using bivariate dependent sorting. At the end of each month, stocks are sorted into quintiles based on average volatility over the past 12 months. Within each volatility quintile, stocks are sorted based on values of *LI*. Each *LI* subquintile is combined across volatility quintiles into a single quintile. The next month's value-weighted average (panel A), equal-weighted average (panel B), and median (panel C) values of return volatility (*RVOL*), systematic volatility (*SVOL*), and idiosyncratic volatility (*IVOL*) expressed as a percentage are computed for each *LI* quintile. See Table A1 for complete variable definitions. The row labeled "5−1" presents the difference in volatility between the highest and lowest quintile portfolios. Newey and West (1987) *t*-statistics are given for the 5−1 portfolio. The sample period is July 1964 to December 2016. * p <.1; ** p <.05; *** p <.01 (significance levels for two-sided tests are indicated).

volatility (about 7% and 7.5% of return volatility for quintile 1 and quintile 5, respectively). However, it is important to recognize that stocks likely carry a high degree of uncertainty that simply cannot be resolved through learning. To evaluate the economic significance of the learning index, we use a multivariate regression approach in the following two sections. This approach allows us to control for other stock characteristics and compare the explanatory power of *LI* with that of several other variables identified in the literature as predictors of volatility.

Altogether, the results from these sorting analyses indicate that learning is associated with a cross-sectional reduction in both (a) the firm-specific and systematic components of risk and (b) total 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 comovement with the market or other common risk factors.

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, systematic, and idiosyncratic volatilities in a multivariate setting. To facilitate interpretation of the coefficient estimates, we standardize all explanatory variables within each month to a mean of zero and standard deviation of one. 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 of the cross-sectional regression coefficient estimates from the first stage.[30] The full cross-sectional model estimated at the end of each month is

$$\begin{aligned} VOL_{i,t+1} = {} & \lambda_0 + \lambda_1 LI_{i,t} + \lambda_2 ROE_{i,t} + \lambda_3 ROEVOL_{i,t} + \lambda_4 AGE_{i,t} \\ & + \lambda_5 DIVD_{i,t} + \lambda_6 LEV_{i,t} + \lambda_7 INVPRC_{i,t} + \lambda_8 R_{i,t+1} \\ & + \lambda_9 SIZE_{i,t} + \lambda_{10} BM_{i,t} + \lambda_{11} MOM_{i,t} + \lambda_{12} STR_{i,t} \\ & + \sum_{j=0}^{11} \gamma_{j,t} VOL_{i,t-j} + \varepsilon_{i,t+1}, \end{aligned} \tag{13}$$

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 their findings, 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's (*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. 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.[31]

---

[30] Results based on Litzenberger and Ramaswamy (1979) precision-weighted time-series averages are qualitatively similar.

[31] We repeat the cross-sectional regression analysis of volatility without standardizing the explanatory variables within each month. While the magnitudes of the coefficient estimates differ from the respective estimates in our main tests, we continue to find that the learning index is negatively and significantly related to future volatility.

Table 7
Cross-sectional regressions of volatility on learning index

Dependent variable:	*RVOL*		*SVOL*		*IVOL*	
	(1)	(2)	(3)	(4)	(5)	(6)
LI	−0.286***	−0.258***	−0.228***	−0.164***	−0.159***	−0.185***
	(−5.35)	(−6.74)	(−5.87)	(−6.04)	(−3.98)	(−6.18)
ROE		−0.210***		−0.151***		−0.185***
		(−8.17)		(−7.51)		(−8.72)
ROEVOL		0.079***		0.052***		0.083***
		(3.17)		(2.65)		(4.07)
AGE		−0.161***		−0.107***		−0.155***
		(−5.58)		(−4.41)		(−6.94)
DIVD		−0.330***		−0.242***		−0.327***
		(−7.22)		(−6.49)		(−8.73)
LEV		0.015		0.079		−0.095
		(0.15)		(1.01)		(−1.43)
INVPRC		0.706***		0.502***		0.639***
		(10.34)		(9.85)		(11.66)
R		0.689***		0.435***		0.543***
		(3.45)		(3.42)		(3.49)
SIZE		−0.028		0.139**		−0.193***
		(−0.43)		(2.04)		(−5.68)
BM		−0.344***		−0.243***		−0.308***
		(−6.94)		(−6.15)		(−8.11)
MOM		0.109		0.237***		−0.043
		(1.16)		(2.75)		(−0.76)
STR		−0.944***		−0.500***		−0.734***
		(−8.10)		(−5.27)		(−9.35)
Lagged volatilities	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R^2	.492	.532	.438	.478	.419	.455

This table presents results from two-stage cross-sectional regressions. At the end of each month, we estimate a cross-sectional regression of the next month's return volatility (*RVOL*), systematic volatility (*SVOL*), or idiosyncratic volatility (*IVOL*) expressed as a percentage on a set of explanatory variables. All explanatory variables are standardized within each month to a mean of zero and standard deviation of one. Each column reports the average slope coefficients for each different regression specification. The explanatory variable of interest is the learning index (*LI*). Control variables include return on equity (*ROE*), volatility of return on equity (*ROEVOL*), firm age (*AGE*), a dividend dummy (*DIVD*), leverage (*LEV*), inverse of stock price (*INVPRC*), the next month's return (*R*), firm size (*SIZE*), book-to-market ratio (*BM*), momentum (*MOM*), short-term reversal (*STR*), and 12 lagged values of volatility. See Table A1 in the appendix for complete variable definitions. 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. This regression analysis is based on 686,245 stock-month observations from December 1974 to December 2016 with no missing values for all variables. $^*p < .1$; $^{**}p < .05$; $^{***}p < .01$ (significance levels for two-sided tests are indicated).

Table 7 reports the cross-sectional regression results for volatility. In the first column, we estimate monthly cross-sectional regressions of total return volatility in the next month on *LI*, while controlling for lagged monthly volatilities over the past year. In this specification, we find a negative and significant relation between the learning index and volatility, supporting the prediction that learning is associated with a reduction in uncertainty in the cross-section. Column 2 of Table 7 contains results from the full regression specification. 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. 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 average coefficient for *LI* is −0.258 (*t*-statistic = −6.74).

Based on coefficient estimates in column 2, variation in the next month's returns, current month's returns, 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 *STR*, *INVPRC*, and *R* are associated with average cross-sectional differences of −0.94%, 0.71%, and 0.69% in the next month's 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 0.26%, 0.34%, 0.33%, and 0.21% on average, holding all other variables constant.

Next, we focus on explaining the systematic and idiosyncratic components of return volatility using a similar cross-sectional multivariate analysis. Table 7 presents average coefficient estimates from regressions of systematic volatility in columns 3 and 4 and idiosyncratic volatility in columns 5 and 6. In column 3, 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 for *LI* is −0.228 and is significant at the 1% level. Column 4 contains results from regressions of the next month's *SVOL* on the full set of explanatory variables. The results for control variables are consistent with those in column 2 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. In column 4, the average coefficient estimate on *LI* is −0.164 (*t*-statistic = −6.04). In terms of economic significance, the explanatory power of *LI* for the systematic component of the next month's volatility is comparable to that of return on equity and firm size.

In column 5 of Table 7, we repeat the cross-sectional regression analyses using idiosyncratic volatility as the dependent variable and lagged values of idiosyncratic volatility as controls. The average coefficient estimate on the learning index measure is again negative and significant at the 1% level. In the final column, the other control variables are added to the regression. Similar to the previous results in Table 7, the results indicate that the control variables exhibit significant explanatory power for cross-sectional variation in the next month's idiosyncratic volatility. We find that firm size is negatively related to *IVOL*. Thus, it appears that combining the negative effects of size on *IVOL* with the positive effects on *SVOL* results in the insignificant relation with total volatility reported in column 2. After controlling for a number of other stock characteristics, we find that the explanatory power of the learning index for

cross-sectional variation in *IVOL* becomes stronger. The average coefficient estimate on the learning index measures −0.185 and is significant at the 1% level. In terms of economic significance, the cross-sectional explanatory power of *LI* for the next month's *IVOL* is comparable to that of firm size, firm age, and return on equity.

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.

5.3 Panel regressions

Similar to our analysis of expected returns, we also estimate panel regressions to further evaluate the relationship between the learning index and the three measures of volatility. The regression specification is similar to Equation (13), except we do not include lagged volatilities as controls. Instead, we used stock fixed effects to control for differences in average volatility across stocks. The panel regression specification also includes month fixed effects and two-way clustered standard errors at the stock-month level. Explanatory variables are standardized to facilitate interpretation of the coefficient estimates.

Table 8 presents results from panel regressions of return volatility on the learning index. In column 1, the regression is estimated with no fixed effects included. The coefficient for *LI* is −0.655 with a *t*-statistic of −6.96. When only stock fixed effects are added in column 2, the coefficient for *LI* is −0.801. When only month fixed effects are added in column 3, the coefficient for *LI* is −0.655. Each of these estimates is significant at the 1% level. In column 4, the regression is estimated with both stock and month fixed effects included. The coefficient for *LI* is −0.800 (*t*-statistic = −13.32). Based on the magnitudes of the coefficient estimates in column 4, the explanatory power of the learning index for future volatility is comparable to that of the dividend dummy, book-to-market ratio, and short-term reversal. The signs of the control variables are consistent with those of the control variables in the cross-sectional regressions in Table 7, with the exception of leverage, firm size, and momentum. In the cross-sectional regressions, these three variables are not statistically significant. In the panel regression with stock and month fixed effects, each of these variables is positively and significantly related to the next month's volatility.

We perform a similar set of panel regressions in columns 5 and 6 using systematic and idiosyncratic volatilities as the dependent variable. The results indicate that *LI* is negatively related to both the systematic and idiosyncratic components of return volatility. When stock and month fixed effects are included, the coefficient for *LI* is −0.746 (*t*-statistic = −14.01) in the *SVOL* regression in column 5 and −0.350 (*t*-statistic = −10.89) in the *IVOL* regression in column 6. Overall, the findings from these panel regressions reinforce our

Table 8
Panel regressions of volatility on learning index

Dependent variable:	*RVOL* (1)	*RVOL* (2)	*RVOL* (3)	*RVOL* (4)	*SVOL* (5)	*IVOL* (6)
LI	−0.655***	−0.801***	−0.655***	−0.800***	−0.746***	−0.350***
	(−6.96)	(−11.99)	(−7.35)	(−13.32)	(−14.01)	(−10.89)
ROE	−1.201***	−0.361***	−1.201***	−0.421***	−0.247***	−0.332***
	(−12.31)	(−5.77)	(−13.39)	(−7.58)	(−5.28)	(−9.40)
ROEVOL	0.671***	0.385***	0.671***	0.297***	0.227***	0.175***
	(6.48)	(3.65)	(7.48)	(3.45)	(3.43)	(3.07)
AGE	−1.576***	−0.414	−1.576***	−12.435***	−4.296***	−12.940***
	(−11.33)	(−0.63)	(−11.74)	(−6.54)	(−2.66)	(−11.23)
DIVD	−3.420***	−0.913***	−3.420***	−1.092***	−0.823***	−0.678***
	(−21.72)	(−6.41)	(−22.76)	(−9.72)	(−9.01)	(−9.58)
LEV	−0.723***	2.872***	−0.723***	2.624***	1.979***	1.606***
	(−4.38)	(7.69)	(−4.64)	(9.22)	(8.43)	(9.69)
INVPRC	3.781***	2.976***	3.781***	2.941***	1.696***	2.331***
	(27.95)	(15.55)	(30.00)	(19.05)	(13.95)	(23.34)
R	0.456***	0.533***	0.456***	0.512***	0.265***	0.465***
	(2.80)	(4.73)	(2.80)	(4.43)	(2.83)	(6.24)
SIZE	−0.227	1.354***	−0.227	1.252***	1.589***	0.056
	(−1.49)	(3.68)	(−1.56)	(4.01)	(6.03)	(0.31)
BM	−1.934***	−0.854***	−1.934***	−0.771***	−0.541***	−0.545***
	(−13.16)	(−5.56)	(−13.74)	(−5.94)	(−5.14)	(−6.59)
MOM	0.909***	0.496***	0.909***	0.447***	0.367***	0.268***
	(5.32)	(4.64)	(5.42)	(4.11)	(3.57)	(5.11)
STR	−0.578***	−0.738***	−0.578***	−0.760***	−0.552***	−0.489***
	(−3.77)	(−7.05)	(−3.77)	(−7.00)	(−5.70)	(−8.63)
Stock FE	No	Yes	No	Yes	Yes	Yes
Month FE	No	No	Yes	Yes	Yes	Yes
R^2	.138	.026	.197	.043	.028	.046

This table presents results from panel regressions of the next month's return volatility (*RVOL*), systematic volatility (*SVOL*), or idiosyncratic volatility (*IVOL*) expressed as a percentage on a set of explanatory variables. Each column reports results for a different regression specification. All explanatory variables are standardized within each month to a mean of zero and standard deviation of one. Explanatory variables include 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*), the next month's return (*R*), firm size (*SIZE*), book-to-market ratio (*BM*), momentum (*MOM*), and short-term reversal (*STR*). See Table A1 in the appendix for complete variable definitions. The within R^2 is reported in the last row. Standard errors are clustered at the stock-month level for all specifications. This regression analysis is based on 686,245 stock-month observations from December 1974 to December 2016 with no missing values for all variables. $^*p < .1$; $^{**}p < .05$; $^{***}p < .01$ (significance levels for two-sided tests are indicated).

conclusion from the previous analyses of a negative relationship between the learning index and future volatility.

6. Support for Interpretation of the Learning Index

6.1 Equilibrium asset prices

As already noted in Section 1, proposition 4 of the VNV internet appendix implies that as investors learn more about an asset, the CAPM beta conditional on the information learned by investors will be lower than the unconditional CAPM beta. This suggests a test of the VNV model using a conditional version of the CAPM where β is a function of lagged values of *LI*. Consider the following conditional specification of the CAPM:

$$R_{i,t} - R_{f,t} = \alpha_{0,i} + \beta_i \left(LI_{i,t-1}\right) * \left(R_{M,t} - R_{f,t}\right) + \varepsilon_{i,t}. \tag{14}$$

We assume a linear specification for $\beta_i\left(LI_{i,t-1}\right)=\beta_{i,0}\left(1+\gamma LI_{i,t-1}\right)$. Theoretically, *LI* should be negatively related to market risk. Accounting for the reduction in market risk should produce better expectations of returns than the unconditional CAPM, especially for assets that investors learn more about, that is, assets with a high *LI* value. In this section, we test whether the coefficient γ is negative and whether the expectations of returns from Equation (14) are closer to the next month's returns than those from the unconditional CAPM and other factor models.

Letting r_t represent excess returns, the reduced form of the specification in Equation (14) can be written as

$$r_{i,t}=\alpha_{0,i}+\beta_{0,i}r_{M,t}+\gamma\beta_{0,i}\left(LI_{i,t-1}r_{M,t}\right)+\varepsilon_{i,t}, \tag{15}$$

which is a linear model with nonlinearity in the parameters. As such, we cannot directly estimate γ and $\beta_{0,i}$. Assuming that Equation (15) holds for each cross-section through time, we first estimate a time series for γ.

We use a 2-year rolling window of weekly returns to estimate the coefficients $b_{0,i}$ and $b_{1,i}$ in the following regression for all stocks in our sample having at least 2 years of data.

$$r_{i,t}=\alpha_{0,i}+b_{0,i}r_{M,t}+b_{1,i}\left(LI_{i,t-1}r_{M,t}\right)+\varepsilon_{i,t} \tag{16}$$

At time t, we use the n-vectors of these betas, B_0 and B_1, to estimate $\hat{\gamma}_t$ as

$$\hat{\gamma}_t=\frac{\left(B_1\Sigma^{-1}B_0\right)}{\left(B_0\Sigma^{-1}B_0'\right)} \tag{17}$$

where Σ is the diagonal of the covariance of residuals from the regressions in Equation (16). Figure 1 plots the time series of $\hat{\gamma}$. Consistent with the theory, the impact of *LI* on CAPM beta is consistently negative with a mean of -0.973 (significant at the 1% level) and a standard deviation of 0.184.

We test if the conditional version of the CAPM in Equation (15) produces better expectations of returns than other models through a forecasting exercise. Our test data are the value-weighted *LI*-sorted quintile portfolios. Using 60-month rolling windows, we estimate the conditional CAPM (CCAPM), the traditional CAPM, and the Fama and French (2018) six-factor model along with the nested three-factor and five-factor specifications. Forecasts for the next month's return are formed under the null of zero alphas using unconditional expected values of the factors for each model.[32] For example, the forecast errors for portfolio p from the traditional CAPM are

$$FECAPM_{p,t}=r_{p,t+1}-\hat{b}_pE\left(r_{M,t+1}\right) \tag{18}$$

[32] Simin (2008) provides a proof of the benefits of estimating the model under the null of zero alpha if alphas are not included in the expectations.

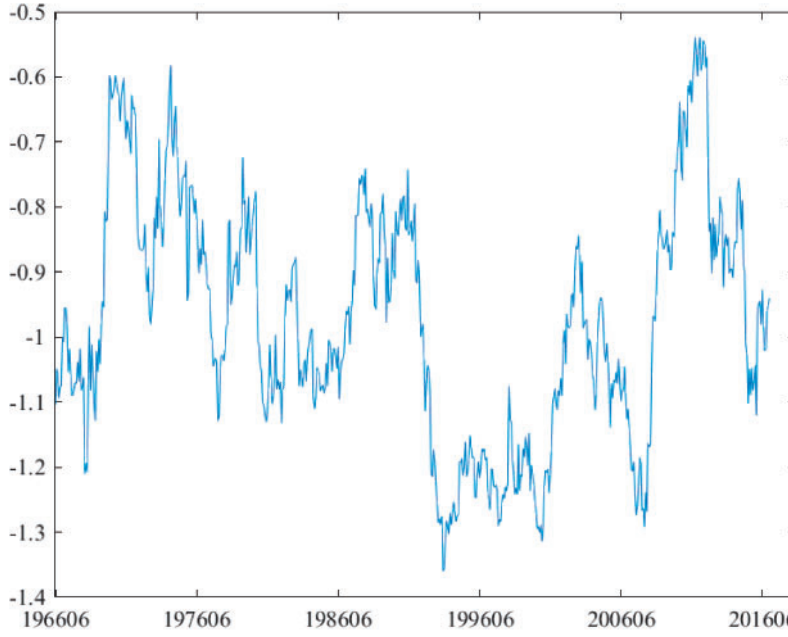

Figure 1
Impact of learning on CAPM beta over time
This figure shows the time-series estimates of γ from the conditional version of the CAPM in Equation (14). We use a 2-year rolling window of weekly returns to estimate the coefficients $b_{0,i}$ and $b_{1,i}$ in the following regression for all stocks in our sample having at least 2 years of data:

$$r_{i,t} = \alpha_{0,i} + b_{0,i} r_{M,t} + b_{1,i} \left(LI_{i,t-1} r_{M,t} \right) + \varepsilon_{i,t}.$$

At time t, we use the n-vectors of these betas, B_0 and B_1, to estimate $\hat{\gamma}_t$ as

$$\hat{\gamma}_t = \frac{\left(B_1 \Sigma^{-1} B_0 \right)}{\left(B_0 \Sigma^{-1} B_0' \right)},$$

where Σ is the diagonal of the covariance of residuals from the regressions. The sample period is July 1964 to December 2016.

and from the conditional version of the CAPM are

$$FECCAPM_{p,t} = r_{p,t+1} - \hat{b}_{0,p} \left(1 + \hat{\gamma}_t \right) \left[LI_{p,t-1} E \left(r_{M,t+1} \right) \right] \tag{19}$$

where $LI_{p,t-1}$ is the average of the lagged learning index across all stocks in portfolio p at time t and $\hat{\gamma}_t$ is from Equation (17).

Table 9 contains the forecast evaluation of the the five linear asset pricing models for the value weighted quintiles of *LI*-sorted portfolios. Panel A contains the average root mean square forecast error for each model. For all the *LI* quintiles, the CCAPM produces expected returns that are closer to the realized returns than any of the other models.

Table 9
Evaluation of expected return forecasts

	*LI*1	*LI*2	*LI*3	*LI*4	*LI*5
		A. Root-mean-square forecast error			
CCAPM	5.371	4.736	4.637	4.465	4.487
CAPM	5.381	4.745	4.660	4.495	4.506
FF3	5.376	4.749	4.662	4.502	4.517
FF5	5.377	4.747	4.658	4.502	4.519
FF6	5.385	4.747	4.656	4.497	4.515
		B. Diebold-Mariano test			
CAPM	0.537	0.329	0.008	0.001	0.006
FF3	0.747	0.230	0.008	0.001	0.001
FF5	0.716	0.285	0.015	0.001	0.001
FF6	0.438	0.294	0.018	0.000	0.000
		C. R^2_{OS}			
CAPM	0.004	0.004	0.010	0.013	0.008
	(0.08)	(0.02)	(0.00)	(0.00)	(0.01)
FF3	0.002	0.005	0.010	0.017	0.013
	(0.07)	(0.01)	(0.00)	(0.00)	(0.00)
FF5	0.002	0.005	0.009	0.016	0.014
	(0.05)	(0.01)	(0.01)	(0.00)	(0.00)
FF6	0.005	0.004	0.008	0.014	0.012
	(0.03)	(0.01)	(0.01)	(0.00)	(0.01)

This table contains the forecast evaluation of five linear asset pricing models for the returns of the value-weighted *LI*-sorted quintile portfolios (*LI*1 through *LI*5). Panel A contains the average root mean square forecast error for each model based on using a 60-month rolling window. Panel B contains the p-values for the Diebold-Mariano test comparing the forecast errors of the conditional CAPM to the other four models. Panel C contains the out-of-sample R^2 values comparing the forecast errors of the conditional CAPM to the other four models with the Clark and West (2007) p-values in parentheses.

Panel B of Table 9 contains the p-values for the Diebold-Mariano (DM) test comparing the forecast errors of the conditional CAPM to the other four models. Define the time t forecast error for model m as u_m and the sample average of the squared forecast error differential of two competing forecasts as $\overline{d} = \frac{1}{T}\sum_{t=1}^{T}\left[u_{1,t}^2 - u_{2,t}^2\right]$. Under the null that $\overline{d} = 0$, the large sample statistic is

$$DM = \frac{\overline{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}} \sim N(0,1) \tag{20}$$

where $\hat{f}_d(0)$ is a consistent estimator of the spectral density (long-run variance) of $\left[u_{1,t}^2 - u_{2,t}^2\right]$ at frequency 0. We estimate $\hat{f}_d(0)$ using the method of Newey and West (1987). This asymptotic test reveals that the CCAPM produces significantly smaller forecast errors relative to the other models for the higher *LI* quintiles. For the two lowest *LI* quintiles, there is no significant difference between forecast errors of the CCAPM and any of the models at the 5% level.

Panel C of Table 9 contains the out-of-sample R^2_{OS} values comparing the forecast errors of the conditional CCAPM to the other four models with the Clark and West (2007) p-values in parentheses. The out-of-sample R^2_{OS} statistic measures the proportional reduction in the mean squared forecast error (*MSFE*)

for the predictive regression forecast relative to another model.

$$R^2_{OS} = 1 - \frac{MSFE_{CCAPM}}{MSFE_m} \tag{21}$$

If $R^2_{OS} > 0$, then the $MSFE_{CCAPM} < MSFE_m$. We use the p-value for the Clark and West (2007) *MSFE*-adjusted statistic to evaluate the statistical significance of R^2_{OS}. To test the null hypothesis $R^2_{OS} \leq 0$ against the alternative hypothesis $R^2_{OS} > 0$, we first calculate

$$\widetilde{d}_m = u^2_m - \left[u^2_{CCAPM} - \left(\hat{r}_m - \hat{r}_{CCAPM}\right)^2\right], \tag{22}$$

where $\hat{r}_m$ is the return forecast from model m, then we regress $\widetilde{d}_m$ on a constant. The *MSFE*-adjusted statistic is the t-statistic corresponding to the constant. In panel C, we see that the CCAPM produces smaller forecast errors than the traditional CAPM and the Fama-French models, particularly for the high *LI* quintiles.

To summarize, we find that the impact of learning on market risk, measured by $\hat{\gamma}$, is negative in every cross-section of our sample and that incorporating the negative influence of learning on CAPM beta improves expectations of returns. The reduction in forecast errors is greater for stocks with higher values of *LI*. We take these results not only as supportive evidence for the VNV model but also as an indication of the importance of accounting for investor learning in asset pricing factor models.

6.2 Long-term predictability

In this section, we examine the cross-sectional explanatory power of the learning index for subsequent months up to 3 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.[33] Figure 2 presents the monthly alphas for the spread portfolio. On a risk-adjusted basis, the value-weighted alpha 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

[33] Results are qualitatively similar based on equal-weighted portfolio average returns.

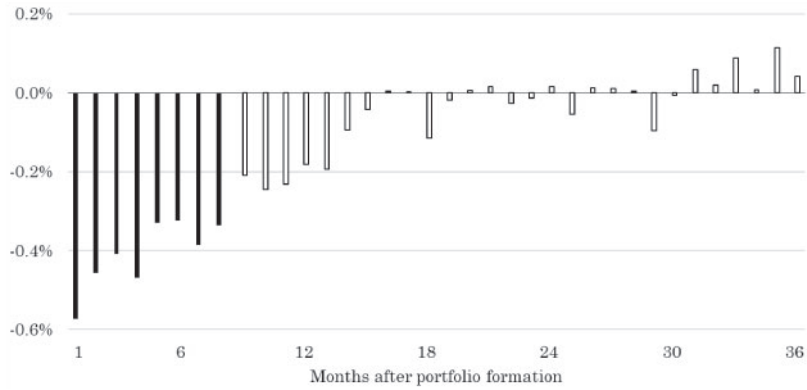

Figure 2
Long-term return predictability using learning index
At the end of each month, we sort stocks into quintiles based on values of the learning index (*LI*) and track the difference in value-weighted returns between the highest *LI* quintile and the lowest *LI* quintile (5 – 1 portfolio) in each of the 36 months following portfolio formation. Stocks are required to have nonmissing return observations for all 36 months to be included in the sample. The figure presents value-weighted monthly risk-adjusted excess returns for the 5 – 1 portfolio. Black bars represent statistical significance at the 10% level. The sample period is July 1964 to December 2016.

declining manner over a period of several months. Furthermore, the differences in alpha between extreme *LI* quintiles are not reversed over the subsequent 3-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 on new information.

Next, we repeat the portfolio sorting analysis and track the difference in portfolio median volatility between the extreme *LI* quintiles over the subsequent 36 months.[34] In Figure 3, the spread in *RVOL* is negative and significant for at least 36 months after portfolio formation. 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 volatility into systematic and idiosyncratic components, we find similar results. Figure 3 shows that the differences in systematic and idiosyncratic volatilities predicted by *LI* do not reverse over the long run.

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.

[34] Conclusions are qualitatively similar based on using equal- or value-weighted portfolio averages, with one exception. Consistent with the portfolio sorting results in Table 6, the spread in value-weighted average idiosyncratic volatility is negative but insignificant over the subsequent 36 months.

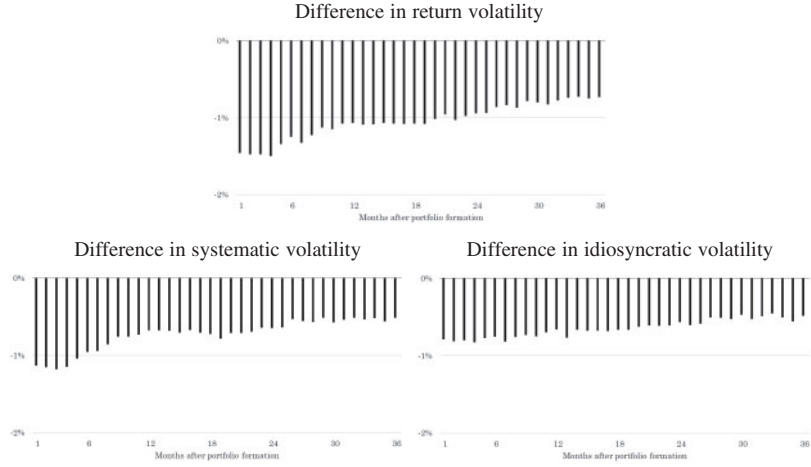

Figure 3
Long-term volatility predictability using learning index
At the end of each month, stocks are sorted into quintiles based on average volatility over the past 12 months. Within each volatility quintile, stocks are sorted based on values of *LI*. Each *LI* subquintile is combined across volatility quintiles into a single quintile. We track the difference in portfolio median return volatility (*RVOL*), systematic volatility (*SVOL*), and idiosyncratic volatility (*IVOL*) 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 nonmissing volatility observations for all 36 months to be included in the sample. Black bars represent 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 appendix for complete variable definitions. The sample period is July 1964 to December 2016.

6.3 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).[35]

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.

[35] Hong, Lim, and Stein (2000) document that firm size is by far the most significant determinant of analyst coverage.

Table 10
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*)	4.696	7.475	2.085	2.284	740	2.789
2	6.924	7.762	2.196	2.822	776	2.943
3	8.490	8.071	2.315	3.591	812	3.019
4	8.822	8.359	2.450	4.047	842	3.037
5 (High *LI*)	7.367	8.612	2.495	4.228	873	3.172
5−1	2.671***	1.136***	0.411***	1.944***	132***	0.383***
t-stat	(3.26)	(6.61)	(5.03)	(4.54)	(3.51)	(6.02)

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 analysts' forecasts ($nFCST$), number of analysts' 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 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 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. $^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$ (significance levels for two-sided tests are indicated).

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 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 6.

Table 10 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 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 7.37% (4.70%) increase in monthly share turnover relative to average monthly turnover during the past year. The difference in abnormal turnover between the extreme quintiles is 2.67% with a *t*-statistic of 3.26. 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, Lim, and Stein (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 et al. (2019) 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 analysts' forecasts of earnings per share (EPS) recorded by I/B/E/S for the nearest fiscal quarter. We also measure the number of analysts' 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 8.61 (7.48) analysts. The difference in coverage is approximately one analyst with a *t*-statistic of 6.61. On average, 2.50 (2.09) analysts covering a high (low) *LI* stock revise their forecasts from the prior month. The average difference in the number of forecast revisions is 0.41 and is significant at the 1% level.

In column 4 of Table 10, 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 4.23% (2.28%). The difference in ΔFA between the extreme *LI* quintiles is 1.94% on average (*t*-statistic = 4.54). 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 analysis, we find qualitatively similar results when we use the median analyst forecast.

The fifth proxy is based on downloads of company filings from the SEC EDGAR database. Drake, Roulstone, and Thornock (2015) study the determinants and consequences of investor information acquisition using EDGAR download activity as a proxy. Although the data are available

beginning in January 2003, these data are prone to significant known issues because of 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. Using 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.[36] 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 (740) times within a month on average. The difference in average EDGAR downloads between these quintiles is approximately 132 with a t-statistic of 3.51. 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, Da, and Israelsen (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, Da, and Israelsen (2017), we measure abnormal attention at the daily frequency using a dummy variable that is equal to one if the Bloomberg daily maximum is a 3 or 4 and zero 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 (2.79) days within a month on average. The difference in abnormal attention days between high and low LI stocks is 0.38 with a t-statistic of 6.02. Overall, the patterns documented in Table 10 serve as supporting evidence of a relationship between the learning index and information flow.[37]

6.4 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

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

[37] 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 10.

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 on their information. Finally, we expect that the negative relation between the learning index and the market reaction to the earnings announcement should be stronger during months when learning is concentrated among fewer earnings announcements as well as for announcements with a larger earnings surprise.

Following the accounting literature, we measure the market reaction to earnings announcements using the magnitude of cumulative abnormal returns. Absolute returns also can 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 2-day period and 5-day period beginning on the announcement date $\left(|CAR|_{d,d+1} \text{ and } |CAR|_{d,d+4}\right)$. 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 2 days after the announcement through 1 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 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.[38] Because earnings announcements occur at various times throughout a month, we estimate the learning index for all firms as of the week prior to 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.[39] 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. Because of data availability, the sample period for this analysis is October 1971 to December 2016.

Panel A of Table 11 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

[38] 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.

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

Table 11
Learning prior to earnings announcements: Portfolios of stocks sorted by learning index controlling for firm size

Quintile	$\|CAR\|_d$	$\|CAR\|_{d,d+1}$	$\|CAR\|_{d,d+4}$	$\|CAR\|_{q+1}$	$ATURN_{w-1}$
			A. Full sample period		
1 (Low *LI*)	2.537	3.944	5.079	12.137	1.333
2	2.448	3.853	4.923	11.948	2.097
3	2.455	3.842	4.910	11.729	3.146
4	2.446	3.855	4.882	11.756	3.923
5 (High *LI*)	2.410	3.757	4.742	11.708	4.591
5−1	−0.127**	−0.188**	−0.337***	−0.429**	3.258***
t-stat	(−2.35)	(−2.28)	(−3.54)	(−2.02)	(5.26)
			B. Months with number of earnings announcements above yearly median		
5−1	−0.071	−0.090	−0.181	−0.120	2.653***
t-stat	(−1.10)	(−0.87)	(−1.43)	(−0.43)	(4.34)
			C. Months with number of earnings announcements below yearly median		
5−1	−0.182**	−0.286**	−0.493***	−0.739**	3.865***
t-stat	(−2.20)	(−2.32)	(−3.72)	(−2.54)	(3.36)

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 $(|CAR|_d)$, during a 2-day period and 5-day period beginning on the announcement date $(|CAR|_{d,d+1}$ and $|CAR|_{d,d+4})$, and during the period beginning 2 days after the earnings announcement date through 1 day after the firm's next quarterly earnings announcement date $(|CAR|_{q+1})$. 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 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 are given for the 5 − 1 portfolio. The full sample period in panel A is October 1971 to December 2016. In panels B and 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). $*p < .1$; $**p < .05$; $***p < .01$ (significance levels for two-sided tests are indicated).

event date is 0.13% smaller for high *LI* stocks compared to low *LI* stocks (*t*-statistic = −2.35). Over a 2-day (5-day) period beginning on the announcement date, the difference in market reaction between the extreme *LI* quintiles is −0.19% (−0.34%). These estimates are significant at the 5% and 1% levels, 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 −0.43% (*t*-statistic = −2.02).

The last column in Table 11 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 4.59% (1.33%) 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 3.26% with a *t*-statistic of 5.26.

In the lower half of Table 11, we examine how the results from the prior analyses vary with the level of earnings announcement activity during the month. If fewer earnings announcements are made during the month, investors' learning efforts may be more concentrated on 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. To test this hypothesis, we count the number of earnings announcements in each month and split the sample 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 subsample groups, Table 11 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 is 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 -0.18%, -0.29%, -0.49%, and -0.74%, with t-statistics of -2.20, -2.32, -3.72, and -2.54, respectively. The fifth column of Table 11 presents differences in abnormal turnover in the week prior to an earnings announcement. The difference in $ATURN_{w-1}$ between extreme *LI* quintiles is 2.65% (t-statistic = 4.34) during high activity months and 3.87% (t-statistic = 3.36) during low activity months.

In Table 12, we investigate the relationship between *LI* and the market reaction for different earnings surprise magnitudes. If investors acquire private information prior to an announcement, they respond less to the earnings announcement. This implies that the negative relation between the learning index and the market reaction should be stronger (more negative) when the surprise in announced earnings is larger in absolute value.

Each month, firms with earnings announcements are sorted into terciles (low, mid, high) according to the earnings surprise magnitude $|SURP|$. We estimate panel regressions of the market reaction to the earnings announcement on the following explanatory variables: learning index (*LI*) in the week prior to the announcement, firm size (*SIZE*) in the week prior to the announcement, number of analysts' forecasts ($nFCST$), and abnormal share turnover in the week prior to the earnings announcement ($ATURN_{w-1}$). All nondummy explanatory variables are standardized within each month to a mean of zero and standard deviation of one. All regressions include stock and month fixed effects, and standard errors are clustered at the stock-month level.

The coefficient for *LI* is negative and significant in all columns in Table 12, indicating that higher values of *LI* are associated with smaller market reactions to earnings announcements. For all three definitions of $|SURP|$, the coefficient

Table 12
Market reaction sensitivity to earnings surprise magnitude: Panel regressions for $|SURP|$ terciles

$\|SURP\|$ =	$\left\|\frac{\text{Actual EPS- Mean EPS}}{\text{Price}}\right\|$			$\left\|\frac{\text{Actual EPS- Mean EPS}}{\text{Mean EPS}}\right\|$			$\left\|\frac{\text{Actual EPS- Mean EPS}}{\text{Actual EPS}}\right\|$		
	Low	Mid	High	Low	Mid	High	Low	Mid	High
LI	−0.086***	−0.137***	−0.178***	−0.106***	−0.136***	−0.173***	−0.107***	−0.129***	−0.173***
	(−2.96)	(−4.05)	(−4.67)	(−3.71)	(−3.64)	(−4.80)	(−3.79)	(−3.64)	(−4.76)
SIZE	−0.305***	−0.501***	−1.663***	−0.458***	−0.785***	−1.309***	−0.482***	−0.750***	−1.294***
	(−2.75)	(−4.11)	(−9.34)	(−3.97)	(−6.23)	(−7.52)	(−4.20)	(−6.13)	(−7.23)
$nFCST$	0.059	0.129*	0.308***	0.135**	0.157**	0.238***	0.123**	0.158**	0.203**
	(0.91)	(1.92)	(3.58)	(2.28)	(2.16)	(3.08)	(2.07)	(2.25)	(2.53)
$ATURN_{w-1}$	−0.068***	−0.102***	0.004	−0.054*	−0.082**	−0.031	−0.053*	−0.080**	−0.033
	(−2.65)	(−3.37)	(0.10)	(−1.94)	(−2.55)	(−0.85)	(−1.92)	(−2.52)	(−0.92)
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	.001	.002	.008	.001	.003	.005	.001	.003	.005
Observations	34,401	34,579	34,504	34,374	34,566	34,474	34,296	34,476	34,388

This table presents results from panel regressions examining the interaction between the learning index and the sensitivity of the market reaction to the magnitude of unexpected earnings. The sample is sorted each month into terciles according to the earnings surprise magnitude ($|SURP|$), measured as the absolute value of the difference between the actual EPS and mean EPS forecast scaled by the stock price at the end of the previous month, the mean EPS forecast, or the actual EPS. The dependent variable is the market reaction to quarterly earnings announcements, defined as the absolute value of the cumulative abnormal return during a 5-day period beginning on the announcement ($|CAR|_{d,d+4}$). Explanatory variables include the learning index in the week prior to the announcement (*LI*), firm size (*SIZE*) in the week prior to the announcement, number of analysts' forecasts ($nFCST$), and abnormal share turnover in the week prior to the earnings announcement ($ATURN_{w-1}$). All explanatory variables are standardized within each month to a mean of zero and standard deviation of one. See Table A1 in the appendix for complete variable definitions. At least five analysts' forecasts are required to be included in the sample. The within R^2 and number of observations are reported at the bottom of the table. All regressions include stock and month fixed effects. Standard errors are clustered at the stock-month level. $^*p<.1$; $^{**}p<.05$; $^{***}p<.01$ (significance levels for two-sided tests are indicated).

for *LI* is more negative and more significant for stocks with larger earnings surprises. These results are consistent with the prediction that if investors acquire more information before the announcement, they respond less to the announcement, and this effect is stronger when there is a large, unexpected component in announced earnings.

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 as well as for earnings announcements containing a large surprise. The findings support the idea that the learning index reflects learning decisions and the flow of information prior to earnings announcements.

6.5 Controlling for information supply and uncertainty shocks

To test whether the explanatory power of the learning index for expected returns is due to the effects on learning intensity of information supply or shocks to uncertainty, we include proxies for these variables as controls in our cross-sectional regression specification.

Following Morck, Yeung, and Yu (2000), we measure the information content in stock prices using the R^2 from the market model regression. A high R^2 indicates a high degree of stock price synchronicity, and therefore a low degree of firm-specific content. We use one minus R^2 to capture the degree of firm-specific information in returns, or stock price asynchronicity (*ASYNC*). To measure shocks to fundamental uncertainty, we first start by measuring earnings shocks following Irvine and Pontiff (2009). We estimate the following pooled cross-sectional time-series model:

$$\begin{aligned} E_{i,q} - E_{i,q-4} = \alpha + \beta_1 (E_{i,q-1} - E_{i,q-5}) + \beta_2 (E_{i,q-2} - E_{i,q-6}) \\ + \beta_3 (E_{i,q-3} - E_{i,q-7}) + e_{i,q}, \end{aligned} \tag{23}$$

where $E_{i,q}$ represents earnings per share for firm i in quarter q. As in Irvine and Pontiff (2009), we estimate this model at the industry level using the Fama-French 49 industry classification, which allows intercepts and slope coefficients to vary by industry. The residuals $e_{i,q}$ represent firm-level innovations in earnings per share. We use the volatility of these residuals to measure fundamental uncertainty $EVOL_{i,q}$. To obtain time-series variation in fundamental uncertainty, we use a 5-year rolling window for estimation. Finally, to measure shocks to fundamental uncertainty, we calculate the quarterly change in earnings volatility $\Delta EVOL$ as $EVOL_{i,q} - EVOL_{i,q-1}$. Because we use a 5-year rolling window to estimate $\Delta EVOL$, the sample period for this analysis begins in July 1969.

Table 13
Cross-sectional return regressions: Controlling for information supply and uncertainty shocks

	A			
	(1)	(2)	(3)	(4)
LI	−0.064***			−0.060***
	(−3.43)			(−3.29)
ASYNC		−0.094**		−0.085**
		(−2.37)		(−2.17)
$\Delta EVOL$			−0.054***	−0.052***
			(−2.73)	(−2.62)
Control variables	Yes	Yes	Yes	Yes
Adj R^2	.089	.091	.089	.092
	B			
	(1)	(2)	(3)	(4)
LI	−0.068***			−0.069***
	(−3.30)			(−3.35)
nFCST		0.110*		0.114**
		(1.93)		(2.02)
ADISP			−0.036	−0.038
			(−1.32)	(−1.44)
Control variables	Yes	Yes	Yes	Yes
Adj R^2	.090	.091	.090	.093

This table presents results from two-stage cross-sectional regressions. At the end of each month, we estimate a cross-sectional regression of the next month's excess stock return on a set of explanatory variables. All explanatory variables are standardized within each month to a mean of zero and standard deviation of one. Each column reports the average slope coefficients for each different regression specification. The explanatory variables of interest are the learning index (*LI*), stock price asynchronicity (*ASYNC*) and quarterly change in earnings volatility ($\Delta EVOL$) in panel A, and number of analysts' forecasts (*nFCST*) and abnormal forecast dispersion (*ADISP*) in panel B. Control variables include 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 appendix for complete variable definitions. The average adjusted R^2 is reported in the last row. The intercept term and coefficient estimates for control variables are not reported for brevity. Newey and West (1987) *t*-statistics are given in parentheses. Regressions in panel A are based on 785,873 stock-month observations from July 1969 to December 2016 with no missing values for all variables. Regressions in panel B are based on 451,501 stock-month observations from January 1989 to December 2016 with no missing values for all variables. *p <.1; **p <.05; ***p <.01 (significance levels for two-sided tests are indicated).

We also consider two additional alternative measures of information production and shocks to uncertainty that are based on analysts' forecast data. To measure the production of information about a firm, we use the number of analysts' forecasts (*nFCST*). To measure shocks to uncertainty, we use a measure of abnormal forecast dispersion (*ADISP*), calculated as the ratio of dispersion in the current month to average dispersion over the past 12 months, minus one and multiplied by 100. Because of the availability of analysts' data, we begin our analysis in January 1989.

Table 13 presents the results. All regressions include the full set of control variables in the main specification (Equation (12)). Panel A contains analyses using *ASYNC* and $\Delta EVOL$. For comparability, we first reestimate our baseline regression from July 1969 to December 2016 and verify that the explanatory power of the learning index is qualitatively similar to that in the full sample period regression results in Section 4. In columns 2 and 3, we estimate

benchmark regressions with either *ASYNC* or $\Delta EVOL$ and the full set of control variables excluding *LI*. The coefficient estimate for *ASYNC* in column 2 is -0.094 (t-statistic$=-2.37$), indicating that a greater degree of firm-specific return variation reflects more efficient stock pricing and a lower expected return. The coefficient estimate for $\Delta EVOL$ in column 3 is -0.054 (t-statistic$=-2.73$). This result is consistent with the idea that large shocks to uncertainty negatively affect stock prices. In column 4, we include all three variables of interest simultaneously (along with the control variables). After controlling for the firm-specific information content in stock prices using *ASYNC* and shocks to fundamental uncertainty using $\Delta EVOL$, the explanatory power of the learning index remains economically and statistically significant.

Panel B of Table 13 contains analyses using *nFCST* and *ADISP*. As before, we first reestimate our baseline regression over the sample period 1989 to 2016 to verify that the explanatory power of the learning index is qualitatively similar to the full sample period regression results. In columns 2 and 3, we estimate a benchmark regressions with either *nFCST* or *ADISP*. The coefficient estimate for *nFCST* in column 2 is 0.110 (t-statistic$=1.93$). This result may reflect the preference of resource-constrained analysts for covering better performing firms. The coefficient estimate for *ADISP* in column 3 is negative but insignificant. After adding analyst coverage and abnormal forecast dispersion to the specification, the explanatory power of the learning index remains economically and statistically significant. Thus, the evidence suggests that the learning index is not merely capturing other variables that are common to models of information choice and affect learning intensity, such as information supply or shocks to uncertainty.

7. Robustness

In this section, we describe additional analyses and robustness checks. The Internet Appendix reports the results. In Table B1, we repeat the sorting analysis of returns using four alternative factor model specifications for risk adjustment. This test allows us to evaluate the robustness of our main results after controlling for exposure to other risk factors identified in the related literature.[40] We first augment the Fama and French (2018) six-factor model by adding the Pastor and Stambaugh (2003) liquidity factor. We consider a further extension of the previous model by adding a short-term reversal factor and a long-term reversal factor. In addition to these two specifications, we also consider the Stambaugh and Yuan (2017) factor model, which contains market, size, and two mispricing factors, and the Hou, Xue, and Zhang (2015) q-factor model,

[40] We obtain liquidity factor data from Lubos Pastor's website (faculty.chicagobooth.edu/lubos.pastor/research), short-term and long-term reversal factor data from Kenneth French's website (mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html), and data for the Stambaugh and Yuan (2017) mispricing factors from Robert Stambaugh's website (finance.wharton.upenn.edu/~stambaug). Kewei Hou provided data for the Hou, Xue, and Zhang (2015) q-factor model.

which contains market, size, profitability, and investment factors.[41] In all cases, the risk-adjusted return spread between high and low learning index stocks is negative and significant at the 1% level.

In addition to the main learning index measure, we also construct a scaled learning index measure by rank-transforming each of the learning index components as well as the sum of the three components to the interval [0,1] in each month. This approach facilitates interpretation and comparability across cross-sections as the number of stocks in each monthly cross-section varies over time. Tables B2 through B7 present the results. The explanatory power of the scaled learning index for future returns and volatility is comparable to that of the unscaled learning index. Based on the scaled learning index measure, spreads in the next month's returns between high and low learning index stocks are slightly larger, while spreads in the next month's volatility are slightly smaller.

To check whether the results are robust to the use of shorter time series of data, we reestimate the learning index using a 1-year (instead of 2-year) rolling window of weekly returns. We then use this measure in our sorting and regression analyses for returns and volatility. Tables B8 through B13 of the Internet Appendix report the results. Using this alternative measure of the learning index, we continue to find evidence of a negative cross-sectional relationship between *LI* and future return and volatility. Compared to the 2-year rolling window *LI*, the 1-year *LI* appears to have weaker explanatory power for the cross-section of returns and stronger explanatory power for the cross-section of volatility.

8. 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 resultant impact on risk and risk premiums across assets. 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 proxy for information flow. Consistent with the model's predictions, we find that the empirical learning

[41] The two mispricing factors (*MGMT* and *PERF*) of Stambaugh and Yuan (2017) capture overpricing or underpricing across 11 well-known anomalies. The *MGMT* factor is constructed based on six anomaly variables that can be directly affected by the decisions of a firm's management: net stock issues, composite equity issues, accruals, net operating assets, asset growth, and investment to assets. The *PERF* factor is constructed based on five anomaly variables related to performance: financial distress, the Ohlson O-score, momentum, gross profitability, and return on assets.

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. Consistent with the predictions of the VNV model, we find that learning more about an asset lowers its market risk, and accounting for this negative impact in a conditional version of the CAPM improves return forecasts. We also 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. The predictive ability of the learning index for the market reaction to an earnings announcement is stronger when learning is concentrated among fewer firms and when the announcement contains a large surprise. Finally, we rule out the alternative explanation that the learning index is capturing the effects on learning intensity of information supply and shocks to uncertainty.

In aggregate, our findings support the theoretical predictions of the Van Nieuwerburgh and Veldkamp (2010) model 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 investigate the role of information choice in asset pricing. More work can be done in the study of information choice to develop theoretical models with more realistic assumptions (e.g., multiple periods, investor heterogeneity, correlated information signals, endogenous noise trading) that can be empirically tested using real-world data.

Data Appendix

Table A1
Variable definitions

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 for a complete description of the variable measurement
β^{MKT}	Market beta, estimated from a regression of excess stock returns on lagged, current, and lead excess market returns using weekly data from the past 2 years within the month.
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 as a percentage from months $t-11$ to $t-1$
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 as a percentage over the past month
LTR	Long-term reversal, defined as the cumulative return as a decimal from months $t-59$ to $t-12$
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
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 one if the firm paid dividends during the most recent fiscal quarter and zero 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.
$ATURN$	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 divided by average weekly turnover over the past 2 months, minus one and multiplied by 100
$nFCST$	Number of analysts' forecasts for the nearest fiscal quarter
$nREV$	Number of analysts' 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 (as a percentage) 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 (in accordance with 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*)

Table A1
(Continued)

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 measured on the earnings announcement date d, $\|CAR\|_{d,d+1}$ and $\|CAR\|_{d,d+4}$ are measured during a 2- and 5-day period beginning on the announcement date, and $\|CAR\|_{q+1}$ is measured during the period beginning 2 days after the earnings announcement date through 1 day after the firm's next quarterly earnings announcement date
$ASYNC$	Stock price asynchronicity, defined as one minus the R^2 from the market model regression used to estimate market beta
$\Delta EVOL$	Quarterly change in earnings volatility. Following Irvine and Pontiff (2009), quarterly earnings shocks are measured by estimating Equation (23) at the Fama-French 49 industry level. Earnings volatility ($EVOL$) is defined as the volatility of residuals from this regression. To obtain time-series variation in earnings volatility, a 5-year rolling window is used to estimate Equation (23). $\Delta EVOL$ is defined as the quarterly change in earnings volatility, or $EVOL_{i,q} - EVOL_{i,q-1}$
$ADISP$	Abnormal forecast dispersion, defined as the ratio of analysts' forecast dispersion in the current month to average analysts' forecast dispersion over the past 12 months, minus one and multiplied by 100
$\|SURP\|$	Earnings surprise magnitude, defined as the absolute value of the difference between the actual EPS and mean EPS forecast scaled by the stock price at the end of the previous month, the mean EPS forecast, or the actual EPS

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

Table A2
Transition probabilities for portfolios sorted by learning index

	$LI1_{t+1}$	$LI2_{t+1}$	$LI3_{t+1}$	$LI4_{t+1}$	$LI5_{t+1}$
A. One-month transition probabilities					
$LI1_t$	72.3	23.1	4.0	0.5	0.1
$LI2_t$	23.0	46.7	24.4	5.2	0.7
$LI3_t$	4.1	24.2	42.9	24.3	4.5
$LI4_t$	0.5	5.3	23.9	46.1	24.1
$LI5_t$	0.1	0.7	4.7	24.0	70.6
	$LI1_{t+6}$	$LI2_{t+6}$	$LI3_{t+6}$	$LI4_{t+6}$	$LI5_{t+6}$
B. Six-month transition probabilities					
$LI1_t$	49.9	25.7	13.8	7.4	3.2
$LI2_t$	25.2	27.3	22.1	16.1	9.3
$LI3_t$	13.2	21.7	24.5	22.7	17.8
$LI4_t$	7.0	15.2	22.1	27.1	28.6
$LI5_t$	3.6	9.5	17.4	27.2	42.3
	$LI1_{t+12}$	$LI2_{t+12}$	$LI3_{t+12}$	$LI4_{t+12}$	$LI5_{t+12}$
C. Twelve-month transition probabilities					
$LI1_t$	32.8	23.2	18.1	14.7	11.2
$LI2_t$	22.9	21.9	20.3	18.4	16.5
$LI3_t$	17.6	20.1	20.9	21.0	20.4
$LI4_t$	13.7	18.1	20.9	23.0	24.4
$LI5_t$	9.4	15.4	20.2	24.6	30.4
	$LI1_{t+24}$	$LI2_{t+24}$	$LI3_{t+24}$	$LI4_{t+24}$	$LI5_{t+24}$
D. Twenty-four-month transition probabilities					
$LI1_t$	21.3	20.5	20.0	19.7	18.5
$LI2_t$	19.2	19.7	20.3	20.6	20.3
$LI3_t$	17.8	19.3	20.5	21.0	21.5
$LI4_t$	17.1	19.0	20.3	21.4	22.3
$LI5_t$	16.1	18.6	20.4	21.8	23.1
	$LI1_{t+36}$	$LI2_{t+36}$	$LI3_{t+36}$	$LI4_{t+36}$	$LI5_{t+36}$
E. Thirty-six-month transition probabilities					
$LI1_t$	20.2	19.8	19.9	20.2	20.0
$LI2_t$	19.1	19.7	20.2	20.7	20.2
$LI3_t$	18.4	19.8	20.6	20.7	20.5
$LI4_t$	18.1	19.8	20.5	20.9	20.8
$LI5_t$	17.3	19.1	20.2	21.1	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.

References

Admati, A. 1985. A noisy rational expectations equilibrium for multi-asset securities markets. *Econometrica* 53:629–58.

An, B.-J., A. Ang, T. G. Bali, and N. Cakici. 2014. The joint cross section of stocks and options. *Journal of Finance* 69:2279–337.

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

Bai, J., and S. Ng. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70:191–221.

Bali, T., R. Engle, and S. Murray. 2016. *Empirical asset pricing: The cross section of stock returns*. Hoboken, NJ: John Wiley & Sons.

Banerjee, S. 2011. Learning from prices and the dispersion in beliefs. *Review of Financial Studies* 24:3025–68.

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* 2007:341–68.

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* 30:3009–47.

Biais, B., P. Bossaerts, and C. Spatt. 2010. Equilibrium asset pricing and portfolio choice under asymmetric information. *Review of Financial Studies* 23:1503–43.

Black Rock. 2017. Viewpoint: Index investing supports vibrant capital markets. Report, New York.

Botosan, C. 1997. Disclosure level and the cost of equity capital. *Accounting Review* 72:323–49.

Brennan, M. J., and P. J. Hughes. 1991. Stock prices and the supply of information. *Journal of Finance* 46:1665–91.

Brown, G. W., and M. T. Cliff. 2005. Investor sentiment and asset valuation. *Journal of Business* 78:405–40.

Burlacu, R., P. Fontaine, S. Jiminez-Garcés, and M. Seasholes. 2012. Risk and the cross section of stock returns. *Journal of Financial Economics* 2012:511–22.

Christie, A. 1982. The stochastic behavior of common stock variances. *Journal of Financial Economics* 10:407–32.

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

Cohen, L., and A. Frazzini. 2008. Economic links and predictable returns. *Journal of Finance* 63:1977–2011.

Corwin, S. A., and J. F. Coughenour. 2008. Limited attention and the allocation of effort in securities trading. *Journal of Finance* 63:3031–67.

Cziraki, P., J. Mondria, and T. Wu. 2020. Asymmetric attention and stock returns. *Management Science*. Advance Access published June 1, 2020, 10.1287/mnsc.2019.3460.

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

De Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98:703–38.

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* 32:1128–61.

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

Easley, D., M. O'Hara, and P. Srinivas. 1998. Option volume and stock prices: Evidence on where informed traders trade. *Journal of Finance* 53:431–65.

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

———. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116:1–22.

———. 2018. Choosing factors. *Journal of Financial Economics* 128:234–52.

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.

Gargano, A., and A. G. Rossi. 2018. Does it pay to pay attention? *Review of Financial Studies* 31:4595–649.

Grossman, S., and J. Stiglitz. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70:393–408.

Harford, J., F. Jiang, R. Wang, and F. Xie. 2019. Analyst career concerns, effort allocation, and firms' information environment. *Review of Financial Studies* 32:2179–224.

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

Hong, C. Y., and F. W. Li. 2019. The information content of sudden insider silence. *Journal of Financial and Quantitative Analysis* 54:1499–1538.

Hong, H., T. Lim, and J. Stein. 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *Journal of Finance* 55:265–95.

Hou, K. 2007. Industry information diffusion and the lead-lag effect in stock returns. *Review of Financial Studies* 20:1113–38.

Hou, K., and T. J. Moskowitz. 2005. Market frictions, price delay, and the cross-section of expected returns. *Review of Financial Studies* 18:981–1020.

Hou, K., C. Xue, and L. Zhang. 2015. Digesting anomalies: An investment approach. *Review of Financial Studies* 28:650–705.

———. 2020. Replicating anomalies. *Review of Financial Studies* 33:2019–133.

Irvine, P., and J. Pontiff. 2009. Idiosyncratic return volatility, cash flows, and product market competition. *Review of Financial Studies* 22:1149–77.

Ivković, Z., C. Sialm, and S. Weisbenner. 2008. Portfolio concentration and the performance of individual investors. *Journal of Financial and Quantitative Analysis* 43:613–55.

Kacperczyk, M., J. Nosal, and L. Stevens. 2019. Investor sophistication and capital income inequality. *Journal of Monetary Economics* 107:18–31.

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.

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* 7:163–95.

Lo, A. W., and J. Wang. 2006. Trading volume: Implications of an intertemporal capital asset pricing model. *Journal of Finance* 61:2805–40.

Menzly, L., and O. Ozbas. 2010. Market segmentation and cross-predictability of returns. *Journal of Finance* 65:1555–80.

Mondria, J. 2010. Portfolio choice, attention allocation, and price comovement. *Journal of Economic Theory* 145:1837–64.

Mondria, J., and C. Quintana-Domeque. 2012. Financial contagion and attention allocation. *Economic Journal* 123:429–54.

Mondria, J., T. Wu, and Y. Zhang. 2010. The determinants of international investment and attention allocation: Using internet search query data. *Journal of International Economics* 82:85 – 95.

Morck, R., B. Yeung, and W. Yu. 2000. The information content of stock markets: why do emerging markets have synchronous stock price movements. *Journal of Financial Economics* 58:215–60.

Newey, W., and K. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55:703–8.

Pan, Y., T. Wang, and M. Weisbach. 2015. Learning about CEO ability and stock return volatility. *Review of Financial Studies* 28:1623–66.

Pastor, L., and R. Stambaugh. 2003. Liquidity risk and expected stock returns. *Journal of Political Economy* 111:642–85.

Pastor, L., and P. Veronesi. 2003. Stock valuation and learning about profitability. *Journal of Finance* 58:1749–89.

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

Simin, T. 2008. The poor predictive performance of asset pricing models. *Journal of Financial and Quantitative Analysis* 43:355–80.

Sims, C. A. 2006. Rational inattention: Beyond the linear-quadratic case. *American Economic Review* 96:158–63.

Spiegel, M. 1998. Stock price volatility in a multiple security overlapping generations model. *Review of Financial Studies* 11:419–47.

Stambaugh, R., and Y. Yuan. 2017. Mispricing factors. *Review of Financial Studies* 30:1270–315.

Stambaugh, R. F., J. Yu, and Y. Yuan. 2015. Arbitrage asymmetry and the idiosyncratic volatility puzzle. *Journal of Finance* 70:1903–48.

Stock, J. H., and M. W. Watson. 2002. Macroeconomic forecasting using diffusion indices. *Journal of Business and Economic Statistics* 20:147–62.

Van Nieuwerburgh, S., and L. Veldkamp. 2009. Information immobility and the home bias puzzle. *Journal of Finance* 64:1187–215.

———. 2010. Information acquisition and under-diversification. *Review of Economic Studies* 77:779–805.

Veldkamp, L. 2011. *Information choice in macroeconomics and finance*. Princeton, NJ: Princeton University Press.

Watanabe, M. 2008. Price volatility and investor behavior in an overlapping generations model with information asymmetry. *Journal of Finance* 63:229–72.

Xu, Y. 2007. Extracting factors with maximum explanatory power. Working paper, University of Texas at Dallas.

Zhang, Y. 2008. Analyst responsiveness and the post-earnings-announcement drift. *Journal of Accounting and Economics* 46:201–15.

Zhao, X. 2017. Does information intensity matter for stock returns? Evidence from Form 8-K filings. *Management Science* 63:1382–404.