Does Diversity Lead to Diverse Opinions? Evidence from Languages and Stock Markets*

Yen-Cheng Chang† Harrison Hong‡ Larissa Tiedens§ Na Wang¶

Bin Zhao‖

First Draft: July 1, 2013

This Draft: December 22, 2014

Abstract

Diversity of opinions among investors plays a crucial role in models of financial market speculation and bubbles. Yet, little is known about the origins of investor disagreement. Using unique data from China, we identify an important cultural and linguistic factor. We show that investors living in linguistically diverse areas express more diverse opinions on stock message boards and trade stocks more actively. We use geographical isolation of an area due to hilly terrain as an instrument for linguistic diversity. We then discriminate in favor of a differential interpretations mechanism and against slow news diffusion due to language barriers.

*We are grateful for the comments of discussants and seminar participants at the 2014 WFA, EFA, EFMA, FMA Asia, NBER Chinese Economy Working Group, 2013 CICF, Second Symposium on China's Financial Markets, 22nd Conference on the Theories and Practices of Securities and Financial Markets, National Chengchi University, Fudan University, Shanghai Advanced Institute of Finance, Soochow University, and Xiamen University.

†Shanghai Advanced Institute of Finance, Shanghai Jiao Tong University (e-mail: ycchang@saif.sjtu.edu.cn)

‡Princeton University, NBER, and China Academy of Financial Research (e-mail: hhong@princeton.edu)

§Stanford Graduate School of Business (e-mail: ltiedens@gsb.stanford.edu)

¶Hofstra University (email: na.wang@hofstra.edu)

‖Shanghai Advanced Institute of Finance, Shanghai Jiao Tong University (e-mail: bzhao@saif.sjtu.edu.cn)

1. Introduction

Diversity of opinions among investors is a crucial ingredient in explanations of financial market speculation and bubbles. For instance, it is used to rationalize the high levels of trading volume in financial markets that would be difficult to reconcile in classical models with homogeneous beliefs and portfolio rebalancing due to consumption needs (Varian (1989), Kandel and Pearson (1995) and Harris and Raviv (1993)). These models argue that investors endowed with heterogeneous priors or models lead them to interpret common public information in different ways, thereby generating diverse opinions and speculation. In conjunction with short-sales constraints, it is used to model the over-valuation of stock markets and speculative bubble episodes as pessimistic investors are forced to the sideline and the prices of stocks are determined by the most optimistic opinions (Miller (1977), Harrison and Kreps (1978), Morris (1996), Chen, Hong, and Stein (2002), Scheinkman and Xiong (2003)). Diversity of opinions also plays a central role in recent models of leverage and financial crises as pessimists lend to optimists, thereby exacerbating asset price bubbles (Geanakoplos (2010)).

While there is growing use of heterogeneous beliefs in modeling financial markets, we actually know relatively little about the forces which create diversity of views among investors in the first place. There are two standard measures of divergence of opinions in the literature for retail investors or households: the extent of investor stock trading (Barber and Odean (2001), Daniel, Hirshleifer, and Subrahmanyam (1998)) and the standard deviation of buy versus sell opinions measured from stock message boards (see, e.g., Antweiler and Frank (2004)).[1] While the literature has shown such measures to be predictive of asset price movements, it is silent on what proximate societal factors might magnify or shape the intensity of these disagreement measures. Yet, recent work on how cultural biases and personal histories (such as Guiso, Sapienza, and Zingales (2004) and Guiso, Sapienza, and Zingales (2009),

[1]The third widely used measure for professional security analysts or investors is the dispersion of their earnings forecasts (see, e.g., Diether, Malloy, and Scherbina (2002)).

Greenwood and Nagel (2009) and Malmendier and Nagel (2011)) and linguistic structure of past versus future tenses (such as Chen (2013)) affect investor risk-taking and savings suggest that these same biases and forces are likely to be important in influencing investor disagreement.

In this paper, we propose a diversity hypothesis for disagreement — namely, that the cultural diversity of a society generally, and linguistic diversity in particular, is likely to amplify disagreement among investors living in that society. Indeed, two main channels have been articulated through which diversity leads to disagreement (see Hong and Stein (2007)). The first mechanism is that diverse populations simply have more heterogeneity in priors or differential interpretations of the same news due to say different cultures or languages placing emphasis on different types of information. In addition, research from psychology in the field of diversity indicates that linguistic diversity in particular should lead to more varied and creative interpretations of signals. For instance, for students who have studied abroad, hearing foreign languages in experiments beforehand leads to better results on creativity tests such as story writing than students who have not studied abroad (Bialystok and Martin (2004), Maddux and Galinsky (2009), and Maddux, Adam, and Galinsky (2010)). Psychology studies of diversity all place emphasis on simply the awareness of linguistic or cultural diversity being stimulating even if people are not speaking the foreign language or from that culture per se.[2]

The second mechanism is that a diverse society speaking different languages means that news will naturally have a harder time traveling across the population, leading thereby to different information sets and opinions. These two channels might also be present using other forms of diversity such as ethnic or religious, though less obviously so.

The key prediction of this diversity hypothesis is that if we take two populations of investors, one in a high and the other in a low linguistic diversity area, we expect to see

[2]Consistent with this view, seven-month old babies exposed to bilingualism at home are better able to anticipate, using eye-ball tracking tests, movements of a puppet across the television screen (Kovacs and Mehler (2009)).

more disagreement among investors in the high diversity area, where an area might be a city, a state or a country. The challenge of course in testing our diversity hypothesis is that diverse populations tend to live in denser, richer and more financially developed areas due to immigration and other selection factors. So investors living in diverse cities might simply be more exposed to financial information and disagree more as a result of this exposure.

As such, we turn to a unique setting in China to test our hypothesis using new linguistic measures of diversity across Chinese provinces and standard measures of investor disagreement. In the context of China, there is little ethnic and religious diversity to begin with and more importantly little variation in these two measures of diversity across provinces.[3] However and fortunately, even though China has the same written language, there is significant linguistic diversity in spoken languages that we can use. Linguists often characterize China as being as linguistically diverse as all the countries in Europe combined (see, e.g., Ramsey (1987)).[4]

Our independent variables of interest are the number of languages spoken across the 30 provinces or across the roughly 300 cities, which we hand-collect using the Language Atlas of China (1987). We then define our linguistic diversity measures for each province as the number of languages spoken in that province (LD). We also construct a variety of diversity measures based on sub-languages or sub-dialects, sub-sub-languages and sub-sub-sub languages (i.e., branches within each language or LD-SUB1, LD-SUB2 and LD-SUB3). Our linguistic diversity measures are meant to capture the native languages spoken in the different provinces. People living in diverse areas like Wenzhou, a city in Zhejiang, just speak one of the many dialects but they are aware of the diversity in languages and cultures. In other words, even for younger adults, that they grew up with people who spoke multiple

[3]According to the 2010 Population Census of China, 91.6% of people in China are Han with little variation of ethnicity proportions across Chinese provinces. According to the China Family Panel Studies by Peking University, 89.56% of people in China respond as having no religious affiliation.

[4]The persistence of local languages at the city or province level is recognized by the Chinese themselves in the saying, "The furthest distance between people is two Shanghainese meeting together but talking to each other in Mandarin." This phrase simultaneously captures both the influence of local languages for social interactions even in today's China and the extent to which linguistic diversity is a good measure for whether a city or province is diverse.

languages would be influential even if they do not speak the languages themselves. The Chinese Hukou system which strictly limits mobility across China tends to then make these diversity differences persist over time.[5]

To construct our first measure of investor disagreement, we have account level trading data from one of the largest brokerage houses in China for roughly 300,000 households across 21 provinces over the period of 2006-2012. We also know the city where each of the households resides and can therefore measure the extent of stock trading by households across different cities or provinces in China. Our measures of stock trading are number of shares traded or value of the trades.

To construct our second measure of investor disagreement, we use one of the largest and most active message boards in the world, guba.eastmoney.com, with close to 24 million messages during the period of 2008-2012. Eastmoney is part of one of the largest brokerage houses in China and covers over 1,500 of the largest stocks in China. Following the literature, we construct a standard measure of disagreement using textual analysis and machine learning (see, e.g., Mehl (2006)) on the messages posted. The opinions in each post is coded with -2, -1, 0, 1, 2, denoting Strong Sell, Sell, Neutral, Buy, and Strong Buy using a Naïve Bayes' method for text classification similar to Antweiler and Frank (2004). We then calculate for each stock its disagreement measure—the standard deviation in these scores across the posts of the stock; this is our first dependent variable of interest. We know the city and province, through a computer's IP address, that the message originates from, and hence we can calculate the disagreement among messages emanating from different cities or provinces. Given the size of the message board, we focus on eight provinces: four with high and four with low linguistic diversity measures.

Our baseline OLS regression specifications regress these two stock market diversity of opinion measures across provinces on the linguistic diversity of provinces. We find econom-

[5]The fact that we can get such linguistic diversity under one legal system also makes China especially interesting to consider how diversity affects various outcomes without having to deal with cross-country heterogeneity.

ically and statistically significant results across a number of specifications that control for a host of household characteristics and demographic factors such as GDP per capita, population density and social capital or trust measures along the lines of Guiso, Sapienza, and Zingales (2004). For instance, for the number of shares traded measure, the coefficient in front of LD-SUB3 is 0.234 with a t-statistic of 2.74. The economic significance is around 5% of a standard deviation of the left-hand side variable for a given one-standard deviation increase in the various measures of diversity. This economic magnitudes are comparable in strength to individual household characteristics such as wealth and demographic factors such as GDP per capita in explaining the cross-section of individual trading behavior. Using stock message board divergence of opinion, we get similar results in that an increase in diversity is associated with an increase in disagreement that is around 10% of its standard deviation. But the statistical significance is weaker, due in part to fewer provinces in our sample of analysis.

What is most interesting about China's linguistic diversity and useful from our perspective of identification is its geographic origins, which is well known to linguists and which we will use as an identification strategy.[6] The north of China, including provinces like Beijing, Shandong, Liaoning, is flat and desert like. Linguists believe the easy travel across the flat lines led to the use of the same language. In contrast, the south of China, including provinces like Fujian and Zhejiang, is hilly and watery and as a result made travel more difficult and so more languages developed.[7]

We then use geographic isolation of an area as an instrument for the linguistic diversity of the province. Our exclusion restriction is that an area's hilliness is uncorrelated with disagreement in stock markets other than through linguistic diversity. We argue below that this assumption is a plausible one since technology and government policies have fostered development of cities both in GDP and population that is independent of the hilliness of an

[6]See for example Chapter 2 of Ramsey (1987) on the geographic origins of the Chinese languages.

[7]Indeed, the Chinese, of course, also have another pithy phrase for the origins of their linguistic diversity: "in the south the boat, in the north the horse".

area. Given the technology to trade and read message boards are available throughout China, there is no reason to think that the hilliness of an area would influence investors' opinions other than through this linguistic diversity channel. We are in essence using geographic isolation which influenced persistent linguistic and cultural heterogeneity (but not financial or population development population) to identify our effect.[8] Indeed, our analysis can also be implemented with Administrative Area fixed effects. Administrative Areas are like US Census Divisions. So our effects are not simply a function of Northern versus Southern China per se as there is heterogeneity in hilliness across different provinces within the same Administrative Area.

The first-stage regression of the various diversity measures on an area's hilliness indeed yields significant coefficients on hilliness. Depending on our samples of analysis (the 21 provinces from the trading activity analysis or the eight provinces from the message board analysis), the t-statistics range anywhere from around 5 to 40. Importantly, hilliness is by far the strongest and consistent explanatory variable for linguistic diversity when compared with other demographic factors such as GDP or population. The second-stage of the 2SLS yields economically and statistically significant coefficients in front of linguistic diversity that are significantly improved from the OLS results. This improvement, typically around 50-75% of the OLS results, is consistent with our linguistic diversity measure having some measurement error, which would tend to bias down the OLS estimates. Then combining or projecting LD on hilliness produces a less noisy measure, which would then increase the size and significance of our estimates.

We then distinguish between the two potential mechanisms for our findings. The set-up in Hong and Stein (1999) offers two testable predictions for the slow news diffusion mechanism. The first is that there should be more drift in prices in the stocks predominantly traded among investors from more diverse areas. The second is that there should be scope for

[8]Linguists have proposed to a lesser degree latitude and temperature as also influencing linguistic diversity. As we discuss in more detail below, these indeed turn out to be less powerful explanatory variables in China compared to geographic isolation (see, e.g., Nichols (1990), Nettle (1996), Stepp, Castaneda, and Cervone (2005)), Maffi (2007)).

trading profits from the slow news diffusion—since news travels slowly and is only gradually incorporated, households who receive the news will have some returns since markets are not efficient. We find neither.

Instead, we find that households in diverse areas are more likely to lose money from their trades — that is they engage in excessive trading a la Odean (1999). Related, we find that households in diverse areas traded much more excessively than households in less diverse areas during the Chinese Stock Market or Olympics Bubble of 2006-2007 (Xiong and Yu (2011)). These results are more consistent with a differential interpretations view especially in light of previous research on trading of Internet stocks. It is not that news travels slowly but there is too much stimulation which in the presence of overconfidence traders leads to excessive trading.[9]

Finally, we discuss alternative explanations and provide an alternative empirical strategy using province fixed effects and city differences. That is, rather than having provinces be the unit of analysis, we focus on cities. This is in the spirit of the Administrative Area fixed effects but now drilling down even more finely within a province. The downside is that data on cities is less clean and we do not have geographic information at the city level. Nonetheless, it is comforting that this alternative within-province identification strategy yields similar results. We get very similar economic effects when we use these two identification strategies though the province fixed effects set-up yields smaller coefficients.

Our paper proceeds as follows. Section 2 describes the data and key variables of interest. In Section 3, we present our OLS and 2SLS results on linguistic diversity of provinces and investor disagreement. In Section 4, we distinguish between competing mechanisms for our findings. In Section 5, we show that our results are robust using an alternative identification strategy using cities and province fixed effects. We conclude in Section 6.

[9]During the Internet Bubble, for example, disproportionately more media attention was paid to internet stocks than to non-internet stocks (Bhattacharya et al. (2009)), which Hong and Stein (2007) argued might have led to more excessive trading in Internet stocks relative to others due to diversity of opinions.

2. Data and Variables

2.1. Linguistic Diversity Measures

We use two alternative sources that employ similar survey methodologies of linguistics across China to measure LD. The first source is the Language Atlas of China (1987, hereafter the Atlas), a collaborative work by the Australian Academy of the Humanities and the Chinese Academy of Social Sciences. The Atlas is the first comprehensive survey of the Chinese languages and has become an authoritative reference for many following studies in Chinese linguistics. The second source is the Linguistic Atlas of Chinese Dialects (2008), a more recent work published by the Beijing Language and Culture University (BLCU). The surveys in the Atlas were conducted from 1983 to 1987, while the BLCU study was done from 2001 to 2007.

In these studies, survey posts are set up at the county level. Phonetic and syntax features are then used to determine the classification of the interviewees' language and to assess whether it represents a stable language at each location. Each language is then given a list of counties in which it occurs, along with examples of its phonetic and syntax features. In this paper we focus primarily on the linguistic variation among ethnic Han, thus our diversity measure includes only the languages in the Sinitic language branch under the Sino-Tibetan language family.

Importantly, the surveys are designed to capture language occurrences among the indigenous population. Typically, the surveys sample senior local residents with age over 60 that have not lived out-of-town extensively. Province-level LDs from these two studies are very similar with a correlation coefficient of 0.95. Across these two studies, province-level LDs only differ slightly in three provinces, Fujian, Guangxi, and Hunan. In our empirical analyses, we focus on LD derived from the Atlas, though in robustness tests (not reported for brevity) we verify the results using LD by BLCU are similar. Even though these suveys were conducted almost 15 years apart, the high correlation points to the persistence of linguistic

diversity. This is likely in part due to the household registration (Hukou) system in China, which greatly restricts demographic mobility. In China, one would not be able to receive housing, health care benefits, or support on children's education if she works outside her registered location. This system has been in effect since 1949 and has only started to go through limited reform in recent years.

The Atlas identifies ten unique languages in the Sinitic language branch. These ten languages are Gan, Guan, Hui, Jin, Kejia, Minyu, Ping, Wu, Xiang, and Yue. Note that there are two other languages, Tu and Xianghua, where the Atlas indicates that it has not been able to properly classify them into the Sinitic language group. We therefore drop these two languages in our analyses. We also do not include minority languages.

Each language has a hierarchical structure and can be further classified into finer sub-languages (or sub-dialects). Following the Atlas, we can further classify each language into level 1, level 2, and level 3 sub-languages. Specifically, the Atlas lists the geographic coverage of each language or sub-language using different administrative levels such as cities, counties, villages, or townships. From the Atlas, we are able to identify 2,010 unique locations.

In order to merge the language-location pairs with our stock market data, we need to merge these 2,010 locations into today's prefecture-level city in China. Therefore, we manually update all location names and identify administrative changes throughout the years. Finally, we are able to identify 329 prefecture-level cities in 30 provinces. We also obtain from the National Bureau of Statistics of China the GDP per capita for the different provinces and cities over the sample period and the population of each city in each province.

In Panel A of Table 1, we report by province the different languages spoken in the 30 provinces of China and the total number of unique languages spoken in each province, denoted by LD, which is our main measure of linguistic diversity across Chinese provinces. The province language is simply the union of the number of languages spoken in the various cities in a province. The results are sorted by log GDP per capita (Log(GDP)). We also show linguistic diversity measures using different levels of sub-languages (LD-SUB1, LD-

SUB2, and LD-SUB3).

We show LD in Panel A of Figure 1 as a heat map of the number of languages spoken in Chinese provinces: the darker the color the more linguistic diverse the province. Guan, which is Mandarin, is spoken in the largest number of provinces. For instance, Beijing only speaks Guan as do a number of other provinces in northeast China such as Jilin, Tianjin and Shandong. These provinces all lie on the northeastern part of China by the coast and as we show below are very developed and relatively prosperous by comparison to provinces in the interior of China. In the southeast part of China, the provinces such as Fujian, Zhejiang, and Guangdong are equally prosperous but speak more languages. Fujian has four languages, while Zhejiang and Guangdong each has three.

The distribution of these provinces along the coast will help us deal with unobserved heterogeneity due to economic or financial development. We are fortunate that government policies have favored development of the eastern coastline as opposed to the interiors of China. As one moves west, there is less and less economic development. We want to make sure that we are not capturing government policies with our linguistic diversity variable. Fortunately, the linguistic diversity of China largely runs north to south. This can be seen in Panel A of Figure 1. Similarly, notice from Panel A of Table 1 that for higher Log(GDP) provinces, there is variation in linguistic diversity from one language to as many as three languages.

Note that Mandarin (Guan) is not included in four provinces (Fujian, Guangdong, Hainan, Shanghai) in both studies. The reason is that China focused on the simplification of written Chinese from 1949 to the early 1980s. The promotion of Guan was not its priority during this period and both efforts were suspended during the Cultural Revolution from 1966 to 1971. Efforts to promote Guan were not in effect until the 1980s. The policy of promoting Guan was officially written into the P.R.C. constitution in 1982 and the official table of Mandarin pronunciation was published in 1985 which provided the founda-

tion of Guan promotion.[10] In particular, for these four provinces that do not include Guan, the State Language and Letters Committee (highest governing body for language reform in China) continue to list them as their top priority for Mandarin promotion into the 1990s. This speaks to the large differences between Guan and southern languages in China, and confirms the accuracy of our LD measures.

2.2. Geographic Isolation and Other Demographic Measures

In Panel A of Table 1, we also report our geographic isolation measure; i.e., the fraction of terrain that is hilly (Hill) in each province, obtained from the Thematic Database of Human-Earth System. We show the percent of hills as a heat map in Panel B of Figure 1: the darker the color the more hills there are. We can see from comparing Panel A and B of Figure 1 that provinces with a darker color in Panel A for the number of languages also have a darker color in Panel B for the percent of hills. This supports the presumption alluded in the introduction that linguistic diversity in China is largely driven by geographic isolation, which we will formally show below.

In addition, Log(GDP), the log GDP per capita, and Log(Pop), the log population, are obtained from the National Bureau of Statistics. To the extent that linguistic diversity is affected by government economic policies or simply a reflection of the size of population in a province, it is important to control for these variables. Branch is the number of brokerage branches divided by the population in each province, provided by the Shanghai Stock Exchange. PLocal is from CSMAR and defined as the number of stocks headquartered in each province divided by the number of all stocks in the country. We use Branch and PLocal to control for the effects of trading technology advantage and local bias on trading activity.[11] Trust is the log of trust index constructed by Zhang and Ke (2003), based on surveys on corporate senior managers to rate the trustworthiness of firms in different provinces. Guiso,

[10] The Authorized Table of Mandarin Words with Variant Pronunciations (1985).

[11] For time-varying variables (Log(GDP), Log(Pop), Branch, PLocaL), we take the average over our brokerage house sample period 2006 to 2012.

Sapienza, and Zingales (2009) show that trust can affect economic exchange. It turns out that the correlation of trust and diversity is slightly negative—so there is as one might expect less trust in linguistically heterogeneous populations—but the variable does not discernably influence our results.[12]

Importantly, from Panel B of Table 1, we can see that the correlation of Hill with these demographic factors is typically low, all less than 0.25 in absolute terms. For instance, the correlation of Hill with Log(GDP) is -0.14 and for Log(POP), it is 0.08. This is consistent with our premise that economic and population development are independent of a province's Hilliness.

We also use Administrative Area dummy variables that group adjacent provinces together with similar geographic and climate features. All provinces are categorized into seven Administrative Areas: Northeast, Northern China, Northwest, Eastern China, Central China, Southwest, and Southern China.[13] These official areas are commonly used for geographic studies, weather forecasts, military districts, etc.

2.3. Measuring Diversity of Opinions with Brokerage House Trading Data

Our first measure of investor disagreement is individual household trading activity. We measure the trading activity of individual investors using brokerage house stock trading data, provided by a nationwide discount broker in China. Around 300,000 individual investors were randomly selected from the broker's existing accounts at the end of 2012. The sampled investors come from 70 cities across 21 provinces in China. The large geographic coverage

[12]One might have two interpretations of trust in the context of the stock market. The first is that more trust means more trade in the stock market. In this instance, including trust would make our results stronger. The other is that more trust means people trade more bilterally but not on an exchange. Including trust would make your results presumably weaker.

[13]Northeast includes provinces of Heilongjiang, Jinlin, and Liaoning; Northern China includes Beijing, Tianjin, Heibei, Shanxi, and Inner Mongolia; Northwest includes Ninxia, Gansu, Shaanxi, Qinghai, and Xinjiang; Eastern China includes Shandong, Jiangsu, Anhui, Shanghai, Zhejiang, Jiangxi, and Fujian; Central China includes Henan, Hubei, and Hunan; Southwest includes Sichuan, Chongqing, Guizhou, and Yunnan; Southern China includes Guangdong, Guangxi, and Hainan.

is particularly opportune since it provides us with enough variation in linguistic diversity in investors' home location.

The data consists of an account file and a trade file. The account file includes the customer code, gender, age, and total account value at the end of the sample period. It also contains the city and province where each of the investors resides. The trade file contains the records of all trades made by these investor accounts from 2006 to 2012. Each trade record has a customer code, a security code, a buy-sell indicator, and the price, quantity, and date of the trade.

To test if linguistic diversity affects a household's trading activity, we compute two trading activity measures for each investor in each year t over our sample period. Log(Share$_{ijt}$) is the log number of shares of all stocks traded by investor i from province j during year t. Similarly, Log(Value$_{ijt}$) is the log trading value (calculated using the buying and selling prices) of all stocks traded.[14]

In Panel A of Table 2, we report the summary statistics of our trading sample. The mean value of Log(Share) is 10.66 and the standard deviation is 4.39. The corresponding numbers for Log(Value) are 12.87 and 4.94. Log(Wealth) is the log total value of an investor's brokerage account at the end of the sample period. The mean Log(Wealth) in our sample is 11.08 Yuan and the standard deviation is 2.88 Yuan. Sex is a dummy variable, which takes the value of 1 for female investors, and the value of 0 for male investors. Our sample consists of roughly 44% female. Log(Age) is the log age of each investor measured at the end of the sample period. The mean Log(Age) is 3.82 with a standard deviation of 0.26.

To better control for the types of stocks investors' trade, we also include the volatility (Vol) and stock size (Size) of trades of each investor. Vol is the value-weighted average volatility of an investor's traded stocks during each year. Size is the value-weighted average firm market capitalization deciles of an investor's traded stocks during each year. The mean values of trading volatility and size are 0.13 and 6.46, respectively.

[14] Alternatively, we also use log number of trades executed by each investor as a measure of trading activity. The results are robust and we omit these for brevity.

Importantly, note that since our brokerage data consists of a smaller set of 21 provinces, it is informative to compare the overall sample in Table 1. Comparing economic development measures, the brokerage sample has an average Log(GDP) and Log(Pop) of 10.86 and 3.75. These numbers are 10.17 and 3.55 in Table 1. Therefore we interpret the brokerage sample as largely representative and comparable to the full sample.

2.4. Measuring Diversity of Opinions with Stock Message Board

Our second measure of disagreement for each stock is the degree to which posts disagree over whether to buy or sell the stock. We download guba.eastmoney.com messages for all firms headquartered in eight provinces: four provinces with LD>1 (Guangdong, Hunan, Fujian, Zhejiang), and four provinces with LD=1 (Shandong, Sichuan, Beijing, Shanghai). For each firm we download the most recent 10 pages of messages. The sample consists of a total 796,809 message posts.[15]

To form a training sample for applying machine learning methods over the whole sample, we select the most recent 20 messages from a random sample of 30 firms from each province. We use standard textual analysis method from social psychology (see, e.g., Mehl (2006)).[16] The opinions in each post are coded manually with -2, -1, 0, 1, 2, denoting Strong Sell, Sell, Neutral, Buy, and Strong Buy.

Based on the training sample, we can use machine learning techniques to systematically classify all messages from the downloaded sample. Similar to Antweiler and Frank (2004), we use a Naïve Bayes' method for text classification using Weka, a machine learning software developed by the University of Waikato, New Zealand. Conceptually, we compute the conditional probability of the direction of each message given the words in a message. According

[15]The structure of guba.eastmoney.com message boards makes downloading the time stamp for each message infeasible. Therefore in following tests we aggregate all message data from 2008 (when Guba message board started) to May of 2012 (when we downloaded the data).

[16]In this procedure, two graduate students are hired to independently classify the messages. Their classifications are 0.8 correlated, which is considered high enough for us to trust their output. We then have them resolve any discrepancies to come up with a final classification which we use as a training sample to calibrate our algorithm.

to Bayes' rule,

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{1}$$

where y is the direction of each message, and x is the message consisting of a sequence of words. We can measure each term on the right-hand-side of Equation (1) using the training sample. $P(y)$ is the unconditional probability of the direction of messages. The Naïve Bayes' method assumes the occurrences of words and phrases are independent of each other. Therefore, $P(x|y)$ is simply the product of the conditional probabilities of each word in x given y, and $P(x)$ is the product of the unconditional probabilities.

Unlike English where the meaning of each word is usually self-contained, Chinese words typically contain different numbers of characters to carry their meaning. Therefore we use a Chinese sentence splitter software *fundannlp* to first retrieve the key words in all the messages. For our analysis, we include text in both the subject line and the content of the message by the original poster. With *fundannlp*, we can also determine the lexical categories of each word. We keep only nouns, verbs, and adjectives. For example, if a message says "Tesla stock is going to rise," the sentence splitter will then keep "Tesla", "stock", and "rise" and assign the lexical categories to each word. The sentence splitter is critical for Chinese because "Tesla", "stock", and "rise" can be expressed with one to up to three characters.

In Table 3, we show the key words with the top five highest conditional probabilities (i.e., $P(y|x)$) from the training sample. The top key words with the highest probabilities of being associated with a Strong Sell message are "bad", "dive", or "to empty ones positions." On the contrary, the top key words being associated with a Strong Buy message are "in a leading position", or "full positions." Applying Bayes' rule, we can then compute the conditional probabilities of each message being categorized as -2, -1, 0, 1, and 2. The predicted value for each message is then the buy/sell signal with the highest conditional probabilities.

Our second dependent variable of interest is then the log standard deviation (Log(STDEV)) in the generated buy/sell signals. Our message board data allows us to pinpoint, for each

message, its originating province and the stock it is discussing. We can then compute Log(STDEV$_{ik}$) for each stock i-province k pair and see if the diversity of opinions is indeed high for messages from high linguistic diversity areas, controlling for the stock in question.

Panel B of Table 2 reports the summary statistics for our second dependent variable of interest. Notice that the standard deviation of Log(STDEV) is 1.82, while the standard deviation of Log(LD) is .597. We will use these two quantities to calculate the economic significance of the relationship between diversity and divergent opinions. We also report the corresponding summary statistics for the other key independent variables using the subsample of provinces corresponding to the message board posts.

3. Linguistic Diversity and Differences of Opinions

3.1. OLS Estimates

We start by regressing household trading activity on linguistic diversity measures. The regression specification is as follows:

$$Trade_{ijt} = \alpha + \beta Log(LD_{ijt}) + \gamma' \mathbf{X} + \epsilon_{ijt}, \tag{2}$$

where Trade$_{ijt}$ is the trading activity of investor i from province j in year t, as measured by Log(Share$_{ijt}$) or Log(Value$_{ijt}$). Log(LD$_{ijt}$) is the log number of languages spoken in investor i's home province j in year t, and $\mathbf{X}$ consists of province and investor characteristics, Administrative Area dummies, and year dummies. The primary coefficient of interest is β, which identifies the effect of linguistic diversity on trading activity. Province characteristics include Trust, Log(GDP), Log(Pop), Branch, and PLocal. We further control for the following investor characteristics, including wealth, sex, age, the volatility of the stocks held and the size of the stock held. We cluster standard errors at the province level.

Table 4 presents our OLS regression results. The key variable of interest is linguistic

diversity and we show the coefficients in the first row. For each linguistic diversity measure, we show first results using Log(Share) as the measure for trading activity and then Log(Value). Column (1) uses Log(Share) as the measure for trading activity and Log(LD) as the linguistic diversity measure. Being in a more linguistically diverse environment increases trading activity as the coefficient in front of the Log(LD) is 0.273 with a t-statistic of 1.65. We get similar result when using Log(Value) as the dependent variable in Column (2). In Columns (3)-(8), we report analogous results using different levels of sub-languages and obtain similar results. For instance, the coefficient of Log(LD-SUB2) in Column (6) is 0.208 with a t-statistic of 2.53. Finally, in columns (9) and (10) we replace Log(LD) with the fraction of hilly area (Hill) in the message originating province. This is essentially the reduced-form regression of the 2SLS that we will run next, where we use geographic isolation as an instrument for linguistic diversity. The coefficient for Hill is a significantly positive 0.63 with a t-statistic of 4.11. In short, being in a more linguistically diverse environment increases trading activity.

The effects of other investor and province characteristics are also worth mentioning. Consistent with findings in the behavioral finance literature, we find that wealthier, male, and older investors are associated with higher levels of trading activity. Stock trading volatility, size, local GDP and brokerage branch density are positively related to our dependent variables. Interestingly, trust comes in with a negative sign. Nevertheless, our coefficients of interest remain significantly positive even after adding all these controls.

To put the economic significance of our coefficient estimate in some context, a one standard deviation increase in linguistic diversity translates to around a 5% increase of the left-hand side variable as a fraction of its standard deviation. As a comparison, this economic magnitude is comparable to those implied from log(GDP) or Branch density as well as a number of the other demographic measures. It is somewhat smaller than household characteristics such as Sex or wealth which get implied economic magnitudes more on the order of 10-15%. Nonetheless, among demographic factors, linguistic diversity plays as big

a role as well-known demographic factors.

We next turn to our analysis of message board disagreement. Our baseline OLS regression specification with LD is as follows:

$$Log(STDEV_{ik}) = \alpha + \beta Log(LD_{ik}) + \gamma' \mathbf{X} + \epsilon_{i,k}, \tag{3}$$

where Log(STDEV$_{ik}$) is the standard deviation in the buy/sell signals on stock i originated from province k. Log(LD$_{ik}$) is the log number of languages spoken in province k. $\mathbf{X}$ consists of province characteristics controls as defined in Panel B of Table 2, including Trust, Log(GDP), Log(POP), Branch, PLocal, Administrative Area dummies and firm fixed effects. The primary coefficient of interest β then identifies if linguistic diversity affects diversity of opinions, controlling for both province-level characteristics and the firms which the messages are discussing. We also cluster standard errors at the province level.

Table 5 reports the OLS coefficients. Column (1) shows that, using Log(LD) as the linguistic diversity measure, the estimate of β is 0.284 with a t-statistic of 2.77. Notice that the standard deviation of Log(STDEV) is 1.82, while the standard deviation of Log(LD) is 0.597. So a one standard deviation increase in our RHS side variable leads to around a 9.3% increase in the LHS variable as a fraction of its standard deviation. Columns (2) to (4) use different sub-language levels and show consistent results. For instance, using Log(LD-SUB1) yields a coefficient estimate of 0.166 with a t-statistic of 2.43. But the coefficients in columns (3)-(4) are not statistically significant. However, this will change when we consider the 2SLS below. Finally, in column (5) we replace Log(LD) with the fraction of hilly area (Hill) in the message originating province. This is essentially the reduced-form regression of the 2SLS that we will run next, where we use geographic isolation as an instrument for linguistic diversity. The coefficient for Hill is a significantly positive 1.023 with a t-statistic of 2.69. Overall, the economic significance from these specifications imply a one standard deviation increase in linguistic diversity leads to a to 9% increase of our left-hand side variable as a

fraction its standard deviation.

It is interesting to then compare these estimates to those in front of the other demographic factors. Log(GDP) does not come in significantly but higher population density is associated with more disagreement as we had suggested in the Introduction. Branch shows significant effects with higher brokerage branch density being associated with more disagreement. In contrast, higher PLocal is associated with less disagreement. But the economic magnitudes of these estimates are similar to our diversity measure, pointing to the importance of diversity in explaining disagreement.

3.2. 2SLS Estimates

In Table 6, we then consider the full instrumental variables estimation for our OLS trading activity regression where we instrument a province's diversity with the hilliness of that province. Our exclusion restriction is that an area's hilliness is uncorrelated with disagreement in stock markets other than through linguistic diversity. This assumption, as we have argued in Section 2, is plausible since the economic or population development of a province is independent of a province's hilliness.

Panel A shows the results from the first stage regression. Here we regress the log of linguistic diversity measures on Hill and include all the controls from the OLS specifications. The coefficients of Hill are significantly positive and all the t-statistics are anywhere from 9.5 to 41. In particular, using Hill as an instrument for Log(LD-SUB2) in Column (3), the coefficient of Hill is 2.382 with a t-statistic as large as 41.58. All regressions in Table 6 cluster standard errors at the province level. Overall, results in Panel A again suggests that Hill is a good instrument.

We then present the 2SLS regression results in Panel B of Table 6. Using Hill as the instrumental variable, the effects of linguistic diversity on trading activity are stronger than our OLS estimates. For example, the coefficient of Log(LD) is 0.655 in Column (1) with a t-statistic of 3.67. In Column (6), the coefficient of Log(LD-SUB2) is 0.276 with a t-

statistic of 4.27. These coefficients are larger with stronger statistical significance than their analogs in the OLS results. The economic effects are around 50 to 75% larger than the OLS numbers. The improvement in the 2SLS figures suggests that measurement error in linguistic diversity measures were biasing down the OLS estimates. So instrumenting with hills would then help ameliorate this measurement error and improve both the economic and statistical significance. The coefficients in front of the other covariates are largely aligned with their OLS figures. So our conclusions from the OLS discussion largely carry over with the added caveat that the economic significance of diversity has increased with 2SLS.

We next turn to the 2SLS estimation of the message board disagreement measure in Table 7. The first stage (Panel A) is log of linguistic diversity measure of the message originating province on the fraction of hilly area in that province, controlling for the same set of other covariates as in the OLS. For Log(LD), the coefficient on Hill is 1.772 with a t-statistic of 3.15. This is the weakest of our first stage regressions. But note that the t-statistics for Hill across all the other diversity measures are all much higher, from 4.25 to 5.61. All standard errors are clustered at the province level. As such, we are not concerned about a weak instruments problem.

The second stage results are presented in Panel B and they all show consistent results for different levels of languages and sub-languages. For example, using Log(LD-SUB1) as the linguistic diversity measure, the β coefficient is 0.303 with a t-statistic of 3.01. Overall, the coefficients of interest in Panel B are larger in magnitude and come with stronger statistical significance than our baseline OLS from Table 5. This improvement is again likely due to the measurement error of LD and variants of LD. The other co-variates generally have similar signs to the OLS set-up though their significance weakens somewhat in these 2SLS regressions. As such, diversity is clearly a very significant demographic factor for explaining disagreement.

4. Distinguishing Between Mechanisms: Differential Interpretations and Slow News Diffusion

4.1. Diversity and Post Earnings Announcement Drift

Now that we have established the causal effects of linguistic diversity on our two divergent opinion measures, we turn to distinguish among the two mechanisms behind our findings. Under the slow news diffusion mechanism, residents from different linguistic groups have different information available to them and this in turn leads to divergence of opinions. One implication of this slow information diffusion mechanism modeled in Hong and Stein (1999) is that stock prices of firms predominantly traded by investors in high linguistic diversity areas, or local firms headquartered there, will react slower to news, leading to serial correlation in prices. We use post-earnings-announcement-drift (PEAD) to test this. PEAD has been documented since at least Bernard and Thomas (1989), who attribute this phenomenon to investors' delayed response.

We report the PEAD results in Table 8. First we use all firms in the message board sample and compute, for each firm, the average LD of all messages. Firms with high LD implies that most of their investors are from high linguistic diversity provinces and news travels slower for these stocks. The slow information diffusion mechanism then implies that these firms will have stronger PEAD. Second, to measure earnings surprise, we independently sort our sample firms by their earnings announcement return (EAR) every year. EAR is measured by the cumulative market-adjusted return over a three-day window around the annual earnings announcement. E1 (E5) is the average cumulative abnormal return of the stocks with the lowest (highest) EAR, and E5-E1 is the average return of the high minus low PEAD strategy. We report various return windows from [2,6] to [2,51]. T-statistics are reported in parentheses. The sample period is from 1998 to 2012.[17]

Comparing PEAD returns (E5-E1) between the low linguistic diversity group and the

[17]The sample period from 2006-2012 yields similar results.

high linguistic diversity group does not show visible differences. We formally report the magnitude and statistical significance of this difference in the last row of Table 8. The differences in PEAD returns are small and the t-statistics are generally less than 1 across all windows. Using the [2,21] window as an example, PEAD of low (high) linguistic diversity stocks is 0.4% (1.3%). The difference in PEAD in this case is 0.9% with a t-statistic of 0.51. We conclude that there is little evidence that PEAD is stronger among firms that are traded predominantly by investors in high linguistic diversity areas. In robustness tests (not reported for brevity), we find that the results are similar using sub-languages, or instead sorting firms by the linguistic diversity of their headquartered provinces.

4.2. Diversity and Excessive Trading

The slow news diffusion view implies profits for trading on slow-moving news (see, e.g., Hong and Stein (2007)). To test this prediction, in each year t, we compute the trading performance of each investor. The performance is defined as the difference of the mean cumulative return in year $t+1$ of the sell portfolio minus the buy portfolio. This "excessive trading" variable is first used in Odean (1999) to proxy for investor overconfidence. The higher the excessive trading, the lower the investor performance. We calculate the number of share-weighted mean returns (Excess.SW) and value-weighted mean returns (Excess.VW). The closing prices in year t are used for value-weighting. The two weighting methods are in parallel to the trading activity measures of Log(Share) and Log(Value). We test whether investors in more diverse provinces or cities exhibit better performance, or equivalently less excessive trading. Specifically, we run the following regression specification of excess trading on linguistic diversity:

$$Excess_{ijt} = \alpha + \beta Log(LD_j) + \gamma' \mathbf{X} + \epsilon_{ijt}, \tag{4}$$

where Excess_{ijt} refers to the share-weighted and value-weighted excessive trading of investor i in year t, LD_j is the number of languages spoken in investor i's home province or city j, and $\mathbf{X}$ includes the same controls as the trading activity analysis.[18]

Panel A of Table 9 reports the summary statistics of our main variables of interest and Panel B reports the OLS regression coefficients. Looking at Excess.SW in Column (1), Log(LD) has a coefficient of 0.071 with a t-statistic of 2.83. Column (2) shows similar results when Excess.VW is the dependent variable. In Column (3)-(8), sub-languages have stronger effects on the level of excessive trading. Again, taking Column (6) as an example, the coefficient of Log(LD-SUB2) is 0.054 and the t-statistic is 3.49. One way to interpret its economic significance is that one standard deviation increase in Log(LD-SUB2) leads to an increase in Excess.VW by 5.8%, which more than doubles the mean Excess.VW of 2%. Alternatively, one standard deviation increase in this diversity measure is associated with 11% increase of one standard deviation of the dependent variable. Column (9)-(10) present the reduced-form regression results using Hill as the independent variables of interest. The results are similar if not stronger statistically than using language and sub-languages, with the t-statistics around 3.4. Similar to previous tests, we have clustered standard errors at the province level. Overall, Panel B of Table 9 shows consistent and robust evidence that being in more diverse or geographic isolated areas increases excessive trading or lowers performance of the households.

Panel C of Table 9 presents further evidence from the second stage of the 2SLS regression using Hill as an instrument variable for linguistic diversity. Controls are the same as in the Panel B. Considering the regression of Excess.SW in Column (1), the coefficient of Log(LD) is 0.158 with a t-statistic as large as 3.29. We get similar and strong statistical significance throughout the rest of the specifications.

Overall, we find that households in diverse areas are not only not earning more profits, which would be the prediction of the slow news diffusion view, but they are more likely to

[18]The excessive trading results also hold after controlling the trading activities, i.e., Log(Share) or Log(Value).

lose money from their trades — that is they engage in excessive trading a la Odean (1999). The latter finding are more consistent with a differential interpretations view especially in light of previous research on trading of Internet stocks. It is not that news travels slowly but there is too much stimulation which in the presence of overconfident traders leads to excessive trading. Households are overconfident when they trade (see, e.g., Odean (1999)). This view implies that the more stimulation they get, the more they trade and as a result the worse their performance.

4.3. Diversity, Differential Interpretations and the 2006-2007 Chinese Stock Market Bubble

During the Internet Bubble, for example, disproportionately more media attention was paid to internet stocks than to non-internet stocks (Bhattacharya et al. (2009)), which Hong and Stein (2007) argued might have led to more excessive trading in Internet stocks relative to others due to diversity of opinions. Related, we find that households in diverse areas traded much more excessively than households in less diverse areas during the Chinese Stock Market or Olympics Bubble of 2006-2007. In Table 10 we split our trading activity and performance findings across the years of our sample to the 2006-2007 years of the Chinese Stock Market Bubble versus the rest of the sample. The Chinese Stock Market finished its share reforms in 2005 which lit a fire in the Chinese stock market in 2006 as the Chinese government made a very generous deal for existing shareholders to receive significant compensation to allow insiders to float their shares. This along with rumors that the Chinese government would not let the stock market fall before the 2008 Olympic led to a massive run-up in the stock market of roughly 200% over 2 short years, between the end of 2005 and the end of 2007. Under the differential interpretations view, we expect that households would be much more aggressive in their stock trading as there would be more debate about the run-up in the stock market. This is indeed what we find when we run our specifications by interacting linguistic diversity measures with whether or not the sample is 2006-2007.

Specifically, we define a dummy variable, Bubble_t, indicating the sample period of 2006 to 2007. We include an interaction term of Bubble with linguistic diversity measures in our trading activity and excess trading regressions. Using LD as the linguistic diversity measure, our regression specifications are as follows:

$$Trade_{ijt} = \alpha + \beta Log(LD_j) + \gamma Log(LD_j) \times Bubble_t + \delta' \mathbf{X} + \epsilon_{ijt}, \tag{5}$$

$$Excess_{ijt} = \alpha + \beta Log(LD_j) + \gamma Log(LD_j) \times Bubble_t + \delta' \mathbf{X} + \epsilon_{ijt}, \tag{6}$$

where the other variables are similarly defined in Equation (2) and (4). Our main coefficient of interest is γ in both equations. Panel A of Table 10 reports the results of the investor trading activity regression; i.e., Equation (5). Using Log(Share) in Column (1), the coefficient of Log(LD)×Bubble is 0.229 with a t-statistic of 5.64. The results using sub-languages are similar. For example, in Column (6), the interaction term has the coefficient of 0.075, with a t-statistic of 2.83. Comparing the coefficients of the interaction terms with that of linguistic diversity, these results imply the diversity effect increases by at least 50% during the bubble period. Panel B shows the excessive trading regression results from Equation (6). The coefficients of the interaction term of diversity and Bubble are all positive with large economic magnitudes. For example in Column (3), Log(LD-SUB1)×Bubble has a coefficient of 0.028 and t-statistic of 2.25. Given the coefficient of Log(LD-SUB1) is 0.030, the effect of linguistic diversity doubles during 2006 to 2007. Overall, we find robust evidence that the effect of linguistic diversity on trading activity and excess trading is stronger during the bubble period, consistent with the differential interpretations view.

5. Alternative Explanations and Alternative Identification Strategy Using City Variation within a Province

We have shown that diversity leads to more disagreement and trading using geographical isolation as an instrument. Linguistics have also proposed that latitude and temperature might also drive diversity (see, e.g., Nichols (1990), Nettle (1996), Stepp, Castaneda, and Cervone (2005)), Maffi (2007)). Our reading of the literature is that geographic isolation is more agreed upon and is typically the first factor in analyses. We have also examined these two other potential instruments but they are far less significant compared to geographic isolation.[19] As a result, we focus on hills, though these other results are available upon request.

While our analysis thus far establishes a causal relationship, which is interesting in its own right, we want to emphasize that a number of our control variables also help rule out a number of alternative explanations. For instance, some scientists believe that the ability of the southern areas of China to grow rice might also contribute to diversity of all kinds, including linguistic (Talhelm et al. (2014)). Here it is important to note that we control for Administrative Areas in all of our earlier regressions. In other words, we can find our effects even when considering provinces in the same Administrative Area, e.g., only provinces in the Northeast Administrative Area for instance are compared to identify the effect. So our geographical isolation effects seem distinct from fertile land effects.

In a similar vein, we can use variations in linguistic diversity across cities within the same province to try to identify our effect. Unfortunately, we cannot measure geographic isolation at the city level. So this analysis cannot use the Hill instrument. On the other hand, we can use a different type of variation in diversity and can more finely control for neighborhood effects using province rather than Administrative Areas. It would be comforting if we could get similar results as using the hills instrument. It is to this strategy that we now turn.

[19] We measure the average latitude of cities in a province and the average temperature of cities in a province and use these as the measure of temperature and latitude for a province.

Up to this point all our tests use province level variation in linguistic diversity, controlling for larger Administrative Area effects. An alternative identification strategy is to exploit city level linguistic diversity variation in a province and re-examine its effect on diversity of opinions and trading activity. However, it is not desirable to directly use the number of languages spoken in each city. Among our sample cities, over 70% of cities have LD equal to 1, so there is little variation in this measure within each province. Furthermore, for many of the small cities in our sample, LD may not fully capture the potential multi-lingual environment offered by nearby cities. To address this issue, we construct a circle with a 75 kilometer radius (45 miles) around each city. The size of a 75 kilometer radius circle will roughly correspond to a large metropolitan area such as Shanghai. Across our sample cities, on average we have 3 cities in each city-circle. Linguistic diversity is then defined as the number of unique languages or sub-languages spoken in the circle. Finally, to focus on city variations within each province, we drop provinces with less than two cities.

In Table 11 we report the summary statistics and OLS regression results of investor trading activity on the city level linguistic diversity. Our regression specification is similar to Equation (2), but now at the city level where we include province fixed effects. Investor characteristics and trading activities are measured similarly as in the province specification. We include as controls decile dummy variables for city GDP per capita, population and percent of city stocks. We do not include Trust and Branch where we only have province level data. For these regressions, we have year fixed effect and clustered standard errors at the city level.

Panel A of Table 11 summaries the regression variables used in the city-level analysis. The mean of Log(LD) is 0.5, and those for the three Log(LD-SUBs) are 1.25, 1.56, and 1.61, respectively. The linguistic diversity measures are generally smaller due to the fact that a smaller region is considered comparing to the province level.

Panel B of Table 11 presents our regression results. Consider Log(Share) in Column (1), the coefficient of Log(LD) is 0.059 with a t-statistic of 0.95. However, in Columns

(3)-(8), linguistic diversity measured using sub-languages have stronger effects on trading activities. All these specifications are significant with t-statistics greater than 3. In Column (6), for instance, the coefficient of Log(LD-SUB2) is 0.176 with a t-statistic of 3.90. The economic significance here implies a 3% standard deviation increase in trading activity from a one standard deviation increase in linguistic diversity. The statistical and economical significance is similar using other linguistic diversity measures but smaller than the province level specifications.

Using the same 75 kilometer radius circle, we now run the message board diversity of opinions on linguistic diversity at the city level with province fixed effects. In Panel A, we observe that city-level linguistic diversity measures have lower mean and standard deviation than those on the province level, as expected. For example, Log(LD-SUB3) has a mean (standard deviation) of 1.44 (0.61) here, and its province-level analog in Table 2 is 1.92 (0.77). Consistent with Table 11, we include city-level controls, province fixed effects, and drop provinces with less than two cities to focus on within province variations across cities. We also include firm fixed effects and cluster standard errors at the city level.

Columns (1) to (4) of Panel B show results using different levels of languages and sub-languages. Consistent with province level regressions, linguistic diversity continues to drive diversity of opinions here. While Log(LD) is not significant, coefficients in front of Log(LD-SUB1), Log(LD-SUB2), and Log(LD-SUB3) are all positive and statistically significant. As an example, Column (3) uses Log(LD-SUB2) and yields a coefficient estimate of 0.243 with a t-statistic of 3.03. Overall, the economic significance is around 4 to 12 percent of one standard deviation from a one standard deviation increase in linguistic diversity measures, comparable to the magnitudes from province level regressions.

6. Conclusion

Diversity of opinions among investors plays a crucial role in models of financial market speculation and bubbles. Most models assume that investors are endowed with heterogeneous priors and beliefs. But where do these differences in priors come from? We propose that linguistic diversity ought to increase such differences in priors. Using unique data from China, we show that investors living in linguistically diverse areas express more diverse opinions on stock message boards and trade stocks more actively. We show that a couple of identification strategies, including using geographical isolation of an area due to hilly terrain as an instrument for linguistic diversity, yield very similar results to our OLS findings. We then show that this relationship between diversity and diverse opinions and speculation is driven by a stimulation-overconfidence channel as opposed to a language barrier channel by verifying several additional predictions on asset price drift and retail investor performance.

Our findings might also be of interest for researchers interested in how diversity affects innovation and productivity. In economics, diverse societies by bringing about a variety of abilities, experiences, and cultures have been found to be correlated with economic activity (see Alesina and Ferrara (2005) for a survey). In sociology, experiments in organizations typically find that more diverse teams with different cultural backgrounds are more active in completing tasks because of the heterogeneity in viewpoints when the problems of communication due to diversity are accounted for (see, e.g., O'Reilly, Williams, and Barsade (1997)). Underlying all of this research is the premise that diversity stimulates a diversity of opinions, resulting in activity that would otherwise not occur. Interestingly, this premise has not been fully tested. Our work fills in this gap. For future research, our linguistic diversity measures and geographical isolation instrument would also be useful for studying how linguistic diversity influences innovation and productivity.

References

Alesina, A., and E. L. Ferrara. 2005. Ethnic diversity and economic performance. *Journal of Economic Literature* 43:762–800.

Antweiler, W., and M. Z. Frank. 2004. Is all that talk just noise? the information content of internet stock message boards. *Journal of Finance* 59:1259–94.

Barber, B. M., and T. Odean. 2001. Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics* 116:261–92.

Bernard, V. L., and J. K. Thomas. 1989. Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research* 27:1–36.

Bhattacharya, U., N. Galpin, R. Ray, and X. Yu. 2009. The role of the media in the internet ipo bubble. *Journal of Financial and Quantitative Analysis* 44:657–82.

Bialystok, E., and M. Martin. 2004. Attention and inhibition in bilingual children: evidence from the dimensional change card sort task. *Journal of Accounting Research* 7:325–39.

Chen, J., H. Hong, and J. C. Stein. 2002. Breadth of ownership and stock returns. *Journal of financial Economics* 66:171–205.

Chen, K. M. 2013. The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *American Economic Review* 103.

Daniel, K., D. Hirshleifer, and A. Subrahmanyam. 1998. Investor psychology and security market under-and overreactions. *the Journal of Finance* 53:1839–85.

Diether, K. B., C. J. Malloy, and A. Scherbina. 2002. Differences of opinion and the cross section of stock returns. *The Journal of Finance* 57:2113–41.

Geanakoplos, J. 2010. The leverage cycle. In *NBER Macroeconomics Annual 2009, Volume 24*, 1–65. University of Chicago Press.

Greenwood, R., and S. Nagel. 2009. Inexperienced investors and bubbles. *Journal of Financial Economics* 93:239–58.

Guiso, L., P. Sapienza, and L. Zingales. 2004. The role of social capital in financial development. *American Economic Review* 94:526–56.

———. 2009. Cultural biases in economic exchange? *Quarterly Journal of Economics* 124:1095–131.

Harris, M., and A. Raviv. 1993. Differences of opinion make a horse race. *Review of Financial Studies* 6:473–506.

Harrison, J. M., and D. M. Kreps. 1978. Speculative investor behavior in a stock market with heterogeneous expectations. *The Quarterly Journal of Economics* 323–36.

Hong, H., and J. C. Stein. 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance* 54:2143–84.

———. 2007. Disagreement and the stock market. *Journal of Economic Perspectives* 21:109–28.

Kandel, E., and N. D. Pearson. 1995. Differential interpretation of public signals and trade in speculative markets. *Journal of Political Economy* 103:831–72.

Kovacs, A. M., and J. Mehler. 2009. Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences* 106:6556–60.

Maddux, W. W., H. Adam, and A. D. Galinsky. 2010. When in rome ... learn why the romans do what they do: How multicultural learning experiences facilitate creativity. *Personality and Social Psychology Bulletin* 36:731–41.

Maddux, W. W., and A. Galinsky. 2009. Cultural borders and mental barriers: The relationship between living abroad and creativity. *Journal of Personality and Social Psychology* 96:1047–61.

Maffi, L. 2007. Biocultural diversity and sustainability. In J. Pretty, A. Ball, T. Benton, J. Guivant, D. R. Lee, D. Orr, M. Pfeffer, and H. Ward, eds., *The SAGE Handbook of Environment Society*. SAGE Publications.

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

Mehl, M. 2006. Quantitative textual analysis. In M. Eid and E. Deiner, eds., *Handbook of Multimethod Measurement in Psychology*. Washington D.C.: American Psychological Association.

Miller, E. M. 1977. Risk, uncertainty, and divergence of opinion. *The Journal of Finance* 32:1151–68.

Morris, S. 1996. Speculative investor behavior and learning. *The Quarterly Journal of Economics* 1111–33.

Nettle, D. 1996. Language diversity in west africa: An ecological approach. *Journal of Anthropological Archaeology* 15:403–38.

Nichols, J. 1990. Linguistic diversity and the first settlement of the new world. *Language* 66:475–521.

Odean, T. 1999. Do investors trade too much. *American Economic Review* 89:1279–98.

O'Reilly, C., K. Y. Williams, and S. G. Barsade. 1997. Demography and group performance. *Unpublished* .

Ramsey, S. R. 1987. *The Languages of China*. Princeton University Press, Princeton NJ.

Scheinkman, J. A., and W. Xiong. 2003. Overconfidence and speculative bubbles. *Journal of political Economy* 111:1183–220.

Stepp, J., H. Castaneda, and S. Cervone. 2005. Mountains and biocultural diversity. *Mountain Research and Development* 25:223–7.

Talhelm, T., X. Zhang, S. Oishi, D. Duan, and S. Kitayama. 2014. Large-scale psychological differences within china explained by rice versus wheat agriculture. *Science* 344:603–8.

Varian, H. R. 1989. Differences of opinion in financial markets. In C. C. Stone, ed., *Financial Risk: Theory, Evidence and Implications: Proceedings of the Eleventh Annual Economic Policy Conference of the Federal Reserve Bank of St. Louis*. Springer Science+Business Media B.V.

Xiong, W., and J. Yu. 2011. The chinese warrants bubble. *American Economic Review* 101:2723–53.

Zhang, W., and R. Ke. 2003. Trust in china: A cross-regional analysis. *Working Paper* .

Table 1: Province GDP, Population, Diversity, and Terrain

This table reports the languages spoken, linguistic diversity measures, log GDP per capita, land statistics, and other characteristics for each province in China, and correlation coefficients of these variables. In Panel A, Log(GDP) is the log GDP per capita. LD is the number of languages spoken in each province. LD-SUB1, LD-SUB2, and LD-SUB3 denote the number of level 1, 2, and 3 sub-languages, respectively. Hill is the fraction of hill areas. Trust is the log province trust index. Log(Pop) is the log population. Branch is the number of brokerage branches divided by the total population in each province. PLocal is the number of local stocks divided by the number of all stocks in the country. The last two rows of the report the means and standard deviations of these variables. Panel B reports the correlation coefficients.

Panel A: Summary by Province

Province	Languages	Log(GDP)	LD	LD-SUB1	LD-SUB2	LD-SUB3	Hill	Trust	Log(Pop)	Branch	PLocal
Shanghai	Wu	11.14	1	1	1	1	0.00	5.39	3.09	21.33	0.09
Beijing	Guan	11.11	1	2	2	2	0.06	5.13	2.91	11.32	0.07
Tianjin	Guan	11.04	1	2	2	4	0.01	3.91	2.51	6.98	0.02
Zhejiang	Guan, Hui, Wu	10.72	3	11	16	16	0.47	4.35	3.97	5.16	0.08
Jiangsu	Guan, Wu	10.72	2	4	7	7	0.05	4.78	4.36	3.39	0.08
Guangdong	Min, Kejia, Yue	10.59	3	11	14	14	0.39	4.76	4.61	5.60	0.12
Neimenggu	Guan, Jin	10.56	2	5	7	7	0.13	2.45	3.20	1.64	0.01
Liaoning	Guan	10.49	1	3	5	5	0.20	3.47	3.77	4.12	0.03
Shandong	Guan	10.49	1	3	8	11	0.07	4.57	4.55	2.08	0.06
Fujian	Min, Gan, Kejia, Wu	10.44	4	12	14	14	0.31	3.19	3.60	4.11	0.04
Jilin	Guan	10.20	1	1	2	5	0.05	2.69	3.31	2.71	0.02
Hebei	Guan, Jin	10.13	2	5	8	14	0.09	3.40	4.26	1.49	0.02
Heilongjiang	Guan	10.09	1	3	5	7	0.11	2.77	3.65	2.48	0.02
Chongqing	Guan	10.06	1	1	1	1	0.18	2.65	3.36	2.95	0.02
Hubei	Guan, Gan	10.05	2	3	6	6	0.22	2.61	4.05	2.41	0.04
Shanxi	Guan, Jin	10.03	2	8	10	12	0.33	2.49	3.55	1.97	0.02
Shaanxi	Guan, Jin	10.02	2	7	8	8	0.41	2.74	3.62	2.18	0.02
Xinjiang	Guan	9.99	1	3	3	3	0.04	2.75	3.07	2.04	0.02
Ningxia	Guan	9.99	1	2	4	4	0.41	1.53	1.83	2.44	0.01
Henan	Guan, Jin	9.96	2	3	8	9	0.16	2.67	4.54	1.11	0.03
Hunan	Guan, Gan, Kejia, Xiang	9.94	4	12	18	18	0.17	2.29	4.17	1.83	0.03
Qinghai	Guan	9.92	1	1	2	2	0.12	1.57	1.72	1.52	0.01
Hainan	Min	9.92	1	1	5	5	0.20	1.41	2.15	3.49	0.01
Jiangxi	Guan, Gan, Kejia, Hui, Wu	9.80	5	12	12	12	0.20	2.00	3.79	2.01	0.02
Sichuan	Guan	9.80	1	1	3	5	0.18	3.30	4.40	2.06	0.04
Anhui	Guan, Gan, Hui, Wu	9.75	4	11	16	16	0.19	2.53	4.10	1.71	0.03
Guangxi	Guan, Min, Kejia, Xiang, Yue, Ping	9.74	6	12	15	15	0.26	2.56	3.86	1.33	0.02
Gansu	Guan	9.55	1	3	5	5	0.20	2.10	3.24	1.80	0.01
Yunnan	Guan	9.55	1	1	4	7	0.09	2.93	3.82	1.11	0.02
Guizhou	Guan	9.33	1	1	5	5	0.08	2.00	3.57	0.75	0.01
Mean		10.17	1.97	4.83	7.20	8.00	0.18	3.03	3.55	3.50	0.03
SD		0.46	1.38	4.15	5.00	4.97	0.13	1.08	0.76	3.98	0.03

Panel B: Correlation Coefficients

	Log(LD)	Log(LD-SUB1)	Log(LD-SUB2)	Log(LD-SUB3)	Hill	Trust	Log(GDP)	Log(Pop)	Branch	Plocal
Log(LD)	1.00	0.88	0.81	0.74	0.46	-0.04	-0.08	0.46	-0.19	0.15
Log(LD-SUB1)	0.88	1.00	0.87	0.79	0.54	0.07	0.05	0.46	-0.20	0.19
Log(LD-SUB2)	0.81	0.87	1.00	0.95	0.55	-0.11	-0.23	0.55	-0.45	0.09
Log(LD-SUB3)	0.74	0.79	0.95	1.00	0.44	-0.08	-0.23	0.60	-0.50	0.04
Hill	0.46	0.54	0.55	0.44	1.00	-0.19	-0.14	0.08	-0.23	0.08
Trust	-0.04	0.07	-0.11	-0.08	-0.19	1.00	0.79	0.38	0.66	0.85
Log(GDP)	-0.08	0.05	-0.23	-0.23	-0.14	0.79	1.00	-0.06	0.72	0.63
Log(Pop)	0.46	0.46	0.55	0.60	0.08	0.38	-0.06	1.00	-0.19	0.43
Branch	-0.19	-0.20	-0.45	-0.50	-0.23	0.66	0.72	-0.19	1.00	0.59
Plocal	0.15	0.19	0.09	0.04	0.08	0.85	0.63	0.43	0.59	1.00

Table 2: Summary Statistics

This table reports summary statistics for all the variables used in the regression analysis. Panel A reports pooled summary statistics for individual investor trading data. In each year t, we compute the trading activity of each investor. Log(Share) and Log(Value) are the log number of shares and trading value of all stocks traded by each investor, respectively. Log(Wealth) is log wealth of the investor. Sex is an indicator variable that equals 0 for male and 1 for female. Log(Age) is log age of each investor. Vol and Size are the value-weighted average volatility and firm market capitalization deciles of each investor's traded stocks. Log(LD), Log(LD-SUB1), Log(LD-SUB2), and Log(LD-SUB3) are the log linguistic diversity measures. Hill is the fraction of hill areas. Trust is the log province trust index. Log(GDP) is the log GDP per capita. Log(Pop) is the log population. Branch is the number of brokerage branches divided by the total population in each province. PLocal is the number of local stocks divided by the number of all stocks in the country. Panel B reports pooled summary statistics for message board data. Naïve Bayes rule is applied to generate predicted buy/sell signal (-2 strong sell, -1 sell, 0 neutral, 1 buy, 2 strong buy) for each message. The sample consists of 10 pages of messages for each firm from eight provinces: Beijing, Fujian, Guangdong, Hunan, Shandong, Shanghai, Sichuan, and Zhejiang. The training sample used to generate probabilities associated with each key word is described in Table 3. For each firm in our sample, we compute log standard deviation of the generated buy/sell signals for messages from each province, Log(STDEV). The other variables are defined similarly as in Panel A.

Panel A: Individual Investor Trading							
	Mean	SD	Min	25%	Median	75%	Max
Log(Share)	10.66	4.39	-4.61	10.09	11.63	12.94	22.29
Log(Value)	12.87	4.94	-4.61	12.39	13.99	15.35	24.86
Log(LD)	0.71	0.50	0.00	0.00	1.10	1.10	1.79
Log(LD-SUB1)	1.65	0.93	0.00	0.69	2.40	2.40	2.48
Log(LD-SUB2)	1.94	1.07	0.00	0.69	2.77	2.77	2.89
Log(LD-SUB3)	1.97	1.05	0.00	0.69	2.77	2.77	2.89
Hill	0.29	0.21	0.00	0.05	0.47	0.47	0.47
Trust	4.54	0.59	0.99	4.35	4.35	5.13	5.39
Log(Wealth)	11.08	2.88	-4.61	10.41	11.47	12.51	21.31
Sex	0.44	0.50	0.00	0.00	0.00	1.00	1.00
Log(Age)	3.82	0.26	2.56	3.64	3.83	4.01	4.58
Vol	0.13	0.07	0.00	0.11	0.13	0.16	8.30
Size	6.46	2.46	0.00	5.47	7.00	8.15	10.00
Log(GDP)	10.86	0.32	9.21	10.63	10.89	11.05	11.42
Log(Pop)	3.75	0.54	2.17	3.16	3.97	4.00	4.66
Branch	8.08	5.75	0.88	3.99	6.38	10.16	23.47
PLocal	0.08	0.02	0.01	0.07	0.08	0.09	0.14

Panel B: Message Board							
	Mean	SD	Min	25%	Median	75%	Max
Log(STDEV)	-0.37	1.82	-9.21	-0.19	0.03	0.18	1.04
Log(LD)	0.55	0.597	0.00	0.00	0.69	1.10	1.79
Log(LD-SUB1)	1.27	0.93	0.00	0.69	1.10	2.40	2.49
Log(LD-SUB2)	1.77	0.82	0.00	1.39	1.95	2.49	2.89
Log(LD-SUB3)	1.92	0.77	0.00	1.61	1.95	2.64	2.89
Hill	0.18	0.12	0.00	0.08	0.18	0.22	0.47
Trust	3.27	1.04	1.41	2.56	2.77	4.35	5.39
Log(GDP)	10.38	0.45	9.55	10.10	10.26	10.71	11.22
Log(Pop)	3.72	0.64	1.73	3.36	3.80	4.18	4.64
Branch	3.97	4.05	0.92	1.97	2.35	4.33	20.73
PLocal	0.04	0.03	0.01	0.02	0.03	0.04	0.13

Table 3: Conditional Probabilities of Key Words

This table reports the top five key words with the highest conditional probabilities for each buy/sell signal; i.e., Pr(Signal | Word). Key Words are written in Chinese with the English translation in parentheses. The training sample consists of the most recent 20 messages from a random sample of 30 firms in each sample province. The sample provinces include Beijing, Fujian, Guangdong, Hunan, Shandong, Shanghai, Sichuan, and Zhejiang.

Buy/Sell Signal	Key Words	Probability
-2	坏 (bad)	0.751
-2	轮渡 (ferry)	0.751
-2	跳水 (dive)	0.715
-2	清仓 (to empty one's positions)	0.693
-2	空仓 (zero position)	0.667
-1	破股 (bad stock)	1.000
-1	下行 (heading down)	1.000
-1	抛压 (selling pressure)	1.000
-1	失望 (disappointment)	1.000
-1	僵尸 (zombie)	1.000
0	基建 (construction)	1.000
0	城 (city)	1.000
0	曙光 (dawn)	1.000
0	指点 (pointers)	1.000
0	油料 (paint)	1.000
1	长线 (long-term)	1.000
1	人人 (everybody)	1.000
1	空调 (AC)	1.000
1	旺季 (busy season)	1.000
1	捡 (pick up)	1.000
2	独领风骚 (in a leading position)	1.000
2	控 (control)	0.801
2	全仓 (full positions)	0.801
2	证 (securities)	0.708
2	立贴 (to write a post)	0.556

Table 4: Linguistic Diversity and Individual Investor Trading

This table reports OLS regression results of individual investor trading and linguistic diversity. In each year t, we compute the trading activity of each investor. Log(Share) and Log(Value) are the log number of shares and trading value of all stocks traded by each investor, respectively. Log(LD), Log(LD-SUB1), Log(LD-SUB2), and Log(LD-SUB3) are the log linguistic diversity measures of each investor's home province. Hill is the fraction of hilly areas. Control variables include Trust, Log(Wealth), Sex, Log(Age), Vol, Size, Log(GDP), Log(Pop), Branch and PLocal, defined as in Table 2. Administrative area and year dummies are included in all specifications. T-statistics are in parentheses and standard errors are clustered by province. The sample period is from 2006 to 2012.

Dependent Variable	(1) Log(Share)	(2) Log(Value)	(3) Log(Share)	(4) Log(Value)	(5) Log(Share)	(6) Log(Value)	(7) Log(Share)	(8) Log(Value)	(9) Log(Share)	(10) Log(Value)
Diversity Proxy	Log(LD)		Log(LD-SUB1)		Log(LD-SUB2)		Log(LD-SUB3)		Hill	
Diversity	0.273	0.289	0.171	0.180	0.198	0.208	0.234	0.245	0.630	0.658
	(1.65)	(1.73)	(2.24)	(2.31)	(2.45)	(2.53)	(2.74)	(2.82)	(4.11)	(4.28)
Trust	-0.402	-0.438	-0.374	-0.410	-0.345	-0.380	-0.254	-0.284	-0.403	-0.441
	(-3.89)	(-4.66)	(-3.35)	(-4.00)	(-2.82)	(-3.35)	(-1.75)	(-2.08)	(-3.89)	(-4.67)
Log(Wealth)	0.165	0.183	0.165	0.182	0.165	0.182	0.165	0.183	0.165	0.182
	(21.76)	(21.18)	(21.75)	(21.18)	(21.75)	(21.18)	(21.85)	(21.27)	(21.80)	(21.23)
Sex	-0.233	-0.209	-0.233	-0.209	-0.233	-0.209	-0.233	-0.209	-0.234	-0.210
	(-10.31)	(-8.58)	(-10.26)	(-8.54)	(-10.23)	(-8.52)	(-10.26)	(-8.54)	(-10.28)	(-8.57)
Log(Age)	1.891	2.019	1.894	2.022	1.895	2.023	1.895	2.023	1.895	2.023
	(32.61)	(31.55)	(32.84)	(31.81)	(32.91)	(31.89)	(32.97)	(31.94)	(33.07)	(32.04)
Vol	19.467	22.299	19.464	22.296	19.464	22.295	19.462	22.294	19.462	22.294
	(19.35)	(19.39)	(19.37)	(19.42)	(19.38)	(19.42)	(19.38)	(19.43)	(19.37)	(19.42)
Size	0.761	0.876	0.761	0.876	0.761	0.876	0.761	0.876	0.761	0.876
	(33.25)	(31.83)	(33.26)	(31.84)	(33.25)	(31.83)	(33.22)	(31.81)	(33.27)	(31.84)
Log(GDP)	0.463	0.506	0.466	0.507	0.418	0.457	0.282	0.315	0.559	0.603
	(2.40)	(2.48)	(2.61)	(2.77)	(2.39)	(2.61)	(1.61)	(1.87)	(3.31)	(3.59)
Log(Pop)	0.129	0.172	0.200	0.244	0.140	0.182	0.115	0.156	0.372	0.424
	(0.97)	(1.17)	(1.41)	(1.61)	(1.21)	(1.48)	(1.19)	(1.57)	(3.30)	(3.63)
Branch	0.054	0.061	0.064	0.071	0.069	0.076	0.071	0.078	0.066	0.073
	(4.10)	(4.26)	(4.26)	(4.43)	(4.48)	(4.63)	(4.90)	(5.08)	(7.77)	(8.24)
PLocal	3.300	4.006	1.673	2.335	1.162	1.796	1.739	2.412	0.515	1.155
	(1.01)	(1.21)	(0.47)	(0.65)	(0.32)	(0.49)	(0.55)	(0.76)	(0.18)	(0.41)
Admin. Area	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year Dum	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
N	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789
Adj. R^2	0.428	0.453	0.428	0.453	0.428	0.453	0.428	0.453	0.428	0.453

Table 5: Linguistic Diversity and Diversity of Opinions

This table reports OLS regression results of diversity of opinions on linguistic diversity using message board data. Naïve Bayes rule is applied to generate predicted buy/sell signal (-2 strong sell, -1 sell, 0 neutral, 1 buy, 2 strong buy) for each message. The sample consists of 10 pages of messages for each firm from eight provinces: Beijing, Fujian, Guangdong, Hunan, Shandong, Shanghai, Sichuan, and Zhejiang. The training sample used to generate probabilities associated with each key word is described in Table 3. For each firm in our sample, we compute log standard deviation of the generated buy/sell signals for messages from each province, Log(STDEV). Log(LD), Log(LD-SUB1), Log(LD-SUB2), and Log(LD-SUB3) are the log linguistic diversity measures. Hill is the fraction of hill areas. Other control variables include Trust, Log(GDP), Log(Pop), Branch, and PLocal, as defined in Table 2. Administrative area and firm dummies are included in all specifications. T-stats are in parentheses and standard errors are clustered by province.

Dependent Variable: Log(STDEV)					
	(1)	(2)	(3)	(4)	(5)
Log(LD)	0.284				
	(2.77)				
Log(LD-SUB1)		0.166			
		(2.43)			
Log(LD-SUB2)			0.118		
			(1.43)		
Log(LD-SUB3)				0.079	
				(0.94)	
Hill					1.023
					(2.69)
Trust	0.249	0.217	0.105	0.069	0.251
	(1.12)	(0.96)	(0.48)	(0.32)	(1.09)
Log(GDP)	0.527	0.537	0.689	0.689	0.498
	(1.35)	(1.34)	(1.67)	(1.65)	(1.28)
Log(Pop)	0.537	0.557	0.666	0.686	0.645
	(2.73)	(2.65)	(3.51)	(3.73)	(3.50)
Branch	0.032	0.040	0.051	0.049	0.043
	(1.81)	(2.34)	(2.48)	(2.16)	(2.28)
PLocal	-8.211	-9.151	-8.742	-7.874	-11.723
	(-3.69)	(-3.67)	(-3.24)	(-2.68)	(-3.88)
Admin. Area	YES	YES	YES	YES	YES
Firm Dum	YES	YES	YES	YES	YES
N	33,775	33,775	33,775	33,775	33,775
Adj. R^2	0.118	0.117	0.116	0.115	0.117

Table 6: Linguistic Diversity and Individual Investor Trading (2SLS)

This table reports 2SLS regression results of individual investor trading activity on linguistic diversity using fraction of hills as the instrumental variable. In each year t, we compute the trading activity of each investor. Log(Share) and Log(Value) are the log number of shares and trading value of all stocks traded by each investor. Linguistic diversity measures are Log(LD), Log(LD-SUB1), Log(LD-SUB2), and Log(LD-SUB3). Panel A reports first stage regressions of log trading activity on the fraction of hills (Hill). Controls include Trust, Log(Wealth), Sex, Log(Age), Vol, Size, Log(GDP), Log(Pop), Branch, and PLocal, as defined in Table 2. Panel B reports the 2SLS results with Hill as the IV. Control variables are the same as the first stage regressions. Administrative area and year dummies are included in all specifications. T-stats are in parentheses and standard errors are clustered by province. The sample period is from 2006 to 2012.

Panel A: First Stage				
Dependent	(1)	(2)	(3)	(4)
Variable	Log(LD)	Log(LD-SUB1)	Log(LD-SUB2)	Log(LD-SUB3)
Hill	0.963	2.547	2.382	2.092
	(9.52)	(30.10)	(41.58)	(17.27)
Trust	-0.230	-0.360	-0.425	-0.740
	(-3.94)	(-4.56)	(-7.20)	(-10.40)
Log(Wealth)	0.000	0.000	0.000	-0.000
	(0.80)	(0.88)	(0.64)	(-1.26)
Sex	-0.001	0.000	0.001	0.000
	(-0.93)	(0.18)	(1.15)	(0.09)
Log(Age)	0.004	-0.001	-0.004	-0.002
	(0.65)	(-0.23)	(-1.46)	(-0.64)
Vol	-0.007	-0.002	0.001	0.006
	(-0.91)	(-0.53)	(0.39)	(1.36)
Size	0.000	0.000	0.000	-0.000
	(1.61)	(1.69)	(1.18)	(-0.96)
Log(GDP)	-0.254	0.035	0.353	0.911
	(-1.95)	(0.27)	(3.24)	(7.47)
Log(Pop)	0.084	0.329	0.692	0.737
	(0.64)	(1.71)	(4.80)	(5.23)
Branch	-0.017	-0.038	-0.050	-0.048
	(-2.01)	(-2.61)	(-4.49)	(-4.51)
PLocal	4.438	5.561	5.445	1.336
	(2.50)	(3.45)	(4.79)	(0.89)
Admin. Area	YES	YES	YES	YES
Year Dum	YES	YES	YES	YES
N	1,091,789	1,091,789	1,091,789	1,091,789
Adj. R^2	0.979	0.990	0.996	0.995

Table 6—*Continued*

Panel B: 2SLS								
Dependent Variable	(1) Log(Share)	(2) Log(Value)	(3) Log(Share)	(4) Log(Value)	(5) Log(Share)	(6) Log(Value)	(7) Log(Share)	(8) Log(Value)
Diversity Proxy	Log(LD)		Log(LD-SUB1)		Log(LD-SUB2)		Log(LD-SUB3)	
Diversity	0.655	0.683	0.247	0.258	0.265	0.276	0.301	0.314
	(3.67)	(3.79)	(4.06)	(4.22)	(4.10)	(4.27)	(4.32)	(4.55)
Trust	-0.252	-0.284	-0.314	-0.348	-0.290	-0.323	-0.180	-0.208
	(-1.62)	(-1.90)	(-2.48)	(-2.94)	(-2.26)	(-2.70)	(-1.21)	(-1.51)
Log(Wealth)	0.165	0.182	0.165	0.182	0.165	0.182	0.165	0.182
	(22.08)	(21.54)	(22.28)	(21.71)	(22.32)	(21.74)	(22.41)	(21.81)
Sex	-0.233	-0.209	-0.234	-0.210	-0.234	-0.210	-0.234	-0.210
	(-10.58)	(-8.81)	(-10.54)	(-8.78)	(-10.53)	(-8.77)	(-10.54)	(-8.78)
Log(Age)	1.893	2.020	1.896	2.024	1.897	2.025	1.896	2.024
	(33.38)	(32.33)	(33.86)	(32.82)	(33.97)	(32.94)	(33.96)	(32.91)
Vol	19.467	22.298	19.463	22.294	19.462	22.293	19.460	22.292
	(19.83)	(19.88)	(19.85)	(19.90)	(19.85)	(19.90)	(19.86)	(19.91)
Size	0.761	0.876	0.761	0.876	0.761	0.876	0.761	0.876
	(34.14)	(32.68)	(34.10)	(32.64)	(34.08)	(32.63)	(34.06)	(32.60)
Log(GDP)	0.725	0.776	0.550	0.594	0.465	0.505	0.284	0.316
	(3.63)	(3.73)	(3.06)	(3.30)	(2.52)	(2.78)	(1.56)	(1.84)
Log(Pop)	0.317	0.366	0.291	0.339	0.189	0.233	0.150	0.192
	(2.18)	(2.43)	(2.41)	(2.70)	(1.75)	(2.12)	(1.64)	(2.15)
Branch	0.077	0.084	0.075	0.082	0.079	0.086	0.080	0.088
	(6.02)	(6.32)	(6.75)	(6.99)	(6.87)	(7.08)	(7.55)	(7.85)
PLocal	-2.390	-1.878	-0.861	-0.281	-0.925	-0.348	0.113	0.735
	(-0.61)	(-0.47)	(-0.27)	(-0.09)	(-0.29)	(-0.11)	(0.04)	(0.26)
Admin. Area	YES	YES	YES	YES	YES	YES	YES	YES
Year Dum	YES	YES	YES	YES	YES	YES	YES	YES
N	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789
Adj. R^2	0.428	0.453	0.428	0.453	0.428	0.453	0.428	0.453

Table 7: Linguistic Diversity and Diversity of Opinions (2SLS)

This table reports 2SLS results of diversity of opinions on linguistic diversity using message board data and fraction of hills as the instrumental variable. Diversity of opinions is measured by Log(STDEV), the log standard deviation of the generated buy/sell signals for each firm-province pair. Linguistic diversity measures are Log(LD), Log(LD-SUB1), Log(LD-SUB2), and Log(LD-SUB3). Panel A reports first stage regressions of log linguistic diversity measures on the fraction of hills (Hill). Controls include Trust, Log(GDP), Log(Pop), Branch, and PLocal, as defined in Table 2. Administrative area and firm dummies are included in all specifications. Panel B reports the 2SLS results with Hill as the IV. Control variables are the same as the first stage regressions. T-stats are in parentheses and standard errors are clustered by province.

Panel A: First Stage				
Dependent	(1)	(2)	(3)	(4)
Variable	Log(LD)	Log(LD-SUB1)	Log(LD-SUB2)	Log(LD-SUB3)
Hill	1.772	3.401	2.242	2.042
	(3.15)	(5.61)	(4.92)	(4.25)
Trust	-0.338	-0.319	0.026	0.245
	(-1.37)	(-1.22)	(0.12)	(0.95)
Log(GDP)	0.190	0.205	-0.599	-0.694
	(0.79)	(0.73)	(-1.36)	(-1.27)
Log(Pop)	0.516	0.738	0.300	0.292
	(2.67)	(3.00)	(1.75)	(1.52)
Branch	0.038	0.013	-0.071	-0.091
	(1.18)	(0.42)	(-4.35)	(-4.70)
PLocal	-4.776	-4.029	1.334	-3.518
	(-0.83)	(-0.65)	(0.37)	(-0.87)
Admin. Area	YES	YES	YES	YES
Firm Dum	YES	YES	YES	YES
N	33,775	33,775	33,775	33,775
Adj. R^2	0.796	0.841	0.887	0.855

Panel B: 2SLS				
Dependent Variable: Log(STDEV)				
	(1)	(2)	(3)	(4)
Log(LD)	0.581			
	(2.59)			
Log(LD-SUB1)		0.303		
		(3.01)		
Log(LD-SUB2)			0.460	
			(2.30)	
Log(LD-SUB3)				0.505
				(2.21)
Trust	0.443	0.343	0.236	0.124
	(1.58)	(1.45)	(0.90)	(0.50)
Log(GDP)	0.400	0.449	0.785	0.861
	(1.03)	(1.20)	(1.94)	(1.92)
Log(Pop)	0.343	0.419	0.505	0.497
	(1.44)	(1.93)	(2.18)	(1.97)
Branch	0.021	0.039	0.076	0.090
	(1.23)	(2.74)	(4.36)	(3.61)
PLocal	-9.245	-10.800	-12.652	-10.255
	(-3.07)	(-3.79)	(-3.55)	(-3.05)
Admin. Area	YES	YES	YES	YES
Firm Dum	YES	YES	YES	YES
N	33,775	33,775	33,775	33,775
Adj. R^2	0.052	0.053	0.049	0.046

Table 8: Linguistic Diversity and Post-Earnings-Announcement Drift

This table reports post-earnings-announcement-drift (PEAD) in provinces with different linguistic diversity levels. Every year, all firms in CSMAR are sorted into quintiles by their cumulative abnormal return around annual earnings announcement date [-1,1]. E1 (E5) are firms with low (high) earnings surprises. We show cumulative abnormal returns over different windows after earnings announcement: [2,6], [2,11], [2,21], [2,31], [2,41], and [2,51]. Firms are independently sorted into four levels of linguistic diversity by the average LD of all messages in the Guba message board data. The sample period is 1998 to 2012. T-statistics are in parentheses.

Linguistic Diversity Low	[2,6]	[2,11]	[2,21]	[2,31]	[2,41]	[2,51]
E1	0.008	0.002	-0.001	-0.006	0.102	0.102
	(1.80)	(0.42)	(-0.10)	(-0.60)	(0.81)	(0.77)
E5	-0.002	-0.004	0.004	-0.003	-0.002	-0.005
	(-0.36)	(-0.68)	(0.38)	(-0.31)	(-0.11)	(-0.36)
E5-E1 (a)	-0.009	-0.007	0.004	0.002	-0.103	-0.108
	(-1.5)	(-0.79)	(0.34)	(0.18)	(-0.81)	(-0.80)
Linguistic Diversity 2	[2,6]	[2,11]	[2,21]	[2,31]	[2,41]	[2,51]
E1	0.004	0.007	0.003	0.006	-0.019	-0.015
	(1.06)	(0.95)	(0.30)	(0.47)	(-1.16)	(-0.88)
E5	-0.002	0.009	0.019	0.007	-0.011	-0.013
	(-0.31)	(1.07)	(1.69)	(0.51)	(-0.72)	(-0.74)
E5-E1	-0.006	0.002	0.016	0.001	0.007	0.002
	(-0.85)	(0.15)	(0.97)	(0.06)	(0.34)	(0.09)
Linguistic Diversity 3	[2,6]	[2,11]	[2,21]	[2,31]	[2,41]	[2,51]
E1	0.004	0.005	-0.007	-0.005	-0.029	-0.037
	(0.92)	(0.76)	(-0.87)	(-0.47)	(-2.62)	(-2.81)
E5	-0.001	0.002	-0.008	-0.006	-0.021	-0.035
	(-0.25)	(0.27)	(-0.83)	(-0.55)	(-1.59)	(-2.47)
E5-E1	-0.005	-0.003	0.000	-0.002	0.008	0.002
	(-0.78)	(-0.29)	(-0.03)	(-0.11)	(0.47)	(0.12)
Linguistic Diversity High	[2,6]	[2,11]	[2,21]	[2,31]	[2,41]	[2,51]
E1	0.007	0.004	-0.007	-0.003	-0.037	-0.031
	(2.00)	(0.87)	(-1.01)	(-0.39)	(-3.89)	(-2.68)
E5	-0.001	0.006	0.007	0.002	-0.014	-0.022
	(-0.11)	(0.8)	(0.66)	(0.13)	(-1.14)	(-1.61)
E5-E1 (b)	-0.008	0.001	0.013	0.005	0.022	0.008
	(-1.26)	(0.14)	(1.10)	(0.33)	(1.41)	(0.46)
(b) - (a)	0.001	0.008	0.009	0.003	0.125	0.116
	(0.14)	(0.63)	(0.51)	(0.12)	(1.05)	(0.92)

Table 9: Linguistic Diversity and Individual Investor Trading Performance

This table reports summary statistics and regression results of individual investor trading performance and linguistic diversity. In each year t, we compute the trading performance of each investor. Performance is defined as the difference of the mean cumulative return in year t+1 of the sell portfolio minus the buy portfolio. The mean returns are number of shares-weighted (Excess.SW) or value-weighted (Excess.VW). Closing prices in year t are used for value-weighting. Log(LD), Log(LD-SUB1), Log(LD-SUB2), and Log(LD-SUB3) are the log linguistic diversity measures of each investor's home province. Hill, Trust, Log(Wealth), Sex, Log(Age), Vol, Size, Log(GDP), Log(Pop), Branch, and PLocal are defined in Table 3. Panel A reports the summary statistics. Panel B reports the OLS regression. Panel C reports the 2SLS results with Hill as the IV. Administrative area and year dummies are included in all specifications in Panel B and Panel C. T-statistics are in parentheses and standard errors are clustered by province. The sample period is from 2006 to 2012.

Panel A: Summary Statistics

	Mean	SD	Min	25%	Median	75%	Max
Excess.SW	0.02	0.54	-8.50	-0.08	0.00	0.20	8.34
Excess.VW	0.02	0.53	-8.50	-0.08	0.00	0.20	8.34
Log(LD)	0.71	0.50	0.00	0.00	1.10	1.10	1.79
Log(LD-SUB1)	1.66	0.93	0.00	0.69	2.40	2.40	2.48
Log(LD-SUB2)	1.94	1.07	0.00	0.69	2.77	2.77	2.89
Log(LD-SUB3)	1.97	1.05	0.00	0.69	2.77	2.77	2.89

Table 9—*Continued*

Panel B: OLS Regression										
Dependent Variable	(1) Excess.SW	(2) Excess.VW	(3) Excess.SW	(4) Excess.VW	(5) Excess.SW	(6) Excess.VW	(7) Excess.SW	(8) Excess.VW	(9) Excess.SW	(10) Excess.VW
Diversity Proxy	Log(LD)		Log(LD-SUB1)		Log(LD-SUB2)		Log(LD-SUB3)		Hill	
Diversity	0.071	0.072	0.046	0.046	0.053	0.054	0.035	0.036	0.145	0.145
	(2.83)	(2.81)	(3.23)	(3.24)	(3.49)	(3.49)	(2.86)	(2.90)	(3.43)	(3.42)
Trust	0.083	0.079	0.101	0.096	0.111	0.106	0.091	0.086	0.091	0.085
	(2.56)	(2.45)	(3.17)	(3.05)	(3.29)	(3.17)	(2.59)	(2.48)	(3.31)	(3.17)
Log(Wealth)	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(-8.59)	(-8.48)	(-8.55)	(-8.45)	(-8.39)	(-8.26)	(-8.08)	(-7.91)	(-8.55)	(-8.45)
Sex	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.24)	(0.26)	(0.17)	(0.20)	(0.15)	(0.18)	(0.18)	(0.21)	(0.13)	(0.16)
Log(Age)	-0.007	-0.010	-0.007	-0.010	-0.007	-0.009	-0.007	-0.010	-0.007	-0.009
	(-1.35)	(-1.94)	(-1.23)	(-1.79)	(-1.18)	(-1.73)	(-1.21)	(-1.77)	(-1.20)	(-1.76)
Vol	0.004	0.003	0.003	0.002	0.003	0.002	0.003	0.002	0.003	0.002
	(0.28)	(0.23)	(0.23)	(0.18)	(0.21)	(0.16)	(0.23)	(0.19)	(0.22)	(0.17)
Size	0.001	-0.001	0.001	-0.001	0.001	-0.001	0.001	-0.001	0.001	-0.001
	(0.98)	(-0.45)	(0.98)	(-0.45)	(0.99)	(-0.45)	(1.01)	(-0.43)	(0.99)	(-0.44)
Log(GDP)	-0.068	-0.064	-0.075	-0.070	-0.090	-0.085	-0.110	-0.106	-0.061	-0.057
	(-1.58)	(-1.51)	(-1.88)	(-1.82)	(-2.35)	(-2.31)	(-2.51)	(-2.49)	(-1.59)	(-1.51)
Log(Pop)	-0.029	-0.025	-0.010	-0.006	-0.025	-0.021	-0.042	-0.038	0.023	0.026
	(-0.93)	(-0.83)	(-0.34)	(-0.21)	(-0.96)	(-0.83)	(-1.43)	(-1.34)	(0.69)	(0.80)
Branch	-0.002	-0.002	0.000	0.001	0.001	0.002	-0.001	-0.001	-0.000	0.000
	(-1.07)	(-0.86)	(0.07)	(0.28)	(0.69)	(0.91)	(-0.52)	(-0.30)	(-0.23)	(0.00)
PLocal	-0.049	-0.046	-0.076	-0.073	-0.085	-0.082	-0.041	-0.038	-0.081	-0.077
	(-2.14)	(-1.99)	(-2.71)	(-2.57)	(-2.90)	(-2.77)	(-1.99)	(-1.82)	(-2.86)	(-2.72)
Admin. Area	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year Dum	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
N	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946
Adj. R^2	0.059	0.060	0.059	0.060	0.059	0.060	0.059	0.060	0.059	0.060

Table 9—*Continued*

Panel C: 2SLS								
Dependent Variable	(1) Excess.SW	(2) Excess.VW	(3) Excess.SW	(4) Excess.VW	(5) Excess.SW	(6) Excess.VW	(7) Excess.SW	(8) Excess.VW
Diversity Proxy	Log(LD)		Log(LD-SUB1)		Log(LD-SUB2)		Log(LD-SUB3)	
Diversity	0.158	0.158	0.060	0.060	0.067	0.067	0.067	0.067
	(3.29)	(3.31)	(3.52)	(3.52)	(3.46)	(3.45)	(2.92)	(2.92)
Trust	0.140	0.135	0.120	0.115	0.130	0.125	0.138	0.133
	(3.45)	(3.37)	(3.53)	(3.42)	(3.45)	(3.34)	(2.96)	(2.87)
Log(Wealth)	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(-9.18)	(-8.88)	(-8.90)	(-8.77)	(-8.67)	(-8.53)	(-8.36)	(-8.18)
Sex	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.26)	(0.28)	(0.16)	(0.19)	(0.13)	(0.16)	(0.14)	(0.17)
Log(Age)	-0.008	-0.010	-0.007	-0.010	-0.007	-0.009	-0.007	-0.009
	(-1.46)	(-2.07)	(-1.24)	(-1.82)	(-1.18)	(-1.74)	(-1.18)	(-1.74)
Vol	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.002
	(0.27)	(0.22)	(0.21)	(0.16)	(0.20)	(0.15)	(0.18)	(0.14)
Size	0.001	-0.001	0.001	-0.001	0.001	-0.001	0.001	-0.001
	(0.96)	(-0.51)	(0.99)	(-0.47)	(1.00)	(-0.46)	(1.02)	(-0.44)
Log(GDP)	-0.038	-0.033	-0.069	-0.065	-0.089	-0.084	-0.125	-0.121
	(-0.73)	(-0.65)	(-1.72)	(-1.65)	(-2.42)	(-2.37)	(-3.14)	(-3.12)
Log(Pop)	0.003	0.007	0.004	0.008	-0.018	-0.014	-0.030	-0.026
	(0.10)	(0.21)	(0.15)	(0.27)	(-0.68)	(-0.55)	(-1.04)	(-0.93)
Branch	0.002	0.002	0.002	0.002	0.003	0.004	0.003	0.003
	(0.66)	(0.83)	(0.85)	(1.04)	(1.32)	(1.49)	(0.94)	(1.08)
PLocal	-0.115	-0.112	-0.101	-0.098	-0.107	-0.104	-0.080	-0.077
	(-2.99)	(-2.90)	(-3.07)	(-2.96)	(-3.03)	(-2.93)	(-2.45)	(-2.34)
Controls	YES	YES	YES	YES	YES	YES	YES	YES
Admin. Area	YES	YES	YES	YES	YES	YES	YES	YES
Year Dum	YES	YES	YES	YES	YES	YES	YES	YES
N	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946
Adj. R^2	0.059	0.060	0.059	0.060	0.059	0.060	0.058	0.060

Table 10: Linguistic Diversity and Investor Trading during the 2006-2007 Chinese Stock Market Bubble

This table reports the results of regressing individual investor trading activity and performance on linguistic diversity during the 2006-2007 Chinese Stock Market Bubble. In each year t, we compute the trading activity and performance of each investor. Log(Share) and Log(Value) are the log number of shares and trading value of all stocks traded by each investor. Performance is defined as the difference of the mean cumulative return in year t+1 of the sell portfolio minus the buy portfolio. The mean returns are number of shares-weighted (Excess.SW) or value-weighted (Excess.VW). Closing prices in year t are used for value-weighting. Linguistic diversity measures are Log(LD), Log(LD-SUB1), Log(LD-SUB2), and Log(LD-SUB3). Bubble is a dummy variable indicating the Chinese stock market bubble period of 2006 to 2007. Panel A reports the results with trading activity and Panel B with performance. Controls include Trust, Log(Wealth), Sex, Log(Age), Vol, Size, Log(GDP), Log(Pop), Branch, and PLocal, as defined in Table 3. Administrative area and year dummies are included in all specifications. T-stats are in parentheses and standard errors are clustered by province. The sample period is from 2006 to 2012.

Panel A: Trading Activity								
Dependent Variable	(1) Log(Share)	(2) Log(Value)	(3) Log(Share)	(4) Log(Value)	(5) Log(Share)	(6) Log(Value)	(7) Log(Share)	(8) Log(Value)
Diversity Proxy	Log(LD)		Log(LD-SUB1)		Log(LD-SUB2)		Log(LD-SUB3)	
Diversity	0.113	0.130	0.119	0.127	0.140	0.148	0.181	0.190
	(0.71)	(0.74)	(1.39)	(1.32)	(1.63)	(1.53)	(2.15)	(2.01)
Diversity×Bubble	0.229	0.226	0.074	0.074	0.072	0.075	0.073	0.076
	(5.64)	(3.86)	(3.02)	(2.05)	(4.21)	(2.83)	(5.80)	(3.70)
Controls	YES	YES	YES	YES	YES	YES	YES	YES
Admin. Area	YES	YES	YES	YES	YES	YES	YES	YES
Year Dum	YES	YES	YES	YES	YES	YES	YES	YES
N	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789	1,091,789
Adj. R^2	0.428	0.453	0.428	0.453	0.428	0.453	0.428	0.453

Panel B: Performance								
Dependent Variable	(1) Excess.SW	(2) Excess.VW	(3) Excess.SW	(4) Excess.VW	(5) Excess.SW	(6) Excess.VW	(7) Excess.SW	(8) Excess.VW
Diversity Proxy	Log(LD)		Log(LD-SUB1)		Log(LD-SUB2)		Log(LD-SUB3)	
Diversity	0.037	0.040	0.030	0.032	0.046	0.048	0.046	0.048
	(1.92)	(1.99)	(2.02)	(2.09)	(2.35)	(2.41)	(2.05)	(2.11)
Diversity×Bubble	0.052	0.050	0.028	0.026	0.023	0.022	0.023	0.021
	(1.69)	(1.65)	(2.25)	(2.15)	(1.95)	(1.86)	(1.84)	(1.76)
Controls	YES	YES	YES	YES	YES	YES	YES	YES
Admin. Area	YES	YES	YES	YES	YES	YES	YES	YES
Year Dum	YES	YES	YES	YES	YES	YES	YES	YES
N	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946	1,023,946
Adj. R^2	0.059	0.060	0.059	0.060	0.059	0.060	0.058	0.060

Table 11: Linguistic Diversity and Individual Investor Trading with City Circles

This table reports the summary statistics and OLS regression results of individual investor trading activity on linguistic diversity on the city circle level. For each city, we construct a circle with 75km radius. Log(LD), Log(LD-SUB1), Log(LD-SUB2), and Log(LD-SUB3) are the log linguistic diversity measures for each city circle. In each year t, we compute the trading activity of each investor. Log(Share) and Log(Value) are the log number of shares and trading value of all stocks, respectively. Log(Wealth) is log wealth of the investor. Sex is an indicator variable that equals 0 for male and 1 for female. Log(Age) is log age of each investor. Vol and Size are the value-weighted average volatility and firm market capitalization deciles of each investor's traded stocks. Panel A reports summary statistics and Panel B reports the OLS regression. All regressions include decile dummies for city GDP, Pop, and PLocal. Province and year dummies are also included in all specifications. T-stats are in parentheses and standard errors are clustered by city. The sample period is from 2006 to 2012.

Panel A: Summary Statistics

	Mean	SD	Min	25%	Median	75%	Max
Log(Share)	10.66	4.39	-4.61	10.09	11.63	12.94	22.29
Log(Value)	12.87	4.94	-4.61	12.39	13.99	15.35	24.86
Log(LD)	0.50	0.52	0.00	0.00	0.00	1.10	1.61
Log(LD-SUB1)	1.25	0.78	0.00	0.69	1.10	2.08	2.30
Log(LD-SUB2)	1.56	0.85	0.00	1.10	1.39	2.40	2.56
Log(LD-SUB3)	1.61	0.83	0.00	1.39	1.39	2.40	2.56
Log(Wealth)	11.08	2.88	-4.61	10.41	11.47	12.51	21.31
Sex	0.44	0.50	0.00	0.00	0.00	1.00	1.00
Log(Age)	3.82	0.26	2.56	3.64	3.83	4.01	4.58
Vol	0.13	0.07	0.00	0.11	0.13	0.16	8.30
Size	6.46	2.46	0.00	5.47	7.00	8.15	10.00

Table 11—*Continued*

Panel B: OLS Regression								
Dependent Variable	(1) Log(Share)	(2) Log(Value)	(3) Log(Share)	(4) Log(Value)	(5) Log(Share)	(6) Log(Value)	(7) Log(Share)	(8) Log(Value)
Diversity Proxy	Log(LD)		Log(LD-SUB1)		Log(LD-SUB2)		Log(LD-SUB3)	
Diversity	0.059	0.058	0.245	0.273	0.156	0.176	0.157	0.177
	(0.95)	(0.84)	(3.92)	(4.13)	(3.57)	(3.90)	(3.51)	(3.85)
Log(Wealth)	0.156	0.172	0.156	0.172	0.156	0.172	0.156	0.172
	(22.21)	(21.14)	(22.18)	(21.09)	(22.22)	(21.13)	(22.22)	(21.14)
Sex	-0.241	-0.215	-0.242	-0.215	-0.241	-0.215	-0.241	-0.215
	(-10.66)	(-8.36)	(-10.62)	(-8.33)	(-10.63)	(-8.34)	(-10.63)	(-8.34)
Log(Age)	1.942	2.075	1.939	2.071	1.940	2.073	1.940	2.073
	(29.62)	(28.99)	(29.29)	(28.65)	(29.41)	(28.77)	(29.41)	(28.77)
Vol	19.089	21.872	19.090	21.873	19.090	21.873	19.090	21.873
	(10.89)	(10.90)	(10.89)	(10.90)	(10.89)	(10.90)	(10.89)	(10.90)
Size	0.784	0.901	0.784	0.901	0.784	0.901	0.784	0.901
	(32.66)	(31.87)	(32.63)	(31.84)	(32.65)	(31.86)	(32.65)	(31.86)
City GDP Dum	YES	YES	YES	YES	YES	YES	YES	YES
City Pop Dum	YES	YES	YES	YES	YES	YES	YES	YES
City PLocal Dum	YES	YES	YES	YES	YES	YES	YES	YES
Province Dum	YES	YES	YES	YES	YES	YES	YES	YES
Year Dum	YES	YES	YES	YES	YES	YES	YES	YES
N	749,602	749,602	749,602	749,602	749,602	749,602	749,602	749,602
Adj. R^2	0.434	0.459	0.434	0.459	0.434	0.459	0.434	0.459

Table 12: Linguistic Diversity and Diversity of Opinions with City Circles

This table reports summary statistics and OLS results of diversity of opinions on linguistic diversity on the city circle level using message board data. For each city, we construct a circle with 75km radius. Log(LD), Log(LD-SUB1), Log(LD-SUB2), and Log(LD-SUB3) are the log linguistic diversity measures for each circle. The dependent variable Log(STDEV) is the log standard deviation of the generated buy/sell signals of messages from each circle. The methodology for constructing the buy/sell signals are described in Table 2 and Table 3. Panel A reports the summary statistics and Panel B reports OLS regression results. All specifications include decile dummies for city GDP, Pop, PLocal. Province and firm dummies are also included in all specifications. T-stats are in parentheses and standard errors are clustered by city.

Panel A: Summary Statistics

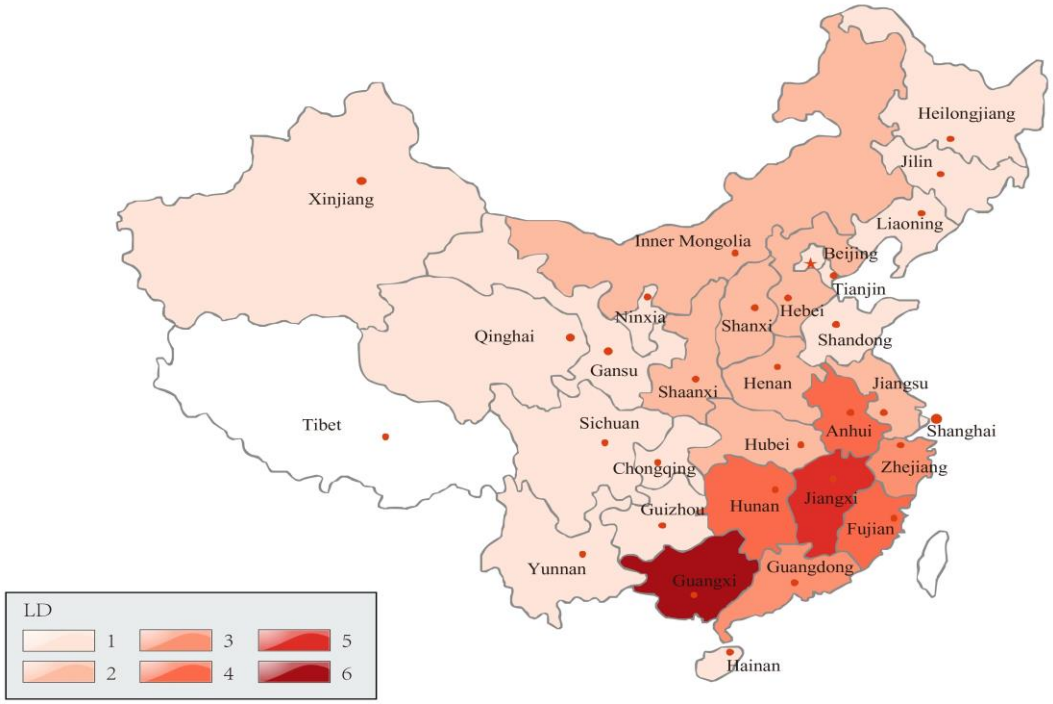
	Mean	SD	Min	25%	Median	75%	Max
Log(STDEV)	-0.33	1.21	-4.61	-0.33	0.01	0.20	1.04
Log(LD)	0.47	0.49	0.00	0.00	0.69	0.69	1.79
Log(LD-SUB1)	1.03	0.70	0.00	0.69	1.10	1.61	2.30
Log(LD-SUB2)	1.39	0.61	0.00	1.10	1.39	1.95	2.56
Log(LD-SUB3)	1.44	0.61	0.00	1.10	1.39	1.95	2.56

Panel B: OLS Regression

Dependent Variable: Log(STDEV)

	(1)	(2)	(3)	(4)
Log(LD)	0.103			
	(0.88)			
Log(LD-SUB1)		0.162		
		(2.02)		
Log(LD-SUB2)			0.243	
			(3.02)	
Log(LD-SUB3)				0.234
				(2.93)
City GDP Dum	YES	YES	YES	YES
City Pop Dum	YES	YES	YES	YES
City PLocal Dum	YES	YES	YES	YES
Province Dum	YES	YES	YES	YES
Firm Dum	YES	YES	YES	YES
N	100,644	100,644	100,644	100,644
Adj. R^2	0.064	0.065	0.066	0.066

Panel A: Heatmap of LD across Chinese Provinces



Panel B: Heatmap of Percentage of Terrain Due to Hills

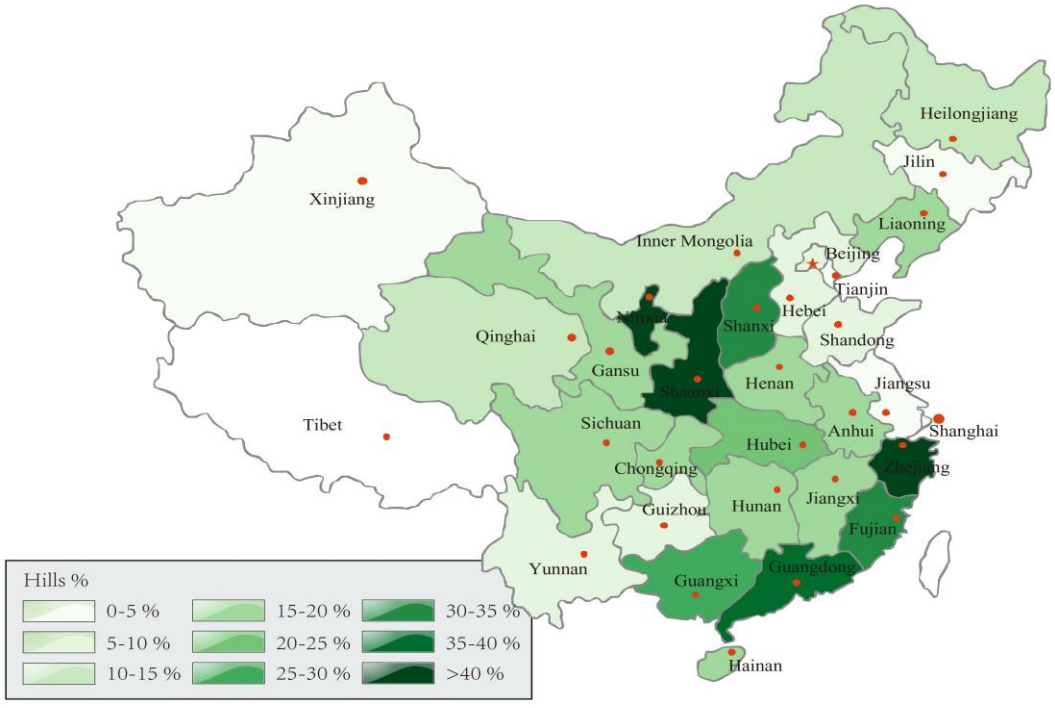


Figure 1 Hills and LD. This figure plots the LD (Panel A) and percentage of hill areas (Panel B) for each province in China. Tibet is excluded from both graphs.