Skill Acquisition and Data Sales

Shiyang Huang, Yan Xiong, and Liyan Yang*

April 29, 2021

Abstract

We develop a data-sales model to study the implications of alternative data for financial markets. Investors acquire skills to process the purchased raw data, and developing such skills is costly and involves considerable uncertainty. The data vendor controls the size of the data sample to influence how precise information investors can extract from the purchased data. Price informativeness is hump-shaped in skill-acquisition costs, while the cost of capital and return volatility are U-shaped in skill-acquisition costs. Similar patterns can arise for skill mean and skill volatility. Our analysis suggests that the funds industry and the data industry foster each other.

Key words: Alternative data, information, data sales, skill acquisition, price informativeness, asset prices

*Huang: The University of Hong Kong; Email: huangsy@hku.hk. Xiong: Hong Kong University of Science and Technology; Email: yanxiong@ust.hk. Yang: Rotman School of Management, University of Toronto; Guanghua School of Management, Peking University, 100871, Peking, China. Email: liyan.yang@rotman.utoronto.ca. This paper was previously circulated with title "Skill Acquisition and Information Sales." We thank Agostino Capponi (editor) and two anonymous referees whose comments and suggestions have greatly improved this manuscript. For helpful discussions and comments, we thank Pat Akey, Kerry Back, Robert Battalio, Snehal Banerjee, Bruno Biais, Philip Bond, Ling Cen, Hui Chen, Alexander Chinco, Peter Christoffersen, David Easley, Thierry Foucault, Diego García, David Hirshleifer, Jennifer Huang, Ron Kaniel, Hong Liu, Nadya Malenko, Tom McCurdy, Joshua Mollner, Christine A. Parlour, Alessio Piccolo, Francesco Sangiorgi, Joel Shapiro, Chester Spatt, Laura Veldkamp, Xavier Vives, Ansgar Walther, Yajun Wang, Yizhou Xiao, Hongjun Yan, Kathy Yuan, Mindy Zhang, and the audiences at the 2016 CMU Credit Ratings Conference, the 2017 China International Conference in Finance (CICF), the 2017 Financial Intermediation Research Society (FIRS) Conference, the 2017 Northern Finance Association Annual Conference (NFA), the 2017 Oxford Financial Intermediation Theory (OxFIT) Conference, the 2017 FTG Summer School on Liquidity in Financial Markets and Institutions, the 2017 SFS Cavalcade Asia-Pacific, the 2020 AFA annual meeting, and the seminar in CUHK Shenzhen. Liyan Yang thanks Social Sciences and Humanities Research Council of Canada (SSHRC) for financial support.

1 Introduction

In financial markets, asset management companies increasingly rely on data to generate profitable trading strategies. Data is useful only after being analyzed and processed. Developing data skills is costly and often involves considerable uncertainty.[1] These observations are particularly relevant in the context of alternative data. By today's standards, alternative data ranges from customer/investor sentiment to crop yields calculated from satellite images, and from app usage to survey data on construction permits. Because such data is often new to investors, investors need to build the appropriate technology infrastructure and hire data scientists to decode the data. In this paper, we develop a data-sales model to examine the implications of alternative data for financial markets.

The first contribution of this paper is to provide a framework that explicitly models how investors extract information from raw data and how the data vendor affects the precision of investors' extracted information. Our framework builds on the traditional information-sales model by Admati and Pfleiderer (1986), but we draw a conceptual difference between data and information. In our setting, the data vendor owns a large amount of data and sells data subsets to investors who exploit these data subsets to trade assets in a financial market. Raw data is not directly usable, and investors must analyze it to extract useful information. We model each raw data point as a signal of the asset payoff in the form of fundamental plus noise (with an infinite variance). Investors' ability to process raw data is further modeled as purging the noise term in this signal.[2] Higher skill levels enable them to purge more noise and result in more precise processed information.

The data vendor uses the size of the data sample to influence the precision of investors' extracted information: a larger data sample allows skilled investors to extract more precise

[1] For instance, as a chief investment officer observes, "there's a gold rush in data mining; most people that went off to prospect for gold came back penniless, but that doesn't mean there wasn't any gold." ("Hedge funds see a gold rush in data mining," *Financial Times*, August 29, 2017). This vividly describes that it remains uncertain whether an investor seeking to develop skills and decode data can succeed in the end.

[2] This modeling approach closely resembles the practical view of how data is transformed into information. See Section 5.1 for detailed discussions.

information about asset fundamentals. Statistically, skilled investors refine each raw data point into a conditionally independent informative signal; thus, the sufficient statistic of a data subset is an average of the refined signals with a precision that is a product of the size of the data sample and investors' skill level. As a result, by controlling the size of the data sample, the vendor influences how much noise remains in skilled investors' processed information. We endogenize the size of the data sample by the vendor's profit-maximization problem, and endogenize investors' skill levels by allowing investors to optimally acquire skills. To the best of our knowledge, our framework is the first to concretize the theoretical notion that the data vendor adds "personalized noise" to the sold information, as commonly assumed in the information-sales literature (e.g., Admati and Pfleiderer, 1986; García and Sangiorgi, 2011).

The second contribution of this paper is to show that our framework delivers novel economic insights. When the data vendor's data source is sufficiently accurate, the vendor chooses to sell independent data subsets with a finite sample size—in the words of Admati and Pfleiderer (1986), to add personalized noise to the sold information. This strategic response by the data vendor, combined with uncertain skill acquisition by investors, yields results that are fundamentally different from those in the existing literature. We explore equilibrium investor composition and performance, as well as the vendor's data selling behavior. In addition, we examine three market variables: price informativeness (the amount of information revealed in the asset price), the cost of capital (the expected difference between the asset's fundamental and its price), and return volatility. We find that price informativeness is hump-shaped in skill-acquisition costs, whereas the cost of capital and return volatility are U-shaped in skill-acquisition costs. Similar comparative statics patterns can arise for the mean and volatility of investor skills.

The intuitions for our results are as follows. When skill-acquisition costs increase, two competing effects arise. The first direct effect is to decrease the equilibrium mass of skilled investors. Since only skilled investors are able to extract information from data, fewer skilled investors imply less informed trading, which directly reduces price informativeness and increases the cost of capital and return volatility. Second, an indirect effect is associated with the data vendor's behavior. As explained by Admati and Pfleiderer (1986), the vendor needs to consider an information leakage effect via the asset price. A less informative price results in a less severe information leakage effect, and the vendor sells larger data subsets accordingly. In turn, the increased size of the data sample implies that each skilled investor can extract more information from the purchased data, which boosts price informativeness and reduces the cost of capital and return volatility. The interaction between these two competing effects leads to nonmonotone patterns of the market variables in our setting.

Our results differ from what standard information-sales models suggest. For instance, in Admati and Pfleiderer (1986), in response to changes in some primitive parameters, the data vendor chooses the optimal amount of noise added into the sold information such that price informativeness remains constant. In our setting, although the vendor can still use the size of the data sample to fully control price informativeness, maintaining a constant level of price informativeness is suboptimal for the vendor. Intuitively, skilled investors have heterogeneous skills, but the vendor cannot fully discriminate them, which makes the vendor's optimal price informativeness dependent on the composition of skilled investors. Therefore, the uncertainty of skill acquisition is crucial in driving our results.

Our mechanism is more relevant to alternative data than traditional data. Alternative data is a dynamic and relative notion. At different times, alternative data may refer to different types of data sets; at any given time, some types of data sets are more alternative than others. For instance, by today's standards, satellite images and sentiments are alternative but can become mainstream in the future. In contrast, stock prices and accounting data were once alternative data decades ago but have become completely traditional today. Thus, every type of data has its own cycle, and when a particular type of data is in its relatively early stage, it is relatively alternative. Given this view, most (although not all) of alternative data is relatively new to investors; thus, interpreting such data requires skills and developing such skills is costly and uncertain. In Section 5.2, we have developed a dynamic version of the baseline model to better illustrate this dynamic notion of alternative data and the role of skill uncertainty in delivering our results. The analysis of our baseline model and its extension in Section 5.2 sheds light on empirical implications of alternative data, such as the connection between the funds industry and the data industry, as well as the performance and market consequences of alternative data funds.

Our paper contributes to three strands of literature. The first is the information-sales literature, such as Admati and Pfleiderer (1986, 1988, 1990), Allen (1990), Fishman and Hagerty (1995), Naik (1997), Cespa (2008), García and Vanden (2009), García and Sangiorgi (2011), Easley, O'Hara, and Yang (2016), and Bimpikis, Crapis, and Tahbaz-Salehi (2019). This literature does not differentiate data from information and assumes that investors have equal ability to process data. In contrast, our paper draws a conceptual distinction between data and information, and endogenizes investors' skills to process data. In our setting, statistically, both raw data and processed information are signals about the asset fundamental; however, the precision of raw data as a signal is exogenous (equal to zero in our case), while the precision of processed information is endogenously determined by the investors' skill level and the sample size of the data subsets. Our framework also concretizes the theoretical notion of the vendor adding personalized noise in the sold information, as is commonly assumed in the existing information-sales literature. In addition, our analysis yields novel theoretical results, such as that price informativeness is hump-shaped in skill-acquisition costs.

The second strand of related literature considers the possibility, albeit in different settings, that investors have different data-processing abilities. For example, Glosten and Milgrom (1985, p.77) state that informed traders "may be individuals who are particularly skillful in processing public information." Indjejikian (1991) and Pagano and Volpin (2012) explicitly analyze models in which investors have different sophistication levels in understanding public disclosures made by firms or asset issuers. Consistently, Engelberg, Reed, and Ringgenberg (2012), Huang, Tan, and Wermers (2020), and Chen, Kelly, and Wu (2020) find empirical evidence that some investors' trading advantage comes from their ability to analyze publicly available data. In our setting, investors have different skills in understanding the data sold by a data vendor. Myatt and Wallace (2012) also consider that players choose how much costly attention to pay to various data sources. The microfounded signal structure in our analysis is similar to the signal structure assumed in Myatt and Wallace (2012) and some other studies such as Angeletos and Pavan (2009) and Colombo, Femminis, and Pavan (2014).

Finally, our paper is related to the strand of literature on institutional investors, informed trading, and market quality. In our paper, the notion of skilled investors is conceptually related to a fund manager's skills, such as market timing and stock picking (e.g., Kacperczyk, Nieuwerburgh, and Veldkamp, 2014). The existing literature has demonstrated that institutional trading significantly impacts price informativeness through various channels (e.g., Kacperczyk, Nosal, and Sundaresan, 2020; Breugem and Buss, 2019; Kacperczyk, Sundaresan, and Wang, 2020; Huang, Qiu, and Yang, 2020; Buss and Sundaresan, 2020). For example, Breugem and Buss (2019) show that the benchmark concerns among institutional investors may harm price efficiency. Huang, Qiu, and Yang (2020) demonstrate that institutionalization could increase the cost of capital and return volatility when there are moral hazard problems between investors and fund managers. Buss and Sundaresan (2020) find that passive ownership can improve price informativeness. Our analysis highlights the interactions between investors' skill acquisition and the data vendor's strategic response that drive the behavior of price informativeness.

2 A Model of Skill Acquisition and Data Sales

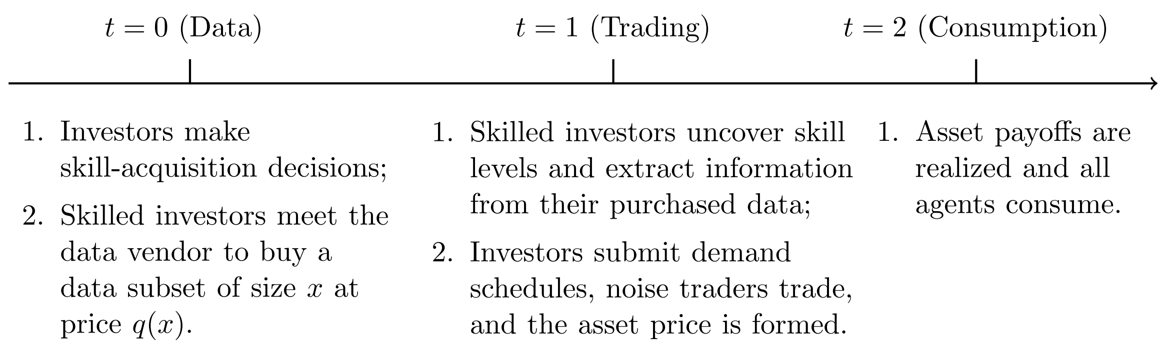


Figure 1: Timeline

We study a data-sales model with one data vendor and a continuum of investors. Our economy has three dates: $t = 0$, 1, and 2. The order of events is described in Figure 1. At date 0, investors make skill-acquisition decisions, and skilled investors purchase data from the data vendor. At date 1, an asset market opens and investors trade. At date 2, the assets pay off and all agents consume.

2.1 Assets and Asset Markets

There are two tradable assets in the date-1 financial market: a risk-free asset and a risky asset. The payoff of the risk-free asset has a constant value of 1 for simplicity and is in unlimited supply. The risky asset is traded at an endogenous price $\tilde{p}$ per unit in the date-1 market, and its total supply is normalized as one share. The risky asset pays an uncertain cash flow $\tilde{\theta}$ at date 2, where $\tilde{\theta} \sim N\left(0, \tau_\theta^{-1}\right)$ with $\tau_\theta \in (0, \infty)$. There is noisy demand $\tilde{u}$ for the risky asset, where $\tilde{u} \sim N\left(0, \tau_u^{-1}\right)$ with $\tau_u \in (0, \infty)$. Noisy trading $\tilde{u}$ is independent of all other random variables in the economy and serves the usual role of preventing private information from being perfectly revealed by the price.

2.2 From Raw Data to Processed Information

The data vendor, such as a big-data company, owns a large amount of raw data and sells subsets of data to investors. To capture the idea that the data vendor's data size is "large", we assume that the data set in its entirety has a continuum of data points.[3] Each data point $j \in [0, 1]$ is a signal about the asset fundamental $\tilde{\theta}$ in the following form:

$$\tilde{\delta}_j = \tilde{\theta} + \tilde{\eta} + \tilde{e}_j, \tag{1}$$

where $(\tilde{\theta}, \tilde{\eta}, \{\tilde{e}_j\})$ are mutually independent. Variable $\tilde{\eta}$ is normally distributed with mean 0 and precision $\tau_\eta \in (0, \infty)$ (i.e., $\tilde{\eta} \sim N\left(0, \tau_\eta^{-1}\right)$). This variable captures the noise inherent in the data vendor's data source and, thus, cannot be further reduced. For instance, if the risky asset is a firm's stock and each data point represents a customer's review about the firm's products, then $\tilde{\eta}$ can be the common bias among customers' reviews in evaluating the firm's fundamental value. Variable $\tilde{e}_j$ follows a uniform distribution on the real line (i.e., $\tilde{e}_j \sim N(0, \infty)$). This zero precision assumption of $\tilde{e}_j$ captures the "raw" feature of data: without processing, each data point is useless in predicting fundamental $\tilde{\theta}$.

Skilled investors can extract information from raw data to guide trading decisions. We follow the literature (e.g., Kim and Verrecchia, 1994; Ganguli and Yang, 2009; Farboodi

[3] We assume a continuum of data points to match the assumption of a continuum of investors. Alternatively, we could have assumed that the entire data set has J_D data points and that there are J_I investors, where J_D and J_I are positive integers. We can then take J_D to ∞ and J_I to ∞ sequentially to capture the feature of a large data set and a large economy. The results are the same under this alternative treatment.

et al., 2020) and model this information-extraction problem as investors purging the noise $\tilde{e}_j$ in raw data. As will be discussed in Section 5.1, this modeling approach also closely resembles the practical view of extracting information from data. Specifically, skilled investor i can produce a noisy signal $\tilde{f}_{j,i}$ about $\tilde{e}_j$ as follows:

$$\tilde{f}_{j,i} = \tilde{e}_j + \tilde{\xi}_{j,i}, \text{ with } \tilde{\xi}_{j,i} \sim N\left(0, \frac{1}{A_D z_i}\right), \tag{2}$$

where $(\{\tilde{e}_j\}, \{\tilde{\xi}_{j,i}\})$ are mutually independent. Here, parameter $A_D > 0$ controls how easy it is to decode raw data, which is specific to the data source. Parameter $z_i \geq 0$ is investor i's skill level of decoding raw data, and a higher value of z_i implies that investor i is more skilled. To see this point, we notice that raw data point $\tilde{\delta}_j$ and noise-purging signal $\tilde{f}_{j,i}$ combined are equivalent to the following signal in predicting fundamental $\tilde{\theta}$:

$$\tilde{s}_{j,i} \equiv \tilde{\delta}_j - E(\tilde{e}_j | \tilde{f}_{j,i}) = \tilde{\theta} + \tilde{\eta} - \tilde{\xi}_{j,i}. \tag{3}$$

Note that if $z_i = 0$ (i.e., if investor i is unskilled), then $\tilde{\xi}_{j,i} \sim N(0, \infty)$, such that $\tilde{s}_{j,i}$ is useless for predicting $\tilde{\theta}$. If the data vendor sells M data points $\left\{\tilde{\delta}_{1,i}, ..., \tilde{\delta}_{M,i}\right\}$ to investor i, then investor i processes each data point to obtain a signal in the form of (3), which implies that the sufficient statistic of the purchased data set is:

$$\tilde{s}_i \equiv \frac{1}{M} \sum\nolimits_{j=1}^{M} \tilde{s}_{j,i} = \tilde{\theta} + \tilde{\eta} + \tilde{\varepsilon}_i, \tag{4}$$

where

$$\tilde{\varepsilon}_i \equiv -\frac{1}{M} \sum\nolimits_{j=1}^{M} \tilde{\xi}_{j,i} \sim N\left(0, \frac{1}{A_D M z_i}\right). \tag{5}$$

The data vendor sells each investor an independent subset of data points[4] and chooses the size M of each data subset. Working with discrete number M complicates the analysis considerably but adds little economic insight. Therefore, we follow the literature (e.g., Farboodi, Matray, and Veldkamp, 2017; Begenau, Farboodi, and Veldkamp, 2018) and consider a quasicontinuous limit. That is, we define $x \equiv A_D M$ and consider the limit of $A_D \to 0$. In

[4]Anecdotal evidence suggests that vendors of alternative data may sell different subsets of data to different clients. For example, some hedge funds are paying a couple of millions a year to obtain a specific data set and these contracts often involve some measure of exclusive usage rights granted to the funds ("The new gold rush? Wall Street wants your data," Matt Turck Blog, January 17, 2017). As will be clear in Lemma 2, in our model, the vendor sells different subsets of data to different clients when τ_η is sufficiently high.

this setup, we have

$$\tilde{\varepsilon}_i \sim N\left(0, \frac{1}{xz_i}\right), \tag{6}$$

where x can be approximated as a continuous variable taking values on $(0, \infty)$.[5] From now on, we work with variable x and refer to it as the "data sample size." In summary, although raw data and information are different economic concepts, statistically, both can be envisioned as signals about asset fundamentals. It is just that raw data's precision is exogenously given, while processed information's precision xz_i is endogenously determined by the vendor's data sample size x and the investor's skill level z_i. In particular, the data vendor can reduce or enhance the precision level of investors' processed information by selling a smaller or larger independent data subset. This interpretation concretizes the abstract theoretical notion in the information-sales literature that data vendors control the size of personalized noise in their sold information to manage information leakage via asset prices (e.g., Admati and Pfleiderer, 1986; García and Sangiorgi, 2011).

In our baseline model, we focus on the symmetric case in which the vendor sells the same number of data points to each investor because skilled investors in the data market have not discovered their skill levels yet; therefore, the data vendor cannot discriminate. This unobservability feature of investor skills turns out to be crucial in driving many of our results. In Section 5.2, we consider an extension to allow the vendor to sell different numbers of data points to different investors after investors discover their skill levels.

2.3 Skill Acquisition and Asset Trading

There exists a continuum of investors, indexed by $i \in [0, 1]$. Investors acquire skills and purchase data at date 0, trade assets at date 1, and consume at date 2. All investors derive expected utility over their date-2 wealth according to a constant absolute risk aversion (CARA) utility with a common risk-aversion coefficient γ.

Skill acquisition is costly and uncertain. For instance, to trade on alternative data, a financial institution needs to hire data specialists and build the relevant infrastructure to analyze the data and develop algorithms, which costs resources. In addition, the developed

[5] For instance, if $A_D = 0.0001$, then x can take values $0.0001, 0.0002, 0.0003, ...$ As variable A_D becomes smaller, the grid can shrink further.

algorithms may or may not work well, which corresponds to uncertainty in decoding alternative data. To capture these two features, we make two assumptions. First, acquiring skills incurs a cost $C(i)$ for investor i, where $C(i)$ is continuous, increasing, and convex in i, $C(0) = 0$, and $C(1) = \infty$. We use λ to denote the fraction of investors who decide to acquire skills. The functional form of $C(\cdot)$ ensures an interior mass λ of skilled investors. Second, after investor i decides to acquire skills, her skill level can be Z_L or Z_H with equal probability, where $0 \leq Z_L \leq Z_H < \infty$. We refer to investors with skill level Z_H as "high-type skilled investors" and to those with skill level Z_L as "low-type skilled investors." We define the mean and volatility of skill levels, respectively, as

$$\bar{Z} \equiv \frac{Z_H + Z_L}{2} \text{ and } \Delta \equiv \frac{Z_H - Z_L}{2}. \tag{7}$$

At date 1, skilled investors uncover their skill levels before trading. Overall, there are a mass $\frac{\lambda}{2}$ of high-type skilled investors and a mass $\frac{\lambda}{2}$ of low-type skilled investors. The remaining mass $1 - \lambda$ of investors are unskilled. Both skilled and unskilled investors trade in the financial market. As standard in the rational-expectations equilibrium literature, all investors submit demand schedules and can condition their trades on prices. Given a slight abuse of notations, a high-type skilled investor i's information set is $\mathcal{F}_{H,i} = \{\tilde{p}, \tilde{s}_{H,i}\}$, where $\tilde{s}_{H,i}$ is given by (4) replacing $\tilde{\varepsilon}_i$ with $\tilde{\varepsilon}_{H,i}$, which has precision xZ_H. This investor chooses demand $D_{H,i}$ for the risky asset to maximize $E\left[-e^{-\gamma D_{H,i}(\tilde{\theta}-\tilde{p})}\middle|\mathcal{F}_{H,i}\right]$. The decision problems of low-type skilled investors and unskilled investors are defined similarly.

2.4 Data Prices and the Vendor's Profit

Both data and skills are scarce resources; thus, both the data vendor and skilled investors should earn economic rents. We follow Gârleanu and Pedersen (2018) and use Nash bargaining as a modeling device to implement profit sharing. Nash bargaining also captures the real-world features of data markets (Anand and Galetovic, 2000); particularly for alternative data markets, the buy-side often hires teams of staff to speak with data owners, and the two sides directly bargain over the price. Each skilled investor and the vendor form a pair to engage in Nash bargaining. We use $\beta \in [0, 1]$ to denote the bargaining power of the data vendor. The data price q depends on the data sample size x and the mass λ of skilled investors, that is, $q = q(x; \lambda)$. The data sample size x affects the usefulness of data (i.e., precision xz_i of the processed information) and, hence, the data price. The mass λ of skilled investors affects the data price through information leakage via asset prices (Admati and Pfleiderer, 1986; García and Sangiorgi, 2011). The profit from selling data is $\pi(x; \lambda) = \lambda q(x; \lambda)$. The vendor takes λ as given and chooses x to maximize profit $\pi(x; \lambda)$: $\max_x \lambda q(x; \lambda)$.

3 Equilibrium Characterization

The economy is defined by a tuple of seven exogenous parameters and one exogenous function, $\mathcal{E} \equiv \{\tau_\theta, \tau_\eta, \tau_u, \gamma, \beta, Z_H, Z_L, C(\cdot)\}$. Note that parameters $\bar{Z}$ and Δ specified in (7) are equivalent to $\{Z_H, Z_L\}$ given that $Z_H = \bar{Z} + \Delta$ and $Z_L = \bar{Z} - \Delta$. For a given economy $\mathcal{E}$, we define an equilibrium in a subgame-perfect sense.

Definition 1 *An equilibrium is characterized by a mass λ^* of skilled investors, a skill-acquisition decision rule $A(i; \lambda^*)$ of investor i, a data sample size x^*, a data price $q(x^*; \lambda^*)$, an asset price function $p(\tilde{\theta} + \tilde{\eta}, \tilde{u})$, and a demand schedule $D(\mathcal{F}_i)$ of investor i, such that:*

(a) *At date 1, (i) demand schedule $D(\mathcal{F}_i)$ maximizes investor i's expected utility conditional on her information set $\mathcal{F}_i$; (ii) the price function $p(\tilde{\theta} + \tilde{\eta}, \tilde{u})$ almost surely clears the asset market.*

(b) *At date 0, (i) the skill-acquisition decision rule $A(i; \lambda^*)$ maximizes investor i's ex ante expected utility (where $A(i; \lambda^*) = 1$ or 0), and $\lambda^* = \int_0^1 A(i; \lambda^*)\, di$; (ii) the data sample size x^* maximizes the vendor's profit $\lambda^* q(x; \lambda^*)$, where the data price $q(x; \lambda^*)$ is generated via Nash bargaining between the data vendor and a typical skilled investor.*

3.1 Financial Market Equilibrium

As is standard in the literature, we consider the following linear price function:

$$\tilde{p} = a_0 + a_\theta(\tilde{\theta} + \tilde{\eta}) + a_u \tilde{u}, \tag{8}$$

where the a-coefficients are endogenously determined in equilibrium. For any investor, the information contained in the price is equivalent to the following signal $\tilde{s}_p$:

$$\tilde{s}_p \equiv \frac{\tilde{p} - a_0}{a_\theta} = \tilde{\theta} + \tilde{\eta} + \frac{\tilde{u}}{\alpha}, \text{ with } \alpha \equiv \frac{a_\theta}{a_u}, \tag{9}$$

where $\tilde{s}_p$ is normally distributed, with mean $\tilde{\theta} + \tilde{\eta}$ and precision $\alpha^2 \tau_u$. The CARA-normal setup implies that investor i's demand function is

$$D(\mathcal{F}_i) = \frac{E(\tilde{\theta}|\mathcal{F}_i) - \tilde{p}}{\gamma Var(\tilde{\theta}|\mathcal{F}_i)}, \tag{10}$$

where $\mathcal{F}_i$ is investor i's information set (including the asset price $\tilde{p}$).

Investors differ in their data-processing abilities and, hence, information sets. As mentioned in Section 2, three types of investors exist depending on their skill levels for processing raw data: (i) high-type skilled investors, labeled H, with data-processing ability Z_H; (ii) low-type skilled investors, labeled L, with data-processing ability Z_L; and (iii) unskilled investors, labeled U, without any data-processing ability. The masses of the three types of investors are $\frac{\lambda}{2}$, $\frac{\lambda}{2}$, and $1 - \lambda$, respectively.

An H-type skilled investor i has information set $\{\tilde{p}, \tilde{s}_{H,i}\}$, which is equivalent to $\{\tilde{s}_p, \tilde{s}_{H,i}\}$. Processed information $\tilde{s}_{H,i}$ has precision xZ_H in predicting $\tilde{\theta} + \tilde{\eta}$. Applying Bayes' rule, we can compute the conditional moments of an H-type skilled investor i as follows:

$$E(\tilde{\theta}|\tilde{p}, \tilde{s}_{H,i}) = \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{\alpha^2 \tau_u \tilde{s}_p + xZ_H \tilde{s}_{H,i}}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha^2 \tau_u + xZ_H}, \tag{11}$$

$$Var(\tilde{\theta}|\tilde{p}, \tilde{s}_{H,i}) = \frac{1}{\tau_\theta + \tau_\eta} + \left(\frac{\tau_\eta}{\tau_\theta + \tau_\eta}\right)^2 \frac{1}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha^2 \tau_u + xZ_H}. \tag{12}$$

An L-type skilled investor i has information set $\{\tilde{p}, \tilde{s}_{L,i}\}$, equivalent to $\{\tilde{s}_p, \tilde{s}_{L,i}\}$, where $\tilde{s}_{L,i}$ has precision xZ_L in predicting $\tilde{\theta} + \tilde{\eta}$. We can compute the conditional moments as follows:

$$E(\tilde{\theta}|\tilde{p}, \tilde{s}_{L,i}) = \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{\alpha^2 \tau_u \tilde{s}_p + xZ_L \tilde{s}_{L,i}}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha^2 \tau_u + xZ_L}, \tag{13}$$

$$Var(\tilde{\theta}|\tilde{p}, \tilde{s}_{L,i}) = \frac{1}{\tau_\theta + \tau_\eta} + \left(\frac{\tau_\eta}{\tau_\theta + \tau_\eta}\right)^2 \frac{1}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha^2 \tau_u + xZ_L}. \tag{14}$$

An unskilled investor can forecast asset fundamentals based only on price $\tilde{p}$. Using Bayes' rule, we have

$$E(\tilde{\theta}|\tilde{p}) = \frac{\tau_\eta}{\tau_\theta+\tau_\eta}\frac{\alpha^2\tau_u\tilde{s}_p}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha^2\tau_u}, \tag{15}$$

$$Var(\tilde{\theta}|\tilde{p}) = \frac{1}{\tau_\theta+\tau_\eta}+\left(\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\right)^2\frac{1}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha^2\tau_u}. \tag{16}$$

By (12), (14), and (16), we can see that $Var(\tilde{\theta}|\tilde{p},\tilde{s}_{H,i}) < Var(\tilde{\theta}|\tilde{p},\tilde{s}_{L,i}) < Var(\tilde{\theta}|\tilde{p})$, which reflects the different data-processing abilities of the three groups of investors.

The market-clearing condition is

$$\int_0^{\lambda/2} D_H\left(\tilde{p},\tilde{s}_{H,i}\right)di+\int_{\lambda/2}^{\lambda} D_L\left(\tilde{p},\tilde{s}_{L,i}\right)di+(1-\lambda)\,D_U\left(\tilde{p}\right)+\tilde{u}=1. \tag{17}$$

To derive the equilibrium price function, we insert the demand functions into the market-clearing condition to solve for the price in terms of $\tilde{\theta}+\tilde{\eta}$ and $\tilde{u}$. We then compare this implied price function with the conjectured price function in equation (8) to obtain and solve a system that characterizes the unknown a-coefficients.

Lemma 1 (Financial market equilibrium) *Given $(\lambda, x; \tau_\theta, \tau_\eta, \tau_u, \gamma, Z_H, Z_L)$, there exists a unique linear equilibrium price function, given by equation (8), where a_0, a_θ, and a_u are specified by equations (A2)-(A4) in the Appendix. The equilibrium is characterized by the ratio $\alpha = \frac{a_\theta}{a_u}$, which is the unique real positive root of the following equation:*

$$2\gamma\alpha - x\lambda\tau_\eta\left(\frac{Z_H}{xZ_H+\tau_u\alpha^2+\tau_\eta}+\frac{Z_L}{xZ_L+\tau_u\alpha^2+\tau_\eta}\right)=0. \tag{18}$$

The endogenous variable α defined by (9) is positively related to price informativeness, which measures how much fundamental information is revealed by the price. Formally, we follow the literature (e.g., Easley et al., 2016; Goldstein and Yang, 2017) and measure price informativeness by $\frac{1}{Var(\tilde{\theta}|\tilde{p})}$. By equation (16), $\frac{1}{Var(\tilde{\theta}|\tilde{p})}$ is determined by four variables: $\tau_\theta, \tau_\eta, \tau_u$, and α. To the extent that parameters τ_θ, τ_η, and τ_u are exogenous, the endogenous variable α uniquely determines price informativeness.[6]

Corollary 1 (Price informativeness) *Given $(x; \tau_\theta, \tau_\eta, \tau_u, \gamma)$, price informativeness $\frac{1}{Var(\tilde{\theta}|\tilde{p})}$ increases with the mass λ of skilled investors and the mean $\bar{Z}$ of skill levels, and decreases with the volatility Δ of skill levels.*

[6] Our price-informativeness measure is a monotonic transformation of the price-informativeness measure adopted by Bond, Edmans, and Goldstein (2012) and Bai, Philippon, and Savov (2016). They measure price informativeness by the forecasting price efficiency (FPE), $Var[E(\tilde{\theta}|\tilde{p})]$. Given that $Var[E(\tilde{\theta}|\tilde{p})] = Var(\tilde{\theta}) - Var(\tilde{\theta}|\tilde{p})$, our price-informativeness measure is isomorphic to the FPE measure.

In Corollary 1, we shut down data sales (i.e., take x as exogenous) and find that price informativeness increases with λ and $\bar{Z}$ but decreases with Δ. As will be shown in Section 4, in the economy with data sales, price informativeness could decrease with λ and $\bar{Z}$ and increase with Δ. Such a contrast highlights the importance of interacting skill acquisition and data sales in driving our results.

3.2 Data Market Equilibrium

At date 0, the data price is determined as an outcome of Nash bargaining between the data vendor and a skilled investor who has not yet discovered her skill type. Recall that the bargaining equilibrium depends on agents' utilities in the event of agreement or no agreement. For the data vendor, if an agreement is reached, its additional profit increases by the data price q. For the investor, we follow Gârleanu and Pedersen (2018) and specify the investor's bargaining objective in terms of certainty-equivalent wealth. That is, the utility when agreeing on data price q is $CE_S - q$, where CE_S is the ex ante certainty equivalent of the skilled investor. If no agreement is reached, the investor's outside option is to invest as the unskilled, yielding a utility of CE_U, where CE_U is the ex ante certainty equivalent of an unskilled investor. Hence, the investor's gain from agreement is $CE_S - CE_U - q$.

The difference $CE_S - CE_U$ captures the total gain from interpreting and trading on the data. We denote this gain by G, that is, $G \equiv CE_S - CE_U$. In the Appendix, we compute the trading gain G as follows:

$$G(x,\alpha) = \frac{1}{\gamma}\ln\left[\frac{\sqrt{Var(\tilde{\theta}|\tilde{p})}}{\frac{1}{2}\sqrt{Var(\tilde{\theta}|\tilde{p},\tilde{s}_{H,i})}+\frac{1}{2}\sqrt{Var(\tilde{\theta}|\tilde{p},\tilde{s}_{L,i})}}\right]. \tag{19}$$

Since $Var(\tilde{\theta}|\tilde{p},\tilde{s}_{H,i})$, $Var(\tilde{\theta}|\tilde{p},\tilde{s}_{L,i})$, and $Var(\tilde{\theta}|\tilde{p})$ in equations (12), (14), and (16) are determined by (x,α), we explicitly express G as a function of (x,α) in (19).

In the Nash-bargaining game, the vendor's gain from agreement is the data price q and the investor's gain from agreement is $G - q$. The vendor has bargaining power $\beta \in [0,1]$. The bargaining outcome maximizes the product of the utility gains from agreement:

$$\max_q q^{\beta}(G-q)^{1-\beta} \Rightarrow q^* = \beta G, \tag{20}$$

where q^* denotes the equilibrium data price. Thus, the total gain G generated from a skilled investor (who interprets and trades on raw data) is divided between the vendor and the investor as follows:

$$G = q + (G - q) = \underbrace{\beta G}_{\text{vendor's rent}} + \underbrace{(1-\beta)\,G}_{\text{skilled investor's rent}}. \tag{21}$$

If we interpret a skilled investor as an active fund, then part of the fund's revenue goes to data providers (as expenses on data purchase), and the remainder is left to the fund (as compensation for trading skills). The division of gains is determined by the bargaining power $(1-\beta)$ of the fund manager in negotiating prices with data providers.

Since G is a function of (x, α) and $q^* = \beta G$, the data price q^* is a function of (x, α) as well. In addition, by equation (18) that characterizes the financial market equilibrium, α is uniquely determined by (x, λ). Thus, we can write the equilibrium data price as $q(x, \alpha) = q(x, \alpha(x, \lambda)) = q(x; \lambda)$. The vendor takes λ as given and chooses x to maximize its profit, $\lambda q(x; \lambda)$ or, equivalently, the data price $q(x; \lambda)$. We find that two types of data sample size x^* arise in equilibrium, as characterized by the following lemma.

Lemma 2 (Data sample size) *There exist two types of equilibrium data sample size x^*:*

(a) *When τ_η is sufficiently low, the vendor sells an infinite amount of data points; that is, $x^* = \infty$.*

(b) *When τ_η is sufficiently high, the vendor sells a finite number of data points; that is, $x^* < \infty$.*

When data accuracy τ_η is low, the vendor sells a data sample including an infinite number of data points, which allows skilled investors to uncover the best signal $\tilde{\theta} + \tilde{\eta}$ that can be delivered by the vendor's entire data set (recall that noise $\tilde{\eta}$ is data-source specific and cannot be removed by any agent). In the language of Admati and Pfleiderer (1986), in this case, the data vendor sells data sets "as is." In contrast, when data accuracy τ_η is high, the vendor sells a data sample with a limited number of data points. In this case, skilled investors cannot uncover the best signal $\tilde{\theta} + \tilde{\eta}$ that could be delivered from the vendor's entire data set. In the language of Admati and Pfleiderer (1986), in this case, the data vendor adds personalized noise into the sold data. Our setting concretizes how this personalized noise is exactly added—that is, the data vendor sells independent data samples to skilled investors to implement personalized noise and then it uses the size of the data sample to control the size of this personalized noise.

Intuitively, increasing the data sample size x has two competing effects. First, other things being equal, more data points imply that skilled investors can extract more accurate information, which directly increases their willingness to pay for the data sample. Second, there is an indirect information-leakage effect via the price. More data points make the price more informative through the trading of skilled investors, which increases the payoff of those who do not purchase data and indirectly decreases the data price. The payoff increases because the data price is determined by the comparison between the payoffs of those purchasing and not purchasing data. When data accuracy is sufficiently high, the vendor is very concerned with information leakage and, thus, chooses a finite level x^*.

3.3 Skill-Acquisition Equilibrium

We finally determine the equilibrium mass λ^* of skilled investors. At date 0, an investor trades off between the benefit and cost of acquiring skills. When making her own skill-acquisition decision, investor i takes as given other investors' skill-acquisition decisions (and, hence, the mass λ^* of skilled investors and the vendor's data sample size x^* as an optimal response to λ^*). If investor i decides to acquire skills, she expects to receive revenue of $(1-\beta)\,G$ resulting from the Nash-bargaining game, but acquiring skills incurs a cost of $C\,(i)$. If $(1-\beta)\,G > C\,(i)$, investor i decides to acquire skills; otherwise, she does not. Thus,

$$A\,(i;\lambda^*) = \begin{cases} 1, & \text{if } (1-\beta)\,G \geq C\,(i)\,, \\ 0, & \text{otherwise,} \end{cases} \tag{22}$$

where $A\,(i;\lambda^*) = 1$ indicates that investor i chooses to acquire skills.

Since $C\,(i)$ monotonically increases from 0 to ∞, there always exists an interior equilibrium mass $\lambda^* \in (0,1)$ of skilled investors. That is, $A\,(i;\lambda^*) = 1$ as long as $i \in [0,\lambda^*]$. The equilibrium mass λ^* is determined by

$$(1-\beta)\,G\,(x^*,\alpha\,(x^*,\lambda^*)) = C\,(\lambda^*)\,, \tag{23}$$

where $G\,(x,\alpha)$ is given by equation (19).

Lemma 3 (Overall equilibrium) *There exists an overall equilibrium such that:*

(a) *At date 0, a mass λ^* of investors decide to acquire skills, and the vendor sells data sample with size x^* at price $q(x^*; \lambda^*)$.*

(b) *At date 1, demand schedules are given by equation (10), and the price function is characterized by Lemma 1.*

4 Skill Acquisition, Data Sales, and Asset Prices

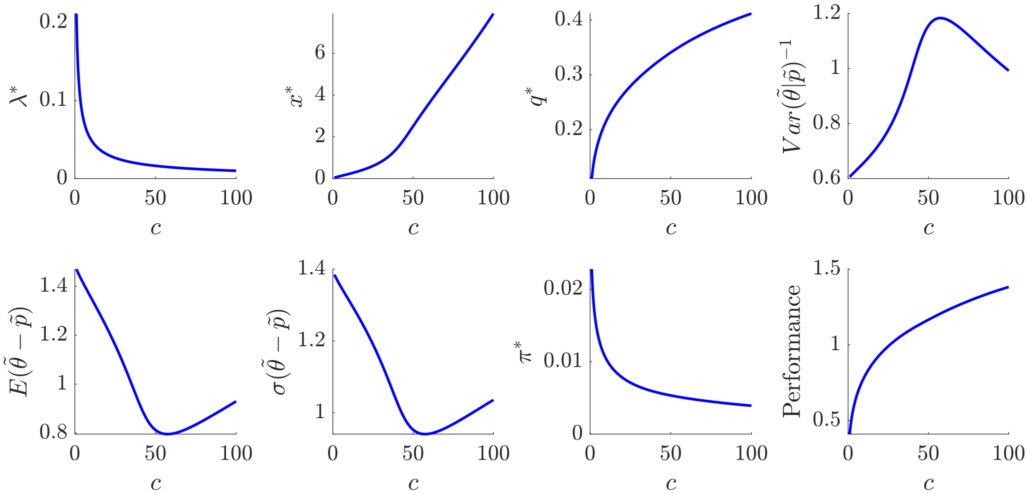
In this section, we examine how skill acquisition affects data sales, financial markets, and the asset management industry by conducting comparative statics with respect to parameters governing the skill-acquisition cost function $C(\cdot)$ and parameters governing skill levels ($\bar{Z}$ and Δ). In the analysis of this section, we focus on the more interesting economy with high data accuracy τ_η because only in this case does the vendor strategically adjust the size of its data sample, leading to novel economic insights.

4.1 Skill-Acquisition Costs

To facilitate the analysis, we specify that the skill-acquisition cost function $C(\cdot)$ is governed by parameter $c > 0$, meaning that an increase in c shifts the entire curve $C(\cdot)$ upward (i.e., $\frac{\partial C(\cdot)}{\partial c} > 0$). We use Figure 2 to graphically illustrate our results for $C(i) = \frac{ci}{1-i}$ with $c > 0$. We choose the following parameter configuration: $\tau_\theta = 0.3, \tau_u = 4, \tau_\eta = 100, \beta = 0.3, \gamma = 1.35, \bar{Z} = 50$, and $\Delta = 40$. Our results are robust to the choice of parameter values.

It is intuitive that as c increases, fewer investors choose to acquire skills (i.e., $\frac{\partial \lambda^*}{\partial c} < 0$). This feeds back to the supply side of the data (data sample size x^* and data price q^*). Because fewer skilled investors exist, the information leakage effect is less of a concern; therefore, the vendor can choose to deliver data samples of a larger size (i.e., $\frac{\partial x^*}{\partial c} > 0$). This larger data sample size implies that skilled investors can extract more precise information from data and, thus, they are willing to pay more (i.e., $\frac{\partial q^*}{\partial c} > 0$).

Price informativeness is hump-shaped in c. This result is driven by two competing forces. First, as c increases, fewer investors acquire skills to interpret and trade on data, which injects



This figure plots the effects of skill-acquisition costs. The skill-acquisition cost function is specified as $c(i) = ci/(1-i)$ with $c > 0$. The other parameters are: $\tau_\theta = 0.3$, $\tau_u = 4$, $\tau_\eta = 100$, $\beta = 0.3$, $\gamma = 1.35$, $\bar{Z} = 50$, and $\Delta = 40$.

Figure 2: Effects of skill-acquisition costs

less information into the price. Second, as c increases, the data vendor increases the size of the data sample, which allows skilled investors to extract more precise information from the purchased data, thereby injecting more information into the price. This second effect is particularly strong when many skilled investors interpret and trade on the data. Thus, for small c (and large λ^*), price informativeness increases with c.

The pattern of price informativeness in our setting differs from that in Admati and Pfleiderer (1986). In Admati and Pfleiderer (1986), all investors have equal data processing ability, and their analysis suggests that price informativeness is held constant in response to changes in some exogenous parameters. This would also be the case in our setting if the data vendor could discriminate between high- and low-skilled investors, such that it could sell them data samples of different sizes and make the two types of investors effectively the same at the trading stage. However, when the data vendor cannot discriminate among investors, setting a constant price-informativeness level is no longer optimal for the vendor. In the extended economy analyzed in Section 5.2, we will make clear these points and argue that the skill uncertainty feature better captures the early stage of the data cycle.

In addition to price informativeness, we also examine two additional asset price variables:

the cost of capital $E(\tilde{\theta}-\tilde{p})$ and return volatility $\sigma(\tilde{\theta}-\tilde{p})$. The cost of capital is the expected difference between the asset fundamental and the asset price. This difference arises from the compensation required to induce investors to hold the risky asset. The return on the risky asset is $\tilde{\theta}-\tilde{p}$; thus, its volatility can be measured by $\sigma(\tilde{\theta}-\tilde{p})$. Similar to price informativeness, the two competing forces from λ^* and x^* make both $E(\tilde{\theta}-\tilde{p})$ and $\sigma(\tilde{\theta}-\tilde{p})$ U-shaped in c. On the one hand, as c increases, there are fewer skilled investors and more unskilled investors, which tends to increase $E(\tilde{\theta}-\tilde{p})$ and $\sigma(\tilde{\theta}-\tilde{p})$ since skilled investors trade more aggressively than unskilled investors. On the other hand, as c increases, the vendor sells larger data samples, which tends to decrease $E(\tilde{\theta}-\tilde{p})$ and $\sigma(\tilde{\theta}-\tilde{p})$ because each skilled investor now trades more aggressively by extracting more accurate information from the purchased data, and unskilled investors can trade more aggressively as the price becomes more informative.

Finally, we report the vendor's profit and skilled investors' performance. The vendor's profit is $\pi^*=\lambda^*q^*$. As c increases, λ^* decreases but q^* increases. Nonetheless, the effect of λ^* dominates; thus, the vendor's profit π^* decreases with c. We follow Gârleanu and Pedersen (2018) and proxy for the performance of skilled investors by the disparity between a skilled investor's certainty equivalent and an unskilled investor's certainty equivalent. Specifically, we compute the performance of a skilled investor $i\in[0,1]$ as $\frac{1}{2\gamma}\ln\left[\frac{Var(\tilde{\theta}|\tilde{p})}{Var(\tilde{\theta}|\tilde{p},\tilde{s}_i)}\right]$, which captures the additional expected utility (outperformance) of a high- or low-skilled investor relative to an unskilled (uninformed) investor.[7] Thus, the performance of a typical skilled investor is the weighted average of the performance of a high- and a low-skilled investor:

$$Performance=\frac{1}{4\gamma}\ln\left[\frac{Var(\tilde{\theta}|\tilde{p})}{Var(\tilde{\theta}|\tilde{p},\tilde{s}_{H,i})}\right]+\frac{1}{4\gamma}\ln\left[\frac{Var(\tilde{\theta}|\tilde{p})}{Var(\tilde{\theta}|\tilde{p},\tilde{s}_{L,i})}\right]. \tag{24}$$

Since trading performance compensates for investors' skill-acquisition costs, performance is higher in financial markets with higher skill-acquisition costs.

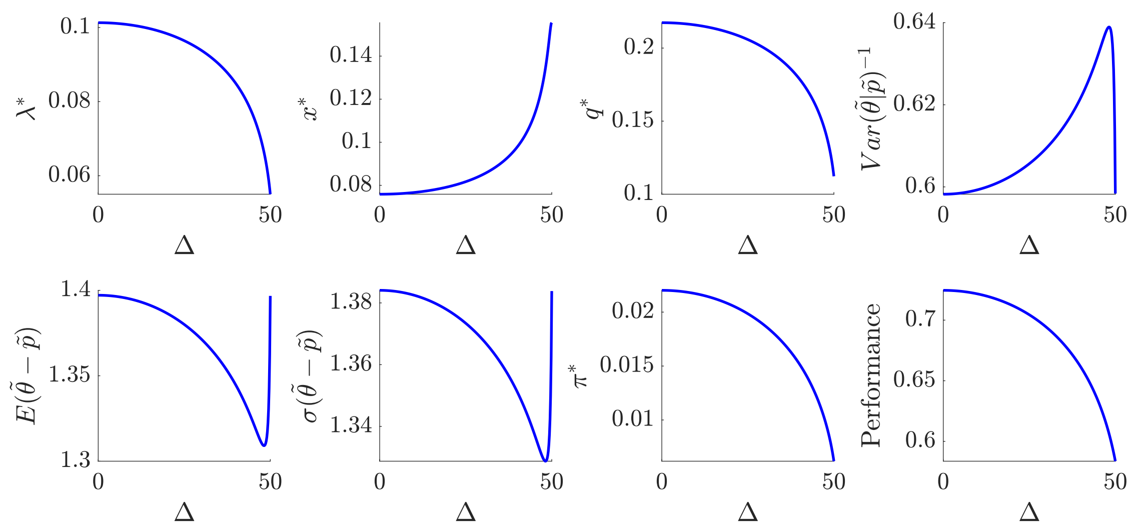
Although we cannot provide a fully analytical proof for these results, we manage to provide a sufficient condition under which the relatively novel results hold, such as price informativeness increasing with c. The idea is to consider a limit of $\tau_\eta\to\infty$ while keeping c finite such that λ^* is bounded below. Formally, we have the following proposition:

[7] As Gârleanu and Pedersen (2018) note, an investor's expected utility is directly linked to her (squared) Sharpe ratio. We do not subtract the data price q or skill-acquisition costs $C(i)$ from performance because, in practice, fund performance is examined without consideration of the expenses assumed by the fund side.

Proposition 1 (Skill-acquisition costs) *Suppose that the data vendor owns sufficiently accurate data (τ_η is sufficiently high) and that c is finite such that λ^* is bounded below. When skill-acquisition cost parameter c increases (meaning that $C(\cdot)$ shifts upward), the population λ^* of skilled investors decreases, the data sample size x^* increases, the data price q^* increases, price informativeness $\frac{1}{Var(\tilde{\theta}|\tilde{p})}$ increases, the cost of capital $E(\tilde{\theta}-\tilde{p})$ decreases, return volatility $\sigma(\tilde{\theta}-\tilde{p})$ decreases, and skilled investors' performance increases.*

4.2 Skill Volatility

We use Figure 3 to examine the implications of skill volatility Δ. In this exercise, we fix the skill mean $\bar{Z}$ at 50 and the skill-acquisition cost parameter c at 4.5. All other parameter values are the same as those in Figure 2.



This figure plots the effects of skill volatility Δ. The skill-acquisition cost function is the same as in Figure 2 with $c = 4.5$. The parameters other than Δ take the same values as those in Figure 2

Figure 3: Effects of skill volatility

The effects of increasing skill volatility are largely similar to those of increasing skill-acquisition costs. For a fixed skill mean $\bar{Z}$, increasing skill volatility Δ means that the skill level Z_H of high-skilled investors increases and the skill level Z_L of low-skilled investors decreases at the trading stage. However, at the skill-acquisition stage, investors are uncertain about their future skill types. Since investors are risk averse, their expected benefit of acquiring skills decreases. As a result, fewer investors choose to acquire skills in response to an increase in Δ (i.e., $\frac{\partial \lambda^*}{\partial \Delta} < 0$). The reduced mass of skilled investors alleviates the data vendor's information leakage concern, which in turn encourages it to sell more data points to investors (i.e., $\frac{\partial x^*}{\partial \Delta} > 0$).

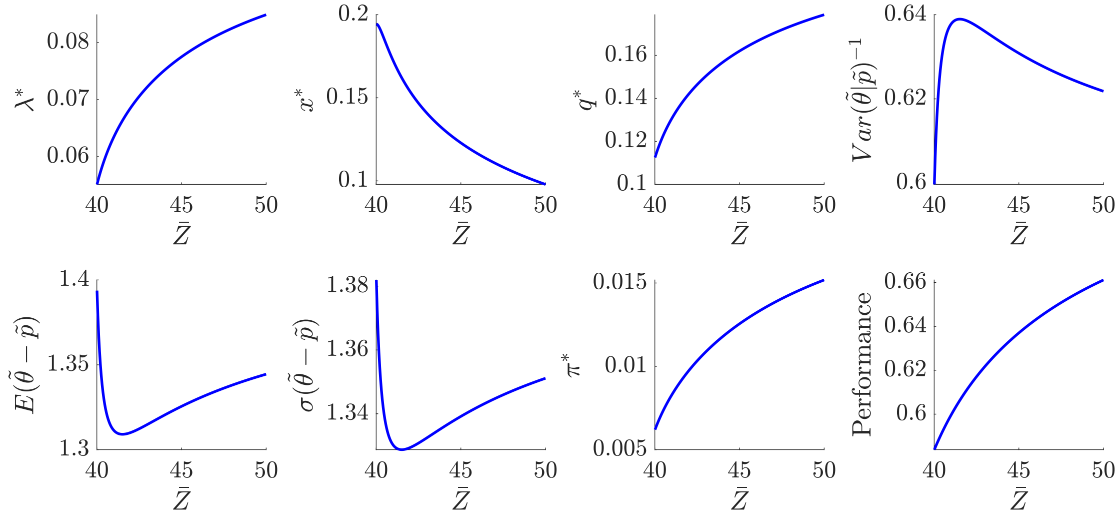
Similar to the case for skill-acquisition costs, the two competing forces from λ^* and x^* cause price informativeness to be hump-shaped in Δ, and the cost of capital and return volatility to be U-shaped in Δ. However, unlike the case for skill-acquisition costs, the data price q^* decreases with Δ because, as previously mentioned, investors are risk averse, and a higher skill volatility lowers their willingness to purchase data. In addition, because both data demand λ^* and data price q^* decrease in Δ, the vendor's profit π^* decreases in Δ. Finally, in Figure 3, the performance of skilled investors decreases in Δ, which is primarily driven by the deteriorating performance of low-skilled investors.

Proposition 2 (Skill volatility) *Suppose that the vendor owns sufficiently accurate data (i.e., τ_η is sufficiently high). When skill volatility Δ increases, the population λ^* of skilled investors decreases, the data sample size x^* increases, and the data price q^* decreases. Price informativeness $\frac{1}{Var(\tilde{\theta}|\tilde{p})}$ initially increases and finally decreases; the cost of capital $E(\tilde{\theta} - \tilde{p})$ and return volatility $\sigma(\tilde{\theta} - \tilde{p})$ initially decrease and finally increase. Moreover, when parameter set Ψ_Δ given by (A23) is nonempty, there exists $\hat{\Delta} \in [0, \bar{Z})$ such that when $\Delta > \hat{\Delta}$, skilled investors' performance decreases with skill volatility Δ.*

4.3 Skill Mean

The mean $\bar{Z}$ of skill levels can be interpreted as the average ability of the asset management industry. In Figure 4, we plot the implications of the skill mean. Here, we fix $c = 4.5$ and $\Delta = 40$. Other parameter values are the same as those in Figure 2.

As $\bar{Z}$ increases, skilled investors expect that, on average, they will extract more accurate information from purchased data and earn higher trading profits, which enhances their incentives to acquire skills and willingness to pay for data. As a result, both the population λ^* of skilled investors and the data price q^* increase in Figure 4 (i.e., $\frac{\partial \lambda^*}{\partial \bar{Z}} > 0$ and $\frac{\partial q^*}{\partial \bar{Z}} > 0$). The increased mass of skilled investors worsens the information leakage concern



This figure plots the effects of skill mean $\bar{Z}$. The skill-acquisition cost function is the same as in Figure 2 with $c = 4.5$. The parameters other than $\bar{Z}$ take the same values as those in Figure 2.

Figure 4: Effects of skill mean

of the data vendor; in response, the vendor optimally decreases the size of the data sample (i.e., $\frac{\partial x^*}{\partial \bar{Z}} < 0$). Again, the different responses of λ^* and x^* generate a hump-shaped pattern for price informativeness and a U-shape for the cost of capital and return volatility. Since both data demand λ^* and data price q^* increase with $\bar{Z}$, the vendor's profit π^* increases with $\bar{Z}$. Finally, in Figure 4, the performance of skilled investors increases with their average skill level $\bar{Z}$.

Proposition 3 (Skill mean) *Suppose that the vendor owns sufficiently accurate data (i.e., τ_η is sufficiently high). When the skill mean $\bar{Z}$ increases, the population λ^* of skilled investors increases, the data sample size x^* decreases, and the data price q^* increases. Price informativeness $\frac{1}{Var(\tilde{\theta}|\tilde{p})}$ initially increases and finally decreases; the cost of capital $E(\tilde{\theta} - \tilde{p})$ and return volatility $\sigma(\tilde{\theta} - \tilde{p})$ initially decrease and finally increase. Moreover, when parameter set $\Psi_{\bar{Z}}$ given by (A24) is nonempty, there exists $\hat{Z} \in (\Delta, \infty)$ such that when $\bar{Z} < \hat{Z}$, skilled investors' performance increases with skill mean $\bar{Z}$.*

5 Discussion, Extension, and Applications

5.1 Data Skill and Alternative Data

In this subsection, we first discuss the key features that make our model different from the existing literature. We then argue that these features are particularly relevant in the context of alternative data because, almost by definition, alternative data is at the early stage of the data cycle. These discussions place our analysis into context and, thus, help in interpreting our results.

Our analysis emphasizes two distinguishing features of trading on data. First, processing and trading on certain types of data require skills. In February 2018, Battlefin organized a data conference that brought together 107 asset managers, 94 data providers, and approximately 100 other industry professionals. One common theme throughout the conference was that access to certain data sources is no longer the main source of alpha; rather, it is the ability to process the data well and reach the best insights most rapidly. As stated by Barry Hurewitz, UBS's global head of Evidence Lab, "[p]eople say data is the new oil, but there is a refiner needed...And we're an integrated refiner of that data."[8] In our framework, we explicitly model this refining process and conceptually draw a difference between raw data (input to be refined) and information (refined output). Specifically, investor i uses signal $\tilde{f}_{j,i}$ in equation (2) to refine the "impurities" $\tilde{e}_j$ in raw data in equation (1), generating a better signal $\tilde{s}_{j,i}$ in equation (3); the effectiveness of investors' refining technology is captured by investor i's skill level z_i. Thus, our theoretical framework closely models the practical view of data and information and can be used in other data-related contexts.

Second, investors can spend resources to improve their data-processing skills, and skill development often involves considerable uncertainty. It is this feature in particular, the element of skill uncertainty, that drives our results which are new to the data-sales literature. As one recent article in New York Times described it, "[e]normous data sets and the computing power to analyze them cost a lot. So do the specialists who must be employed to clean the data, ensuring the information's validity and reliability and removing errant bits of it, and then interpret what's left. Fund providers need to spread the costs over a large asset base to

[8] "Wall Street analysts are now selling more data, less analysis," *Wall Street Journal*, November 7, 2018.

make it worthwhile."[9] In real markets, although many funds have invested substantially in data and hiring data scientists, the possibility of successfully processing the data and rapidly developing the best insights is low (De Prado, 2018); in particular, the success in decoding data as well as the performance of these quantitative funds can be uncertain in nature, i.e., unpredictable based on observable fund characteristics such as size. Therefore, investing in data-processing technology can be risky. In a broad sense, information extraction can be viewed as a production function, which combines data and skill as inputs to produce information as the output. As long as data and skill do not exhaust all of the relevant input factors, the production function cannot always be deterministic.

Although these two features do not apply to all types of data, they are particularly relevant in the context of alternative data. Alternative data has no clear definition but is usually referred to as any nontraditional data.[10] This negative definition of alternative data is elusive because it requires defining what traditional data is in the first place. Nonetheless, it suggests that alternative data constitutes a dynamic concept, and its landscape is evolving over time.[11] That is, every data has its cycle, and some data that is alternative from today's perspectives will become traditional in the future; similarly, any traditional data today might have been alternative in the past. Alternative data is at the early stage of the data cycle, and traditional data is at the late stage of the data cycle.

Because alternative data is new to investors, interpreting and processing many (although not all) types of alternative data require skills. The more difficult that it is to process data,

[9] "Will 'Big Data' Restore Active Managers' Mojo?" *New York Times*, April 16, 2020.

[10] Some definitions of alternative data are as follows: (1) "'Alternative data' refers to any data that are not 'traditional.' We use 'alternative' in a descriptive rather than normative sense and recognize there may not be an easily definable line between traditional and alternative data." ("Request for Information Regarding Use of Alternative Data and Modeling Techniques in the Credit Process," February 21, 2017, Consumer Financial Protection Bureau) (2) "Defining alternative data is not easy... In the simplest sense, alternative data can be regarded as everything that doesn't fall within the realm of traditional financial and economic data." ("Casting the net: How hedge funds are using alternative data," Alternative Investment Management Association) (3) "Alternative data refers to data used by investors to evaluate a company or investment that is not within their traditional data sources (financial statements, SEC filings, management presentations, press releases, etc.)." ("What is Alternative Data?" AlternativeData.org) (4) "Alternative data in finance refers to proxy metrics or information originating from unofficial or noncompany sources that individuals can use to gain insight into an investment. Such information can be the difference between making a profitable and unprofitable bet." ("Alternative Data – WSJ," *Wall Street Journal*, December 9, 2019)

[11] The report by Alternative Investment Management Association echoes this point: "As a concept, alternative data is not new; indeed, it is quite ancient..." https://www.aima.org/educate/aima-research/casting-the-net.html.

the more uncertain that the skill-acquisition process is. For instance, ranked based on the level of skills needed (from the highest to the lowest), alternative data, defined based on the current standard of financial markets, includes satellite images, public data such as weather data, social/sentiment data, and app usage. Analyzing these types of data requires skills, such as atmospheric physics, textual analysis, and natural language processing. The total spending on alternative data, including outlays for data sources, data science, IT infrastructure, data management and systems development, is estimated at \$7 billion in 2020; if we subtract \$1.7 billion in expenses for data alone, the remaining \$5.3 billion is the skill-acquisition cost for alternative data and is economically significant.[12] At the same time, success in developing profitable trading algorithms comes with considerable uncertainty.

5.2 An Extension with Multiple Rounds of Data Transactions and Asset Trading

In this subsection, we consider an extended economy in which skilled investors purchase and trade on data twice. Analyzing this extended economy helps to determine the role of skill uncertainty in driving our results. In addition, this extension parsimoniously illustrates the concept of the data cycle mentioned in the previous subsection.

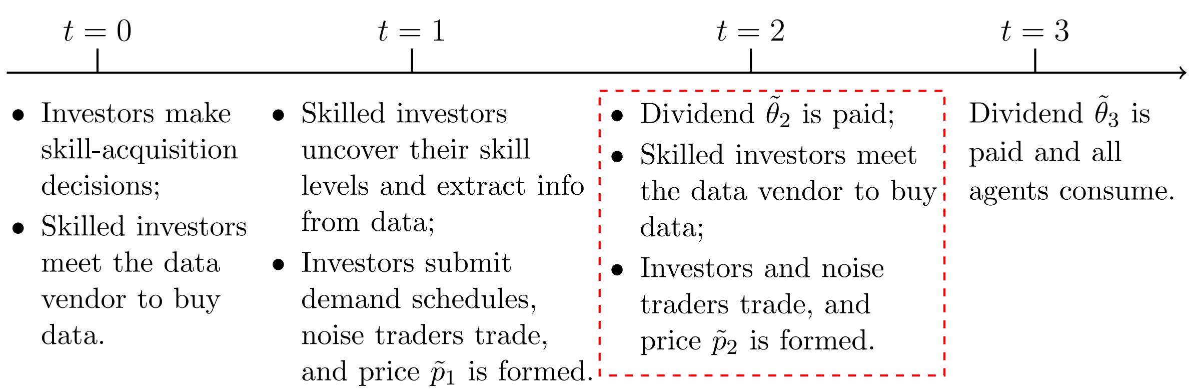


Figure 5: Timeline of the extended economy

[12]See "The explosion of 'alternative' data gives regular investors access to tools previously employed only by hedge funds," *MarketWatch*, September 19, 2019; "Will 'Big Data' Restore Active Managers' Mojo?" *New York Times*, April 16, 2020.

5.2.1 Setting

The extended economy has four dates, $t = 0, 1, 2$, and 3. Dates $0, 1$, and 3 correspond to the baseline economy in Section 2. The new events happen at date 2. Figure 5 describes the timeline of the extended economy and puts the new date in a box. Skilled investors buy data from the data vendor at dates 0 and 2. As in the baseline model, in the date-0 data transactions, neither the data vendor nor skilled investors themselves know investors' skill levels. In the date-2 data transactions, investors' skill levels become known to investors and the data vendor. Investors trade at dates 1 and 2 and consume at the final date 3.

To fix ideas, let us assume that the data is about dividends. Accordingly, we modify the risky asset's payoffs as two random dividends, $\tilde{\theta}_2$ and $\tilde{\theta}_3$, paid respectively at dates 2 and 3. As before, the dividends are independent and identically distributed (i.i.d.), $\tilde{\theta}_t \sim N\left(0, \tau_\theta^{-1}\right)$ with $\tau_\theta \in (0, \infty)$, for $t = 2, 3$. The vendor sells dividend data in the form of data points, $\tilde{\delta}_{t,j} \equiv \tilde{\theta}_t + \tilde{\eta}_t + \tilde{e}_{t,j}$, where $\tilde{\eta}_t \sim N\left(0, \tau_\eta^{-1}\right)$ with $\tau_\eta \in (0, \infty)$, $\tilde{e}_{t,j} \sim N(0, \infty)$, and $\left(\tilde{\theta}_2, \tilde{\theta}_3, \tilde{\eta}_2, \tilde{\eta}_3, \{\tilde{e}_{2,j}\}, \{\tilde{e}_{3,j}\}\right)$ are mutually independent. Parameter τ_η controls the vendor's data quality; we still focus on the case with sufficiently high τ_η. Similar to the baseline model, data processing is modeled as skilled investors reducing noise $\tilde{e}_{t,j}$ and translating data into a signal in the form of equations (4)–(6).

At date 0, dividend data is new (and "alternative") to the economy. Investors spend resources to develop interpreting and trading skills for alternative data about dividends $\tilde{\theta}_2$. As in the baseline model, this skill-acquisition investment is risky, generating either high skill Z_H or low skill Z_L with equal probability. At date 1, skilled investors discover their skill type and trade on dividend data. Given this date-1 trading experience, dividend data becomes old (and "traditional") to skilled investors, and their skills with old dividend data about $\tilde{\theta}_2$ can still be applied to the new dividend data about $\tilde{\theta}_3$.[13] When skilled investors purchase data on $\tilde{\theta}_3$, their skill levels become common knowledge.

Financial markets open at dates 1 and 2. Let $\tilde{p}_t$ denote the ex-dividend price of the risky

[13] As emphasized in the previous subsection, whether data is alternative or traditional is a relative concept. To complete the data cycle, we could add another date, $t = 2.5$. The events at $t = 2.5$ are similar to those at $t = 2$, except that all investors at $t = 2.5$ have a skill level Z_H for free. This captures that dividend data becomes completely standard at $t = 2.5$, such that most investors can interpret it well. The economy at this new date is the classical setting analyzed by Admati and Pfleiderer (1986); therefore, we omit it.

asset. There is noisy demand $\tilde{u}_t$ for the risky asset at the date-t market, where $\tilde{u}_t \sim N(0, \tau_u^{-1})$ with $\tau_u \in (0, \infty)$ and $\tilde{u}_t$ is independent of each other and of other random variables. There are a mass $\frac{\lambda}{2}$ of high-type skilled investors, a mass $\frac{\lambda}{2}$ of low-type skilled investors, and a mass $1 - \lambda$ of unskilled investors in markets for both dates. All investors submit demand schedules to maximize their expected utilities; they are forward looking when making trading decisions. In the subsequent analysis, we compute the equilibrium backward.

5.2.2 Equilibrium at Date 2

At date 2, since investors' skill levels are observable, the data vendor discriminates across the two types of skilled investors by choosing two different data sample sizes, denoted by $x_{H,2}$ and $x_{L,2}$. A high-type skilled investor i interprets the purchased data subset as a signal about dividend $\tilde{\theta}_3$, $\tilde{s}_{H,i,2} = \tilde{\theta}_3 + \tilde{\eta}_3 + \tilde{\varepsilon}_{H,i,2}$, with $\tilde{\varepsilon}_{H,i,2} \sim N\left(0, \frac{1}{x_{H,2}Z_H}\right)$. She chooses $D_{H,i,2}$ to maximize $E\left[-e^{-\gamma\left[D_{H,i,1}(\tilde{\theta}_2+\tilde{p}_2-\tilde{p}_1)+D_{H,i,2}(\tilde{\theta}_3-\tilde{p}_2)\right]}|\tilde{p}_2, \tilde{s}_{H,i,2}; \tilde{p}_1, \tilde{\theta}_2, \tilde{s}_{H,i,1}\right]$, where $D_{H,i,1}$ is her position established at the date-1 financial market and $\tilde{s}_{H,i,1}$ is the signal about dividend $\tilde{\theta}_2$ that she extracts from the data sample purchased at date 0. Because dividends $\tilde{\theta}_2$ and $\tilde{\theta}_3$ are mutually independent, we can compute $D_{H,2}\left(\tilde{p}_2, \tilde{s}_{H,i,2}\right) = \frac{E(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{H,i,2}) - \tilde{p}_2}{\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{H,i,2})}$. Similarly, we can compute the demand functions of low-type skilled investor i and of an unskilled investor as follows: $D_{L,2}\left(\tilde{p}_2, \tilde{s}_{L,i,2}\right) = \frac{E(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{L,i,2}) - \tilde{p}_2}{\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{L,i,2})}$ and $D_{U,2}\left(\tilde{p}_2\right) = \frac{E(\tilde{\theta}_3|\tilde{p}_2) - \tilde{p}_2}{\gamma Var(\tilde{\theta}_3|\tilde{p}_2)}$.

Following similar steps as the proof for Lemma 1, we can show that there exists a unique linear price function in the date-2 financial market,

$$\tilde{p}_2 = a_{0,2} + a_{\theta,2}(\tilde{\theta}_3 + \tilde{\eta}_3) + a_{u,2}\tilde{u}_2, \tag{25}$$

where $a_{0,2}$, $a_{\theta,2}$, and $a_{u,2}$ are given respectively by (A36), (A37), and (A38) in the Appendix. The equilibrium is characterized by the ratio $\alpha_2 = \frac{a_{\theta,2}}{a_{u,2}}$, which is the unique real positive root of the following equation:

$$2\gamma\alpha_2 - \lambda\tau_\eta\left(\frac{x_{H,2}Z_H}{\tau_\eta + \alpha_2^2\tau_u + x_{H,2}Z_H} + \frac{x_{L,2}Z_L}{\tau_\eta + \alpha_2^2\tau_u + x_{L,2}Z_L}\right) = 0. \tag{26}$$

Again, parameter α_2 is positively related to price informativeness $\frac{1}{Var(\tilde{\theta}_3|\tilde{p}_2)}$ of $\tilde{p}_2$.

Next, we solve for the data vendor's problem. The vendor charges different prices to different types of skilled investors, since it sells them data samples of different sizes. The data prices, $q_{H,2}\left(x_{H,2}, x_{L,2}; \lambda\right)$ and $q_{L,2}\left(x_{H,2}, x_{L,2}; \lambda\right)$ are still determined by Nash-bargaining equilibrium: $q_{H,2}\left(x_{H,2},x_{L,2};\lambda\right)=-\frac{\beta}{2\gamma}\ln\left[\frac{Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{H,i,2})}{Var(\tilde{\theta}_3|\tilde{p}_2)}\right]$ and $q_{L,2}\left(x_{H,2},x_{L,2};\lambda\right)=-\frac{\beta}{2\gamma}\ln\left[\frac{Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{L,i,2})}{Var(\tilde{\theta}_3|\tilde{p}_2)}\right]$. The vendor takes λ as given and chooses $\left(x_{H,2},x_{L,2}\right)$ to maximize its date-2 profit, $\pi_2\left(x_{H,2},x_{L,2};\lambda\right)=\frac{\lambda}{2}q_{H,2}\left(x_{H,2},x_{L,2};\lambda\right)+\frac{\lambda}{2}q_{L,2}\left(x_{H,2},x_{L,2};\lambda\right)$. We can show that the vendor chooses $\left(x_{H,2},x_{L,2}\right)$ such that $x_{H,2}Z_H=x_{L,2}Z_L$, which means that both types of skilled investors end up with the same precision from interpreting the purchased data packages. By doing so, the data vendor can control the information leakage effect most effectively because investors' willingness to pay is concave in precision. Specifically, the vendor sets $\left(x^*_{H,2},x^*_{L,2}\right)$ such that α^*_2 is fixed at a constant,

$$\alpha^*_2=\sqrt{\frac{\tau_\eta\tau_\theta}{\tau_u(\tau_\eta+\tau_\theta)}}. \tag{27}$$

As a result, the cost of capital and return volatility are also constant,

$$E(\tilde{\theta}_3-\tilde{p}_2)=\frac{\gamma\left(\tau_u\alpha^{*2}_2+\tau_\eta\right)}{\tau_\theta\left(\tau_u\alpha^{*2}_2+\tau_\eta\right)+\tau_\eta\alpha^*_2\left(\gamma+\tau_u\alpha^*_2\right)}, \tag{28}$$

$$\sigma(\tilde{\theta}_3-\tilde{p}_2)=\sqrt{\frac{\left(\tau_\eta+\alpha^{*2}_2\tau_u\right)\left(\tau_\theta\tau_u\left(\tau_\eta+\alpha^{*2}_2\tau_u\right)+\tau_\eta\left(\gamma+\alpha_2\tau_u\right)^2\right)}{\tau_u\left(\alpha^{*2}_2\tau_\theta\tau_u+\tau_\eta\left(\tau_\theta+\alpha^*_2\left(\gamma+\alpha^*_2\tau_u\right)\right)\right)^2}}. \tag{29}$$

Proposition 4 *Suppose the data vendor owns sufficiently accurate data in the extended economy. Then, at date 2:*

(a) The vendor sets the data sample sizes $x^*_{H,2}$ and $x^*_{L,2}$ such that $x^*_{H,2}Z_H=x^*_{L,2}Z_L$. Specifically, $x_{H,2}=\frac{\gamma\alpha^*_2(\tau_u\alpha^{*2}_2+\tau_\eta)}{Z_H(\lambda\tau_\eta-\gamma\alpha^*_2)}$ and $x_{L,2}=\frac{\gamma\alpha^*_2(\tau_u\alpha^{*2}_2+\tau_\eta)}{Z_L(\lambda\tau_\eta-\gamma\alpha^*_2)}$, where α^*_2 is given by (27).

(b) As the skill-acquisition cost parameter c, skill mean $\bar{Z}$, or skill volatility Δ changes, price informativeness, the cost of capital, and return volatility remain unchanged.

5.2.3 Equilibrium at Dates 0 and 1

Dates 0 and 1 resemble our baseline model in Section 2. In particular, in the date-0 data transactions, the vendor does not know skilled investors' types and, thus, cannot discriminate, which implies that it chooses one data sample size x_1. We first compute the date-1 financial market equilibrium and then go back to date 0 to characterize the data market equilibrium and skill-acquisition equilibrium. The derivations are very similar to the baseline model.

In the date-1 financial market, investors compute their demands for the risky asset, in anticipation of their future optimal demands in the date-2 financial market. For instance, a high-type skilled investor i interprets the purchased data as a signal about dividend $\tilde{\theta}_2$, $\tilde{s}_{H,i,1} = \tilde{\theta}_2 + \tilde{\eta}_2 + \tilde{\varepsilon}_{H,i,1}$, with $\tilde{\varepsilon}_{H,i,1} \sim N\left(0, \frac{1}{x_1 Z_H}\right)$. She has information set $\{\tilde{p}_1, \tilde{s}_{H,i,1}\}$ and $D_{H,i,1}$ to maximize $E\left[-e^{-\gamma\left[D_{H,i,1}(\tilde{\theta}_2+\tilde{p}_2-\tilde{p}_1)+D_{H,2}\left(\tilde{p}_2,\tilde{s}_{H,i,2}\right)(\tilde{\theta}_3-\tilde{p}_2)\right]}|\tilde{p}_1, \tilde{s}_{H,i,1}\right]$. We can compute her demand as $D_{H,1}\left(\tilde{p}_1, \tilde{s}_{H,i,1}\right) = \frac{E(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{H,i,1})+a_{0,2}-c_1-\tilde{p}_1}{\gamma\left[Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{H,i,1})+c_2\right]}$, where c_1 and c_2 are two constants given in the Appendix. Similarly, we can compute the demand functions of a low-type skilled investor and an unskilled investor as $D_{L,1}\left(\tilde{p}_1, \tilde{s}_{L,i,1}\right) = \frac{E(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{L,i,1})+a_{0,2}-c_1-\tilde{p}_1}{\gamma\left[Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{L,i,1})+c_2\right]}$ and $D_{U,1}\left(\tilde{p}_1\right) = \frac{E(\tilde{\theta}_2|\tilde{p}_1)+a_{0,2}-c_1-\tilde{p}_1}{\gamma\left[Var(\tilde{\theta}_2|\tilde{p}_1)+c_2\right]}$. Following similar steps as in the baseline model, we can use the demand functions and the market clearing condition to show that there exists a unique linear price function for the date-1 financial market,

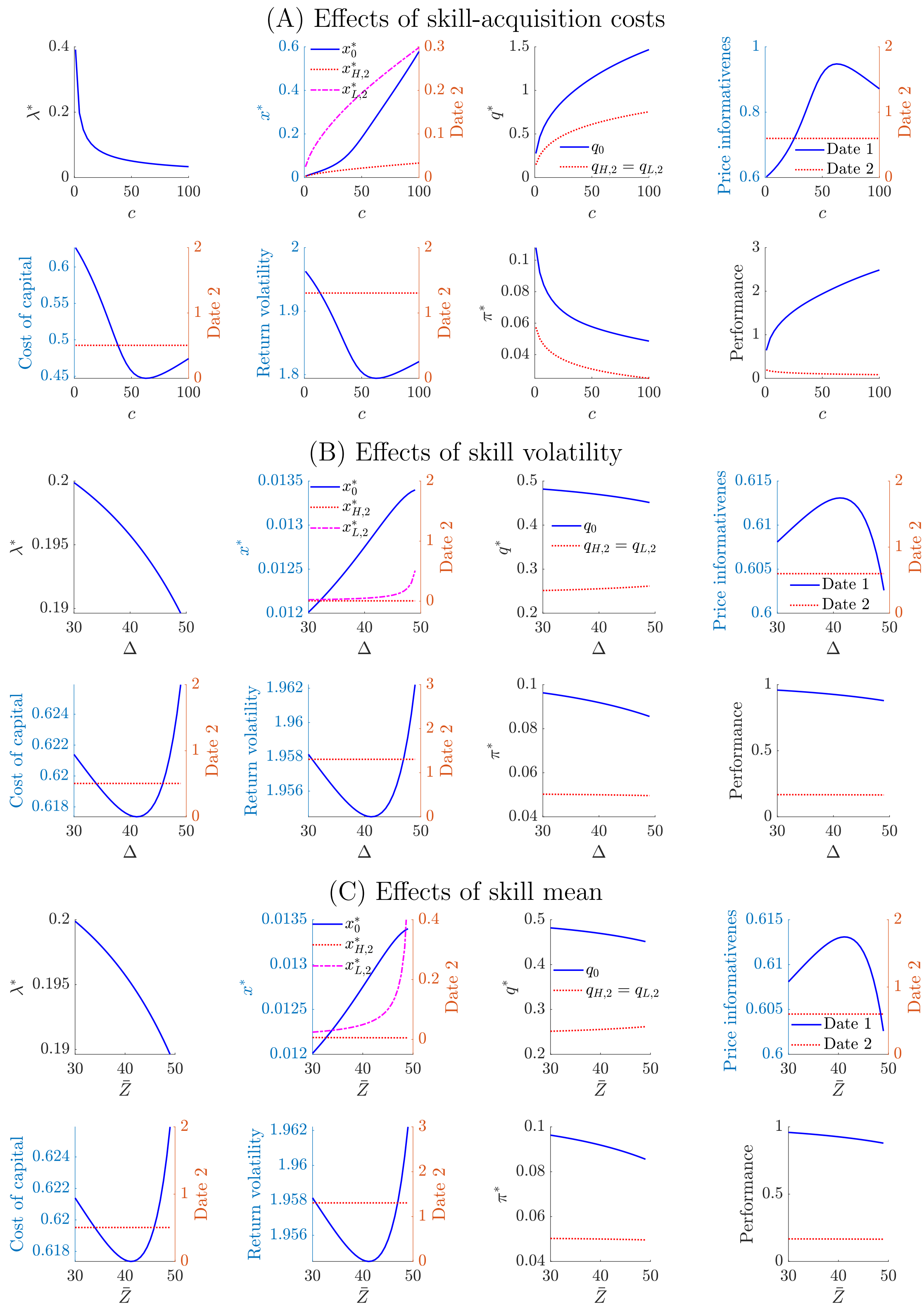
$$\tilde{p}_1 = a_{0,1} + a_{\theta,1}(\tilde{\theta}_2 + \tilde{\eta}_2) + a_{u,1}\tilde{u}_1, \tag{30}$$

where $a_{0,1}, a_{\theta,1}$, and $a_{u,1}$ are given respectively by (A47), (A48), and (A49) in the Appendix. The equilibrium is characterized by the ratio $\alpha_1 = \frac{a_{\theta,1}}{a_{u,1}}$, which is the unique real positive root of the following equation:

$$2\gamma\alpha_1 - \lambda\tau_\eta \left(\begin{array}{c} \frac{x_1 Z_H}{\tau_\eta+(\alpha_1^2\tau_u+x_1Z_H)[1+c_2(\tau_\theta+\tau_\eta)]+\tau_\theta\tau_\eta c_2} \\ +\frac{x_1 Z_L}{\tau_\eta+(\alpha_1^2\tau_u+x_1Z_L)[1+c_2(\tau_\theta+\tau_\eta)]+\tau_\theta\tau_\eta c_2} \end{array} \right) = 0, \tag{31}$$

where α_1 is positively related to the date-1 price informativeness, $\frac{1}{Var(\tilde{\theta}_2|\tilde{p}_1)}$. The cost of capital and return volatility in the date-1 market are given by $E(\tilde{\theta}_2 + \tilde{p}_2 - \tilde{p}_1)$ and $\sigma(\tilde{\theta}_2 + \tilde{p}_2 - \tilde{p}_1)$, respectively.

We now go back to date 0 to examine the vendor's problem and investors' skill-acquisition decisions. In the Appendix, we compute the trading gain of skilled investors, $G_1(x_1, \alpha_1, x_{H,2}, x_{L,2}, \alpha_2)$. The data price q_1^* is given by the Nash-bargaining equilibrium, $q_1 = \beta G_1$. The vendor's profit from selling data about $\tilde{\theta}_2$ is $\pi_1 = \lambda q_1$. The vendor chooses x_1 to maximize profits, taking as given the mass λ of skilled investors and the date-2 equilibrium. Because dividends are i.i.d. over time, the vendor's data sales problems are independent; therefore, it chooses x_1 to maximize π_1 (or equivalently q_1). Finally, the equilibrium mass of skilled investors λ^* is still de-



This figure plots the effects of skill-acquisition costs, skill volatility, and skill mean in the extended economy. The equilibrium outcomes for dates 0 and 1 are plotted in solid curves, while the equilibrium outcomes for date 2 are plotted in dashed curves. The skill-acquisition cost function and all other parameter values are the same as those for Figures 2-4 except $\gamma = 0.35$.

Figure 6: The extended economy

termined by the indifference condition, $(1-\beta)G_1(x_1^*, \alpha_1^*(x_1^*, \lambda^*), x_{H,2}^*, x_{L,2}^*, \alpha_1^*(x_{H,2}^*, x_{L,2}^*, \lambda^*))$ $= C(\lambda^*)$.

5.2.4 Comparative Statics

Given the complexity of the equilibrium at dates 0 and 1, we rely on numerical analysis to examine the effects of skill-acquisition costs, skill mean, and skill volatility. Figure 6 shows the results. The skill-acquisition cost function and all other parameter values are the same as those for Figures 2-4 except that $\gamma = 0.35$. The equilibrium outcomes for dates 0 and 1 are plotted in solid curves, while the equilibrium outcomes for date 2 are plotted in dashed curves. Comparing Figure 6 with Figures 2-4, we find that our results in the baseline model continue to hold at dates 0 and 1 in the extended economy, when the data is at the early stage of the data cycle and, thus, truly alternative. Consistent with Proposition 4, once the data enters its late stage of the data cycle and becomes more traditional, the economy is more similar to the classical settings of Admati and Pfleiderer (1986). This exercise highlights the crucial role of skill uncertainty in driving our results about alternative data.

The distinction between dates 1 and 2 is slightly striking in the extended economy because date 2 in our current extension corresponds to the end of the life of given alternative data, at which point investors know the data well, and their skills in decoding the data are high. Alternatively, we could consider a more realistic variation, in which new investors enter the market at date 2 and buy alternative data for the first time. Investors who acquired skills and purchased data at date 0 know their types, while the new investors remain uncertain about their skills at date 2. Skill uncertainty still (partially) prevails at date 2; thus, our results would remain valid but with an attenuated magnitude. This model variant would further strengthen the notion of alternative data becoming traditional over time.

5.3 Applications and Predictions

Data Industry and Funds Industry Our setting illustrates a generic message that the asset management industry and the data industry tend to foster one another. We can interpret skilled investors as active asset management firms; thus, the population λ^* of skilled investors can be a measure of the size of the active funds. The data seller's profit π^* can be a proxy for the size of the data industry. Examining Figures 2–4, we find that λ^* and π^* move in the same direction in response to changes in exogenous parameters, suggesting that the two industries tend to foster each other. This observation can shed light on the debate over how alternative data changes the landscape of active funds. A practical view is that the rapid growth of alternative data sets may shrink the active funds industry in the future because some asset management firms will prove unable to adapt to the radically changing landscape.[14] Our analysis suggest that this view is incomplete and that the active funds industry may actually grow with the development of alternative data.

Fund Performance For alternative data, it is often argued that "there's the concern the vendor will be selling that data set to a wide range of investors, which means its advantage could soon be arbitraged away."[15] This argument associates a larger mass λ^* of investors purchasing and trading on alternative data with a lower performance of these investors. In our analysis, whether this argument holds depends on the underlying reason driving the increase in λ^*. If the increase in λ^* is driven by a decrease in skill-acquisition costs, this argument is indeed valid (see Proposition 1). However, if the increase in λ^* is driven by a decrease in skill volatility or an increase in skill mean, then the opposite of this argument holds true (see Propositions 2 and 3). For instance, in recent decades, increasingly more talent has been drawn to the financial sector (e.g., Philippon and Reshef, 2012; Glode and Lowery, 2016), which corresponds to an increase in skill mean $\bar{Z}$ over time. This trend implies that the performance of those funds trading on alternative data, which, again, is a dynamic concept and may refer to different data sets in different times, has been increasing over the past decades.

[14] For instance, a recent Opimas article wrote the following: "The explosion of alternative data will require hedge funds and other asset managers to make large investments to acquire the necessary skills and infrastructure to leverage these sources of information. We expect that alternative data will contribute significantly to a further shrinkage in the hedge fund population, as firms unable to exploit the information needed to compete effectively in the new world of intelligent investing will fall behind." ("AI and alternative data: Moving to trading's next model," Axel Pierron, July 24, 2017, http://www.opimas.com/research/267/detail/)

[15] "The explosion of alternative data gives regular investors access to tools previously employed only by hedge funds," *MarketWatch*, September 19, 2019.

Price Informativeness and Asset Prices Our analysis predicts nonmonotonic price informativeness over the life cycle of a data set. For instance, for a given alternative data set, over time, the costs to crunch the data decrease, and the average of investors' data analysis skills increases. Both lead to an increasing number of skilled investors trading on alternative data, which in our setting is interpreted as an increase in the equilibrium mass λ^* of skilled investors. Figures 2 and 4 indicate hump-shaped price informativeness. In this sense, our theory predicts that, over the life cycle of certain alternative data, other things being equal, price informativeness initially increases and eventually decreases. Similarly, the cost of capital and return volatility can exhibit nonmonotonic patterns as well. At the same time, we caution that our theory is silent about the trend of price informativeness for very long time periods, covering multiple alternative data's life cycles.

In addition to the time-series prediction, we might also associate heterogeneous skill characteristics across different markets. For instance, a given data set, such as Tweets, can be used to trade in both bond and equity markets. Investor bases are different across these markets, with the former being dominated by institutional investors. Thus, the former market can arguably be associated with higher skill mean and possibly lower skill volatility. Our theory hence suggests that, through the data-sales channel, such different investor compositions can result in different levels of price informativeness.

6 Conclusion

In financial markets, data vendors sell data to sophisticated investors (such as hedge funds or active mutual funds) that, in turn, analyze and trade on the purchased data to generate trading profits. Extracting useful information from raw data requires skills, and developing such skills is costly and involves considerable uncertainty. This view is particularly relevant for alternative data, which is at the early stage of the data cycle and, thus, new to investors. In this paper, we study a data-sales model with endogenous data skill acquisition by investors. We draw a conceptual difference between data and information, and explicitly model how skilled investors extract information from raw data, and how the data vendor affects the precision of skilled investors' information by controlling the size of the data sample. We find that price informativeness is hump-shaped in skill-acquisition costs, while both the cost of capital and return volatility are U-shaped in skill-acquisition costs. Similar comparative statics patterns can arise for skill mean and skill volatility. Our analysis suggests that the funds industry and the data industry tend to foster each other, which speaks to the implications of alternative data for financial markets.

References

Admati, A. R. and P. Pfleiderer (1986). A monopolistic market for information. *Journal of Economic Theory 39*(2), 400–438.

Admati, A. R. and P. Pfleiderer (1988). A theory of intraday patterns: Volume and price variability. *Review of Financial Studies 1*(1), 3–40.

Admati, A. R. and P. Pfleiderer (1990). Direct and indirect sale of information. *Econometrica 58*(4), 901–928.

Allen, F. (1990). The market for information and the origin of financial intermediation. *Journal of Financial Intermediation 1*(1), 3–30.

Anand, B. N. and A. Galetovic (2000). Information, nonexcludability, and financial market structure. *Journal of Business 73*(3), 357–402.

Angeletos, G.-M. and A. Pavan (2009). Policy with dispersed information. *Journal of the European Economic Association 7*(1), 11–60.

Bai, J., T. Philippon, and A. Savov (2016). Have financial markets become more informative? *Journal of Financial Economics 122*(3), 625–654.

Begenau, J., M. Farboodi, and L. Veldkamp (2018). Big data in finance and the growth of large firms. *Journal of Monetary Economics 97*, 71–87.

Bimpikis, K., D. Crapis, and A. Tahbaz-Salehi (2019). Information sale and competition. *Management Science 65*(6), 2646–2664.

Bond, P., A. Edmans, and I. Goldstein (2012). The real effects of financial markets. *Annual Review of Financial Economics 4*(1), 339–360.

Breugem, M. and A. Buss (2019). Institutional investors and information acquisition: Implications for asset prices and informational efficiency. *Review of Financial Studies 32*(6), 2260–2301.

Buss, A. and S. Sundaresan (2020). More risk, more information: How passive ownership can improve informational efficiency. *Working Paper*, INSEAD, Imperial College London.

Cespa, G. (2008). Information sales and insider trading with long-lived information. *Journal of Finance 63*(2), 639–672.

Chen, Y., B. Kelly, and W. Wu (2020). Sophisticated investors and market efficiency: Evidence from a natural experiment. *Journal of Financial Economics*, Forthcoming.

Colombo, L., G. Femminis, and A. Pavan (2014). Information acquisition and welfare. *Review of Economic Studies 81*(4), 1438–1483.

De Prado, M. L. (2018). The 10 reasons most machine learning funds fail. *Journal of Portfolio Management 44*(6), 120–133.

Easley, D., M. O'Hara, and L. Yang (2016). Differential access to price information in financial markets. *Journal of Financial and Quantitative Analysis 51*(4), 1071–1110.

Engelberg, J. E., A. V. Reed, and M. C. Ringgenberg (2012). How are shorts informed?: Short sellers, news, and information processing. *Journal of Financial Economics 105*(2), 260–278.

Farboodi, M., A. Matray, and L. Veldkamp (2017). Where has all the big data gone? *Working Paper*, MIT, Princeton University, Columbia University, and New York University.

Farboodi, M., A. Matray, and L. Veldkamp (2020). Long-run growth of financial data technology. *American Economic Review 110*(8), 2485–2523.

Fishman, M. J. and K. M. Hagerty (1995). The incentive to sell financial market information. *Journal of Financial Intermediation 4*(2), 95–115.

Ganguli, J. V. and L. Yang (2009). Complementarities, multiplicity, and supply information. *Journal of the European Economic Association 7*(1), 90–115.

García, D. and F. Sangiorgi (2011). Information sales and strategic trading. *Review of Financial Studies 24*(9), 3069–3104.

García, D. and J. M. Vanden (2009). Information acquisition and mutual funds. *Journal of Economic Theory 144*(5), 1965–1995.

Gârleanu, N. and L. H. Pedersen (2018). Efficiently inefficient markets for assets and asset management. *Journal of Finance 73*(4), 1663–1712.

Glode, V. and R. Lowery (2016). Compensating financial experts. *Journal of Finance 71*(6), 2781–2808.

Glosten, L. R. and P. R. Milgrom (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics 14*(1), 71–100.

Goldstein, I. and L. Yang (2017). Information disclosure in financial markets. *Annual Review of Financial Economics 9*, 101–125.

Huang, A. G., H. Tan, and R. Wermers (2020, 01). Institutional Trading around Corporate News: Evidence from Textual Analysis. *Review of Financial Studies 33*(10), 4627–4675.

Huang, S., Z. Qiu, and L. Yang (2020). Institutionalization, delegation, and asset prices. *Journal of Economic Theory 186*, 104977.

Indjejikian, R. J. (1991). The impact of costly information interpretation on firm disclosure decisions. *Journal of Accounting Research 29*(2), 277–301.

Kacperczyk, M., S. V. Nieuwerburgh, and L. Veldkamp (2014). Time-varying fund manager skill. *Journal of Finance 69*(4), 1455–1484.

Kacperczyk, M. T., J. B. Nosal, and S. Sundaresan (2020). Market power and price informativeness. *Working Paper*, Imperial College London, Boston College.

Kacperczyk, M. T., S. Sundaresan, and T. Wang (2020). Do foreign institutional investors improve price efficiency? *Review of Financial Studies*, Forthcoming.

Kim, O. and R. E. Verrecchia (1994). Market liquidity and volume around earnings announcements. *Journal of Accounting and Economics 17*(1-2), 41–67.

Myatt, D. P. and C. Wallace (2012). Endogenous information acquisition in coordination games. *Review of Economic Studies 79*(1), 340–374.

Naik, N. Y. (1997). Multi-period information markets. *Journal of Economic Dynamics and Control 21*(7), 1229–1258.

Pagano, M. and P. Volpin (2012). Securitization, transparency, and liquidity. *Review of Financial Studies 25*(8), 2417–2453.

Philippon, T. and A. Reshef (2012). Wages and human capital in the us finance industry: 1909–2006. *Quarterly Journal of Economics 127*(4), 1551–1609.

Appendix: Proofs

Proof of Lemma 1

Inserting the conditional moments (11)–(16) into the demand functions for different types of investors, we can rearrange the market-clearing condition (17) and compute the implied price as follows:

$$\tilde{p} = \frac{A_p \tilde{s}_p + B_p(\tilde{\theta} + \tilde{\eta}) + \tilde{u} - 1}{\frac{\lambda}{2\gamma Var(\tilde{\theta}|\tilde{p},\tilde{s}_{H,i})} + \frac{\lambda}{2\gamma Var(\tilde{\theta}|\tilde{p},\tilde{s}_{L,i})} + \frac{1-\lambda}{\gamma Var(\tilde{\theta}|\tilde{p})}}, \tag{A1}$$

where

$$\begin{aligned}
A_p &= \frac{\lambda}{2\gamma Var(\tilde{\theta}|\tilde{p},\tilde{s}_{H,i})} \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{\alpha^2 \tau_u}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha^2 \tau_u + x Z_H} \\
&\quad + \frac{\lambda}{2\gamma Var(\tilde{\theta}|\tilde{p},\tilde{s}_{L,i})} \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{\alpha^2 \tau_u}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha^2 \tau_u + x Z_L} \\
&\quad + \frac{1-\lambda}{\gamma Var(\tilde{\theta}|\tilde{p})} \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{\alpha^2 \tau_u}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha^2 \tau_u}, \\
B_p &= \frac{\lambda}{2\gamma Var(\tilde{\theta}|\tilde{p},\tilde{s}_{H,i})} \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{x Z_H}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha^2 \tau_u + x Z_H} \\
&\quad + \frac{\lambda}{2\gamma Var(\tilde{\theta}|\tilde{p},\tilde{s}_{L,i})} \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{x Z_L}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha^2 \tau_u + x Z_L}.
\end{aligned}$$

From equation (A1), we directly have $\alpha = B_p$. Using the expression of B_p above, we can simplify equation $\alpha = B_p$ to equation (18) in Lemma 1. Next we show the unique existence of α in equilibrium. We use $H(\alpha)$ to denote the left-hand-side (LHS) of equation (18). It is clear that $H(\alpha)$ is negative when $\alpha = 0$, and $H(\alpha)$ is positive when α is large enough. Thus, by the intermediate value theorem, there exists a positive root to equation (18). In addition, $H(\alpha)$ is monotonically increasing in α and thus, there is a unique positive solution to equation (18).

After some algebra, we can compute the price coefficients as follows:

$$a_0 = -\frac{1}{\frac{\lambda}{2\gamma Var\left(\tilde{v}|\tilde{p},\tilde{s}_{H,i}\right)} + \frac{\lambda}{2\gamma Var\left(\tilde{v}|\tilde{p},\tilde{s}_{L,i}\right)} + \frac{1-\lambda}{\gamma Var(\tilde{v}|\tilde{p})}}, \tag{A2}$$

$$a_\theta = \frac{A_p + B_p}{\frac{\lambda}{2\gamma Var\left(\tilde{v}|\tilde{p},\tilde{s}_{H,i}\right)} + \frac{\lambda}{2\gamma Var\left(\tilde{v}|\tilde{p},\tilde{s}_{L,i}\right)} + \frac{1-\lambda}{\gamma Var(\tilde{v}|\tilde{p})}}, \tag{A3}$$

$$a_u = B_p^{-1} a_\theta. \tag{A4}$$

QED.

Proof of Corollary 1

Since λ, $\bar{Z}$, and Δ affect price informativeness only through varaible α (see equation (16)), it suffices to prove that α increases with λ and $\bar{Z}$ but decreases with Δ, where α is determined by equation (18). Replacing $Z_H = \bar{Z} + \Delta$ and $Z_L = \bar{Z} - \Delta$ in (18), we obtain the following equation:

$$2\gamma\alpha - x\lambda\tau_\eta \left(\frac{\bar{Z}+\Delta}{x(\bar{Z}+\Delta)+\tau_u\alpha^2+\tau_\eta} + \frac{\bar{Z}-\Delta}{x(\bar{Z}-\Delta)+\tau_u\alpha^2+\tau_\eta} \right) = 0. \tag{A5}$$

Note that the LHS of equation (A5) is increasing in α and decreasing in λ. Thus, when λ increases, α must increase to maintain equation (A5). So, we have $\frac{\partial \alpha}{\partial \lambda} > 0$, which implies that $\frac{\partial}{\partial \lambda}\frac{1}{Var(\tilde{\theta}|\tilde{p})} > 0$.

Taking the derivative of H (the LHS of equation (18)) with respect to $\bar{Z}$ and Δ respectively yields the following:

$$\frac{\partial H}{\partial \bar{Z}} = -\frac{2\lambda x\tau_\eta\left(\tau_\eta+\alpha^2\tau_u\right)\left(\left(\tau_\eta+\alpha^2\tau_u\right)\left(\tau_\eta+\alpha^2\tau_u+2x\bar{Z}\right)+x^2\left(\Delta^2+\bar{Z}^2\right)\right)}{\left(\tau_\eta+\alpha^2\tau_u+x(\bar{Z}-\Delta)\right)^2\left(\tau_\eta+\alpha^2\tau_u+x(\Delta+\bar{Z})\right)^2} < 0,$$

$$\frac{\partial H}{\partial \Delta} = \frac{4\Delta\lambda x^2\tau_\eta\left(\tau_\eta+\alpha^2\tau_u\right)\left(\tau_\eta+\alpha^2\tau_u+x\bar{Z}\right)}{\left(\tau_\eta+\alpha^2\tau_u+x(\bar{Z}-\Delta)\right)^2\left(\tau_\eta+\alpha^2\tau_u+x(\Delta+\bar{Z})\right)^2} > 0.$$

Thus, the LHS of equation (A5) is decreasing in $\bar{Z}$. When $\bar{Z}$ increases, α must increase to maintain equation (A5). So, we have $\frac{\partial \alpha}{\partial \bar{Z}} > 0$ and $\frac{\partial}{\partial \bar{Z}}\frac{1}{Var(\tilde{\theta}|\tilde{p})} > 0$ given $(x;\tau_\theta,\tau_\eta,\tau_u,\gamma)$. Similarly, we can show that $\frac{\partial \alpha}{\partial \Delta} < 0$ and $\frac{\partial}{\partial \Delta}\frac{1}{Var(\tilde{\theta}|\tilde{p})} < 0$. QED.

Proof of Equation (19)

Following Grossman and Stiglitz (1980), we can compute the ex-ante expected utilities of a high-type and a low-type skilled investor respectively as follows:

$$V_H = -\sqrt{\frac{Var(\tilde{\theta}|\tilde{p},\tilde{s}_{H,i})}{Var(\tilde{\theta}-\tilde{p})}}\exp\left(-\frac{\left[E(\tilde{\theta}-\tilde{p})\right]^2}{2Var(\tilde{\theta}-\tilde{p})}\right),$$

$$V_L = -\sqrt{\frac{Var(\tilde{\theta}|\tilde{p},\tilde{s}_{L,i})}{Var(\tilde{\theta}-\tilde{p})}}\exp\left(-\frac{\left[E(\tilde{\theta}-\tilde{p})\right]^2}{2Var(\tilde{\theta}-\tilde{p})}\right),$$

where we have ignored the skill-acquisition cost, since this cost is sunk at the stage of Nash bargaining as we have discussed in the text. Given that a skilled investor can be of a high type with probability $\frac{1}{2}$ and a low type with probability $\frac{1}{2}$, the ex-ante expected utility of a skilled investor is $V_S = \frac{1}{2}V_H + \frac{1}{2}V_L$. The certainty equivalent of a skilled investor is

$$\begin{aligned} CE_S &\equiv -\frac{1}{\gamma}\log\left(-V_S\right) \\ &= -\frac{1}{\gamma}\ln\left[\frac{\frac{1}{2}\sqrt{Var(\tilde{\theta}|\tilde{p},\tilde{s}_{H,i})}+\frac{1}{2}\sqrt{Var(\tilde{\theta}|\tilde{p},\tilde{s}_{L,i})}}{\sqrt{Var(\tilde{\theta}-\tilde{p})}}\right]+\frac{\left[E(\tilde{\theta}-\tilde{p})\right]^2}{2\gamma Var(\tilde{\theta}-\tilde{p})}. \end{aligned}$$

Similarly, we can compute the certainty equivalent of an unskilled investor as follows:

$$CE_U = -\frac{1}{\gamma}\ln\left[\frac{\sqrt{Var(\tilde{\theta}|\tilde{p})}}{\sqrt{Var(\tilde{\theta}-\tilde{p})}}\right]+\frac{\left[E(\tilde{\theta}-\tilde{p})\right]^2}{2\gamma Var(\tilde{\theta}-\tilde{p})}.$$

The expression of the trading gain G follows immediately. QED.

Proof of Lemma 2

The data vendor's optimization problem is to choose x to maximize the data price, since it takes λ as given. Given that the data price is proportional to the trading gain (see equation (20)), the data vendor should maximize the trading gain (19), which is rewritten as $G = -\frac{1}{\gamma}\ln(\frac{1}{2}\sqrt{f_H}+\frac{1}{2}\sqrt{f_L})$, where

$$f_H \equiv \frac{Var(\tilde{\theta}|\tilde{p},\tilde{s}_{H,i})}{Var(\tilde{\theta}|\tilde{p})} = \frac{\frac{1}{\tau_\theta+\tau_\eta}+\left(\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\right)^2\frac{1}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha^2\tau_u+xZ_H}}{\frac{1}{\tau_\theta+\tau_\eta}+\left(\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\right)^2\frac{1}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha^2\tau_u}}, \tag{A6}$$

$$f_L \equiv \frac{Var(\tilde{\theta}|\tilde{p},\tilde{s}_{L,i})}{Var(\tilde{\theta}|\tilde{p})} = \frac{\frac{1}{\tau_\theta+\tau_\eta}+\left(\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\right)^2\frac{1}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha^2\tau_u+xZ_L}}{\frac{1}{\tau_\theta+\tau_\eta}+\left(\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\right)^2\frac{1}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha^2\tau_u}}. \tag{A7}$$

The optimal data sample size x is thus determined by the following first-order condition:

$$\frac{\partial G}{\partial x} = -\frac{1}{2\gamma}\frac{\frac{1}{\sqrt{f_H}}\frac{\partial f_H}{\partial x}+\frac{1}{\sqrt{f_L}}\frac{\partial f_L}{\partial x}}{\sqrt{f_H}+\sqrt{f_L}} = 0.$$

We next discuss the two cases $\tau_\eta \to 0$ and $\tau_\eta \to \infty$ separately.

Case 1: $\tau_\eta \to 0$. We first show that when $\tau_\eta \to 0$, the optimal data sample size $x^* \to \infty$. It suffices to prove that when $\tau_\eta \to 0$, $\frac{\partial f_H}{\partial x} < 0$ and $\frac{\partial f_L}{\partial x} < 0$, and thus $\frac{\partial G}{\partial x} > 0$. Since the analysis of $\frac{\partial f_L}{\partial x}$ is similar to that of $\frac{\partial f_H}{\partial x}$, we hereafter only focus on $\frac{\partial f_H}{\partial x}$. Taking derivative of f_H with respect to x yields the following, which is always negative as $\tau_\eta \to 0$:

$$\frac{\partial f_H}{\partial x} = -\left(\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\right)^2\frac{1}{\left(\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha^2\tau_u+xZ_H\right)^2}\frac{1}{\left(\frac{\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha^2\tau_u\frac{1}{\tau_\theta+\tau_\eta}\right)^2}\tau_\eta^2\times f_{H1} < 0,$$

where

$$f_{H1} = \begin{bmatrix} \left(\frac{\tau_\theta}{\tau_\theta+\tau_\eta} + \frac{\alpha^2\tau_u}{\tau_\eta}\right)\left(\frac{1}{\tau_\theta+\tau_\eta} + \alpha^2\tau_u\frac{1}{(\tau_\theta+\tau_\eta)\tau_\eta}\right)Z_H \\ -2Z_H\tau_u\frac{\alpha}{\tau_\eta}\left(\frac{\tau_\eta}{\tau_\theta+\tau_\eta} + \alpha^2\tau_u\frac{1}{(\tau_\theta+\tau_\eta)}\right)\frac{x}{\tau_\eta}\frac{\partial\alpha}{\partial x} \\ -2Z_H\tau_u\frac{1}{\tau_\theta+\tau_\eta}\frac{\alpha}{\tau_\eta}\left(\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta} + \alpha^2\tau_u + xZ_H\right)\frac{x}{\tau_\eta}\frac{\partial\alpha}{\partial x} \end{bmatrix}$$

$$> \begin{bmatrix} \left(\frac{\tau_\theta}{\tau_\theta+\tau_\eta} + \frac{\alpha^2\tau_u}{\tau_\eta}\right)\left(\frac{1}{\tau_\theta+\tau_\eta} + \alpha^2\tau_u\frac{1}{(\tau_\theta+\tau_\eta)\tau_\eta}\right)Z_H \\ -2Z_H\tau_u\frac{\alpha}{\tau_\eta}\left(\frac{\tau_\eta}{\tau_\theta+\tau_\eta} + \alpha^2\tau_u\frac{1}{(\tau_\theta+\tau_\eta)}\right)\lambda(\tau_u\alpha^2+\tau_\eta)\cdot\kappa \\ -2Z_H\tau_u\frac{1}{\tau_\theta+\tau_\eta}\frac{\alpha}{\tau_\eta}\left(\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta} + \alpha^2\tau_u\right)\lambda(\tau_u\alpha^2+\tau_\eta)\cdot\kappa \\ -2Z_H\tau_u\frac{1}{\tau_\theta+\tau_\eta}\frac{\alpha}{\tau_\eta}\left(\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta} + \alpha^2\tau_u\right)\lambda(\tau_u\alpha^2+\tau_\eta)\frac{1}{2\gamma}\left(\frac{1}{Z_H}+\frac{1}{Z_L}\right) \end{bmatrix} \to \frac{Z_H}{\tau_\theta} > 0,$$

where κ is a positive constant. In the above expressions, the limit is obtained by imposing $\tau_\eta \to 0$ and the first inequality follows due to the following facts:

- Fact 1: $\frac{\alpha}{\tau_\eta}$ is limited and we always have $\frac{\alpha}{\tau_\eta} \leq \frac{\lambda}{\gamma}$. This follows equation (18) directly.
- Fact 2: $\lim_{\tau_\eta\to 0}\alpha = 0$. Since $\alpha \leq \frac{\lambda\tau_\eta}{\gamma}$ (see Fact 1), as $\tau_\eta \to 0$, it is clear that $\alpha \to 0$.
- Fact 3: $\lim_{\tau_\eta\to 0}\frac{\partial\alpha}{\partial x} = 0$ and $\lim_{\tau_\eta\to 0}\frac{\partial\alpha}{\partial x}\frac{1}{\tau_\eta} = 0$. This is obvious as we apply implicit function theorem to (18) and obtain the following:
$$\frac{\partial\alpha}{\partial x} = \tau_\eta\frac{\lambda(\tau_u\alpha^2+\tau_\eta)\left[\frac{Z_H}{(xZ_H+\tau_u\alpha^2+\tau_\eta)^2} + \frac{Z_L}{(xZ_L+\tau_u\alpha^2+\tau_\eta)^2}\right]}{2\gamma + \lambda\tau_\eta\frac{xZ_H}{(xZ_H+\tau_u\alpha^2+\tau_\eta)^2}2\alpha\tau_u + \lambda\tau_\eta\frac{xZ_L}{(xZ_L+\tau_u\alpha^2+\tau_\eta)^2}}.$$
- Fact 4: When $\tau_\eta \to 0$, $\frac{x}{\tau_\eta}\frac{\partial\alpha}{\partial x}$ and $\frac{x^2}{\tau_\eta}\frac{\partial\alpha}{\partial x}$ are finite. Specifically,
$$\frac{x}{\tau_\eta}\frac{\partial\alpha}{\partial x} < \lambda\left(\tau_u\alpha^2+\tau_\eta\right)\frac{1}{2\gamma}\begin{pmatrix} \frac{Z_H x}{(xZ_H+\tau_u\alpha^2+\tau_\eta)^2} \\ +\frac{Z_L x}{(xZ_L+\tau_u\alpha^2+\tau_\eta)^2} \end{pmatrix} < \lambda(\tau_u\alpha^2+\tau_\eta)\cdot\kappa,$$
$$\frac{x^2}{\tau_\eta}\frac{\partial\alpha}{\partial x} < \lambda(\tau_u\alpha^2+\tau_\eta)\frac{1}{2\gamma}\begin{pmatrix} \frac{Z_H x^2}{(xZ_H+\tau_u\alpha^2+\tau_\eta)^2} \\ +\frac{Z_L x^2}{(xZ_L+\tau_u\alpha^2+\tau_\eta)^2} \end{pmatrix} < \lambda(\tau_u\alpha^2+\tau_\eta)\frac{1}{2\gamma}\cdot\left(\frac{1}{Z_H}+\frac{1}{Z_L}\right).$$

Case 2: $\tau_\eta \to \infty$. We now show that when $\tau_\eta \to \infty$, the optimal data sample size x^* is finite. We first prove that in the case of $\tau_\eta \to \infty$, $\frac{\partial G}{\partial x} < 0$ when x is higher than a certain value. Again, we focus on $\frac{\partial f_H}{\partial x}$. Taking derivative of f_H with respect to x yields the following:

$$\frac{\partial f_H}{\partial x} = \left(\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\right)^2\frac{Z_H}{[\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta} + \alpha^2\tau_u + xZ_H]^2}\frac{1}{\frac{\tau_\eta}{\tau_\theta+\tau_\eta} + \alpha^2\tau_u\frac{1}{\tau_\theta+\tau_\eta}} \times f_{H2},$$

where

$$\begin{aligned} f_{H2} &= 2\tau_u x\alpha\frac{\partial\alpha}{\partial x}\left(1+\frac{1}{\tau_\theta+\tau_\eta}\right)-\left(\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha^2\tau_u\right) \\ &> x\cdot 2\alpha\tau_u\cdot\frac{\partial\alpha}{\partial x}-\left(\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha^2\tau_u\right) \\ &> \left(\frac{\lambda}{2\gamma}\right)^2(Z_H+Z_L)^2x^2\tau_u-1. \end{aligned}$$

The last inequality follows due to the following fact, which follows Lemma 1:

- Fact 5: $\lim_{\tau_\eta\to\infty}\alpha=\frac{\lambda}{2\gamma}(Z_H+Z_L)x$ and $\lim_{\tau_\eta\to\infty}\frac{\partial\alpha}{\partial x}=\frac{\lambda}{2\gamma}(Z_H+Z_L)$.

The analysis is similar for $\frac{\partial f_L}{\partial x}$. This suggests that when $x>\frac{2\gamma}{\lambda(Z_H+Z_L)\sqrt{\tau_u}}$, $\frac{\partial G}{\partial x}<0$. Now we can prove that x^* is finite by contradiction. Assume that $x^*=\infty$. Since when $x>\frac{2\gamma}{\lambda(Z_H+Z_L)\sqrt{\tau_u}}$, $\frac{\partial G}{\partial x}<0$, the data vendor can always increase its profit by decreasing x^* from ∞ to a finite value that is slightly higher than $\frac{2\gamma}{\lambda(Z_H+Z_L)\sqrt{\tau_u}}$. This contradicts the assumption that the optimal x^* is ∞. Therefore, in the equilibrium, x^* must be finite when $\tau_\eta\to\infty$. QED.

Proof of Lemma 3

The equilibrium mass of skilled investors λ^* is determined by equation (23). Thus, to prove the existence of an overall equilibrium, we need to show there exists a positive solution to (23). Define $F(\lambda)\equiv(1-\beta)\,G(\lambda)-C(\lambda)$. Next we will show that $F(0)>0$ and $F(1)<0$. Then, by the intermediate value theorem, there must exists $\lambda^*\in(0,1)$ such that $F(\lambda)=0$; that is, there must exist an equilibrium.

First, we prove that $F(0)>0$. When $\lambda=0$, $C(\lambda)=0$. Further, according to Lemma 2, when $\lambda=0$, $\alpha=0$. Thus, when $\lambda=0$, $f_H(0)=1-\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\frac{xZ_H}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+xZ_H}<1$, and similarly, $f_L(0)=1-\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\frac{xZ_L}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta+xZ_L}}<1$, where f_H and f_L are given by (A6) and (A7), respectively. Therefore, $F(0)>0$.

Second, we prove that $F(1)<0$. Given that $C(1)=+\infty$, the key to prove $F(1)<0$ is to show that $G(1)$ is finite. Since $f_H\geq 1-\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\frac{xZ_H}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+xZ_H}\geq\frac{\tau_\theta}{\tau_\theta+\tau_\eta}$ and similarly, $f_L\geq\frac{\tau_\theta}{\tau_\theta+\tau_\eta}$, we know that $G\leq-\frac{1}{2\gamma}\ln\left[\frac{\tau_\theta}{\tau_\theta+\tau_\eta}\right]$. This suggests that $G(1)$ is finite, and thus $F(1)<0$. QED.

Proof of Proposition 1

We prove our results in the limiting economy in which $\tau_\eta \to \infty$. To ease the exposition we define $w \equiv \tau_\eta^{-1}$, insert $Z_H = \bar{Z} + \Delta$ and $Z_L = \bar{Z} - \Delta$ into equations (18) and (23), and rewrite the two equations as follows:

$$2\gamma\alpha - x\lambda\left(\frac{\bar{Z}+\Delta}{1+wx\left(\bar{Z}+\Delta\right)+w\tau_u\alpha^2} + \frac{\bar{Z}-\Delta}{1+wx\left(\bar{Z}-\Delta\right)+w\tau_u\alpha^2}\right) = 0, \tag{A8}$$

$$\frac{1}{2}\exp\left(\frac{\gamma C(\lambda)}{1-\beta}\right)\left(\sqrt{f_H}+\sqrt{f_L}\right) = 1, \tag{A9}$$

where

$$f_H = 1 - \frac{x\left(\bar{Z}+\Delta\right)}{\left(1+w\tau_u\alpha^2\right)\left(\tau_\theta+\tau_u\alpha^2\left(1+w\tau_\theta\right)+x\left(1+\tau_\theta w\right)\left(\bar{Z}+\Delta\right)\right)},$$

$$f_L = 1 - \frac{x\left(\bar{Z}-\Delta\right)}{\left(1+w\tau_u\alpha^2\right)\left(\tau_\theta+\tau_u\alpha^2\left(1+w\tau_\theta\right)+x\left(1+\tau_\theta w\right)\left(\bar{Z}-\Delta\right)\right)}.$$

Note that f_H and f_L are equivalent to (A6) and (A7), respectively. Recall that for a given λ, the vendor's problem is to choose data sample size x to maximize profit $\pi(x;\lambda) = \lambda q(x;\lambda)$. Since the data price is proportional to the trading gain as shown by (20), the vendor effectively maximizes the trading gain (19): $G(x,\alpha) = -\frac{1}{\gamma}\ln\left(\frac{1}{2}\sqrt{f_H}+\frac{1}{2}\sqrt{f_L}\right)$. Further, the equilibrium mass of skilled investors λ^* is determined by the indifference condition (A9). We next examine the effects of skill-acquisition costs in this limiting economy.

1. The effect of c on λ^* We proceed in the following three steps. First, we work with equation (A8) and apply Taylor series to expand x around $w = 0$. In this way, we obtain x as approximated by a function of α for a given λ and parameter w: $x\left(\alpha,\lambda;w\right)$. Second, we take λ as given and insert $x\left(\alpha,\lambda;w\right)$ into (19), and then the data vendor's problem becomes to choose α to maximize the trading gain. We can characterize the optimal α, apply Taylor series to expand α around $w = 0$, and then obtain the approximate value of α for a given λ and parameter w: $\alpha\left(\lambda;w\right)$. Third, inserting $\alpha\left(\lambda;w\right)$ into skilled investors' indifference condition (A9), we can obtain an equation that determines the equilibrium λ^* and conduct the comparative statics of λ^* with respect to skill-acquisition costs. The detailed process is as follows.

First, applying the implicit function theorem to equation (A8) to obtain the derivative of x with respect to w, we use the first two terms of the Taylor series expansion of $x(\alpha,\lambda;w)$

centered at $w = 0$ to approximate x: $x(\alpha, \lambda; w) = x(\alpha, \lambda; 0) + \frac{\partial x}{\partial w}(\alpha, \lambda; 0) w + o(w)$, or equivalently,

$$x(\alpha, \lambda; w) = \frac{\alpha\gamma}{\lambda\bar{Z}} + \frac{\gamma\alpha^2\left(\gamma(\bar{Z}^2 + \Delta^2) + \bar{Z}^2\lambda\tau_u\alpha\right)}{\lambda^2\bar{Z}^3} w + o(w), \tag{A10}$$

where $o(w)$ denotes higher orders.

Second, we solve the data vendor's profit-maximization problem, which is equivalent to choosing α to maximize the trading gain (19) for a given λ. Maximizing the trading gain with respect to α yields

$$\frac{\partial G}{\partial \alpha} = -\frac{1}{2\gamma}\frac{\frac{1}{\sqrt{f_H}}\frac{\partial f_H}{\partial \alpha} + \frac{1}{\sqrt{f_L}}\frac{\partial f_L}{\partial \alpha}}{\sqrt{f_H} + \sqrt{f_L}}, \tag{A11}$$

where

$$\frac{1}{\sqrt{f_H}}\frac{\partial f_H}{\partial \alpha} = -\frac{(\bar{Z}+\Delta)\begin{pmatrix} \frac{\partial x}{\partial \alpha}\left[\left(1 + w\tau_u\alpha^2\right)\left(\tau_\theta + \tau_u\alpha^2(1 + w\tau_\theta) + x(1 + \tau_\theta w)(\bar{Z}+\Delta)\right)\right] \\ -x\begin{pmatrix} 2w\tau_u\alpha\left(\tau_\theta + \tau_u\alpha^2(1 + w\tau_\theta) + x(1 + \tau_\theta w)(\bar{Z}+\Delta)\right) \\ +\left(1 + w\tau_u\alpha^2\right)\left(2\tau_u(1 + w\tau_\theta)\alpha + \left(\bar{Z}+\Delta\right)(1 + \tau_\theta w)\frac{\partial x}{\partial \alpha}\right)\end{pmatrix}\end{pmatrix}}{2f_H\left[\left(1 + w\tau_u\alpha^2\right)\left(\tau_\theta + \tau_u\alpha^2(1 + w\tau_\theta) + x(1 + \tau_\theta w)(\bar{Z}+\Delta)\right)\right]^2},$$

$$\frac{1}{\sqrt{f_L}}\frac{\partial f_L}{\partial \alpha} = -\frac{(\bar{Z}-\Delta)\begin{pmatrix} \frac{\partial x}{\partial \alpha}\left[\left(1 + w\tau_u\alpha^2\right)\left(\tau_\theta + \tau_u\alpha^2(1 + w\tau_\theta) + x(1 + \tau_\theta w)(\bar{Z}-\Delta)\right)\right] \\ -x\begin{pmatrix} 2w\tau_u\alpha\left(\tau_\theta + \tau_u\alpha^2(1 + w\tau_\theta) + x(1 + \tau_\theta w)(\bar{Z}-\Delta)\right) \\ +\left(1 + w\tau_u\alpha^2\right)\left(2\tau_u(1 + w\tau_\theta)\alpha + \left(\bar{Z}-\Delta\right)(1 + \tau_\theta w)\frac{\partial x}{\partial \alpha}\right)\end{pmatrix}\end{pmatrix}}{2f_L\left[\left(1 + w\tau_u\alpha^2\right)\left(\tau_\theta + \tau_u\alpha^2(1 + w\tau_\theta) + x(1 + \tau_\theta w)(\bar{Z}-\Delta)\right)\right]^2},$$

and x and $\frac{\partial x}{\partial \alpha}$ are computed based on equation (A10). According to the first-order condition $\frac{\partial G}{\partial \alpha} = 0$, we can obtain the Taylor series of $\alpha(\lambda; w)$ around $w = 0$: $\alpha(\lambda; w) = \alpha(\lambda; 0) + \frac{\partial \alpha}{\partial w}(\lambda; 0) w + o(w)$, which can be simplified as:

$$\alpha(\lambda; w) = \sqrt{\frac{\tau_\theta}{\tau_u}} + \alpha_w w + o(w), \tag{A12}$$

where

$$\alpha_w = -\frac{\sqrt{\frac{\tau_\theta}{\tau_u}}\left(\frac{(\bar{Z}+\Delta)\left(\lambda\bar{Z}^2\tau_\theta + 2\gamma\Delta(\bar{Z}-\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}\right)}{\left(\gamma(\bar{Z}+\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}} + 2\lambda\bar{Z}\tau_\theta\right)^{3/2}} + \frac{(\bar{Z}-\Delta)\left(\lambda\bar{Z}^2\tau_\theta - 2\gamma\Delta(\bar{Z}+\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}\right)}{\left(\gamma(\bar{Z}-\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}} + 2\lambda\bar{Z}\tau_\theta\right)^{3/2}}\right)}{2\lambda\bar{Z}^2\left(\frac{\bar{Z}-\Delta}{\left(\gamma(\bar{Z}-\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}} + 2\lambda\bar{Z}\tau_\theta\right)^{3/2}} + \frac{\bar{Z}+\Delta}{\left(\gamma(\bar{Z}+\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}} + 2\lambda\bar{Z}\tau_\theta\right)^{3/2}}\right)}.$$

Third, we insert (A10) and (A12) into equation (A9). Imposing $w \to 0$, we obtain the following equation that uniquely characterizes the equilibrium λ^*:

$$\frac{1}{2}\exp\left(\frac{\gamma C(\lambda)}{1-\beta}\right)\left(\frac{1}{\sqrt{1+\frac{\gamma(\bar{Z}+\Delta)}{2\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}}}}+\frac{1}{\sqrt{1+\frac{\gamma(\bar{Z}-\Delta)}{2\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}}}}\right)=1. \tag{A13}$$

It is straightforward that the LHS of equation (A13) is increasing in both λ and c. Thus, as c increases, λ must decrease to maintain the equation, i.e., $\frac{\partial\lambda^*}{\partial c}<0$.

2. The effect of c on α^* and market variables Since c affect α only through λ, we first examine how λ affects α. Taking the derivative of $\alpha(\lambda;w)$ in (A12) with respect to λ yields the approximate value as follows (note that as $w\to 0$, the effect of λ on α through the higher order $o(w)$ can be neglected):

$$\frac{\partial\alpha}{\partial\lambda}\approx\frac{\partial\alpha_w}{\partial\lambda}w=-\frac{2\gamma\Delta^2\left(\bar{Z}^2-\Delta^2\right)\tau_\theta^2}{\bar{Z}^2\lambda^2\tau_u^2}w$$
$$\times\frac{\left(\begin{array}{c}-8\lambda^3\bar{Z}^3\tau_\theta^2\tau_u+4\lambda^2\bar{Z}^2\tau_u\tau_\theta\sqrt{\gamma(\bar{Z}-\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta}\sqrt{\gamma(\bar{Z}+\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta}\\ \left(\gamma^2\left(\bar{Z}^2-\Delta^2\right)+10\gamma\lambda\bar{Z}^2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)\sqrt{\gamma(\bar{Z}-\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta}\sqrt{\gamma(\bar{Z}+\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta}\\ +\left(3\lambda\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)2\gamma^2\bar{Z}\left(\bar{Z}^2-\Delta^2\right)\end{array}\right)}{\left((\bar{Z}+\Delta)\left(\gamma(\bar{Z}-\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta\right)^{3/2}+(\bar{Z}-\Delta)\left(\gamma(\bar{Z}+\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta\right)^{3/2}\right)^2}<0,$$

where the inequality follows because

$$\begin{aligned}&-8\lambda^3\bar{Z}^3\tau_\theta^2\tau_u+4\lambda^2\bar{Z}^2\tau_u\tau_\theta\sqrt{\gamma(\bar{Z}-\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta}\sqrt{\gamma(\bar{Z}+\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta}\\ > \;&-8\lambda^3\bar{Z}^3\tau_\theta^2\tau_u+4\lambda^2\bar{Z}^2\tau_u\tau_\theta\times 2\lambda Z\tau_\theta=0.\end{aligned}$$

Together with the result that $\frac{\partial\lambda^*}{\partial c}<0$, we can get that as $w\to 0$, $\frac{\partial\alpha^*}{\partial c}=\frac{\partial\alpha^*}{\partial\lambda}\frac{\partial\lambda}{\partial c}>0$. According to equation (16), skill-acquisition costs affect price informativeness $\frac{1}{Var(\tilde{\theta}|\tilde{p})}$ only through α, and the price informativeness is increasing with α. Thus, as $w\to 0$, $\frac{\partial}{\partial c}\left(\frac{1}{Var(\tilde{\theta}|\tilde{p})}\right)=\frac{\partial}{\partial\alpha}\left(\frac{1}{Var(\tilde{\theta}|\tilde{p})}\right)\frac{\partial\alpha}{\partial c}>0$.

Next, the cost of capital is $E(\tilde{\theta}-\tilde{p})=-a_0$, where a_0 is given by (A2). Inserting (A10) and (A12) into (A2), we approximate $E(\tilde{\theta}-\tilde{p})$ using the first two terms of the Taylor series expansion as follows:

$$E(\tilde{\theta}-\tilde{p})=\frac{\gamma}{2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}}+\frac{\gamma\left(-\alpha_w\left(\gamma+2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)+\tau_\theta^2+\gamma\tau_\theta\sqrt{\frac{\tau_\theta}{\tau_u}}\right)}{\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^2}w+o\left(w\right). \qquad \text{(A14)}$$

Taking the derivative of $E(\tilde{\theta}-\tilde{p})$ with respect to λ (note that α_w is a function of λ) yields $\frac{\partial E(\tilde{\theta}-\tilde{p})}{\partial\lambda}\approx-\frac{\gamma\left(\gamma+2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)}{\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^2}\frac{\partial\alpha_w}{\partial\lambda}w>0$, where the inequality follows because $\frac{\partial\alpha_w}{\partial\lambda}<0$. Thus, as $w\to0$, $\frac{\partial E(\tilde{\theta}-\tilde{p})}{\partial c}=\frac{\partial E(\tilde{\theta}-\tilde{p})}{\partial\lambda}\frac{\partial\lambda}{\partial c}<0$.

By equation (8), return volatility is $\sigma(\tilde{\theta}-\tilde{p})=\sqrt{\frac{(1-a_\theta)^2}{\tau_\theta}+\frac{a_\theta^2}{\tau_\eta}+\frac{a_u^2}{\tau_u}}$, where a_θ and a_u are specified by equations (A3) and (A4), respectively. Similarly, with (A10) and (A12), we can approximate $\sigma^2(\tilde{\theta}-\tilde{p})$ using the first two terms of the Taylor series expansion:

$$\begin{aligned}\sigma^2(\tilde{\theta}-\tilde{p}) &= \frac{\gamma^2+2\tau_u\left(\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)}{\tau_u\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^2}\\ &\quad +w\frac{\begin{pmatrix}\tau_\theta\left(\gamma\tau_\theta\left(\gamma^2+3\tau_\theta\tau_u\right)-2\alpha_w\tau_u\left(3\gamma^2+2\tau_\theta\tau_u\right)\right)\\ +2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\left(\tau_\theta^2\left(\gamma^2+\tau_\theta\tau_u\right)-\alpha_w\left(\gamma^3+3\gamma\tau_\theta\tau_u\right)\right)\end{pmatrix}}{\tau_u^2\sqrt{\frac{\tau_\theta}{\tau_u}}\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^3}+o\left(w\right). \qquad \text{(A15)}\end{aligned}$$

Taking the derivative of $\sigma^2(\tilde{\theta}-\tilde{p})$ with respect to λ (noting that α_w is a function of λ) yields $\frac{\partial\sigma^2(\tilde{\theta}-\tilde{p})}{\partial\lambda}\approx-\frac{2\sqrt{\frac{\tau_\theta}{\tau_u}}\left(\gamma^3\sqrt{\frac{\tau_\theta}{\tau_u}}+3\gamma\tau_\theta\left(\gamma+\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)+2\tau_\theta^2\tau_u\right)}{\tau_\theta\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^3}\frac{\partial\alpha_w}{\partial\lambda}w>0$, where the inequality follows because $\frac{\partial\alpha_w}{\partial\lambda}<0$. Thus, as $w\to0$, $\frac{\partial\sigma(\tilde{\theta}-\tilde{p})}{\partial c}=\frac{\partial\sigma(\tilde{\theta}-\tilde{p})}{\partial\lambda}\frac{\partial\lambda}{\partial c}<0$.

3. The effect of c on x^* and q^* With (A12), we can approximate x in (A10) as follows:

$$x=\frac{\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}}{\lambda\bar{Z}}+\frac{\gamma\left(\lambda\bar{Z}^2\tau_u\left(\alpha_w+\tau_\theta\sqrt{\frac{\tau_\theta}{\tau_u}}\right)+\gamma\tau_\theta\left(\bar{Z}^2+\Delta^2\right)\right)}{\lambda^2\bar{Z}^3\tau_u}w+o\left(w\right). \qquad \text{(A16)}$$

As $w\to0$, $\frac{\partial x^*}{\partial\lambda}\approx-\frac{\gamma}{\lambda^2\bar{Z}}\sqrt{\frac{\tau_\theta}{\tau_u}}<0$. Thus, $\frac{\partial x^*}{\partial c}=\frac{\partial x^*}{\partial\lambda}\frac{\partial\lambda}{\partial c}>0$.

Further, with (A12), we can approximate q in (20) as follows:

$$\begin{aligned}q &= -\frac{\beta}{\gamma}\ln\left(\frac{1}{2\sqrt{1+\frac{\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\left(\bar{Z}-\Delta\right)}{\bar{Z}\lambda}}}+\frac{1}{2\sqrt{1+\frac{\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\left(\bar{Z}+\Delta\right)}{\bar{Z}\lambda}}}\right)\\ &\quad -\frac{\beta\tau_\theta^2\left(\frac{(\bar{Z}+\Delta)\left(\lambda\bar{Z}^2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}+2\gamma\Delta(\bar{Z}-\Delta)\right)}{\left(\gamma(\bar{Z}+\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta\right)^{3/2}}+\frac{(\bar{Z}-\Delta)\left(\lambda\bar{Z}^2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}-2\gamma\Delta(\bar{Z}+\Delta)\right)}{\left(\gamma(\bar{Z}-\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta\right)^{3/2}}\right)}{4\bar{Z}\tau_u\left(\frac{1}{\sqrt{\gamma(\bar{Z}+\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta}}+\frac{1}{\sqrt{\gamma(\bar{Z}-\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta}}\right)}w+o\left(w\right) \qquad \text{(A17)}\end{aligned}$$

It is straightforward to show that as $w \to 0$, $\frac{\partial q^*}{\partial \lambda} < 0$. Thus, $\frac{\partial q^*}{\partial c} = \frac{\partial q^*}{\partial \lambda}\frac{\partial \lambda}{\partial c} > 0$.

4. The effect of c on skilled investors' performance When $w \to 0$, equation (24) can be simplified to the following:

$$\begin{aligned} Performance =& \frac{1}{4\gamma}\ln\left[\left(1+\frac{\gamma\left(\bar{Z}-\Delta\right)}{2\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}}\right)\left(1+\frac{\gamma\left(\bar{Z}+\Delta\right)}{2\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}}\right)\right] \\ &-\frac{\Delta\left(\gamma^2\left(\bar{Z}^2-\Delta^2\right)\sqrt{\frac{\tau_\theta}{\tau_u}}+\lambda\tau_\theta\left(\lambda\bar{Z}^2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}+2\gamma\left(\bar{Z}^2-\Delta^2\right)\right)\right)}{4\lambda\bar{Z}\left(4\lambda\bar{Z}^2\tau_u\left(\lambda\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)+\gamma^2\left(\bar{Z}^2-\Delta^2\right)\right)}w+o\left(w\right). \quad \text{(A18)} \end{aligned}$$

It is straightforward to show that as $w \to 0$, $\frac{\partial Performance}{\partial \lambda} < 0$. Thus, $\frac{\partial Performance}{\partial c} = \frac{\partial Performance}{\partial \lambda}\frac{\partial \lambda}{\partial c} > 0$. QED.

Proof of Proposition 2

Similar to the proof of Proposition 1, we examine the economy in which $\tau_\eta \to \infty$, or equivalently, $w \to 0$.

1. The effect of Δ on λ^*, x^*, and q^* Following the same steps as in the proof of Proposition 1, when $w \to 0$, we obtain the approximate values of x and α as given by (A10) and (A12) respectively, and the equilibrium λ^* is uniquely determined by equation (A13). Define the LHS of equation (A13) as K. We can show

$$\frac{\partial K}{\partial \Delta}=\frac{\gamma\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}e^{\frac{\gamma C}{1-\beta}}}{2}\left(\begin{array}{c}\frac{1}{\left(\gamma\left(\bar{Z}-\Delta\right)+2\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}\right)^{3/2}\sqrt{\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}}} \\ -\frac{1}{\left(\gamma\left(\bar{Z}+\Delta\right)+2\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}\right)^{3/2}\sqrt{\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}}}\end{array}\right) > 0.$$

Together with $\frac{\partial K}{\partial \lambda} > 0$, we immediately have $\frac{\partial \lambda^*}{\partial \Delta} < 0$ as $w \to 0$. Moreover, we also approximate x^* as in (A16). As $w \to 0$, $\frac{\partial x^*}{\partial \Delta} \approx -\frac{\gamma}{\lambda^2\bar{Z}}\sqrt{\frac{\tau_\theta}{\tau_u}}\frac{\partial \lambda}{\partial \Delta} > 0$, where the inequality follows because $\frac{\partial \lambda}{\partial \Delta} < 0$.

Moreover, based on equations (20) and (23), $q = \frac{\beta}{1-\beta}C\left(\lambda\right)$, which suggests that $\frac{\partial q^*}{\partial \lambda} > 0$ since $C(\cdot)$ is an increasing function. Combined with $\frac{\partial \lambda^*}{\partial \Delta} < 0$, $\frac{\partial q^*}{\partial \Delta} = \frac{\partial q}{\partial \lambda}\frac{\partial \lambda}{\partial \Delta} < 0$ immediately follows.

2. The effect of Δ on α^* and market variables Taking the derivative of α in (A12) with respect to Δ (noting λ is a function of Δ), we obtain the following:

$$\frac{\partial\alpha}{\partial\Delta} \approx \frac{\partial\alpha_w}{\partial\Delta}w = -\frac{2\gamma\Delta\tau_\theta^2}{\bar{Z}^2\lambda^2\tau_u^2}w$$
$$\times \frac{\left(\begin{array}{c} \sqrt{\gamma(\bar{Z}-\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta}\sqrt{\gamma(\bar{Z}+\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta} \\ \times\left(\begin{array}{c} 2\lambda\bar{Z}^2\tau_u\left(\begin{array}{c} \gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\left(7\Delta^2\lambda+5\Delta\left(\bar{Z}^2-\Delta^2\right)\lambda'(\Delta)\right) \\ +2\Delta\lambda\tau_\theta\left(2\Delta\lambda+\left(\bar{Z}^2-\Delta^2\right)\lambda'(\Delta)\right) \end{array}\right) \\ +\gamma^2\left(\bar{Z}^2-\Delta^2\right)\left(2\Delta^2\lambda+\Delta\left(\bar{Z}^2-\Delta^2\right)\lambda'(\Delta)-3\lambda\bar{Z}^2\right) \end{array}\right) \\ +\bar{Z}\left(\begin{array}{c} \gamma^3\left(\bar{Z}^2-\Delta^2\right)^2\left(2\Delta\lambda'(\Delta)-3\lambda\right)\sqrt{\frac{\tau_\theta}{\tau_u}} \\ -8\Delta\lambda^3\bar{Z}^2\tau_\theta^2\tau_u\left(2\Delta\lambda+\left(\bar{Z}^2-\Delta^2\right)\lambda'(\Delta)\right) \\ -6\gamma\lambda\tau_\theta\left(\bar{Z}^2-\Delta^2\right)\left(2\lambda^2\bar{Z}^2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}+\gamma\left(\bar{Z}^2-\Delta^2\right)\left(2\lambda-\Delta\lambda'(\Delta)\right)\right) \end{array}\right) \end{array}\right)}{\left((\bar{Z}+\Delta)\left(\gamma(\bar{Z}-\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta\right)^{3/2}+(\bar{Z}-\Delta)\left(\gamma(\bar{Z}+\Delta)\sqrt{\frac{\tau_\theta}{\tau_u}}+2\lambda\bar{Z}\tau_\theta\right)^{3/2}\right)^2}. \quad \text{(A19)}$$

We next discuss two extreme cases to show the non-monotonic effect of Δ on α.

Case 1: $\Delta \to 0$. When $\Delta \to 0$, equation (A19) can be approximated by the following (note that according to equation (A13), $\lambda'(\Delta)$ is finite, so $\lim_{\Delta\to 0}\Delta\lambda'(\Delta) = 0$):

$$\frac{\partial\alpha}{\partial\Delta} \approx \frac{2\gamma\Delta\tau_\theta^2}{\bar{Z}^2\lambda\tau_u^2}\frac{3\gamma\left(\gamma^2\sqrt{\frac{\tau_\theta}{\tau_u}}+4\lambda\tau_\theta\left(\gamma+\lambda\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)\right)}{2\left(2\lambda\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^3}w > 0. \quad \text{(A20)}$$

Thus, as $\Delta \to 0$, $\frac{\partial\alpha}{\partial\Delta} > 0$, and $\frac{\partial}{\partial\Delta}\left(\frac{1}{Var(\tilde{\theta}|\tilde{p})}\right) = \frac{\partial}{\partial\alpha}\left(\frac{1}{Var(\tilde{\theta}|\tilde{p})}\right)\frac{\partial\alpha}{\partial\Delta} > 0$ immediately follows.

Similarly, we can can approximate the cost of capital $E(\tilde{\theta}-\tilde{p})$ and return variance $\sigma^2(\tilde{\theta}-\tilde{p})$ respectively as in (A14) and (A15). Taking the derivative of $E(\tilde{\theta}-\tilde{p})$ with respect to Δ yields $\frac{\partial E(\tilde{\theta}-\tilde{p})}{\partial\Delta} \approx -\frac{\gamma\left(\gamma+2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)}{\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^2}\frac{\partial\alpha_w}{\partial\Delta}w < 0$, where the inequality follows because $\frac{\partial\alpha_w}{\partial\Delta} > 0$ when $\Delta \to 0$ (see (A20)). Similarly, $\frac{\partial\sigma^2(\tilde{\theta}-\tilde{p})}{\partial\Delta} \approx -\frac{2\sqrt{\frac{\tau_\theta}{\tau_u}}\left(\gamma^3\sqrt{\frac{\tau_\theta}{\tau_u}}+3\gamma\tau_\theta\left(\gamma+\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)+2\tau_\theta^2\tau_u\right)}{\tau_\theta\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^3}\frac{\partial\alpha_w}{\partial\Delta}w < 0$.

Case 2: $\Delta \to \bar{Z}$. When $\Delta \to \bar{Z}$, (A19) can be approximated by the following (note that according to equation (A13), $\lambda'(\Delta)$ is finite, so $\lim_{\Delta\to\bar{Z}}\left(\bar{Z}-\Delta\right)\lambda'(\Delta) = 0$):

$$\frac{\partial\alpha}{\partial\Delta} \approx -2\gamma\Delta\tau_\theta^2 w\frac{\left(\lambda\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)\sqrt{\lambda\tau_\theta\left(\lambda\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)}-\lambda^2\tau_\theta^2}{2\lambda^3\bar{Z}^2\tau_\theta^3\tau_u} < 0.$$

Thus, as $\Delta \to \bar{Z}$, $\frac{\partial \alpha}{\partial \Delta} < 0$. And $\frac{\partial}{\partial \Delta}\left(\frac{1}{Var(\tilde{\theta}|\tilde{p})}\right) = \frac{\partial}{\partial \alpha}\left(\frac{1}{Var(\tilde{\theta}|\tilde{p})}\right)\frac{\partial \alpha}{\partial \Delta} < 0$ immediately follows. Similar to the proof in case 1, we can show that as $\Delta \to \bar{Z}$, $\frac{\partial E(\tilde{\theta}-\tilde{p})}{\partial \Delta} > 0$ and $\frac{\partial \sigma(\tilde{\theta}-\tilde{p})}{\partial \Delta} > 0$.

3. The effect of Δ on skilled investors' performance Similar to the proof of Proposition 1, we can approximate skilled investors' performance in (A18). For the proof of $\frac{\partial Performance}{\partial \Delta} < 0$, we rely on the indifference condition for the marginal skilled investor (A9), which can be simplified to (A13) when $w \to 0$. Taking the square of both sides of (A13) and rearranging yields $\sqrt{\left(1+\frac{\gamma\left(\bar{Z}-\Delta\right)}{2\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}}\right)\left(1+\frac{\gamma\left(\bar{Z}+\Delta\right)}{2\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}}\right)} = \frac{1}{4}\Omega_1$, where

$$\Omega_1 = e^{\frac{\gamma C(\lambda)}{1-\beta}}\left(e^{\frac{\gamma C(\lambda)}{1-\beta}} + \sqrt{e^{\frac{2\gamma C(\lambda)}{1-\beta}} + 8\left(1+\frac{\gamma}{2\lambda\sqrt{\tau_\theta\tau_u}}\right)}\right). \tag{A21}$$

Note that Ω_1 is not directly affected by Δ. Therefore, compared to the approximate performance in (A18), to examine the effect of Δ on it, we only need to investigate how Δ affects Ω_1 through λ. Taking the derivative of Ω_1 with respect to Δ yields the following:

$$\frac{\partial \Omega_1}{\partial \Delta} = -\frac{\partial \lambda}{\partial \Delta}\frac{2\gamma e^{\frac{\gamma C(\lambda)}{1-\beta}}}{(1-\beta)\lambda^2\sqrt{\tau_\theta\tau_u}\sqrt{e^{\frac{2\gamma C(\lambda)}{1-\beta}} + 8\left(1+\frac{\gamma}{2\lambda\sqrt{\tau_\theta\tau_u}}\right)}}\Omega_2(\lambda),$$

where

$$\Omega_2(\lambda) = \left(\begin{array}{c} 2\gamma\lambda C'(\lambda) - (1-\beta) \\ +\lambda^2 C'(\lambda)\sqrt{\tau_\theta\tau_u}\left(4 + e^{\frac{\gamma C(\lambda)}{1-\beta}}\sqrt{e^{\frac{2\gamma C(\lambda)}{1-\beta}} + 8\left(1+\frac{\gamma}{2\lambda\sqrt{\tau_\theta\tau_u}}\right)} + e^{\frac{2\gamma C(\lambda)}{1-\beta}}\right)\end{array}\right). \tag{A22}$$

Now, if we can find a sufficient condition for $\Omega_2(\lambda) < 0$, coupled with the fact that $\frac{\partial \lambda^*}{\partial \Delta} < 0$, we can ensure that $\frac{\partial Performance}{\partial \Delta} < 0$ as $w \to 0$. Since $C(\lambda)$ is a convex function of λ, $\Omega_2(\lambda)$ is increasing in λ. Let $\hat{\lambda}$ be a value such that $\Omega_2(\hat{\lambda}) = 0$. Note that the equilibrium mass of skilled investors $\lambda^*\left(\Delta, \bar{Z}\right)$ is uniquely solved from equation (A13) for given Δ and $\bar{Z}$. Since $\Delta \in (0, \bar{Z}]$ and $\frac{\partial \lambda^*}{\partial \Delta} < 0$, λ^* must lie within the range $[\lambda^*(\bar{Z}, \bar{Z}), \lambda^*\left(0, \bar{Z}\right))$. Thus, if $\hat{\lambda} \in [\lambda^*(\bar{Z}, \bar{Z}), \lambda^*\left(0, \bar{Z}\right))$, there must exist $\hat{\Delta} \in [0, \bar{Z})$ such that when $\Delta > \hat{\Delta}$, $\lambda < \hat{\lambda}$, $\Omega_2(\lambda) < 0$, and thus $\frac{\partial Performance}{\partial \Delta} < 0$. In sum, the sufficient condition for the existence of $\frac{\partial Performance}{\partial \Delta} < 0$ is that Ψ_Δ is a non-empty set, where Ψ_Δ is a parameter set defined as follows:

$$\Psi_\Delta \equiv \{\hat{\lambda} \mid \hat{\lambda} \in [\lambda^*(\bar{Z},\bar{Z}), \lambda^*\left(0,\bar{Z}\right)),$$
$$\text{where } \hat{\lambda} \text{ is the unique root of equation } \Omega_2\left(\lambda\right) = 0,$$
$$\text{and } \lambda^*(\Delta,\bar{Z}) \text{ is the unique root of equation (A13)}\}. \tag{A23}$$

QED.

Proof of Proposition 3

Again, we focus on the economy in which $\tau_\eta \to \infty$, or equivalently, $w \to 0$.

1. The effect of $\bar{Z}$ on λ^* As in the proof of Proposition 1, when $w \to 0$, we can obtain the approximate value of $x = \frac{\alpha\gamma}{\lambda \bar{Z}} + O\left(w\right)$ as in (A10) and that of $\alpha = \sqrt{\frac{\tau_\theta}{\tau_u}} + O\left(w\right)$ as in (A12), where $O\left(w\right)$ denotes terms of order w. Inserting them into equation (A9) and imposing $w \to 0$, we obtain that λ^* is uniquely characterized by (A13). Define the LHS of equation (A13) as K. We can show

$$\frac{\partial K}{\partial \bar{Z}} = \frac{\gamma \Delta e^{\frac{\gamma C}{1-\beta}}}{2\sqrt{2}\lambda \bar{Z}^2\sqrt{\tau_\theta\tau_u}}\left(\begin{array}{c}\left(\frac{\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}}{\bar{Z}\left(\gamma+2\lambda\sqrt{\tau_\theta\tau_u}\right)+\gamma\Delta}\right)^{3/2}\\ -\left(\frac{\lambda\bar{Z}\sqrt{\tau_\theta\tau_u}}{\bar{Z}\left(\gamma+2\lambda\sqrt{\tau_\theta\tau_u}\right)-\gamma\Delta}\right)^{3/2}\end{array}\right) < 0,$$

Thus, combined with $\frac{\partial K}{\partial \lambda} > 0$, we immediately have $\frac{\partial \lambda^*}{\partial \bar{Z}} > 0$ as $w \to 0$.

2. The effect of $\bar{Z}$ on x^*, q^*, and skilled investors' performance Similar to the proof of Proposition 1, as $w \to 0$, we can approximate x as $x = \frac{\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}}{\lambda \bar{Z}} + O\left(w\right)$. Thus, $\frac{\partial x^*}{\partial \bar{Z}} \approx -\frac{\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\left(\bar{Z}\lambda'(\bar{Z})+\lambda\right)}{\bar{Z}^2\lambda^2} < 0$, where the inequality follows from $\frac{\partial \lambda^*}{\partial \bar{Z}} > 0$. Moreover, as in the proof of Proposition 2, we can show that $\frac{\partial q^*}{\partial \bar{Z}} = \frac{\partial q}{\partial \lambda}\frac{\partial \lambda}{\partial \bar{Z}} > 0$.

Similar to the proof of skilled investors' performance in Proposition 2, taking the derivative of Ω_1 in (A21) with respect to $\bar{Z}$ yields

$$\frac{\partial \Omega_1}{\partial \bar{Z}} = -\frac{\partial \lambda}{\partial \bar{Z}} \frac{2\gamma e^{\frac{\gamma C(\lambda)}{1-\beta}}}{(1-\beta)\lambda^2\sqrt{\tau_\theta\tau_u}\sqrt{e^{\frac{2\gamma C(\lambda)}{1-\beta}} + 8\left(1+\frac{\gamma}{2\lambda\sqrt{\tau_\theta\tau_u}}\right)}}\Omega_2\left(\lambda\right),$$

where Ω_2 is given by (A22). Again, if we can find a sufficient condition for $\Omega_2(\lambda) < 0$, coupled with the fact that $\frac{\partial \lambda^*}{\partial \bar{Z}} > 0$, we can ensure that $\frac{\partial Performance}{\partial \bar{Z}} > 0$. Based on equation (A13) we can solve for the unique equilibrium $\lambda^*\left(\Delta,\bar{Z}\right)$ for given parameters Δ and $\bar{Z}$. Since

$\bar{Z} \in [\Delta, \infty)$ and $\frac{\partial \lambda^*}{\partial \bar{Z}} > 0$, the equilibrium λ^* must lie within the range $[\lambda^*(\Delta, \Delta), \lambda^*(\Delta, \infty))$. Thus, if $\hat{\lambda} \in [\lambda^*(\Delta, \Delta), \lambda^*(\Delta, \infty))$, there must exist $\hat{Z} \in (\Delta, \infty)$ such that when $\bar{Z} < \hat{Z}$, $\lambda < \hat{\lambda}$, $\Omega_2(\lambda) < 0$, and thus $\frac{\partial Performance}{\partial \bar{Z}} > 0$. To sum up, the sufficient condition for the existence of $\frac{\partial Performance}{\partial \bar{Z}} > 0$ is that $\Psi_{\bar{Z}}$ is a non-empty set, where $\Psi_{\bar{Z}}$ is a parameter set defined as follows.

$$\begin{aligned}\Psi_{\bar{Z}} \equiv \; & \{\hat{\lambda} \mid \hat{\lambda} \in [\lambda^*(\Delta, \Delta), \lambda^*(\Delta, \infty)), \\ & \text{where } \hat{\lambda} \text{ is the unique root of equation } \Omega_2(\lambda) = 0, \\ & \text{and } \lambda^*(\Delta, \bar{Z}) \text{ is the unique root of equation (A13) } \}. \end{aligned} \tag{A24}$$

3. The effect of $\bar{Z}$ on α^* and market variables To prove the non-monotonic effects of $\bar{Z}$ on α and market variables, we further discuss two extreme cases: $\bar{Z} \to \Delta$ and $\bar{Z} \to \infty$. Note that unlike the proof of Proposition 2, $\bar{Z}$ affects x even when $w = 0$ (see (A10)). Thus, we need to do the Taylor series of x and α in two variables w and $\bar{Z}$ in this proposition.

Case 1: $w \to 0$ and $\bar{Z} \to \Delta$. We follow the similar three steps as in the proof of Proposition 1 and first express the data package size x as a function of α. Based on the financial-market equilibrium condition (A8), we can derive the Taylor series of $x(\alpha, \lambda; w, \bar{Z})$ around $w = 0$ and $\bar{Z} = \Delta$ as $x\left(\alpha; w, \bar{Z}\right) \approx x(\alpha; 0, \Delta) + \frac{\partial x}{\partial w}(\alpha; 0, \Delta) w + \frac{\partial x}{\partial \bar{Z}}(\alpha; 0, \Delta)\left(\bar{Z} - \Delta\right)$ (note that as $w \to 0$ the higher order effects $o(w)$ and $o(\bar{Z} - \Delta)$ can be neglected), which can be simplified as follows:

$$x\left(\alpha; w, \bar{Z}\right) \approx \frac{\alpha\gamma}{\lambda\Delta} + \frac{\alpha^2\gamma\left(2\gamma + \alpha\lambda\tau_u\right)}{\lambda^2\Delta} w - \frac{\alpha\gamma}{\lambda\Delta^2}\left(\bar{Z} - \Delta\right). \tag{A25}$$

Next, inserting the expression of x into the skilled investors' trading gain (19), maximizing it with respect to α and setting the FOC to zero, we obtain an equation that implicitly determines α for a given λ. We can derive the Taylor series of $\alpha\left(\lambda; w, \bar{Z}\right)$ around $w = 0$ and $\bar{Z} = \Delta$ as: $\alpha\left(\lambda; w, \bar{Z}\right) \approx \alpha(\lambda; 0, \Delta) + \frac{\partial \alpha}{\partial w}(\lambda; 0, \Delta) w + \frac{\partial \alpha}{\partial \bar{Z}}(\lambda; 0, \Delta)\left(\bar{Z} - \Delta\right) + \frac{1}{2}\frac{\partial^2 \alpha}{\partial w^2}(\lambda; 0, \Delta) w^2 + \frac{1}{2}\frac{\partial^2 \alpha}{\partial \bar{Z}^2}(\lambda; 0, \Delta)\left(\bar{Z} - \Delta\right)^2 + \frac{\partial^2 \alpha}{\partial w \partial \bar{Z}}(\lambda; 0, \Delta)\left(\bar{Z} - \Delta\right) w$, which can be simplified as follows:

$$\alpha\left(\lambda; w, \bar{Z}\right) \approx \sqrt{\frac{\tau_\theta}{\tau_u}} + \alpha_{w1} w, \tag{A26}$$

where

$$\alpha_{w1} = -\frac{\tau_\theta}{2}\sqrt{\frac{\tau_\theta}{\tau_u}} - w\frac{\left(\frac{\tau_\theta}{\tau_u}\right)^{5/2}\left(64\gamma^4 + \gamma^2\lambda\tau_u\left(157\lambda\tau_\theta + 176\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right) - 3\lambda^3\tau_\theta\tau_u^2\left(\lambda\tau_\theta - 14\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)\right)}{4\lambda^2\left(\lambda\tau_\theta + \gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^2}$$

$$+\left(\bar{Z} - \Delta\right)\left(\begin{array}{c} \frac{\tau_\theta}{\Delta\lambda^3\tau_u^3\left(\lambda\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^2} \\ \times\left(\begin{array}{c} \sqrt{\frac{\lambda\tau_\theta}{\lambda\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}}}\left(\begin{array}{c} \gamma^5 + 2\gamma^3\lambda\tau_u\left(3\lambda\tau_\theta + 2\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right) \\ +\gamma\lambda^3\tau_\theta\tau_u^2\left(\lambda\tau_\theta + 4\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)\end{array}\right) \\ +\lambda^2\tau_\theta\tau_u\left(3\gamma^3 + \gamma\lambda\tau_u\left(7\lambda\tau_\theta + 8\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right) + 2\lambda^3\tau_u^3\left(\frac{\tau_\theta}{\tau_u}\right)^{3/2}\right)\end{array}\right)\end{array}\right). \quad \text{(A27)}$$

When $w \to 0$ and $\bar{Z} \to \Delta$, we obtain the following (note that according to equation (A13), $\lambda'\left(\bar{Z}\right)$ is finite, so $\lim_{\bar{Z}\to\Delta}\left(\bar{Z}-\Delta\right)\lambda'\left(\bar{Z}\right) = 0$ and $\lim_{w\to 0} w\lambda'\left(\bar{Z}\right) = 0$):

$$\frac{\partial\alpha}{\partial\bar{Z}} \approx \frac{\partial\alpha_{w1}}{\partial\bar{Z}}w = \frac{w\left(\frac{\lambda\tau_\theta}{\lambda\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}}\right)^{5/2}}{\Delta\lambda^5\tau_u^4\left(\lambda\tau_\theta + \gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^2}$$

$$\times\left(\begin{array}{c} \gamma^7 + 3\gamma^5\lambda\tau_u\left(5\lambda\tau_\theta + 2\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right) \\ +5\gamma^3\lambda^3\tau_\theta\tau_u^2\left(3\lambda\tau_\theta + 4\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right) + \gamma\lambda^5\tau_\theta^2\tau_u^3\left(\lambda\tau_\theta + 6\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right) \\ +\lambda\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\sqrt{\frac{\lambda\tau_\theta}{\lambda\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}}}\left(\begin{array}{c} 3\gamma^6 + \gamma^4\lambda\tau_u\left(40\lambda\tau_\theta + 17\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right) \\ +5\gamma^2\lambda^3\tau_\theta\tau_u^2\left(7\lambda\tau_\theta + 10\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right) \\ +\lambda^5\tau_\theta^2\tau_u^3\left(2\lambda\tau_\theta + 13\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)\end{array}\right)\end{array}\right) > 0. \quad \text{(A28)}$$

Thus, as $w \to 0$ and $\bar{Z} \to \Delta$, α is increasing in $\bar{Z}$, and $\frac{\partial}{\partial\bar{Z}}\left(\frac{1}{Var(\tilde{\theta}|\tilde{p})}\right) = \frac{\partial}{\partial\alpha}\left(\frac{1}{Var(\tilde{\theta}|\tilde{p})}\right)\frac{\partial\alpha}{\partial\bar{Z}} > 0$ immediately follows.

Next, inserting (A25) and (A26) into the cost of capital, we can approximate $E(\tilde{\theta}-\tilde{p})$ as in the same form in (A14) replacing α_w with α_{w1}, where α_{w1} is given in (A27). Taking the derivative of $E(\tilde{\theta}-\tilde{p})$ with respect to $\bar{Z}$ (noting that α_{w1} is a function of $\bar{Z}$) yields $\frac{\partial E(\tilde{\theta}-\tilde{p})}{\partial\bar{Z}} \approx -\frac{\gamma\left(\gamma+2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)}{\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^2}\frac{\partial\alpha_{w1}}{\partial\bar{Z}}w < 0$, where the inequality follows because $\frac{\partial\alpha_{w1}}{\partial\bar{Z}} > 0$ as given by (A28). Similarly, $\sigma^2(\tilde{\theta}-\tilde{p})$ can be approximated as in the same form in (A15) replacing α_w with α_{w1}, where α_{w1} is given in (A27). Taking the derivative of $\sigma^2(\tilde{\theta}-\tilde{p})$ with respect to $\bar{Z}$ (noting that α_{w1} is a function of $\bar{Z}$) yields $\frac{\partial\sigma^2(\tilde{\theta}-\tilde{p})}{\partial\bar{Z}} \approx -\frac{2\sqrt{\frac{\tau_\theta}{\tau_u}}\left(\gamma^3\sqrt{\frac{\tau_\theta}{\tau_u}}+3\gamma\tau_\theta\left(\gamma+\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)+2\tau_\theta^2\tau_u\right)}{\tau_\theta\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^3}\frac{\partial\alpha_{w1}}{\partial\bar{Z}}w < 0$.

Case 2: $w \to 0$ and $\bar{Z} \to \infty$. To ease expression, we denote $\upsilon \equiv \frac{1}{\bar{Z}}$. Thus $\bar{Z} \to \infty$ is equivalent to $\upsilon \to 0$. Similarly, we can derive the Taylor series of $x(\alpha, \lambda; w, \upsilon)$ around $w = 0$

and $\upsilon = 0$ as $x(\alpha; w, \upsilon) \approx x(\alpha; 0, 0) + \frac{\partial x}{\partial w}(\alpha; 0, 0)\, w + \frac{\partial x}{\partial \upsilon}(\alpha; 0, 0)\, \upsilon$, which can be simplified as follows:

$$x(\alpha; w, \upsilon) \approx \frac{\alpha\gamma}{\lambda}\upsilon. \tag{A29}$$

Next, inserting this expression of $x(\alpha; w, \upsilon)$ into the skilled investors' trading gain (19) and maximizing it with respect to α, we obtain a function that implicitly determines $\alpha(\lambda; w, \upsilon)$. Then, the Taylor series of $\alpha(\lambda; w, \upsilon)$ around $w = 0$ and $\upsilon = 0$ is $\alpha(\lambda; w, \upsilon) \approx \alpha(\lambda; 0, 0) + \frac{\partial \alpha}{\partial w}(\lambda; 0, 0)\, w + \frac{\partial \alpha}{\partial \upsilon}(\lambda; 0, 0)\, \upsilon + \frac{1}{2}\frac{\partial^2 \alpha}{\partial w^2}(\lambda; 0, 0)\, w^2 + \frac{1}{2}\frac{\partial^2 \alpha}{\partial \upsilon^2}(\lambda; 0, 0)\, \upsilon^2 + \frac{\partial^2 \alpha}{\partial w \partial \upsilon}(\lambda; 0, 0)\, w\upsilon$, which can be simplified as follows:

$$\alpha(\lambda; w, \upsilon) \approx \sqrt{\frac{\tau_\theta}{\tau_u}} + \alpha_{w2} w, \tag{A30}$$

where

$$\alpha_{w2} = -\frac{1}{2}\tau_\theta\left(\frac{2\gamma}{\lambda\tau_u} + 5\sqrt{\frac{\tau_\theta}{\tau_u}}\right) + \frac{\gamma\tau_\theta\Delta}{2\lambda\tau_u}\upsilon + \frac{\left(\frac{\tau_\theta}{\tau_u}\right)^{3/2}\left(20\gamma^2 + 3\lambda\tau_u\left(41\lambda\tau_\theta + 32\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)\right)}{8\lambda^2} w. \tag{A31}$$

Taking the derivative of α with respect to υ (note that λ is a function of υ) yields the following:

$$\begin{aligned} \frac{\partial \alpha}{\partial \upsilon} &\approx w\frac{\partial \alpha_{w2}}{\partial \upsilon} = w\frac{\gamma\tau_\theta\left(\Delta\lambda^2 - \lambda'(\upsilon)\left(\Delta\lambda\upsilon - 2\lambda + 10\gamma w\sqrt{\frac{\tau_\theta}{\tau_u}} + 24\lambda w\tau_\theta\right)\right)}{2\lambda^3\tau_u} \\ &\approx w\frac{\gamma w\tau_\theta\left(\Delta\lambda + 2\lambda'(0)\right)}{2\lambda^2\tau_u} > 0, \end{aligned} \tag{A32}$$

where the second approximation follows by imposing $w \to 0$ and $\upsilon \to 0$ and the inequality holds because as $\upsilon \to 0$, $\lambda'(\upsilon) = 0$. Specifically, recall that the equilibrium λ^* is determined by equation (A13) when $w \to 0$. Applying implicit function theorem to (A13) we obtain that

$$\lambda'(\upsilon) = -\frac{\frac{\gamma\Delta}{4\lambda\sqrt{\tau_\theta\tau_u}}\left(\frac{1}{\left(1+\frac{\gamma(1-\Delta\upsilon)}{2\lambda\sqrt{\tau_\theta\tau_u}}\right)^{3/2}} - \frac{1}{\left(1+\frac{\gamma(1+\Delta\upsilon)}{2\lambda\sqrt{\tau_\theta\tau_u}}\right)^{3/2}}\right)}{\left(\begin{array}{c}\frac{\gamma C'(\lambda)\left(\frac{1}{\sqrt{1+\frac{\gamma(1+\Delta\upsilon)}{2\lambda\sqrt{\tau_\theta\tau_u}}}} + \frac{1}{\sqrt{1+\frac{\gamma(1-\Delta\upsilon)}{2\lambda\sqrt{\tau_\theta\tau_u}}}}\right)}{(1-\beta)} \\ + \frac{\gamma}{4\lambda^2\sqrt{\tau_\theta\tau_u}}\left(\frac{1+\Delta\upsilon}{\left(\frac{\gamma(1+\Delta\upsilon)}{2\lambda\sqrt{\tau_\theta\tau_u}}+1\right)^{3/2}} + \frac{1-\Delta\upsilon}{\left(1+\frac{\gamma(1-\Delta\upsilon)}{2\lambda\sqrt{\tau_\theta\tau_u}}\right)^{3/2}}\right)\end{array}\right)},$$

which goes to 0 as $\upsilon \to 0$, that is, $\lambda'(0) = 0$. Therefore, as $\upsilon \to 0$ (i.e., $\bar{Z} \to \infty$), $\frac{\partial \alpha}{\partial \upsilon} > 0$, which suggests that $\frac{\partial \alpha}{\partial \bar{Z}} < 0$. And $\frac{\partial}{\partial \bar{Z}}\left(\frac{1}{Var(\tilde{\theta}|\tilde{p})}\right) = \frac{\partial}{\partial \alpha}\left(\frac{1}{Var(\tilde{\theta}|\tilde{p})}\right)\frac{\partial \alpha}{\partial \bar{Z}} < 0$ immediately follows.

Next, inserting (A29) and (A30) into $E(\tilde{\theta}-\tilde{p})$, we approximate it as follows:

$$E(\tilde{\theta}-\tilde{p}) \approx \frac{\gamma}{2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}} + \frac{\gamma\left(\begin{array}{c} \tau_\theta\left(\gamma^2+\lambda\tau_u\left(\tau_\theta+2\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)\right) \\ -\alpha_{w2}\lambda\tau_u\left(\gamma+2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)\end{array}\right)}{\lambda\tau_u\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^2}w,$$

where α_{w2} is given in (A31). Taking the derivative of $E(\tilde{\theta}-\tilde{p})$ with respect to υ (note that both λ and α_{w2} are functions of υ) yields

$$\begin{aligned} \frac{\partial E(\tilde{\theta}-\tilde{p})}{\partial \upsilon} &\approx \frac{\gamma\left(-w\alpha_{w2}'(\upsilon)\left(\gamma+2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)-\frac{\gamma^2 w\tau_\theta\lambda'(\upsilon)}{\lambda^2\tau_u}\right)}{\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^2} \\ &\approx -\frac{\gamma^2\Delta\tau_\theta\left(\gamma+2\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)}{2\lambda\tau_u\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^2}w<0, \end{aligned}$$

where the second approximation follows by imposing $\upsilon \to 0$. Thus, $\frac{\partial E(\tilde{\theta}-\tilde{p})}{\partial \bar{Z}} > 0$ as $\bar{Z} \to \infty$. Similarly, $\sigma^2(\tilde{\theta}-\tilde{p})$ can be approximated as follows:

$$\begin{aligned} \sigma^2(\tilde{\theta}-\tilde{p}) &\approx \frac{\gamma^2+2\tau_u\left(\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)}{\tau_u\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^2} \\ &\quad + \frac{\left(\begin{array}{c} \tau_\theta^2\left(\gamma^3(\lambda-2)+3\gamma\lambda\tau_\theta\tau_u+2\lambda\tau_u^3\left(\frac{\tau_\theta}{\tau_u}\right)^{3/2}\right) \\ -2\lambda\alpha_{w2}\tau_u\left(\gamma^3\sqrt{\frac{\tau_\theta}{\tau_u}}+3\gamma\tau_\theta\left(\gamma+\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)+2\tau_\theta^2\tau_u\right)\end{array}\right)}{\lambda\tau_u^2\sqrt{\frac{\tau_\theta}{\tau_u}}\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^3}w. \end{aligned}$$

Taking the derivative of $\sigma^2(\tilde{\theta}-\tilde{p})$ with respect to υ (note that both λ and α_{w2} are functions of υ) yields

$$\begin{aligned} \frac{\partial \sigma^2(\tilde{\theta}-\tilde{p})}{\partial \upsilon} &\approx -w\frac{2\left(\lambda^2\tau_u\alpha_{w2}'(\upsilon)\left(\begin{array}{c}\gamma^3\sqrt{\frac{\tau_\theta}{\tau_u}}+2\tau_\theta^2\tau_u \\ +3\gamma\tau_\theta\left(\gamma+\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)\end{array}\right)-\gamma^3\tau_\theta^2\lambda'(\upsilon)\right)}{\lambda^2\tau_u^2\sqrt{\frac{\tau_\theta}{\tau_u}}\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^3} \\ &\approx -\frac{\gamma\Delta w\sqrt{\frac{\tau_\theta}{\tau_u}}\left(\gamma^3\sqrt{\frac{\tau_\theta}{\tau_u}}+3\gamma\tau_\theta\left(\gamma+\tau_u\sqrt{\frac{\tau_\theta}{\tau_u}}\right)+2\tau_\theta^2\tau_u\right)}{\lambda\tau_u\left(2\tau_\theta+\gamma\sqrt{\frac{\tau_\theta}{\tau_u}}\right)^3}<0, \end{aligned}$$

where the second approximation follows by imposing $\upsilon \to 0$. Thus, $\frac{\partial \sigma(\tilde{\theta}-\tilde{p})}{\partial \bar{Z}} > 0$ as $\bar{Z} \to \infty$. QED.

Proof of Proposition 4

We take three steps to prove Proposition 4. In the first step, we calculate the equilibrium price at date 2 and pin down α_2. In the second step, we prove that the data vendor sets data sample sizes $x_{H,2}$ and $x_{L,2}$ such that $x_{H,2}Z_H = x_{L,2}Z_L$. In the third step, we calculate the equilibrium α_2 and prove that market variables (i.e., price informativeness, the cost of capital, and return volatility) remain unchanged with respect to skill-acquisition costs, skill mean, and skill volatility.

Step 1 Taking similar steps as in Lemma 1, we pin down $a_{0,2}$, $a_{\theta,2}$, and $a_{u,2}$ in the price function (25) as follows. Applying Bayes' rule, we can compute the conditional moments for an H-type skilled investor i as follows:

$$E(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{H,i,2}) = \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{\alpha_2^2 \tau_{u,2} \tilde{s}_{p,2} + x_{H,2} Z_H \tilde{s}_{H,i,2}}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_{u,2} + x_{H,2} Z_H},$$
$$Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{H,i,2}) = \frac{1}{\tau_\theta + \tau_\eta} + \left(\frac{\tau_\eta}{\tau_\theta + \tau_\eta}\right)^2 \frac{1}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_u + x_{H,2} Z_H}. \tag{A33}$$

Similarly, the conditional moments for an L-type skilled investor i are as follows:

$$E(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{L,i,2}) = \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{\alpha_2^2 \tau_u \tilde{s}_{p,2} + x_{L,2} Z_L \tilde{s}_{L,i,2}}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_u + x_{L,2} Z_L},$$
$$Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{L,i,2}) = \frac{1}{\tau_\theta + \tau_\eta} + \left(\frac{\tau_\eta}{\tau_\theta + \tau_\eta}\right)^2 \frac{1}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_u + x_{L,2} Z_L}. \tag{A34}$$

The conditional moments for uninformed investors are as follows:

$$E(\tilde{\theta}_3|\tilde{p}_2) = \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{\alpha_2^2 \tau_u \tilde{s}_{p,2}}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_{u,2}},$$
$$Var(\tilde{\theta}_3|\tilde{p}_2) = \frac{1}{\tau_\theta + \tau_\eta} + \left(\frac{\tau_\eta}{\tau_\theta + \tau_\eta}\right)^2 \frac{1}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_u}. \tag{A35}$$

Inserting these conditional moments into demand functions and further into the market-clearing condition, $\int_0^{\lambda/2} D_{H,i,2}\left(\tilde{p}_2, \tilde{s}_{H,i,2}\right) di + \int_{\lambda/2}^{\lambda} D_{L,i',2}\left(\tilde{p}_2, \tilde{s}_{L,i',2}\right) di' + (1-\lambda) D_{U,2}\left(\tilde{p}_2\right) + \tilde{u}_2 = 1$, we can obtain the date-2 price as follows:

$$\tilde{p}_2 = \frac{A_2 \tilde{s}_{p,2} + B_2(\tilde{\theta}_3 + \tilde{\eta}_3) + \tilde{u}_2 - 1}{\frac{\lambda}{2\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{H,i,2})} + \frac{\lambda}{2\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{L,i,2})} + \frac{1-\lambda}{\gamma Var(\tilde{\theta}_3|\tilde{p}_2)}},$$

where

$$
\begin{aligned}
A_2 &= \frac{\lambda}{2\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{H,i,2})} \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{\alpha_2^2 \tau_u}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_{u,2} + x_{H,2} Z_H} \\
&\quad + \frac{\lambda}{2\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{L,i,2})} \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{\alpha_2^2 \tau_u}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_{u,2} + x_{L,2} Z_L} \\
&\quad + \frac{1-\lambda}{\gamma Var(\tilde{\theta}_3|\tilde{p}_2)} \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{\alpha_2^2 \tau_u}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_{u,2}}, \\
B_2 &= \frac{\lambda}{2\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{H,i,2})} \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{x_{H,2} Z_H}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_{u,2} + x_{H,2} Z_H} \\
&\quad + \frac{\lambda}{2\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{L,i,2})} \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{x_{L,2} Z_L}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_{u,2} + x_{L,2} Z_L}.
\end{aligned}
$$

That is,

$$
a_{0,2} = -\frac{1}{\frac{\lambda}{2\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{H,i,2})} + \frac{\lambda}{2\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{L,i,2})} + \frac{1-\lambda}{\gamma Var(\tilde{\theta}_3|\tilde{p}_2)}}, \tag{A36}
$$

$$
a_{\theta,2} = \frac{A_2 + B_2}{\frac{\lambda}{2\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{H,i,2})} + \frac{\lambda}{2\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{L,i,2})} + \frac{1-\lambda}{\gamma Var(\tilde{\theta}_3|\tilde{p}_2)}}, \tag{A37}
$$

$$
a_{u,2} = B_2^{-1} a_{\theta,2}. \tag{A38}
$$

From the above equation, we directly have $\alpha_2 = B_2$. Using the expression of B_2 above, we can rearrange equation $\alpha_2 = B_2$. After some algebra, we can compute the price coefficients as follows:

$$
\begin{aligned}
\alpha_2 &= \frac{\lambda}{2\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{H,i,2})} \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{x_{H,2} Z_H}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_u + x_{H,2} Z_H} \\
&\quad + \frac{\lambda}{2\gamma Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{L,i,2})} \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{x_L Z_L}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_u + x_{L,2} Z_L}.
\end{aligned}
$$

Inserting the expressions of $Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{H,i,2})$ and $Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{L,i,2})$ into α_2, we obtain equation (26). Similar to Lemma 1, we can prove that there is a unique positive solution to this equation.

Step 2 In this step, we prove that in equilibrium the data vendor sets data sample sizes $x_{H,2}$ and $x_{L,2}$ such that $x_{H,2} Z_H = x_{L,2} Z_L$. Denote the total precision (given both signals and prices) attained by investor i as r_i. That is, the total precision for the high- and low-type investors are respectively $r_H \equiv Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{H,i,2})^{-1}$ and $r_L \equiv Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{L,i,2})^{-1}$, and the total precision for the uninformed investors is $r_U \equiv Var(\tilde{\theta}_3|\tilde{p}_2)^{-1}$, which only depends on α_2.

To show that $x_{H,2}Z_H = x_{L,2}Z_L$, it suffices to show the following auxiliary lemma.

Lemma A1 *Let $\mathcal{B}$ be an admissible allocation of information where individual investor's signals are conditionally independent given $\tilde{\theta}_3$. Let $\overline{r} = \frac{r_H + r_L}{2}$ be the average total precision of skilled investors attained. If $x_{H,2}Z_H \neq x_{L,2}Z_L$, then there exists another allocation $\mathcal{B}'$ where signals are conditionally independent and identically distributed given $\tilde{\theta}_3$ and such that the vendor has more profits under allocation $\overline{r} = \frac{r_H + r_L}{2}$.*

We prove Lemma A1 as follows. We first show that the ratio α_2 only depends on the average information precision across skilled investors. That is, for any two personalized allocation $\mathcal{B}$ and $\mathcal{B}'$, as long as $\overline{r}(\mathcal{B}) = \overline{r}(\mathcal{B}') = \frac{r_H + r_L}{2}$, uninformed investors have the same amount of information in both allocations. Then we show that for a given personalized allocation $\mathcal{B}$, if $r_H \neq r_L$, the data vendor can always earn higher profits by allocating symmetric personalized noise to all skilled investors.

Given the definitions of r_H, r_L, and r_U, we can express the r_H, r_L, and α_2 as follows:

$$\begin{aligned}
\frac{1}{r_H} &= \frac{1}{\tau_\theta + \tau_\eta} + \left(\frac{\tau_\eta}{\tau_\theta + \tau_\eta}\right)^2 \frac{1}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_u + x_{H,2}Z_H}, \\
\frac{1}{r_L} &= \frac{1}{\tau_\theta + \tau_\eta} + \left(\frac{\tau_\eta}{\tau_\theta + \tau_\eta}\right)^2 \frac{1}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_u + x_{L,2}Z_L}, \\
\alpha_2 &= \frac{\lambda}{2\gamma} \frac{x_{H,2}Z_H \tau_\eta}{\tau_\eta + \alpha_2^2 \tau_u + x_{H,2}Z_H} + \frac{\lambda}{2\gamma} \frac{x_{L,2}Z_L \tau_\eta}{\tau_\eta + \alpha_2^2 \tau_u + x_{L,2}Z_L}.
\end{aligned}$$

Based on the expression of r_H and r_L we can solve $x_{H,2}Z_H$ and $x_{L,2}Z_L$ as follows:

$$x_{H,2}Z_H = \frac{\left(\frac{\tau_\eta}{\tau_\theta + \tau_\eta}\right)^2}{\frac{1}{r_H} - \frac{1}{\tau_\theta + \tau_\eta}} - \frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} - \alpha_2^2 \tau_u \text{ and } x_{L,2}Z_L = \frac{\left(\frac{\tau_\eta}{\tau_\theta + \tau_\eta}\right)^2}{\frac{1}{r_L} - \frac{1}{\tau_\theta + \tau_\eta}} - \frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} - \alpha_2^2 \tau_u.$$

Plugging $x_{H,2}Z_H$ and $x_{L,2}Z_L$ into α_2, we know α_2 is implicitly solved from the following:

$$\alpha_2 = \frac{\lambda}{2\gamma} \frac{\alpha_2^2 \tau_u + \tau_\eta}{\tau_\eta} (r_H + r_L) - \frac{\lambda}{\gamma} \frac{\tau_\eta \tau_\theta + \alpha_2^2 \tau_u (\tau_\eta + \tau_\theta)}{\tau_\eta},$$

which shows that for a personalized allocation $\mathcal{B}$, $\alpha(\mathcal{B})$ depends on $\mathcal{B}$ only through $r_H + r_L$. Now consider an allocation $\mathcal{B}'$ in which high-type and low-type investors have the same precision $r_H(\mathcal{B}') = r_L(\mathcal{B}') = \frac{r_H + r_L}{2}$, it is clear that $\alpha_2(\mathcal{B})$ and $\alpha_2(\mathcal{B}')$ are the same. In this sense, we have $r_U(\mathcal{B}) = r_U(\mathcal{B}')$. Next we show that the vendor's profits in $\mathcal{B}'$ are no lower than that in $\mathcal{B}$.

Given the expressions of $q_{H,2}$ and $q_{L,2}$, the vendor's profit is the following:

$$\frac{\beta\lambda}{2\gamma} \left[-\ln r_U(\mathcal{B}) + \frac{1}{2} \ln r_H(\mathcal{B}) + \frac{1}{2} \ln r_L(\mathcal{B}) \right] \leq \frac{\beta\lambda}{2\gamma} \ln \left[\frac{\frac{1}{2} r_H(\mathcal{B}) + \frac{1}{2} r_L(\mathcal{B})}{r_U(\mathcal{B})} \right],$$

where the inequality follows from Jensen's inequality. These results suggest that the vendor can make more profits by setting the same effective precision among both types of investors. That is, $x_{H,2}Z_H = x_{L,2}Z_L$.

Step 3 In this step, we calculate the equilibrium α_2 and prove that the market variables (i.e., price informativeness, the cost of capital, and return volatility) remain unchanged with respect to the skill-acquisition cost, skill mean, and skill volatility. Note that Lemma A1 suggests that $x_{H,2}Z_H = x_{L,2}Z_L$. When $x_{H,2}Z_H = x_{L,2}Z_L$, equation (26) can be simplified to the following: $\gamma\alpha_2\left(x_{H,2}Z_H + \tau_u\alpha_2^2 + \tau_\eta\right) - \lambda\tau_\eta x_{H,2}Z_H = 0$. Using this equation, coupled with that $x_{H,2}Z_H = x_{L,2}Z_L$, we obtain the following:

$$x_{H,2} = \frac{\gamma\alpha_2(\tau_u\alpha_2^2 + \tau_\eta)}{Z_H(\lambda\tau_\eta - \gamma\alpha_2)} \text{ and } x_{L,2} = \frac{\gamma\alpha_2(\tau_u\alpha_2^2 + \tau_\eta)}{Z_L(\lambda\tau_\eta - \gamma\alpha_2)}. \tag{A39}$$

Inserting $x_{H,2}$ and $x_{L,2}$ into data prices $q_{H,2}$ and $q_{L,2}$ and maximizing the profit $\frac{\lambda}{2}(q_{H,2}+q_{L,2})$ yields the following:

- If $\lambda\tau_\eta > \gamma\sqrt{\frac{\tau_\eta\tau_\theta}{\tau_u(\tau_\eta+\tau_\theta)}}$, $\alpha_2^* = \sqrt{\frac{\tau_\eta\tau_\theta}{\tau_u(\tau_\eta+\tau_\theta)}}$, and x_H^* and x_L^* are obtained by equations (A39).
- Otherwise, $\alpha_2^* = \frac{\lambda\tau_\eta}{\gamma}$, and $x_{H,2}^* = x_{L,2}^* = \infty$.

This suggests that the data vendor sells a finite number of data points only when τ_η is sufficiently high. As in the main text, we focus on the case with sufficiently high τ_η and conduct comparative statics for price informativeness, the cost of capital, and return volatility. With $x_{H,2}^*$, $x_{L,2}^*$, α_2^* and equations (A36)–(A38), we obtain the cost of capital as follows:

$$E\left(\tilde{\theta}_3 - \tilde{p}_2\right) = \frac{\gamma\left(\tau_u\alpha_2^2 + \tau_\eta\right)}{\tau_\theta\left(\tau_u\alpha_2^2 + \tau_\eta\right) + \tau_\eta\alpha_2\left(\gamma + \tau_u\alpha_2\right)}.$$

Return volatility $\sigma(\tilde{\theta}_3 - \tilde{p}_2) = \sqrt{\left(1 - a_{\theta,2}\right)^2 \frac{1}{\tau_\theta} + a_{\theta,2}^2\frac{1}{\tau_\eta} + a_{u,2}^2\frac{1}{\tau_u}}$ can be simplified as follows:

$$\sigma\left(\tilde{\theta}_3 - \tilde{p}_2\right) = \sqrt{\frac{\left(\tau_\eta + \alpha_2^2\tau_u\right)\left(\tau_\theta\tau_u\left(\tau_\eta + \alpha_2^2\tau_u\right) + \tau_\eta\left(\gamma + \alpha_2\tau_u\right)^2\right)}{\tau_u\left(\alpha_2^2\tau_\theta\tau_u + \tau_\eta\left(\tau_\theta + \alpha_2\left(\gamma + \alpha_2\tau_u\right)\right)\right)^2}}.$$

It is straightforward to show that both $E(\tilde{\theta}_3 - \tilde{p}_2)$ and $\sigma(\tilde{\theta}_3 - \tilde{p}_2)$ depend only on α_2, not on λ. Note that price informativeness $Var^{-1}(\tilde{\theta}_3|\tilde{p}_2)$ is the reverse of (A35), which depends only on α_2 as well. Since for sufficiently high τ_η, α^* remains a constant as given by (27), these three market variables remain unchanged as the skill-acquisition cost, skill mean, or skill volatility changes. QED.

Extended Economy: Equilibrium at Date 0 and 1

Calculation of the date-1 price

We take two steps to pin down the equilibrium price at date 1 as given by (30). Again, let $\tilde{s}_{p,1} \equiv \tilde{\theta}_2 + \tilde{\eta}_2 + \alpha_1^{-1}\tilde{u}_1$, with $\alpha_1 \equiv \frac{a_{\theta,1}}{a_{u,1}}$. We first calculate the demand functions of investors at date 1. Then we use the market-clearing condition to pin down the equilibrium price at date 1, particularly α_1.

Step 1 In this step, we calculate investors' demand functions. For skilled investors, given the result that the private signals of both types of skilled investors at date 2 effectively have the same precision, without loss of generality, we can focus on calculating the expected utility of a generic skilled investor. For skilled investor i, her terminal wealth is $W_{i,2} = D_{i,1}(\tilde{\theta}_2 + \tilde{p}_2 - \tilde{p}_1) + D_{i,2}(\tilde{\theta}_3 - \tilde{p}_2)$. Inserting her date-2 demand $D_{i,2} = \frac{E(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2})-\tilde{p}_2}{\gamma Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2})}$ into the wealth, we can compute the expected utility of the skilled investor conditional on her information set at date 2 (i.e., $\tilde{p}_2$, $\tilde{s}_{i,2}$) as follows:

$$
\begin{aligned}
&E\left[-e^{-\gamma W_{i,2}} \mid \tilde{p}_2, \tilde{s}_{i,2}\right] \\
&= E\left[-\exp\left[-\gamma D_{i,1}\left(\tilde{\theta}_2 + \tilde{p}_2 - \tilde{p}_1\right) - \frac{E(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2}) - \tilde{p}_2}{Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2})}\left(\tilde{\theta}_3 - \tilde{p}_2\right)\right] \mid \tilde{p}_2, \tilde{s}_{i,2}\right] \\
&= -\exp\left[-\gamma D_{i,1}\left(\tilde{\theta}_2 + \tilde{p}_2 - \tilde{p}_1\right) - \frac{z_1^2}{2}\right],
\end{aligned}
$$

where $z_1 \equiv \frac{E(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2})-\tilde{p}_2}{\sqrt{Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2})}}$. Conditional on $\tilde{p}_2$, z_1 follows a normal distribution with mean $\mu_1 = \frac{E(\tilde{\theta}_3|\tilde{p}_2)-\tilde{p}_2}{\sqrt{Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2})}}$ and variance $\sigma_1^2 = \frac{Var\left(E(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2})|\tilde{p}_2\right)}{Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2})}$. Thus, using the property of z_1, the skilled investor's expected utility conditional on $\tilde{p}_2$ can be computed as follows:

$$
\begin{aligned}
&E\left[-e^{-\gamma W_{i,2}} \mid \tilde{p}_2\right] = \\
&\qquad -\frac{\sqrt{Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2})}}{\sqrt{Var(\tilde{\theta}_3|\tilde{p}_2)}}\exp\left[-\gamma D_{i,1}(\tilde{\theta}_2 + \tilde{p}_2 - \tilde{p}_1) - \frac{\left(E(\tilde{\theta}_3|\tilde{p}_2) - \tilde{p}_2\right)^2}{2Var(\tilde{\theta}_3|\tilde{p}_2)}\right].
\end{aligned} \tag{A40}
$$

Let

$$z_2 \equiv E(\tilde{\theta}_3|\tilde{p}_2) - \tilde{p}_2. \tag{A41}$$

Since z_2 is independent of all random variables at date 1 (e.g., $\tilde{p}_1$, $\tilde{s}_{H,i,1}$, $\tilde{s}_{L,i,1}$), conditional on all information sets at date 1, z_2 is a normal distribution with mean and variance given respectively by:

$$\mu_2 = -a_{0,2} \text{ and } \sigma_2^2 = (b_2 - a_{\theta,2})^2 \left(\frac{1}{\tau_\theta} + \frac{1}{\tau_\eta} + \frac{1}{\alpha_2^2 \tau_u} \right), \text{ where } b_2 \equiv \frac{\tau_\eta}{\tau_\theta + \tau_\eta} \frac{\alpha_2^2 \tau_u}{\frac{\tau_\theta \tau_\eta}{\tau_\theta + \tau_\eta} + \alpha_2^2 \tau_u}.$$

With the definition of z_2, it is straightforward to show that $\tilde{p}_2$ can be expressed as a function of z_2: $\tilde{p}_2 = \frac{a_{\theta,2}}{b_2 - a_{\theta,2}} z_2 + \frac{b_2}{b_2 - a_{\theta,2}} a_{0,2}$. Inserting this expression of $\tilde{p}_2$ into equation (A40) and using the property of z_2, we can compute the skilled investor's expected utility conditional on her information set at date 1 (i.e., $(\tilde{p}_1, \tilde{s}_{i,1})$) as follows:

$$\begin{aligned}
& E\left[-e^{-\gamma W_{i,2}} \mid \tilde{p}_1, \tilde{s}_{i,1}\right] \\
& = -\frac{1}{\sqrt{Var(\tilde{\theta}_3|\tilde{p}_2)/Var(\tilde{\theta}_3|\tilde{p}_2, \tilde{s}_{i,2})}} \frac{1}{\sqrt{1 + \frac{\sigma_2^2}{Var(\tilde{\theta}_3|\tilde{p}_2)}}} \\
& \times \exp\left[-\gamma D_{i,1}(E(\tilde{\theta}_2|\tilde{s}_{p,1}, \tilde{s}_{i,1}) + \frac{b_2}{b_2 - a_{\theta,2}} a_{0,2} - \tilde{p}_1) + \frac{1}{2}\gamma^2 D_{i,1}^2 Var(\tilde{\theta}_2|\tilde{s}_{p,1}, \tilde{s}_{i,1})\right] \\
& \times \exp\left[-\gamma D_{i,1} \frac{a_{\theta,2}}{b_2 - a_{\theta,2}} \mu_2 - \frac{\mu_2^2}{2Var(\tilde{\theta}_3|\tilde{p}_2)} + \frac{1}{2} \frac{\sigma_2^2 (\gamma D_{i,1} \frac{a_{\theta,2}}{b_2 - a_{\theta,2}} + \frac{\mu_2}{Var(\tilde{\theta}_3|\tilde{p}_2)})^2}{1 + \frac{\sigma_2^2}{Var(\tilde{\theta}_3|\tilde{p}_2)}}\right].
\end{aligned} \tag{A42}$$

Note that the conditional moments $E(\tilde{\theta}_2|\tilde{s}_{p,1}, \tilde{s}_{i,1})$ and $Var(\tilde{\theta}_2|\tilde{s}_{p,1}, \tilde{s}_{i,1})$ take the same forms as in (11)–(14) replacing α with α_1 and x with x_1. Maximizing this expected utility with respect to $D_{i,1}$ yields the optimal demand of the skilled investor at date 1 as follows:

$$D_{i,1}(\tilde{s}_{p,1}, \tilde{s}_{i,1}) = \frac{E(\tilde{\theta}_2|\tilde{s}_{p,1}, \tilde{s}_{i,1}) + a_{0,2} - c_1 - \tilde{p}_1}{\gamma \left[Var(\tilde{\theta}_2|\tilde{s}_{p,1}, \tilde{s}_{i,1}) + c_2\right]}, \tag{A43}$$

where c_1 and c_2 are two constants given below:

$$c_1 \equiv \frac{\sigma_2^2 (\frac{\mu_2}{Var(\tilde{\theta}_2|\tilde{p}_2)})}{1 + \frac{\sigma_2^2}{Var(\tilde{\theta}_2|\tilde{p}_2)}} \frac{a_{\theta,2}}{b_2 - a_{\theta,2}} \text{ and } c_2 \equiv \frac{\sigma_2^2 (\frac{a_{\theta,2}}{b_2 - a_{\theta,2}})^2}{1 + \frac{\sigma_2^2}{Var(\tilde{\theta}_2|\tilde{p}_2)}} = \frac{\tau_u \left(\tau_u \tau_\theta + \gamma(\tau_\eta + \tau_\theta)\sqrt{\frac{\tau_\eta \tau_\theta}{\tau_u(\tau_\eta + \tau_\theta)}}\right)^2}{\tau_\eta \tau_\theta^2 (\tau_\eta + \tau_\theta) \left(\gamma^2 + 2\tau_u(\tau_\theta + \gamma\sqrt{\frac{\tau_\eta \tau_\theta}{\tau_u(\tau_\eta + \tau_\theta)}})\right)}, \tag{A44}$$

and the second equation of c_2 is obtained by inserting the equilibrium α_2^* in (27).

Similarly, for an unskilled investor, her terminal wealth is $W_{U,2} = D_{U,1}(\tilde{\theta}_2 + \tilde{p}_2 - \tilde{p}_1) + D_{U,2}(\tilde{\theta}_3 - \tilde{p}_2)$. Inserting her date-2 demand $D_{U,2} = \frac{E(\tilde{\theta}_3|\tilde{p}_2) - \tilde{p}_2}{\gamma Var(\tilde{\theta}_3|\tilde{p}_2)}$ into the wealth, we obtain the unskilled investor's expected utility conditional on her information set at date 2 (i.e., $\tilde{p}_2$) as follows:

$$E\left[-e^{-\gamma W_{U,2}} \mid \tilde{p}_2\right] = -\exp\left[-\gamma D_{U,1}(\tilde{\theta}_2 + \tilde{p}_2 - \tilde{p}_1) - \frac{z_2^2}{2Var(\tilde{\theta}_3|\tilde{p}_2)}\right],$$

where z_2 is given by (A41). Using the property of z_2, we can compute the unskilled investor's expected utility conditional on her information set at date 1 (i.e., $\tilde{p}_1$) as follows:

$$\begin{aligned}
&E\left[-e^{-\gamma W_{U,2}} \mid \tilde{p}_1\right] \\
=&E\left[-\exp\left[-\gamma D_{U,1}(\tilde{\theta}_2+\frac{b_2}{b_2-a_{\theta,2}}a_{0,2}-\tilde{p}_1)-\gamma D_{U,1}\frac{a_{\theta,2}}{b_2-a_{\theta,2}}z_2-\frac{z_2^2}{2Var(\tilde{\theta}_3|\tilde{p}_2)}\right] \mid \tilde{p}_1\right] \\
=&-\frac{1}{\sqrt{1+\frac{\sigma_2^2}{Var(\tilde{\theta}_3|\tilde{p}_2)}}}\times\exp\left[-\gamma D_{U,1}\left(E(\tilde{\theta}_2|\tilde{s}_{p,1})+\frac{b_2}{b_2-a_{\theta,2}}a_{0,2}-\tilde{p}_1\right)+\frac{1}{2}\gamma^2 D_{U,1}^2 Var(\tilde{\theta}_2|\tilde{s}_{p,1})\right] \\
&\times\exp\left[-\gamma D_{U,1}\frac{a_{\theta,2}}{b_2-a_{\theta,2}}\mu_2-\frac{\mu_2^2}{2Var(\tilde{\theta}_3|\tilde{p}_2)}+\frac{1}{2}\frac{\sigma_2^2\left(\gamma D_{U,1}\frac{a_{\theta,2}}{b_2-a_{\theta,2}}+\frac{\mu_2}{Var(\tilde{\theta}_2|\tilde{p}_2)}\right)^2}{1+\frac{\sigma_2^2}{Var(\tilde{\theta}_3|\tilde{p}_2)}}\right], \qquad \text{(A45)}
\end{aligned}$$

where the second equation uses the property of z_2. Note that the conditional moments $E(\tilde{\theta}_2|\tilde{s}_{p,1})$ and $Var(\tilde{\theta}_2|\tilde{s}_{p,1})$ take the same forms as in (15)–(16) replacing α with α_1 and x with x_1. Maximizing the expected utility with respect to $D_{U,1}$ yields the optimal demand of the unskilled investor as follows:

$$D_{U,1}(\tilde{p}_1)=\frac{E(\tilde{\theta}_1|\tilde{s}_{p,1})+a_{0,2}-c_1-\tilde{p}_1}{\gamma\left[Var(\tilde{\theta}_1|\tilde{s}_{p,1})+c_2\right]}, \qquad \text{(A46)}$$

where c_1 and c_2 are given by (A44).

Step 2 In this step, we pin down the equilibrium at date 1. With the demand functions (A43) and (A46) for different types of investors, we can rearrange the market-clearing condition at date 1 and compute the implied date-1 price as follows:

$$\tilde{p}_1=a_{0,2}-c_1+\frac{A_1\tilde{s}_{p,1}+B_1(\tilde{\theta}_2+\tilde{\eta}_2)+\tilde{u}_1-1}{\frac{\lambda}{2\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{H,i,1})+c_2\right]}+\frac{\lambda}{2\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{L,i,1})+c_2\right]}+\frac{1-\lambda}{\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1})+c_2\right]}},$$

where

$$\begin{aligned}
A_1 \;=\;& \frac{\lambda}{2\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{H,i,1})+c_2\right]}\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\frac{\alpha_1^2\tau_u}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha_1^2\tau_u+x_1Z_H} \\
&+\frac{\lambda}{2\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{L,i,1})+c_2\right]}\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\frac{\alpha_1^2\tau_u}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha_1^2\tau_u+x_1Z_L} \\
&+\frac{1-\lambda}{\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1})+c_2\right]}\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\frac{\alpha_1^2\tau_u}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha_1^2\tau_u}, \\
B_1 \;=\;& \frac{\lambda}{2\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{H,i,1})+c_2\right]}\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\frac{x_1Z_H}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha_1^2\tau_u+x_1Z_H} \\
&+\frac{\lambda}{2\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{L,i,1})+c_2\right]}\frac{\tau_\eta}{\tau_\theta+\tau_\eta}\frac{x_1Z_L}{\frac{\tau_\theta\tau_\eta}{\tau_\theta+\tau_\eta}+\alpha_1^2\tau_u+x_1Z_L}.
\end{aligned}$$

That is:

$$a_{0,1} = a_{0,2} - c_1 - \frac{1}{\frac{\lambda}{2\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{H,i,1})+c_2\right]} + \frac{\lambda}{2\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{L,i,1})+c_2\right]} + \frac{1-\lambda}{\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1})+c_2\right]}}, \tag{A47}$$

$$a_{\theta,1} = \frac{A_1 + B_1}{\frac{\lambda}{2\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{H,i,1})+c_2\right]} + \frac{\lambda}{2\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{L,i,1})+c_2\right]} + \frac{1-\lambda}{\gamma\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1})+c_2\right]}}, \tag{A48}$$

$$a_{u,1} = \frac{a_{\theta,1}}{\alpha_1}. \tag{A49}$$

Thus, pinning down the equilibrium is equivalent to pinning down α_1. Again, $\alpha_1 = B_1$. After simplification, α_1 is implicitly determined by equation (31). It is clear that we have unique solution of α_1.

Calculation of G_1

We take two steps to compute the trading gain of skilled investors. In the first step, we calculate the ex-ante expected utility of investors. In the second step, we calculate skilled investors' trading gain.

Step 1 In this step, we first calculate the expected utility of investors conditional on $\tilde{p}_1$. For an unskilled investor, inserting $D_{U,1}$ in (A46) into the utility (A45) yields the unskilled investor's expected utility conditional on $\tilde{p}_1$ as follows:

$$E\left[-e^{-\gamma W_{U,2}} \mid \tilde{p}_1\right] = -\frac{1}{\sqrt{1+\frac{\sigma_2^2}{Var(\tilde{\theta}_3|\tilde{p}_2)}}} \times \exp\left[-\frac{1}{2}\frac{\mu_2^2}{Var(\tilde{\theta}_3|\tilde{p}_2)+\sigma_2^2}\right]$$
$$\times \exp\left[-\frac{1}{2}\frac{\left[E(\tilde{\theta}_2|\tilde{s}_{p,1}) + a_{0,2} - c_1 - \tilde{p}_1\right]^2}{Var(\tilde{\theta}_2|\tilde{s}_{p,1}) + c_2}\right]. \tag{A50}$$

Taking unconditional expectation of (A50) yields the ex-ante expected utility of the unskilled investor (V_U) as follows:

$$\begin{aligned} V_U &= -\frac{1}{\sqrt{1+\frac{\sigma_2^2}{Var(\tilde{\theta}_3|\tilde{p}_2)}}} \times \exp\left[-\frac{1}{2}\frac{\mu_2^2}{Var(\tilde{\theta}_3|\tilde{p}_2)+\sigma_2^2}\right] \\ &\quad \times E\left(\exp\left[-\frac{1}{2}\frac{\left[E(\tilde{\theta}_2|\tilde{s}_{p,1}) + a_{0,2} - c_1 - \tilde{p}_1\right]^2}{Var(\tilde{\theta}_2|\tilde{s}_{p,1}) + c_2}\right]\right). \end{aligned} \tag{A51}$$

For a skilled investor i, we insert her demand function $D_{i,1}$ in (A43) into the utility (A42) and obtain her utility conditional on her information set at date 1 (i.e., $(\tilde{p}_1,\ \tilde{s}_{i,1})$) as follows:

$$
\begin{aligned}
E\left[-e^{-\gamma W_{i,2}} \mid \tilde{p}_1, \tilde{s}_{i,1}\right] &= -\frac{1}{\sqrt{Var(\tilde{\theta}_3|\tilde{p}_2)/Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2})}}\frac{1}{\sqrt{1+\frac{\sigma_2^2}{Var(\tilde{\theta}_3|\tilde{p}_2)}}} \\
&\quad \times \exp\left[-\frac{1}{2}\frac{\mu_2^2}{Var(\tilde{\theta}_3|\tilde{p}_2)+\sigma_2^2}-\frac{1}{2}\frac{\left[E(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{i,1})+a_{0,2}-c_1-\tilde{p}_1\right]^2}{\left[Var(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{i,1})+c_2\right]}\right].
\end{aligned}
$$

Let $z_3=\frac{E(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{i,1})+a_{0,2}-c_1-\tilde{p}_1}{\sqrt{Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{i,1})+c_2}}$. Conditional on $\tilde{p}_1$, z_3 is a normal distribution with mean $\mu_3=\frac{E(\tilde{\theta}_2|\tilde{p}_1)+a_{0,2}-c_1-\tilde{p}_1}{\sqrt{Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{i,1})+c_2}}$ and variance $\sigma_3^2=\frac{Var\left(E(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{i,1})|\tilde{p}_1\right)}{Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{i,1})+c_2}=\frac{Var(\tilde{\theta}_2|\tilde{p}_1)-Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{i,1})}{Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{i,1})+c_2}$. Therefore, conditional on $\tilde{p}_1$, we have the following expression:

$$
\begin{aligned}
& E\left(\exp\left[-\frac{1}{2}\frac{\left[E(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{i,1})+a_{0,2}-c_1-\tilde{p}_1\right]^2}{Var(\tilde{\theta}_2|\tilde{s}_{p,1},\tilde{s}_{i,1})+c_2}\right] \mid \tilde{p}_1\right) \\
&= \frac{1}{\sqrt{\frac{Var(\tilde{\theta}_2|\tilde{p}_1)+c_2}{Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{i,1})+c_2}}}\exp\left[-\frac{1}{2}\frac{\left[E(\tilde{\theta}_2|\tilde{s}_{p,1})+a_{0,2}-c_1-\tilde{p}_1\right]^2}{Var(\tilde{\theta}_2|\tilde{s}_{p,1})+c_2}\right].
\end{aligned}
$$

Thus, the ex-ante expected utility of a skilled investor can be simplified as follows:

$$
\begin{aligned}
V_i &= -\frac{\sqrt{\frac{Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{i,1})+c_2}{Var(\tilde{\theta}_2|\tilde{p}_1)+c_2}}}{\sqrt{\frac{Var(\tilde{\theta}_3|\tilde{p}_2)}{Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2})}}}\frac{1}{\sqrt{1+\frac{\sigma_2^2}{Var(\tilde{\theta}_3|\tilde{p}_2)}}}\times\exp\left[-\frac{1}{2}\frac{\mu_2^2}{Var(\tilde{\theta}_3|\tilde{p}_2)+\sigma_2^2}\right] \\
&\quad \times E\left(\exp\left[-\frac{1}{2}\frac{\left[E(\tilde{\theta}_2|\tilde{s}_{p,1})+a_{0,2}-c_1-\tilde{p}_1\right]^2}{Var(\tilde{\theta}_2|\tilde{s}_{p,1})+c_2}\right]\right).
\end{aligned}
$$

Further, we can pin down the relations between the ex-ante expected utility of skilled investors (i.e., V_H for high-type skill investors and V_L for low-type skill investors) as follows:

$$
V_H=\frac{1}{\sqrt{Var(\tilde{\theta}_3|\tilde{p}_2)/Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{H,i,2})}}\frac{1}{\sqrt{\frac{Var(\tilde{\theta}_2|\tilde{p}_1)+c_2}{Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{H,i,1})+c_2}}}V_U,
$$

$$
V_L=\frac{1}{\sqrt{Var(\tilde{\theta}_3|\tilde{p}_2)/Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{L,i,2})}}\frac{1}{\sqrt{\frac{Var(\tilde{\theta}_2|\tilde{p}_1)+c_2}{Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{L,i,1})+c_2}}}V_U,
$$

where V_U is given by (A51).

Step 2 In this step, we show how the data vendor chooses the optimal date sample size x_1. Given that a skilled investor can be of a high type with probability $\frac{1}{2}$ and a low type with probability $\frac{1}{2}$, her ex-ante expected utility is $V_S=\frac{1}{2}V_H+\frac{1}{2}V_L$. The certainty equivalent of a skilled investor is thus:

$$
\begin{aligned}
CE_S &= -\frac{1}{\gamma}\ln(-V_S) \\
&= -\frac{1}{\gamma}\ln\left(\begin{array}{c} \frac{1}{\sqrt{Var\left(\tilde{\theta}_3|\tilde{p}_2\right)}\sqrt{Var\left(\tilde{\theta}_2|\tilde{p}_1\right)+c_2}} \\ \times\frac{1}{2}\left(\begin{array}{c} \sqrt{Var\left(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{H,i,2}\right)}\sqrt{Var\left(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{H,i,1}\right)+c_2} \\ +\sqrt{Var\left(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{L,i,2}\right)}\sqrt{Var\left(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{L,i,1}\right)+c_2} \end{array}\right) \end{array}\right) - \frac{1}{\gamma}\ln(-V_U).
\end{aligned}
$$

Following the analysis in the baseline model, the trading gain $G_1 \equiv CE_S - CE_U$. Since we have proved that $x_{H,2}Z_H = x_{L,2}Z_L$ at date 2, $Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{H,i,2}) = Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{L,i,2})$, G_1 can thus be simplified as follows:

$$
G_1 = -\frac{1}{\gamma}\ln\left(\frac{\sqrt{Var(\tilde{\theta}_3|\tilde{p}_2,\tilde{s}_{i,2})}}{\sqrt{Var(\tilde{\theta}_3|\tilde{p}_2)}}\right) - \frac{1}{\gamma}\ln\left(\frac{\frac{1}{2}\sqrt{Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{H,i,1})+c_2}+\frac{1}{2}\sqrt{Var(\tilde{\theta}_2|\tilde{p}_1,\tilde{s}_{L,i,1})+c_2}}{\sqrt{Var(\tilde{\theta}_2|\tilde{p}_1)+c_2}}\right),
$$

where c_2 does not depend on the data-sale choice of the data vendor at date 1 and is given by equation (A44).