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

Yen-Cheng Chang[†] Harrison Hong[‡] Larissa Tiedens[§] Bin Zhao[¶]

First Draft: July 1, 2013

This Draft: Dec 26, 2013

Abstract

An oft-cited premise for why diverse societies, be it ethnic, linguistic or religious, can grow faster than homogeneous ones is that they bring about diverse opinions, which foster problem solving and creativity. We provide evidence for this premise using a linguistic measure of diversity across Chinese provinces and stock market measures of diverse opinions. This cross-province variation in linguistic diversity is correlated with the extent of hilly terrain in a province but is uncorrelated with financial development in that province. Households in provinces with more linguistic diversity have more diverse opinions as measured by greater trading of and disagreement on stock message boards about local stocks. Linguistic diversity is also correlated with small private enterprise diversity. An analogous cross-country regression suggests that our conclusion extrapolates beyond China.

[*] We are grateful for the comments of seminar participants at 2013 CICF, Second Symposium on China's 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)

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

1. Introduction

Diversity (whether it be measured by ethnicity, language, or religion) is generally thought to hinder economic growth. Economists typically find that measures of diversity, usually a Herfindahl-based index of ethno-linguistic fractionalization, are associated with less economic growth in cross-country regressions (Easterly and Levine (1997)). Fragmentation especially explains the poor economic performance of Africa, where a history of colonization left unstable ethnic compositions and low political rights (Collier and Gunning (1999)). Cross-county regressions in the US also find that racial fractionalization is correlated with less population growth (Glaeser, Scheinkman, and Shleifer (1995) and Alesina and Ferrara (2005)). Diversity leads to racism, prejudices, conflicts of preferences and in some of these cases civil wars, which stifle economic development.

Recently, there is emerging evidence of a bright side of diversity for economic development. Alesina, Harnoss, and Rapoport (2013) find that birthplace diversity, using immigration data from 195 countries, is uncorrelated with ethnic and linguistic fractionalization and is positively correlated with economic growth. This finding is consistent with some earlier suggestive evidence on cultural and immigration diversity in the US being correlated with economic progress (see, e.g., Ottaviano and Peri (2006)). Ashraf and Galor (2013) document an inverted U-shaped relationship between genetic diversity within a population and productivity using extensive data on migratory patterns from the pre-historic period when humans first left Africa. One interpretation of these recent findings is that these diversity measures, when purged of the selection bias from colonization and associated policies of segregation and low political rights, better capture the positive effect of diversity for a society's production possibilities frontier.

The key mechanism often discussed in the literature for why this might come about is that diversity brings about a variety of abilities, experiences, and cultures that may be productive and may lead to innovation (see Alesina and Ferrara (2005) for a review of this literature). The linchpin of this pro-diversity argument is that a diverse society leads to a diversity of opinions, which is good for problem solving and creativity. On the theoretical front, Hong and Page (2001), for instance, show that a more diverse group of people with cognitive limitations can often outperform a more homogeneous group of smarter problem solvers if an individual's likelihood of improving decisions depends more on her having a different perspective from other group members than on her own smarts. Experiments in organizations typically find that more diverse teams do better 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)).

In other words, underlying all the studies that associate diversity measures to economic efficiency is the premise that diversity is good because inhabitants in diverse regions get more stimuli from viewpoints different from their own. We explore in this paper whether this premise is true. We attempt to verify whether diverse societies in fact lead to a diversity of opinions by using a linguistic-based measure of diversity across Chinese provinces and stock market-based measures of diversity of opinions.

To begin with, China is one of the most linguistically diverse countries in the world with at least ten spoken languages. Most provinces still speak in their local language when communicating with friends, families and even local associates and only use Mandarin in very formal business settings. These languages are sometimes referred to as dialects, but in fact they are so mutually unintelligible that linguists refer to them as languages. Linguists often characterize China as being as linguistically diverse as all the countries in Europe combined (see, e.g., Ramsey (1987)). The persistence of local languages 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 cultural diversity.

Using the Language Atlas of China (1987), we calculate the number of languages spoken in different provinces and in different cities in a province. We then define our linguistic diversity measures for each province as the number of languages spoken in that province (LD) and a Herfindahl-based measure of the fraction of the population in the province speaking each language (HDI). 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).

Our linguistic diversity measures are meant to capture the native languages spoken in the different provinces. As we describe below, for older adults in their 50's who are the ones most likely to participate in the stock market, our linguistic diversity measure is an ideal measure of the type of linguistically diverse environment they grew up in. The Chinese Hukou system which strictly limits mobility across China tends to make this diversity differences persistent over time. In other words, even for younger adults, that they grew up with grandparents who spoke multiple languages would be influential even if they do not speak the languages themselves. Indeed, psychologists have shown that just being in an environment where there are multiple languages spoken is enough to foster creativity (Bialystok and Martin (2004), Maddux and Galinsky (2009), Maddux, Adam, and Galinsky (2010), and Kovacs and Mehler (2009)). This creativity presumably might be one channel through which diverse linguistic provinces have more diverse opinions. In this paper, we simply want to causally measure the reduced form association between diversity of languages and diversity of opinions.

What is most interesting about China's linguistic diversity is its geographic origins, which is well known to linguists.[1] 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. 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".

We show below that linguistic diversity is indeed correlated with the terrain of China,

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

using data from the Thematic Database for Human-Earth System. Our measures based on the number of languages typically line up better with the terrain of China and also seem the most robust in the analysis below, though our Herfindahl-based measures are also correlated and yield similar results. So we use as our baseline linguistic diversity measure our simplest measure which is the number of languages spoken in the province. One reason is that the estimates of the fraction of the population actually speaking each language is noisy whereas the number of languages spoken is measured precisely. Perhaps more importantly, both sets are uncorrelated with financial development measures like province GDP per capita. The reason is that province GDP measures are heavily influenced by government policies that explicitly target GDP for government official promotions and these have favored the capital provinces of Beijing and the east coast of China at the expense of the interiors of China. Since the north is flat and the south hilly, the provinces along the east coast of China, which are among the richest in China in terms of GDP per capita, have both flat and hilly provinces.

Hence, we view this cross-province measure as being a good proxy for cultural diversity in different provinces and exogenous enough to be suitable for use as a right-hand side variable in our regressions to explain the diversity of opinions in different provinces. It is similar in spirit to the birthplace diversity measure of Alesina, Harnoss, and Rapoport (2013) and the genetic diversity measure of Ashraf and Galor (2013) in that we view them as having been formed in the far past but which have persistent influences on economic behavior even today. Their measures like ours, are “deeper-rooted” measures of both cultural and ethnic diversity in societies.

Since we are especially interested in opinions regarding economic matters, the stock market would seem a natural place to try to measure this type of diversity of views. So it is surprising that this methodology has not yet been attempted. More precisely, our measure of diversity of opinions builds on two robust findings from the behavioral finance literature on investor behavior. The first is home or local bias of investors: both retail and institutional investors over-weigh stocks with headquarters located near them (Huberman (2001), Coval and Moskowitz (1999), and Grinblatt and Keloharju (2001)). Households especially do not diversify but rather hold concentrated positions in stocks headquartered within 60 miles of where they live. This local bias can be driven by familiarity bias or other informational frictions, whereby investors have both a small radius with which they search for investment opportunities and feel most comfortable with investments they know first hand. This local bias was originally discovered and characterized by French and Poterba (1991) as the international home bias puzzle for the lack of international diversification by equity investors. This local bias of investors is helpful as most of the trading of smaller stocks are done by locals as these stocks are mostly held by locals.

The second finding is that investors trade local stocks because they have different opinions about the future prospects of those stocks (Varian (1989), Harris and Raviv (1993) and Kandel and Pearson (1995), Odean (1999)). A large body of research shows that trading volume can then be used as a market proxy for divergent opinions (see Hong and Stein (2007) for a review of this body of evidence). We can also verify below that the local bias assumption for trading is a good one as we have a random sample of trades emanating from the Shanghai Stock Exchange (SSE) from 2001-2002. This SSE database allows us to know which province the trades emanated from. As a result, we can calculate the analog of our stock turnover measure but just using the trades of locals. We can then assess whether there is more divergence of opinion in different provinces by measuring the trading done in stocks in different regions.

We also develop a second measure of divergent opinions using disagreement on stock message boards, which has been shown to be correlated with share turnover (see, e.g., Antweiler and Frank (2004)). While this message board disagreement measure is more direct than share turnover, it is harder to obtain and more subject to noise compared to share turnover. Hence, we use both measures in our analysis below. Fortunately, China has the second largest stock market in the world, valued at four trillion dollars, and has on average one thousand public firms listed during our sample period. China also has one of the world's largest stock message boards. Thus, we can measure diversity of opinions in each province using the trading volume of stocks headquartered in that province, or as we call them "local stocks," as well as stock messages on these stocks.

Using data from 1998-2012, we regress the log of a stock's share turnover (shares traded to shares outstanding each quarter averaged over the sample period) on the various measures of linguistic diversity in the province where the stock is located. In this pure cross-sectional regression, we control carefully for the market capitalization of each stock and firm news using the firm's stock price volatility. These controls pick up heterogeneity in stock types. Importantly, we also control for GDP per capita of each province since GDP per capita in China is heavily influenced by government development policies as we alluded to earlier. In other words, residual share turnover controlling for stock characteristics and province economic development is our measure of the diversity of opinions in a province.

We regress this abnormal share turnover variable on the linguistic diversity of provinces. For most of our measures, we find an economically meaningful and statistically significant effect with t-statistics of around 2 where we have conservatively clustered standard errors by province. For instance, a one standard deviation increase in our simplest measure Log(LD), the log of the number of languages spoken in the province where the firm is headquartered, is associated with an increase in the stock's log share turnover that is around 7% of the standard deviation of the left-hand side variable. We show that this result is robust to population density controls and dropping extremely developed provinces like Beijing and Shanghai which have people from all provinces working there. These controls typically make our results somewhat stronger as we are more finely measuring linguistic diversity as a result.

We then address more subtle identification worries. The first is that we are not doing a good enough job controlling for economic development in this baseline regression specification. To deal with this issue, we consider a difference-in-difference identification strategy in which we then test to see whether this economic effect is stronger in provinces that are linguistically less segregated. We expect stronger effects where there is true diversity and mixing of languages as opposed to a more segregated experience in which the languages of the province are largely superficial. We measure linguistic segregation by the fraction of the population in each city within each province that speaks the languages of that province. For instance, Zhejiang province has three languages but the inhabitants of the cities typically speak only one of the three languages. In contrast, Fujian has four languages but the inhabitants of the cities in Fujian typically speak two of the four languages. We find that our economic effect is twice as big in provinces with less linguistic segregation, consistent with the importance of diversity as opposed to a purely spurious correlation with GDP per capita. The t-statistics on these estimates are also around 2 to 3.

Our second identification worry is that we are not controlling finely enough for stock characteristics. Our analysis thus far has assumed that there is local bias in terms of the trades of investors but perhaps the correlation with province local stock turnover to the linguistic diversity of that province is due to the type of stocks that locate in that province. Namely, stocks in high linguistic diversity provinces might be more well known nationally and hence they have greater investor participation from around the country and so have greater trading volume or more liquidity. In other words, we need to characterize the investor base of the stocks and see that our turnover effect is coming from the investor base being more linguistically diverse as opposed to simply being more diverse in terms of having more investors.

To rule out this alternative liquidity rationale by measuring a stock's investor base, we follow an earlier work by Wu and Qiu (2012) by using 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. We know the city and province, through a computer's IP address, from where the message originates. For each quarter, we take all the stocks and compute a Herfindahl index calculated from the fraction of messages due to each language. We call this the linguistic diversity of the stock's investor base. At the same time, we also compute a Herfindahl index calculated from the fraction of messages coming from each city (which we call the city diversity of the stock's investor base) and a Herfindahl index of the fraction of messages coming from one of five tiers of provinces defined by GDP per capita (the GDP diversity of a stock's investor base). If our baseline regression is well identified, we expect that it is the linguistic diversity of a stock's investor base driving our turnover findings and not the city or GDP diversities of the stock's investor base.[2]

As expected, we show that the number of languages spoken in the province where the stock is headquartered is strongly positively correlated with the linguistic diversity of the stock's investor base. The more the languages spoken in a province the higher the linguistic diversity of the investors. The t-statistic is around 10 without clustering of standard errors by province and is around 7 when we cluster standard errors of the first stage regression. We then consider the full instrumental variables estimation where we use this previous regression, the linguistic diversity of a stock's investor base on the number of languages spoken in the firm's headquarter province, as the first stage regression. The second stage is then log turnover on the fitted value of the linguistic diversity of a stock's investor base. We find economic and statistical significance. This 2SLS estimation then identifies the effect of diversity in a province to the trading of stocks in that province and hence the diversity of opinions in that province.

And as we mentioned above, we also verify our baseline turnover on linguistic diversity results using our own measure of trading by locals using the SSE data from 2001-2002. The Shanghai Stock Exchange is only one of the two main stock exchanges in China and so it does not capture all the trading done in China. This is why we prefer the entire turnover data as our baseline result. But it is comforting to find that we get similar results for a subset of trades which we know emanate directly from households living in a given province.

[2] Our findings that local message board activity explains local stock trading activity is consistent with the limited attention literature as in DellaVigna and Pollet (2007), Hirshleifer and Teoh (2003), Odean and Barber (2008).

In addition, we also use the stock message board data directly to develop measures of disagreement using textual analysis and machine learning (see, e.g., Mehl (2006)) on the messages posted. We randomly sample stocks from both high linguistically diverse and low linguistically diverse provinces. We then download recent posts for each stock. Posts on stocks typically offer a buy or sell recommendation. For each firm we download the most recent 10 pages of messages. The sample consists of a total 796,809 message posts.

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)). The opinions in each post is coded by two graduate students independently with -2, -1, 0, 1, 2, denoting Strong Sell, Sell, Neutral, Buy, and Strong Buy. 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, to categorize all the messages.

We then calculate for each stock its disagreement measure which is the standard deviation in these scores across the posts of the stock. We use this disagreement measure instead of turnover and find similar results. Note that this divergence of opinion measure at the stock level using textual analysis is correlated with the turnover in that stock as we expected given earlier work.

Finally, the literature on diversity suggests that diversity of opinions ought to be correlated with some measures of real economic activity or productivity. Of course, GDP per capita is a bad measure since it is influenced by government policies. Indeed, even publicly traded firms are heavily influenced by government policies as the government decides how many companies can go public in any given year. The only part of the Chinese economy that is arguably less affected are the private enterprises in China. These private enterprises are much smaller than public firms or state-owned enterprises. We have a unique dataset with over one million of such private enterprises. We construct a measure of private enterprise activity for each province and run the analogs of our earlier regressions. Controlling for the usual co-variates, we find that linguistically diverse provinces are composed of a greater fraction of private enterprises, consistent with the premise that diversity of opinions has some positives for a society's production function. The literature also speaks to the role of government favoritism for some groups (here in our context Beijing and the northern provinces close to the capital city) versus other regions as emphasized in Alesina and Ferrara (2005). So the fact that we find a thriving small entrepreneurial sector in the diverse linguistic Southern provinces may be a lower bound on the positive effects of diversity for economic activity.

Our contribution is to find a well-identified empirical design with which to study the premise that diversity brings about diverse opinions. But one worry with well identified empirical designs is the concern of extrapolation beyond that design. As such, we extend our baseline regression specification into an international sample of forty-one countries to see if greater language diversity in a country is correlated with higher stock market turnover in that country. In our analysis, we control for a host of country characteristics including the economic development, the size of the stock market and other institutional controls. This regression is less well-identified. So our point here is merely to suggest that our conclusions might extrapolate beyond China. We have no identification strategy here but we think it is somewhat informative to see if the correlations match our more causal analysis. It turns out that the economic effects are quite significant—a one standard deviation move in the linguistic diversity across countries leads to a 20% of a standard deviation move in the country's stock market turnover—but the t-statistics given the limited sample size are marginally significant.

Our results are related to a growing body of work on the importance of social interaction, culture and language in economic exchange. A number of papers have shown that social interaction and networks influence a range of economics outcomes from welfare participation, retirement investing and stock market participation (Bertrand, Luttmer, and Mullainathan (2000), Duflo and Saez (2003), Madrian and Shea (2001), Hong, Kubik, and Stein (2004)).

Social capital and culture (Guiso, Sapienza, and Zingales (2004) and Guiso, Sapienza, and Zingales (2009)) have also been shown to influence economic exchange. Another literature is the impact of the structure of language on economic behavior. Notably, Chen (2013) finds that languages through the strength of the association of present to the future influences savings behavior through a discount rate framing mechanism. These structural differences might also drive our effects. Such a structural language mechanism might also lead to variation of beliefs across languages. Our diversity findings might be interpreted as examining the linkages between these various strands of literature.

Our paper proceeds as follows. We describe our datasets in Section 2. We present the main results in Section 3. In Section 4, we show results from text analyses of stock message boards. We present the small private enterprise activities in Section 5 and international sample results in Section 6. We conclude in Section 7.

2. Linguistic Diversity, Geography and GDP

Our primary measure of linguistic diversity (LD) captures the cultural diversity of a location that lasts for generations and could be traced back to the geographic features of each province. We use two alternative sources that employ similar survey methodologies of geographic linguistics to measure LD. First 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 linguistics and other social sciences. 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 is done from 2001 to 2007.

In these studies, survey posts are in general 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. Not surprisingly, the province level LDs from these two studies are very similar with a correlation coefficient of 0.95, speaking to our presumption that LD captures long-lasting variations in diversity. Across these two studies, province-level LDs are different in only three provinces[3]. 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.

Note that Mandarin (Guan) is not included in four provinces (Fujian, Guangdong, Hainan, Shanghai) in both studies though one may argue that younger residents in these provinces do speak it due to the promotion of Guan over the past two to three decades. However, below we show that the promotion of Guan did not start until the 1980s and thus we believe do not affect our goal of capturing diversity of indigenous residents.

The current regime in 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 foundation of Guan promotion[4]. In particular, for these four provinces that do not include Guan, the State Language and Letters Committee

[3] Fujian, Guangxi, and Hunan have LDs equal to 3, 4, and 5 respectively.

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

(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 our belief that our measure LD is immune to the promotion of Guan.

Finally, the household registration (Hukou) system in China greatly restricts demographic mobility. In China, one would not be able to receive housing and health care benefits 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. This special feature in China is helpful in our context since linguistic diversity will be stable over long periods of time.

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 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. We will use the sub-languages in each province, which are denoted by LD-SUB1, LD-SUB2, LD-SUB3, in the robustness section.

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. These provinces typically speak Guan but also a number of local languages. Fujian has four languages, while Zhejiang and Guangdong each has three.

The distribution of these provinces along the coast will help us below to 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. Notice from Table 1 that for higher Log(GDP) provinces, there is variation in linguistic diversity from one language to as much as three languages.

In addition to the number of languages spoken, we also calculate a Herfindahl-based measure of linguistic diversity that takes into account the population speaking the languages in each province. We do not have actual estimates of the population who speaks each language. We only know what languages are spoken in each city. However, we can yet create a Herfindahl-based measure if we assume that the city population speaks all the languages attributed to that city. Then we can calculate a province level linguistic diversity measure HDI that is simply one minus the Herfindahl index based on share of the province population who speaks each language. The measures for each province are reported also in Table 1 and they go from 0 (for everyone speaks the same language) to 1 (if everyone speaks different languages). This HDI measure and our LD measure are highly correlated at 0.91. HDI-SUB1, HDI-SUB2, HDI-SUB3 are analogs for the various sub-languages in cities for each province that we will consider in the robustness section.

In Table 1, we also report the percent of the terrain that is hilly, mountainy and watery in the different provinces, where the data on hills, mountains and water come from the Thematic Database of Human-Earth System. The data reports the fraction of the area in that province that is occupied by hills, mountains and water. 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.

In Table 2 column (1), we test this association formally by regressing the province-level linguistic diversity measure on the percentage of hills in each province (H%) and Log(GDP). In the regressions in Table 2, we have dropped the three provinces (Yunnan, Sichuan and Chongqing) bordering and including Mount Everest since Mount Everest takes up such a huge part of the land area that a regression including these provinces is uninformative. Mount Everest is so large and so far west that it has little population density. The coefficient in front of H% is 3.966 with a t-statistic of 1.97. The R^2 of the regression is 0.107. Notice that the coefficient on Log(GDP) is statistically insignificant. The coefficient is -3.718 but only has a t-statistic of -0.72.

In column (2), we regress the same left-hand side variable on the sum of H% and M% (HM%). We find in column (2) that adding mountains really does not increase the explanatory power of terrain for languages spoken in a province. If anything, it lowers it since large mountains presumably inhibit the number of inhabitants. The coefficient is still positive and economically significant but the t-statistic is only 1.49. In column (3), we add in water area of a province as well (HMW%) and our results improve. The coefficient is 2.182 with a t-statistic of 1.61. So it appears that hills seem to drive much of the explanatory power for diversity and that water adds some incremental explanatory power. But in all these regression specifications, Log (GDP) plays no role in explaining the linguistic diversity of a province.

In other words, linguistic diversity is correlated with geography but is uncorrelated with Log(GDP). As we argued in the introduction, Log(GDP) really picks up economic policies of governments which have heavily favored the east coast of China at the cost of interiors of China. It turns out that the terrain of the east coast of China has both flat provinces in the northeast and hilly provinces in the southeast. As a result, we are lucky that our linguistic diversity measure is uncorrelated with government policies which might affect our inference. We will still control for Log(GDP) in our regression specifications below but we are comfortable in thinking of our linguistic diversity measure LD as exogenous, very much in the spirit of recent papers in the literature such as Alesina, Harnoss, and Rapoport (2013) and Ashraf and Galor (2013).

In Panel B, we use our HDI measure as the left-hand side variable rather than LD and we find that LD is more explained by terrain than HDI. The R^2's of the regressions for LD are larger than for HDI. Also, given that HDI depends on certain assumptions we make, we use LD as our preferred measure of diversity. It turns out that the results do not differ too much in any event.

3. Linguistic Diversity of Province and Local Stock Share Turnover

Our measure of diversity of opinions in each province is the average trading volume of stocks headquartered in each province. We collect our stock trading volume and market capitalization variables from CSMAR for each quarter in the period of 1998 to 2012. More precisely, our diversity of opinions measure is share turnover, which is defined as number of shares traded each quarter divided by total number of tradable shares. In addition, we restrict our baseline sample to provinces in the top four quintiles in terms of GDP per capita and omit stocks in the lowest market capitalization decile. We believe these stocks are ex ante illiquid and so share turnover might be less informative about disagreement. We show that results are robust to alternative sample cuts below.

The summary statistics for our baseline sample is given in Table 3. We report in Panel A statistics by province. In particular, we sort the provinces by the number of languages (LD). We also report the HDI in each province again for convenience. Next to HDI, we calculate for each province the median of the fraction of the languages in the province spoken by cities in that province. This variable is called CS or city share. We will think of a higher CS as being a province that is linguistically less segregated and hence genuinely diverse. For instance, take Fujian which has four languages but a CS of just .25. This means that a city in Fujian province typically just speaks one of the four languages. In contrast, Hunan, which also has four languages, has a city share CS of 0.5. As we explain below our city share variable will help us deal with certain identification issues by asking whether Fujian or Hunan has more diversity of opinions.

We then report Turn which is the average turnover of stocks located in each province. Notice that even in these simple summary statistics one can see our baseline effect. The provinces with the three highest average turnover are Henan at 2.05 (or 205% per quarter), Jiangsu at 1.77, and Zheijiang at 1.72. Henan has 2 languages, Jiangsu has 2 languages and Zhejiang has 3 languages. The provinces with the three lowest average turnover are Heilongjiang, Liaoning, and Tianjin and all have only one language. Shanghai is tied for third lowest and also has one language.

We also show the average firm market capitalization in each province. Beijing has by far the largest stocks but otherwise there are not any ostensible patterns related to LD. We then report the average volatility of stocks in each province (VOL). There are only small differences in these averages across provinces.

In Panel B, we report the summary statistics for the pooled sample which constitutes the basis for our baseline regression. We report mean, standard deviation, and various percentiles for our key variables of interest.

In Table 4, we then regress log turnover on the log number of languages spoken in the province where the stock is located:

$$Log(Turn_i) = \alpha + \beta Log(LD_i) + \gamma' \mathbf{X} + \epsilon_i \tag{1}$$

where the dependent variable, Log(Turn) is the log of mean quarterly turnover over the sample period, LD is the number of languages spoken in a firm's home province, and Log(LD) is the log of this variable. The control variables **X** include market capitalization, GDP, and volatility (VOL) decile dummies over the sample period. Panel A reports results from 1998 to 2012, and Panel B from 2008 to 2012. Standard errors are clustered by province.

From Panel A, columns (1) and (2) report the baseline results with and without VOL controls. The coefficient is 0.06 with a t-statistic of 1.81 when there are no VOL controls. It is .065 with a t-statistic of 1.89 when there are a full set of VOL controls. In other words, our results get stronger when we include volatility controls. We worry that there is somehow heterogeneity in the amount of news as opposed to the level of disagreement that might be driving share turnover since news typically triggers trading. We are comforted then to find that even controlling for news our baseline result gets stronger. Here we find an economically meaningful effect. The coefficient of the log of the number of language, when multiplied by a standard deviation of this independent variable of interest, yields an increase in share turnover for stocks in that province that is 7% of a standard deviation of Log(Turn). This is a non-trivial economic effect. Likewise, we get similar results when we use LD instead of the log of this variable. The statistical and economic significance are around 10% weaker but we are assured that our results are robust to different regression specifications.

In Panel B, we report the results for the most recent period of 2008-2012. The reason we also focus on this sub-sample is that one of our identification strategies below relies on message board data that is only available in this sub-sample. As such, we want to verify that our baseline regression results are robust across different sub-periods. Turning to the results in columns (1) and (2), we find largely similar effects. The coefficient of interest is 0.052 with a t-statistic of 1.98 in column (1) and 0.053 with a t-statistic of 1.93 in column (2). The statistical significance is actually stronger and so is the economic significance than over the whole sample period. But this difference is not too large, suggesting that our estimates in Panel A are quite robust. Moreover, we view the recent sample as being more informative since the Chinese market in the early sample period has fewer stocks than the recent sample.

Panel C of Figure 1 shows a heatmap of turnover across the provinces of China: the darker the color the higher the turnover in the province. We see that if we compare Panels A, B and C, the diversity, the hilly terrain and the higher turnover all line up together in same provinces. This is particularly true as we look up and down the east coast of China, where economic development is fairly similar. Figure 1 reinforces graphically our baseline results in Tables 2 and 4.

In Table 5, we add in an additional control for economic development, which is the population of a province. As Glaeser, Scheinkman, and Shleifer (1995) point out, for a country in which there is potentially more labor mobility than across countries, population might be a better measure of economic development than GDP per capita. In the case of China, the worry is that Chinese government policy also heavily influences population in provinces through the enforcement of residency and work permits. Indeed, the worry we have is that population is then naturally correlated with linguistic diversity to the extent that more population means more people speaking different languages. In our sample, linguistic diversity is positively correlated with population. As such, we want to see if population explains our results. In Table 5, we add in log population as an additional covariate to our baseline regressions. We see that our effects are only affected slightly. The coefficients on linguistic diversity are largely unchanged. The t-statistics are weaker in Panel A, the full sample, but actually stronger in Panel B the more recent sample. Indeed, in the recent sample which we view as more informative since the stock market now has many more stocks and liquidity, we see a much stronger effect for all our specifications. For the baseline specifications in columns (1) and (2), the t-statistics are now above 2. Moreover, in the LD specification, we even get t-statistics now close to 2.

In Table 6, we drop Beijing and Shanghai from our analysis. Beijing and Shanghai are special cities in that there are people from all over China living in these two cities. Since they are provinces with few official languages, we worry that they might bias our inference. We see from both Panels A and B that the coefficients in front of Log(LD) remains largely the same compared to Table 4. Our Log(GDP) controls in Table 4 appears to largely control for potential confounds associated with provinces like Beijing and Shanghai.

As such, we conclude from Tables 4, 5 and 6 that we have made a reasonable effort to address worries about omitted variables related to economic development which is in the control of government policies and not necessarily related to linguistic diversity.

3.1. Identification Strategy 1: Baseline Results by Linguistic Segregation or Integration

Nevertheless, we try to improve on this effort in two ways. The first is that we can do an even better job controlling for economic development in this baseline regression specification by considering a difference-in-difference identification strategy in which we then test to see whether this economic effect is stronger in provinces that are linguistically less segregated. If linguistic diversity is spuriously correlated to turnover through an omitted variable, we can use this auxiliary prediction of diversity to tease out identification. Our proposal is that if linguistic diversity is directly causing more turnover and not spuriously correlated with turnover, then we expect that we should find a stronger effect in provinces where inhabitants of diverse language backgrounds live close to each other in a city, as opposed to where homogeneous inhabitants clustering in separate cities.

We measure linguistic segregation by the fraction of the population in each city within each province that speaks the languages of that province. This is our CS variable reported in Table 3. For instance, Zhejiang province has three languages but the inhabitants of the cities typically speak only one of the three languages. So its CS is 0.33. In contrast, Hunan has four languages but the inhabitants of the cities in Hunan typically speak two of the four languages. So its CS is 0.5. We expect stronger effects where there is true diversity and mixing of languages as opposed to a more segregated experience in which the languages of the province are largely superficial.

To measure this channel, we estimate the following regression specification:

$$Log(Turn_i) = \alpha + \beta_1 Log(LD_i) + \beta_2 CS_i + \beta_3 Log(LD_i) * CS_i + \gamma' \mathbf{X} + \epsilon_i \qquad (2)$$

where we interact Log(LD) with CS so that our coefficient of interest is β_3. In other words, we expect that the coefficient of β_3 to be positive if integration and true diversity in a province matters for stock trading in that province. An alternative specification which we also estimate simply breaks up the baseline coefficient into an effect for high linguistic provinces and an effect for low linguistic provinces. More specifically, the regression specification is given by

$$Log(Turn_i) = \alpha + \mu_1 Log(LD_i) * CS.High_i + \mu_2 Log(LD_i) * CS.Low_i + \gamma' \mathbf{X} + \epsilon_i \qquad (3)$$

where CS.High is a dummy variable which equals one if CS is greater than 0.4, and zero otherwise; and CS.Low is a dummy variable which equals one if CS is lower than or equal to 0.4, and zero otherwise. The rest of these two regression specifications are similar to the baseline one in terms of sample, control variables, and clustering of standard errors.

The results are reported in Table 7. Panel A has the results for the full sample. In column (1), we report the results for the Log(LD) specifications. First, the interaction coefficient of interest in column (1) is 0.33 with a t-statistic of 2.03. This says that the effect of linguistic diversity on turnover is indeed stronger for less linguistically segregated provinces. It is worth dwelling on what this regression is doing. By multiplying Log(LD) with CS, we are essentially re-weighting the Log(LD), whereby provinces with high CS effectively get a higher diversity score. For instance, a province with two languages but with a CS score of 0.5 effectively gets treated as a province with 1 language. In essence, we are comparing more extreme provinces in terms of linguistic diversity scores.

In column (2), we split the baseline effect into high versus low CS scored provinces. We choose the cut-off of 0.4 to get enough provinces into the low CS group. Here, the coefficient for high CS provinces is 0.129 with a t-statistic of 2.98. The coefficient in front of low CS provinces is 0.056 with a t-statistic of 2.05. We can interpret this as the effect of linguistic diversity on turnover for high CS provinces is roughly twice as large as for low CS provinces. Columns (3) and (4) report the results for LD and we see similar effects.

In Panel B, we report the results for the sub-sample of 2008-2012 and we find similar effects. As we pointed out, while we feel that we are fortunate that linguistic diversity is fairly exogenous and hence makes a good right-hand side variable, it is still comforting that this diff-in-diff strategy yields confirming results to our baseline ones.

3.2. Identification Strategy 2: Linguistic Diversity of Local Investor Base and Local Stock Share Turnover

The second identification worry we have is that we are not controlling for enough stock characteristics. We have focused on market capitalization and stock price volatility. Both are introduced as covariates for different reasons but it still might be the case that there are missing stock characteristics that might bias our inference. Namely, stocks in high linguistic diversity provinces might be more well known nationally and hence they have greater investor participation from around the country and so have greater trading volume. In other words, we need to characterize the investor base of the stocks and see that our turnover effect is coming from the investor base being more linguistically diverse as opposed to simply being more diverse in terms of having more investors. To deal with this issue, we consider an instrumental variables technique where we estimate the relationship between share turnover for a local stock and the linguistic diversity of the investor base of that stock.

We measure the linguistic diversity of a stock's investor base using the guba.eastmoney.com message board. Over the period of 2008-2012, we track using guba.eastmoney.com the number of messages and the city origin of each message for each stock in the CSMAR universe.[5] Specifically, we use the IP addresses of original posts to obtain the city origin with the QQ IP address geo-mapping database. Since in this paper we are focusing on the language diversity in Mainland China, we drop all posts that can be traced to overseas origins. Finally, to include only meaningful posts, we drop posts with less than or equal to 5 replies from the users. Using these messages, the language Herfindahl for stock i is calculated as follows:

$$H_i^{Lan} = \Sigma_{l=1}^{L}(\frac{n_i^l}{N_i})^2 \tag{4}$$

[5] We download all messages posted between 2008 (when guba.eastmoney.com started) and May 2013 (when we did the dowload of their site). We do not know the dates of these posts and as such we are simply measuring linguistic diversity of the stocks using the cumulative posts on these message boards. The message board company might randomly take down some posts or might delete some older posts. Our turnover data ends in 2012 and hence our sample for the dependent variable of interest in this analysis is 2008-2012.

where $n_i^l = \Sigma_{m=1}^{M}(N_{i,m} \times Prob(l)_m)$, and $Prob(l)_m = \frac{1}{Lm}$. Here N_i is the number of message board posts, and n_i^l is the sum of the posts by speakers of language l across all provinces. $N_{i,m}$ is the number of posts for stock i from province m. L_m is the total number of languages spoken in province m. $Prob(l)_m$ is the probability that a post is posted by speakers of language l in province m. If a province m has only one language, then $Prob(l)_m$ is 1. If a province has more than one language, then $Prob(l)_m = \frac{1}{L_m}$. For example, Hubei province has two languages: Guan and Gan. In this case, $Prob(Guan)_{Hubei} = Prob(Gan)_{Hubei} = 0.5$.

The City Herfindahl for stock i is calculated as follows:

$$H_i^{city} = \Sigma_{j=1}^{C}\left(\frac{n_i^j}{N_i}\right)^2 \tag{5}$$

where N_i is the number of message board posts for stock i, C is the number of cities in China in our sample, and n_i^j is the number of posts originated from city j for stock i.

The GDP Herfindahl for stock i is calculated as follows:

$$H_i^{GDP} = \Sigma_{j=1}^{5}\left(\frac{n_i^j}{N_i}\right)^2 \tag{6}$$

where N_i is the number of message board posts for stock i, and n_i^j is the number of posts originated from tier j provinces for stock i.

The idea is that we estimate the effect of a stock's language Herfindahl on turnover while controlling for its city and GDP Herfindahls. The latter two pick up stock characteristic pertaining to the investor base such as whether the stock is known nationally or known more in richer provinces. We then instrument for a stock's language Herfindahl using our linguistic diversity measure. The first check is that linguistically diverse provinces should have stocks with more linguistically diverse investor bases. This is the first stage regression. The second stage regression is then to run turnover on the predicted value of a stock's language Herfindahl where the prediction comes from the linguistic diversity of the province. This 2SLS strategy will always include as covariates the city and GDP Herfindahls. So we should be able to adequately address all remaining concerns on omitted stock characteristics driving our results.

Before we present these findings, we report messages posed by locals on stocks located near them is highly correlated with turnover of those stocks. In contrast, messages posted on the same stocks by non-locals is not very correlated with turnover in these stocks. This differential correlation pattern reflects the fact that there is local bias in trading and that the turnover we observe is largely driven by locals and local attention. This local bias in trading attention is also implicitly captured in our identification strategy.

Table 8 presents the summary statistics for our sample. To be consistent with our baseline regressions, we restrict our sample to Tier 1 to Tier 4 provinces and omit the smallest decile stocks. In Panel A, the summary statistics are sorted by the GDP per capita of the province. We also report for convenience the number of languages. It is also easy to see that Zhejiang, Fujian and Jiangsu have lower language Herfindahl index compared to Jilin, Beijing, and Shanghai. We also report the GDP Herfindahl and city Herfindahl along with the number of posts and the number of firms in each province. In Panel B, we report the pooled summary statistics of the variables we use in the following empirical analyses.

Our expectation is that stocks headquartered in linguistically diverse provinces will have a lower language Herfindahl. In Table 9, we regress the log of average quarterly language Herfindahls for a stock on the number of languages spoken in the province where the stock is headquartered. The regression specification is given by

$$Log(H_i^{Lan}) = \alpha + \beta_1 Log(LD_i) + \beta_2 Log(H_i^{City}) + \beta_3 Log(H_i^{GDP}) + \gamma' \mathbf{X} + \epsilon_i \qquad (7)$$

where $Log(H_i^{Lan})$ is the log of language Herfindahl, and $Log(LD_i)$ is the log number of languages spoken in firm i's headquarter province. The other control variables are identical to earlier regression specifications.

The estimate of β_1 is -0.109 with a t-statistic of -7.25. We further consider two alternative measures for number of languages. First, we take simply the number of languages, LD_i. In this case, the coefficient is -0.049 with a t-statistic of -6.35. Second, we use a dummy variable, $LD.Dum_i$, which equals 1 if the firm's headquarter province has more than one language, and zero otherwise. The coefficient is -0.089 with a t-statistic of -8.21. The results suggest that language is a strong instrument for a stock's language Herfindahl.

In Table 10, we then consider the full instrumental variables estimation where the first stage is log Herfindahl of language associated with a stock on the number of languages spoken in the firm's headquarter province. The second stage is log turnover on the fitted values of log language Herfindahl.

The regression specification is given by:

$$Log(Turn_i) = \alpha + \beta_1 \widehat{Log(H_i^{Lan}})+ \beta_2 Log(H_i^{City}) + \beta_3 Log\left(H_i^{GDP}\right) + \gamma'\mathbf{X} + \epsilon_i \qquad (8)$$

where $\widehat{Log(H_i^{Lan}})$ is the fitted value from each of the first-stage regressions, and the other covariates are listed in the captions of Table 10. All three specifications of our instrumental variables give consistent estimates. The implied economic effect, for instance from the first specification, for log turnover on log Herfindahl of language is 0.073 of a standard deviation of the left-hand side. The t-statistic is also a highly significant 2.57.

3.3. Alternative Linguistic Diversity Measures and other Robustness Checks

In addition to the number of languages spoken, we also calculate a Herfindahl-based measure of linguistic diversity that takes into account the population speaking the languages in each province. We do not have actual estimates of the population who speaks each language. We only know what languages are spoken in each city. So we assume that the city population speaks all the languages attributed to that city. We then calculate a province level linguistic diversity measure HDI that is simply one minus the Herfindahl index based on share of the province population who speaks each language. The measures for each province are reported also in Table 1 and they go from 0 (for everyone speaks the same language) to 1 (if everyone speaks different languages).

In Table 11, we then re-run our earlier baseline regressions using the right-hand side variable Log(1+HDI). Panel A reports the results for the full sample. Panel B reports the results for the sub-sample period of 2008-2012. Notice that we get similar qualitative effects. Some estimates are statistically weaker while others are statistically stronger than the baseline ones using LD. But the conclusions we draw are similar.

We next use linguistic diversity measures based on sub-languages. In Table 12, we report the distribution of our various sub-language linguistic diversity measures. For convenience, we report LD for each province again and then LD-SUB1, LD-SUB2, and LD-SUB3. These sub-language measures are the number of unique sub-languages in each province. These sub-languages take into account dialects within different languages. For instance, we see for Shanghai that LD is 1 (its language is Wu) and all the LD-SUBs are also 1. This means Shanghai's language has no variations. But if we look at Beijing, its LD is 1 but its LD-SUBs are 2, which means there are two types of Guan spoken in Beijing. If we look at Zhejiang, its LD is 3 but its LD-SUB1 is 11. This means that among the 3 languages in Zhejiang, there are a large number of dialect variations in these 3 languages. Its LD-SUB2 and LD-SUB3 are 16, which means that there is even variations in dialects among these sub-languages.

The thing to note from this discussion and Panel B is that LD is not perfectly correlated with LD-SUB's. As a result, it will be interesting to see if these sub-language measures also give us information about turnover. Given each of these LD-SUBs, we calculate the population that speaks each of these sub-dialects to get the analogs for HDI-SUB1, HDI-SUB2 and HDI-SUB3.

In Table 13, we replicate our baseline results from Table 4 using the various sub-language measures. It is easy to see that the results are largely consistent with Table 4. The economic magnitudes vary. In some cases, we get stronger results relative to Log(LD), such as in the case using Log(LD-SUB2). In others, we get slightly weaker results, as in Log (LD-SUB1). In the case of Log (1+HDI), we get much stronger results using the sub-languages.

Finally, we have also performed a battery of robustness tests for our baseline regressions (Table 4), identification strategy 1 with city share (Table 7), and identification strategy 2 with language Herfindahls (Tables 9 and 10). Instead of our current sample including the 24 richest provinces, we can focus on the top 12 richest provinces (or top two GDP per capita quintile). Most of these provinces lie along the east coast and this should further assure us results are not confounded by economic development. Similarly, another alternative is to focus on provinces with at least 30 listed firms. We also winsorized extreme turnover to alleviate concerns that these firms might be driving our results. Our results (not reported for brevity) are robust to these different empirical considerations and are all qualitatively similar to those presented in Section 3.

3.4. Local Trading Using Shanghai Stock Exchange Data

Our analysis linking the investor base of a stock using the message board data to turnover is already quite sufficient in vetting the premise behind our baseline results of turnover of a stock regressed on the linguistic diversity of the province where that stock is located. Ideally, we would have liked to simply calculate the trading done by locals in a province of their local stocks. While we cannot do this, we do this partially using a random sample of Shanghai Stock Exchange (SSE) trading data from April 2001 to August 2002 [6]. We can verify that the local bias assumption for trading is a good one as we have trades emanating from the Shanghai Stock Exchange (SSE) from 2001-2002. This SSE database allows us to know which province the trades emanated from. As a result, we can calculate our analog of stock turnover measure but just using the trades of locals. The Shanghai Stock Exchange is only one of the two main stock exchanges in China and so it does not capture all the trading done in China. This is why we prefer the entire turnover data as our baseline result. But it

[6] We obtain this data from Ng and Wu (2010).

would be comforting to find that we get similar results for a subset of trades which we know emanate directly from households living in a given province.

These results are presented in Table 14. In total we have 36,917,571 trades over this time period, representing roughly 25% of the trading volume of SSE-listed stocks. Log(Turn^L) is the number of shares bought and sold by households in its home province divided by the number of outstanding shares. Panel A presents the summary statistics of the key variables for our regression. In Panel B, we then regress this turnover due to locals on our linguistic diversity measure, along with the same control variables we have in our baseline regressions. Note that in addition to the same sample restriction of Tier 1 to Tier 4 provinces in GDP per capita and omitting the smallest decile firms, we also restrict to provinces with at least 15 stocks. With the data covering only SSE-listed stocks from 2001 to 2002, there are provinces with only a handful of listed firms. This is less of an issue in our baseline sample where we have firms from both exchanges covering more recent time periods. With this restriction we have a total of 15 provinces and 464 stocks. We find, consistent with our earlier baseline turnover results, that linguistically diverse provinces have greater local trading of local stocks than less linguistically diverse provinces. The coefficients of Log(LD) and LD are both statistically significant with t-stats over 7 and economic significance over 35%. As such, we conclude that our turnover findings are consistent with our hypothesis that households from more linguistic diverse provinces have greater divergence of opinion.

4. Linguistic Diversity and Diversity of Opinions on Stock Message Boards

In addition to working with turnover, we also employ machine learning methods to measure disagreement of the messages posted for each stock. Our 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.

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)). The opinions in each post is coded by two graduate students independently with -2, -1, 0, 1, 2, denoting Strong Sell, Sell, Neutral, Buy, and Strong Buy.

4.1. Baseline Results with Training Sample

Before generating buy/sell signals for the full sample, we can first use the student coded sample and run a regression analogous to our baseline result in Table 4. We compute STDEV for each stock, which is the standard deviation of the human generated scores across the posts. We have three measures of each firm. STDEV1, STDEV2, and STDEVM are the STDEV using the two students' codings and their average. These will be our dependent variables of interest. The summary statistics are reported in Panel A of Table 15.

Note in passing that the correlation between TURN from our earlier analysis and students STDEV are 0.129 and 0.137 for student 1 and student 2, respectively. Its 0.158 with the STDEV using the student average. In other words, we find consistent with earlier work on internet message boards that diversity in opinions about a stocks future performance is indeed correlated with that stocks turnover (see, e.g., Antweiler and Frank (2004)).

Further, we expect more linguistically diverse provinces to have more disagreement about stocks headquartered there in terms of these standard deviation measures. We test this hypothesis in Panel B. Log(LD) is the log number of languages spoken in a firms home province and our independent variable of interest. We find indeed that more linguistically diverse provinces have greater disagreement in terms of these STDEV measures. For instance, in column (1), we have the measure using the first student's score. The coefficient of interest is 0.028 with a t-statistic of 2.6. In column (2), we show the results using the second students score. The coefficient is similar though the t-statistic is smaller around 1.17. In column (3), we simply take the average the two student's scores and run the regression using this average. The coefficient of interest is 0.04 with a t-statistic of 1.59. So we get marginally significant effects consistent with our turnover results.

4.2. Baseline Results with Full Sample

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 to Bayes rule,

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{9}$$

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 using the training sample. The Naïve Bayes method assumes the occurrences of words and phrases are independent of each other. Therefore, the probability of x is simply the product of the probabilities of each word in x.

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 15, 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. In Table 16 we show the summary statistics of the out-of-sample results by province. There are a total of 796,809 messages. STDEV is the average firm-level standard deviation using the Bayes predicted codes in each province. RATIO 2, RATIO 1, RATIO 0, RATIO -1, and RATIO -2 are the proportions of messages being classified as Strong Buy, Buy, Neutral, Sell, and Strong Sell, respectively. The last two columns report the average firm TURN and correlation between TURN and STDEV in each province. From here we can see a positive correlation between our linguistic diversity measure LD and STDEV. Fujian, with the highest LD of 4 in this sample, has the highest STDEV 1.11. On the other hand, Beijing and Shanghai have relatively low STDEV 1.08 and both have LD of 1. In addition, we see consistent positive correlations between STDEV and TURN, which suggest TURN is indeed measuring diversity of opinions which speaks to the validity of our results in Section 3.

We now formally run analogous baseline regressions where we use STDEV instead of TURN as our dependent variable. We restrict to firms with at least 100 messages to eliminate outliers (less than 1%) that are only scarcely discussed on the message boards. The results are reported in column (1) and (2) of Table 17. We control for the same Size, GDP, and VOL dummies as in our baseline regressions. The coefficient for Log(LD) is 0.0045 with a t-statistic of 2.70. The economic significance is 3.3% of a standard deviation of the LHS variable for a one standard deviation move of the RHS variable, roughly half of that using TURN in our baseline results. Alternatively, in column (3) and (4) we use log of daily average standard deviation as our dependent variable. Here we have stronger statistical significance with a t statistic of over 4 and economic significance around 5.3%. Overall, we show that results using direct measures of diversity of opinions from stock message boards are consistent with those in Section 3 where we use turnover.

5. Linguistic Diversity and Small Private Enterprises

Up to this point, we have argued that our measure of linguistic diversity is plausibly exogenous and hence makes a good right hand side variable. In particular, we have shown that it is uncorrelated with province GDP which is heavily influenced by government policies and hence might naturally influence diversity of opinions, particularly since our diversity of opinion measures come from the stock market. At the same time, the literature on diversity suggests that diversity of opinion ought to be correlated with some measures of real economic activity or productivity. So it would be comforting if we could find some plausible measures of productivity which is uninfluenced or relatively less influenced by government policies. The only part of the Chinese economy that is arguably less affected are the small private enterprises in China. Data on small private businesses is hard to obtain even in the US.

But we are fortunate to have a unique dataset for such private enterprises. This database is extremely comprehensive and includes also public firms. As such, we can construct measures of small enterprise activity for each province and run the analogs of our earlier regressions. Our financial report data of private firms are collected by the National Bureau of Statistics of China, who started tracking manufacturing firms in China since 1998. Our sample is from 1999 to 2005 and includes all SOEs and private firms with more than five million (approximately US$830,000) Chinese Yuan in annual sales. The sample includes

1,236,054 firm-year observations. The mean size of the private companies, as measured by total asset, is only 64,202 Chinese Yuan. This stands in contrast to public manufacturing companies, which has a mean size of about 2.43 billion Chinese Yuan.

We propose several measures of small business activity in Table 18. The first is simply the log number of private firms in a province (Log(NUM)). In essence, we are attempting to measure whether linguistically diverse provinces are composed of a greater fraction of small enterprises, consistent with the premise that diversity of opinions has some positives for a society's production function. Similarly, a second measure is the log number of new private firms in a province (Log(NUM.NEW)). A third measure of small business activity is the fraction of employees hired in private enterprises compared to the sum of private and publicly traded firms (RATIO.EMP). A fourth measure is the assets of private enterprises to the sum of private and publicly traded firms (RATIO.ASSET).

The results of our regressions are reported in Table 19, while controlling for GDP per capita of the province. We find economically and statistically significant effects for our first two measures, Log(NUM) and Log(NUM.NEW), regressed on the linguistic diversity of that province. In column (1), the coefficient in front of Log(LD) is 0.797 with a t-statistic of 2.1. From column (2), we see again that the effect is somewhat larger in high CS provinces. Results using Log(NUM.NEW) in column (3) and (4) are similar.

In columns (5) and (6), the dependent variable is RATIO.EMP. The coefficient in front of Log(LD) is 2.452 with a t-statistic of 1.66 in the univariate specification. The interaction of Log(LD) with CS yields a statistically significant coefficient of 1.531 with a t-statistic of 7.41 in column (6).

In columns (7) and (8), notice that we get a positive coefficient for the univariate specification but it is statistically insignificant. And the other specification with interaction term is not significant. So RATIO.ASSET does not seem to be higher for higher linguistic provinces. This is due in part to the assets of large public companies being so much bigger than small private enterprises that the public assets dominate the analysis making it difficult to measure the contributions of the private enterprises. But nonetheless, we conclude overall that we have evidence that linguistic diversity of a province is related to measures that relate to the productive efficiency of that province.

6. International Evidence

In this section we further extend our empirical evidence to an international setting. Our motivation is to show that results in Section 3 suggest that countries with high linguistic diversity should have high stock market turnover, everything else equal. To measure a nation's linguistic diversity, we first obtain the International Linguistic Diversity Index (IHDI) published by the United Nations' Educational, Scientific and Cultural Organization (UNESCO). The IHDI for each country is computed by taking one minus the language Herfindahl measure based on the population of each language as a proportion of the nations total population[7]. Countries with the top three linguistic diversities are India, Nigeria, and South Africa, while countries with the lowest diversities are South Korea, Portugal, and Venezuela.

The dependent variable of interest is stock market turnover. Other control variables include GDP per capita, size of the stock market, and investor protection indexes. We obtain the time-series average of the median monthly stock market turnover (Turn) and stock market capitalization (MktCap) from Hong and Yu (2009), and GDP per capita (GDPPC) from World Bank's online database. Both MktCap and GDPPC are measured in current U.S dollars. Investor protection indexes include antidirector rights (AntiDir) and judicial efficiency (JudEff) from La Porta, Lopez-de Silanes, Shleifer, and Vishny (1998). AntiDir is an index that ranges from zero to six, indicating the number of criterion a country satisfies in terms of shareholder rights protection. JudEff ranges from zero to ten that measures the efficiency and integrity of a country's legal environment.

Summary statistics of the international variables described above are reported in Table

[7] Hong Kong and Taiwan are not included in the UNESCO report. For these two markets, we follow Greenberg (1956) and compute their IHDI. Hong Kong and Taiwan's IHDI are 0.43 and 0.20, respectively.

20. Note that as opposed to the other variables of interest, Turn, GDPPC, and MktCap are much more non-linear. Therefore, in the following regression analyses we use the log version of these variables. We have a total of 41 countries in our final sample. Table 21 reports the regression results. Our benchmark specification (1) is as follows:

$$Log(Turn_i) = \alpha + \beta_1 IHDI_i^{Lan} + \gamma' \mathbf{X} + \epsilon_i \tag{10}$$

where $\mathbf{X}$ are indicator variables that equals one if country i's $MktCap$ and $GDPPC$ are in the lth or mth quintile, respectively. The coefficient on the $IHDI_i^{Lan}$ is a marginally significant 0.897, consistent with our hypothesis that higher linguistic diversity leads to more trading. This implies that a one standard deviation increase in linguistic diversity leads to a 27.17% of a standard deviation decrease in a country's turnover. In specification (2) and (3) we control for shareholder protection variables. Our economic effect improves slightly as a result of these additional controls. To further assess the robustness of the result, in specification (4) we use decile dummies instead of quintiles, and in specification (5) we directly control for Log(GDPPC) and Log(MktCap). The coefficients on $IHDI_i^{Lan}$ are consistently economically significant with t-stats ranging from 1.58 to 1.89. Overall, the results in Table 21 are consistent with the firm-level results in China that higher linguistic diversity leads to higher stock turnover.

7. Conclusion

The question of the effect of diversity on economic outcomes, which has long been an interesting question in the social sciences, has become even more relevant with globalization. Understanding the mechanisms that guide the trade-offs of diversity has potentially relevant policy implications as societies deal with diversity in both developing and developed countries. In this paper, we try to contribute to this vibrant literature by providing evidence for a much discussed but little studied mechanism which argues that diversity can expand a society's production possibilities frontier because diverse societies bring about diverse opinions, which fosters problem solving and creativity. We provide evidence for the premise that diversity leads to diverse opinions using a linguistic measure of diversity across China and stock market measures of diversity of opinions.

Our contributions are two-fold. To provide a design whereby one can plausibly argue that diversity is exogenous as a right-hand side variable. We show that linguistic diversity across provinces in China reasonably meets this threshold. But perhaps the more original contribution is to link the diversity literature to stock market measures of diverse opinions. International evidence, while less well-identified, shows that our empirical design has some extrapolative value beyond China. As far as we know, this analysis is new. We show that there is a strong causal link of linguistic diversity to stock market measures of diversity of opinions. This paper hence provides new micro-evidence on incoming studies which are beginning to find that diversity, which has long been shown to lead to stagnating economic growth, may also be good for growth under certain circumstances.

The limitation of our study is that we have not spoken to the mechanisms through which diversity of languages leads to diversity of opinions. We alluded to some studies in psychology which point to a creativity channel perhaps for diversity of opinions. Inhabitants knowing that there are different languages stimulates them to also express different opinions. Pinning down such channels would be a very interesting agenda for future work.

References

Alesina, A., and E. L. Ferrara, 2005, "Ethnic Diversity and Economic Performance," *Journal of Economic Literature*, 43(3), 762–800.

Alesina, A., J. Harnoss, and H. Rapoport, 2013, "Birthplace Diversity and Economic Prosperity," *Working Paper*.

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(3), 1259–1294.

Ashraf, Q., and O. Galor, 2013, "The Out of Africa Hypothesis, Human Genetic Diversity, and Comparative Economic Development," *American Economic Review*, 103(1), 1–46.

Bertrand, M., E. F. P. Luttmer, and S. Mullainathan, 2000, "Network Effects And Welfare Cultures," *The Quarterly Journal of Economics*, 115(3), 1019–1055.

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, 325–339.

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

Collier, P., and J. W. Gunning, 1999, "Explaining African Economic Performance," *Journal of Economic Literature*, 37(1), 64–111.

Coval, J. D., and T. J. Moskowitz, 1999, "Home Bias at Home: Local Equity Preference in Domestic Portfolios," *Journal of Finance*, 54(6), 2045–2073.

DellaVigna, S., and J. M. Pollet, 2007, "Demographics and Industry Returns," *American Economic Review*, 97(5), 1667–1702.

Duflo, E., and E. Saez, 2003, "The Role Of Information And Social Interactions In Retirement Plan Decisions: Evidence From A Randomized Experiment," *The Quarterly Journal of Economics*, 118(3), 815–842.

Easterly, W., and R. Levine, 1997, "Africa's Growth Tragedy: Policies and Ethnic Devisions," *The Quarterly Journal of Economics*, 112(4), 1203–1250.

French, K. R., and J. M. Poterba, 1991, "Investor Diversification and International Equity Markets," *American Economic Review*, 81(2), 222–26.

Glaeser, E., J. Scheinkman, and A. Shleifer, 1995, "Economic Growth in a Cross-Section of Cities," *Journal of Monetary Economics*, 36(1), 117–143.

Greenberg, J. H., 1956, "The measurement of linguistic diversity," *Language*, 32, 109–115.

Grinblatt, M., and M. Keloharju, 2001, "How Distance, Language and Culture Influence Stockholdings and Trade," *Journal of Finance*, 56(3), 1053–1073.

Guiso, L., P. Sapienza, and L. Zingales, 2004, "The Role of Social Capital in Financial Development," *American Economic Review*, 94(3), 526–556.

——— , 2009, "Cultural biases in economic exchange?," *Quarterly Journal of Economics*, 124, 1095–1131.

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

Hirshleifer, D., and S. H. Teoh, 2003, "Limited Attention, Information Disclosure, and Financial Reporting," *Journal of Accounting and Economics*, 36(1-3), 337–386.

Hong, H., J. D. Kubik, and J. C. Stein, 2004, "Social Interaction and Stock-Market Participation," *Journal of Finance*, 59(1), 137–163.

Hong, H., and J. C. Stein, 2007, "Disagreement and the Stock Market," *Journal of Economic Perspectives*, 21(2), 109–128.

Hong, H., and J. Yu, 2009, "Gone fishing: Seasonality in trading activity and asset prices," *Journal of Financial Markets*, 12, 672–702.

Hong, L., and S. E. Page, 2001, "Problem Solving by Heterogeneous Agents," *Journal of Economic Theory*, pp. 123–163.

Huberman, G., 2001, "Familiarity Breeds Investment," *Review of Financial Studies*, 14(3), 659–80.

Kandel, E., and N. D. Pearson, 1995, "Differential interpretation of public signals and trade in speculative markets," *Journal of Political Economy*, 103, 831–872.

Kovacs, A. M., and J. Mehler, 2009, "Cognitive gains in 7-month-old bilingual infants," *Proceedings of the National Academy of Sciences*, 106, 6556–6560.

La Porta, R., F. Lopez-de Silanes, A. Shleifer, and R. W. Vishny, 1998, "Law and finance," *Journal of Political Economy*, 106, 1113–1155.

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, 731–741.

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, 1047–1061.

Madrian, B. C., and D. F. Shea, 2001, "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior," *The Quarterly Journal of Economics*, 116(4), 1149–1187.

Mehl, M., 2006, “Quantitative Textual Analysis,” in *Handbook of Multimethod Measurement in Psychology*, ed. by M. Eid, and E. Deiner. American Psychological Association, Washington D.C.

Ng, L., and F. Wu, 2010, “Peer Effects in the Trading Decisions of Individual Investors,” *Financial Management*, 39(2), 807–831.

Odean, T., 1999, “Do investors trade too much,” *American Economic Review*, 89, 1279–1298.

Odean, T., and B. M. Barber, 2008, “All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors,” *Review of Financial Studies*, 21(2), 785–818.

O’Reilly, C., K. Y. Williams, and S. G. Barsade, 1997, “Demography and Group Performance,” *Unpublished*.

Ottaviano, G. I., and G. Peri, 2006, “The Economic Value of Cultural Diversity: Evidence from U.S. Cities,” *Journal of Economic Geography*, 6(1), 9–44.

Ramsey, R., 1987, “The Languages of China,” pp. Princeton University Press, Princeton NJ.

Varian, H. R., 1989, *Survey evidence on the diffusion of interest and information among investors*. Kluwer Academic Publishers, Boston, MA.

Wu, Z., and H. Qiu, 2012, “Local bias of investor attention: Evidence from Chinas internet stock message boards,” *Working Paper*.

Table 1: Province GDP, Diversity, and Terrain

This table reports the languages spoken, linguistic diversity, log GDP per capita, and land statistics for each province in China. The languages spoken in each province are obtained from the Language Atlas of China (1988). LD is the number of languages spoken in each province. HDI is the Herfindahl-based linguistic diversity measure, defined as 1 minus the language Herfindahl index in each province. Province language Herfindahl is measured by the fraction of population speaking each language by aggregating language speakers from all cities. H%, M%, and W% are the fraction of hills, mountains and water areas in each province. Land statistics data was gathered in 1991 and obtained from the Thematic Database for Human-Earth System by the Institute of Geographic Sciences and Natural Resources Research. (Chongqing was part of Sichuan in 1991). The last two rows of the table report the means and standard deviations of these variables.

Province	Languages	Log(GDP)	LD	HDI	H%	M%	W%
Shanghai	Wu	11.14	1	0.00	0.00	0.00	0.18
Beijing	Guan	11.00	1	0.00	0.06	0.45	0.01
Tianjin	Guan	10.83	1	0.00	0.01	0.02	0.03
Zhejiang	Guan, Hui, Wu	10.56	3	0.41	0.47	0.22	0.01
Jiangsu	Guan, Wu	10.50	2	0.47	0.05	0.00	0.14
Guangdong	Min, Kejia, Yue	10.46	3	0.64	0.39	0.22	0.04
Shandong	Guan	10.30	1	0.00	0.07	0.06	0.01
Inner Mongolia	Guan, Jin	10.26	2	0.50	0.13	0.20	0.00
Liaoning	Guan	10.24	1	0.00	0.20	0.27	0.01
Fujian	Min, Gan, Kejia, Wu	10.22	4	0.52	0.31	0.48	0.00
Jilin	Guan	9.95	1	0.00	0.05	0.37	0.01
Hebei	Guan, Jin	9.95	2	0.44	0.09	0.18	0.00
Heilongjiang	Guan	9.88	1	0.00	0.11	0.31	0.01
Shanxi	Guan, Jin	9.80	2	0.45	0.33	0.43	0.00
Xinjiang	Guan	9.79	1	0.00	0.04	0.36	0.00
Hubei	Guan, Gan	9.78	2	0.32	0.22	0.39	0.02
Henan	Guan, Jin	9.75	2	0.24	0.16	0.13	0.01
Chongqing	Guan	9.72	1	0.00			
Shaanxi	Guan, Jin	9.70	2	0.23	0.41	0.42	0.00
Ningxia	Guan	9.69	1	0.00	0.41	0.15	0.01
Hunan	Guan, Gan, Kejia, Xiang	9.67	4	0.72	0.17	0.44	0.02
Hainan	Min	9.66	1	0.00	0.20	0.18	0.00
Qinghai	Guan	9.64	1	0.00	0.12	0.50	0.02
Sichuan	Guan	9.53	1	0.00	0.18	0.74	0.00
Jiangxi	Guan, Gan, Kejia, Hui, Wu	9.52	5	0.69	0.20	0.48	0.03
Guangxi	Guan, Min, Kejia, Xiang, Yue, Ping	9.49	6	0.80	0.26	0.47	0.01
Anhui	Guan, Gan, Hui, Wu	9.47	4	0.49	0.19	0.10	0.00
Yunnan	Guan	9.33	1	0.00	0.09	0.83	0.00
Gansu	Guan	9.30	1	0.00	0.20	0.35	0.00
Guizhou	Guan	8.96	1	0.00	0.08	0.88	0.00
Mean		9.94	1.97	0.23	0.18	0.35	0.02
Stdev		0.52	1.38	0.28	0.13	0.24	0.04

Table 2: Linguistic Diversity and Terrain

This table reports the OLS regression results of linguistic diversity on percentage of hill areas (H%), percentage of hill plus mountain areas (HM%), and percentage of hill, mountain and water areas (HMW%), and log of province GDP per capita (Log(GDP)). In Panel A the dependent variable is the number of languages (LD). The dependent variable in Panel B is Herfindahl-based linguistic diversity measure (HDI), defined as 1 minus the language Herfindahl index in each province. Province language Herfindahl is measured by the fraction of population speaking each language by aggregating language speakers from all cities. The sample excludes provinces that are adjacent to Mt. Everest: Chongqing, Sichuan, and Yunnan. T-stats are in parentheses.

Panel A			
Dependent Variable: LD			
	(1)	(2)	(3)
H%	3.966		
	(1.97)		
HM%		1.935	
		(1.49)	
HMW%			2.182
			(1.61)
Log(GDP)	-3.718	-0.484	-0.750
	(-0.72)	(-0.08)	(-0.13)
#Obs	27	27	27
Adj. R^2	0.107	0.051	0.064

Panel B			
Dependent Variable: HDI			
	(1)	(2)	(3)
H%	0.827		
	(2.05)		
HM%		0.377	
		(1.44)	
HMW%			0.432
			(1.59)
Log(GDP)	-0.141	0.464	0.427
	(-0.14)	(0.37)	(0.35)
#Obs	27	27	27
Adj. R^2	0.088	0.013	0.029

Table 3: Summary Statistics

This table reports summary statistics of key variables in this paper. Panel A reports mean statistics by province. LD is the number of languages spoken in each province. HDI is defined as 1 minus the language Herfindahl index in each province. Province language Herfindahl is measured by the fraction of population speaking each language by aggregating language speakers from all cities. Turn is firm average quarterly turnover over the sample period. MV is firm average quarter-end market capitalization over the sample period, in billions RMB. VOL is firm monthly average volatility over the sample period. CS is median city share, which is the median number of languages spoken in cities of each province divide by LD. Panel B reports pooled summary statistics. The sample includes Tier 1˜Tier 4 provinces and omit firms with MV in the lowest decile. The sample period is from 1998 to 2012.

Panel A

Province	LD	HDI	CS	Turn	MV	VOL
Shanghai	1	0	1	1.30	5.58	0.14
Beijing	1	0	1	1.46	23.14	0.13
Tianjin	1	0	1	1.30	5.21	0.14
Shandong	1	0	1	1.64	2.58	0.14
Liaoning	1	0	1	1.30	2.46	0.15
Jilin	1	0	1	1.44	2.31	0.19
Heilongjiang	1	0	1	1.19	2.19	0.19
Xinjiang	1	0	1	1.63	3.37	0.15
Chongqing	1	0	1	1.37	1.84	0.14
Ningxia	1	0	1	1.50	1.62	0.15
Hainan	1	0	1	1.32	2.35	0.16
Qinghai	1	0	1	1.66	3.98	0.18
Sichuan	1	0	1	1.59	2.67	0.15
Jiangsu	2	0.47	0.50	1.77	2.48	0.13
Inner Mongolia	2	0.50	0.50	1.70	3.43	0.14
Hebei	2	0.44	0.50	1.45	2.89	0.15
Shanxi	2	0.45	0.50	1.35	8.83	0.14
Hubei	2	0.32	0.50	1.38	2.08	0.14
Henan	2	0.24	0.50	2.05	2.88	0.14
Shaanxi	2	0.23	0.50	1.54	2.65	0.14
Zhejiang	3	0.41	0.33	1.72	2.18	0.13
Guangdong	3	0.64	0.33	1.58	4.56	0.14
Fujian	4	0.52	0.25	1.58	4.11	0.15
Hunan	4	0.72	0.50	1.64	2.84	0.14

Panel B

	Mean	Stdev	Min	25%	50%	75%	Max
LD	1.94	0.99	1.00	1.00	2.00	3.00	4.00
Log(LD)	0.53	0.51	0.00	0.00	0.69	1.10	1.39
HDI	0.27	0.26	0.00	0.00	0.32	0.47	0.72
Log(1+HDI)	0.22	0.21	0.00	0.00	0.27	0.39	0.54
Turn	1.55	0.84	0.38	1.07	1.34	1.78	10.73
Log(Turn)	0.34	0.43	-0.96	0.06	0.29	0.58	2.37
MV	5.03	25.63	0.63	1.07	1.66	2.96	622.46
VOL	0.14	0.64	0.04	0.12	0.14	0.15	1.85
CS	0.67	0.30	0.25	0.33	0.50	1.00	1.00

Table 4: Linguistic Diversity and Turnover

This table reports OLS regression results of turnover on linguistic diversity. For each stock, variables are averaged across the sample period. The dependent variable is log of mean quarterly turnover over the sample period. LD (Log(LD)) is the (log) number of languages spoken in each province. Size, GDP, and VOL decile dummies are based on firm's average market capitalization, home province GDP per capita, and volatility over the sample period. The sample includes Tier 1˜Tier 4 provinces and omit firms with size in the lowest decile. Panel A reports results from 1998 to 2012, and Panel B from 2008 to 2012. T-stats are in parentheses. Standard errors are clustered by province.

Panel A: 1998-2012				
Dependent Variable: Log(Turn)				
	(1)	(2)	(3)	(4)
Log(LD)	0.060	0.065		
	(1.81)	(1.89)		
LD			0.023	0.026
			(1.57)	(1.64)
Size Dum	YES	YES	YES	YES
GDP Dum	YES	YES	YES	YES
VOL Dum	NO	YES	NO	YES
Adj. R^2	0.220	0.252	0.219	0.251
# Obs.	1,772	1,772	1,772	1,772
Panel B: 2008-2012				
Dependent Variable: Log(Turn)				
	(1)	(2)	(3)	(4)
Log(LD)	0.052	0.053		
	(1.98)	(1.93)		
LD			0.019	0.021
			(1.79)	(1.73)
Size Dum	YES	YES	YES	YES
GDP Dum	YES	YES	YES	YES
VOL Dum	NO	YES	NO	YES
Adj. R^2	0.353	0.415	0.353	0.415
# Obs.	1,717	1,717	1,717	1,717

Table 5: Linguistic Diversity and Turnover with Population Control

This table reports OLS regression results of turnover on linguistic diversity controlling for province population. For each stock, variables are averaged across the sample period. The dependent variable is log of mean quarterly turnover over the sample period. LD (Log(LD)) is the (log) number of languages spoken in each province. Log(Pop) is the log of average province population over the sample period. Size, GDP, and VOL decile dummies are based on firm's average market capitalization, home province GDP per capita, and volatility over the sample period. The sample includes Tier 1˜Tier 4 provinces and omit firms with size in the lowest decile. The sample period is from 1998 to 2012. T-stats are in parentheses. Standard errors are clustered by province.

Panel A: 1998-2012

Dependent Variable: Log(Turn)

	(1)	(2)	(3)	(4)
Log(LD)	0.053	0.057		
	(1.64)	(1.64)		
LD			0.020	0.022
			(1.41)	(1.42)
Log(Pop)	0.020	0.029	0.022	0.031
	(0.77)	(1.05)	(0.82)	(1.11)
Size Dum	YES	YES	YES	YES
GDP Dum	YES	YES	YES	YES
VOL Dum	NO	YES	NO	YES
Adj. R^2	0.220	0.252	0.220	0.251
# Obs.	1,772	1,772	1,772	1,772

Panel B: 2008-2012

Dependent Variable: Log(Turn)

	(1)	(2)	(3)	(4)
Log(LD)	0.058	0.057		
	(2.26)	(2.13)		
LD			0.021	0.023
			(1.99)	(1.88)
Log(Pop)	-0.022	-0.011	-0.020	-0.009
	(-1.46)	(-0.71)	(-1.23)	(-0.56)
Size Dum	YES	YES	YES	YES
GDP Dum	YES	YES	YES	YES
VOL Dum	NO	YES	NO	YES
Adj. R^2	0.353	0.415	0.353	0.415
# Obs.	1,717	1,717	1,717	1,717

Table 6: Linguistic Diversity and Turnover—Drop Beijing/Shanghai

This table reports OLS regression results of turnover on linguistic diversity, not including provinces with large immigrant population (Beijing and Shanghai). For each stock, variables are averaged across the sample period. The dependent variable is log of mean quarterly turnover over the sample period. Log(LD) is the log number of languages spoken in a firm's home province. Size, GDP, and VOL decile dummies are based on firm's average market capitalization, home province GDP per capita, and volatility over the sample period. The sample includes Tier 1~Tier 4 provinces and omit firms with MV in the lowest decile. Panel A reports results from 1998 to 2012, and Panel B from 2008 to 2012. T-stats are in parentheses. Standard errors are clustered by province.

Panel A: 1998-2012		
	(1)	(2)
	Drop Beijing	Drop Beijing and Shanghai
Log(LD)	0.064	0.065
	(1.84)	(1.89)
Size Dum	YES	YES
GDP Dum	YES	YES
VOL Dum	YES	YES
Adj. R^2	0.245	0.224
# Obs.	1,622	1,465

Panel B: 2008-2012		
	(1)	(2)
	Drop Beijing	Drop Beijing and Shanghai
Log(LD)	0.052	0.049
	(1.89)	(1.82)
Size Dum	YES	YES
GDP Dum	YES	YES
VOL Dum	YES	YES
Adj. R^2	0.402	0.370
# Obs.	1,570	1,413

Table 7: Linguistic Diversity, Segregation, and Turnover

This table reports OLS regression results of turnover on linguistic diversity, and city share. For each stock, variables are averaged across the sample period. The dependent variable is log of mean quarterly turnover. LD (Log(LD) is the (log) number of languages spoken in the firm's home province. CS is the median number of languages spoken in cities of each province divide by LD. CS.High is a dummy variable which equals one if CS is greater than 0.4, and zero otherwise. CS.Low is a dummy variable which equals one if CS is lower than or equal to 0.4, and zero otherwise. Size, GDP, and VOL decile dummies are based on firm's average MV, home province GDP per capita, and volatility over the sample period. The sample includes Tier 1˜Tier 4 provinces and omit firms with MV in the lowest decile. Panel A reports results from 1998 to 2012, and Panel B from 2008 to 2012. T-stats are in parentheses. Standard errors are clustered by province.

Panel A: 1998-2012				
Dependent Variable: Log(Turn)				
	(1)	(2)	(3)	(4)
Log(LD)	-0.048			
	(-0.54)			
LD			-0.049	
			(-1.16)	
CS	-0.017		-0.272	
	(-0.14)		(-1.70)	
Log(LD)*CS	0.330			
	(2.03)			
LD*CS			0.177	
			(2.17)	
Log(LD)*CS.High		0.129		
		(2.98)		
Log(LD)*CS.Low		0.056		
		(2.05)		
LD*CS.High				0.066
				(3.38)
LD*CS.Low				0.034
				(2.92)
Size Dum	YES	YES	YES	YES
GDP Dum	YES	YES	YES	YES
VOL Dum	YES	YES	YES	YES
Adj. R^2	0.252	0.253	0.252	0.253
# Obs.	1,772	1,772	1,772	1,772

Table 7—*Continued*

Panel B: 2008-2012				
Dependent Variable: Log(Turn)				
	(1)	(2)	(3)	(4)
Log(LD)	-0.027			
	(-0.34)			
LD			-0.027	
			(-0.73)	
CS	-0.033		-0.182	
	(-0.39)		(-1.42)	
Log(LD)*CS	0.199			
	(1.48)			
LD*CS			0.102	
			(1.54)	
Log(LD)*CS.High		0.097		
		(3.19)		
Log(LD)*CS.Low		0.047		
		(2.15)		
LD*CS.High				0.050
				(3.50)
LD*CS.Low				0.028
				(3.01)
Size Dum	YES	YES	YES	YES
GDP Dum	YES	YES	YES	YES
VOL Dum	YES	YES	YES	YES
Adj. R^2	0.415	0.416	0.415	0.416
# Obs.	1,772	1,772	1,772	1,772

Table 8: Summary Statistics—Language Herfindahls

This table reports summary statistics of variables for results related to the Guba Eastmoney message board. The sample includes all firms on the Guba Eastmoney message board located in Tier 1 to Tier 4 provinces, excluding firms in the smallest size decile. Message posts with less than or equal to five replies are dropped. Panel A reports the summary statistics by province. LD is the number of languages spoken in each province. Turn is total stock turnover of firms over the sample period. H^{Lan}, H^{GDP}, and H^{City} are the Herfindahl indexes based on language, GDP, and city diversity of Guba message board posts. # Posts is the number of original posts. # Firms is the total number of firms in each province. Panel B reports the pooled summary statistics. The sample period is from 2008 to 2012.

Panel A

Province	LD	Turn		H^{Lan}		H^{GDP}		H^{City}		# Posts	# Firms
		Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev		
Shanghai	1	5.29	2.97	0.26	0.02	0.33	0.04	0.04	0.02	605,232	169
Beijing	1	5.73	3.47	0.27	0.03	0.32	0.03	0.03	0.01	490,035	155
Tianjin	1	5.38	2.44	0.27	0.03	0.33	0.03	0.03	0.01	121,287	31
Shandong	1	6.80	2.95	0.28	0.04	0.31	0.03	0.03	0.01	513,957	132
Liaoning	1	6.44	3.49	0.27	0.03	0.31	0.03	0.03	0.01	268,858	63
Jilin	1	6.34	3.16	0.27	0.04	0.31	0.03	0.03	0.01	92,706	41
Heilongjiang	1	5.80	2.56	0.27	0.03	0.31	0.03	0.03	0.01	126,213	33
Xinjiang	1	6.47	2.62	0.27	0.03	0.30	0.02	0.03	0.01	38,350	35
Chongqing	1	5.93	2.34	0.29	0.03	0.30	0.04	0.03	0.01	128,539	34
Ningxia	1	6.58	1.94	0.29	0.03	0.29	0.02	0.02	0.01	15,572	12
Hainan	1	7.52	3.47	0.26	0.02	0.31	0.02	0.02	0.01	41,153	26
Qinghai	1	6.97	3.48	0.26	0.02	0.31	0.03	0.02	0.00	7,867	11
Sichuan	1	6.63	2.86	0.28	0.03	0.30	0.03	0.03	0.03	322,742	89
Jiangsu	2	6.84	3.09	0.27	0.03	0.32	0.03	0.03	0.01	613,631	191
Inner Mongolia	2	6.75	3.11	0.26	0.02	0.30	0.02	0.03	0.01	30,380	23
Hebei	2	5.96	2.68	0.26	0.02	0.31	0.03	0.03	0.01	223,352	47
Shanxi	2	5.56	2.50	0.25	0.02	0.30	0.03	0.03	0.01	89,377	29
Hubei	2	6.27	2.64	0.26	0.02	0.30	0.03	0.03	0.01	296,090	70
Henan	2	7.11	3.24	0.25	0.03	0.30	0.03	0.03	0.02	295,135	58
Shaanxi	2	6.35	2.64	0.26	0.02	0.30	0.03	0.03	0.01	193,183	32
Zhejiang	3	7.26	3.42	0.25	0.02	0.32	0.04	0.03	0.01	628,601	198
Guangdong	3	6.61	3.39	0.24	0.03	0.33	0.03	0.03	0.01	1,349,423	288
Fujian	4	7.09	2.87	0.25	0.02	0.31	0.03	0.03	0.01	343,760	81
Hunan	4	6.96	2.75	0.25	0.03	0.29	0.02	0.03	0.01	247,116	61

Table 8—*Continued*

Panel B							
	Mean	Stdev	Min	25%	50%	75%	Max
Turn	6.49	3.14	1.43	4.31	5.91	8.16	17.66
Log(Turn)	1.75	0.50	0.35	1.46	1.78	2.10	2.87
H^{Lan}	0.26	0.03	0.17	0.24	0.26	0.28	0.45
Log(H^{Lan})	-1.35	0.12	-1.75	-1.42	-1.35	-1.28	-0.81
H^{GDP}	0.03	0.01	0.02	0.02	0.03	0.03	0.26
Log(H^{GDP})	-3.62	0.29	-4.14	-3.82	-3.68	-3.50	-1.36
H^{City}	0.31	0.04	0.23	0.29	0.31	0.33	0.53
Log(H^{City})	-1.16	0.11	-1.47	-1.24	-1.17	-1.10	-0.64
LDI	1.99	1.03	1.00	1.00	2.00	3.00	6.00
Log(LD)	0.55	0.52	0.00	0.00	0.69	1.10	1.79

Table 9: Linguistic Diversity of Investor Base and Turnover (First Stage)

This table reports regression results of language Herfindahl index on number of languages of a firm's home province. The sample includes all firms on the Guba Eastmoney message board located in Tier 1 to Tier 4 provinces, excluding firms in the smallest size decile. The dependent variable is log language Herfindahl index Log(H^{Lan}). The independent variables are (1) Log(LD): log number of languages of the firm's home province, (2) LD: number of languages of the firm's home province, and (3) LD.Dum: dummy variable which equals one if the firm's home province speaks more than one language, and zero otherwise. Other control variables are Log(H^{City}), Log(H^{GDP}), and size and GDP decile dummies. H^{Lan}, H^{GDP}, and H^{City} are the Herfindahl indexes based on language, GDP, and city diversity of Guba message board posts. Size decile dummies and GDP decile dummies are based on sorts using average total market capitalization and the GDP per capita of the home province of each stock. The t-stats are in parentheses. Standard errors are clustered by province. The sample period is from 2008 to 2012.

Dependent Variable: Log(H^{Lan})			
	(1)	(2)	(3)
Log(LD)	-0.109		
	(-7.25)		
LD		-0.049	
		(-6.35)	
LD.Dum			-0.089
			(-8.21)
Log(H^{City})	0.073	0.073	0.074
	(3.35)	(3.35)	(3.32)
Log(H^{GDP})	-0.362	-0.362	-0.367
	(-6.54)	(-6.56)	(-6.37)
Size Dum	YES	YES	YES
GDP Dum	YES	YES	YES
Adj. R^2	0.238	0.234	0.210
# Obs	1,715	1,715	1,715

Table 10: Linguistic Diversity of Investor Base and Turnover (2SLS)

This table reports 2SLS regression results of language Herfindahl index on number of languages of a firm's home province. The sample includes all firms on the Guba Eastmoney message board located in Tier 1 to Tier 4 provinces, excluding firms in the smallest size decile. The dependent variable is log turnover over the sample period. The instruments for language Herfindahl index are (1) Log(LD): log number of languages of the firm's home province, (2) LD: number of languages of the firm's home province, and (3) LD.Dum: dummy variable which equals one if the firm's home province speaks more than one language, and zero otherwise. Log(H^{Lan1}), Log(H^{Lan2}), Log(H^{Lan3}) and are language Herfindahl index instrumented by (1), (2), and (3) above, respectively. Other control variables are Log(H^{City}), Log(H^{GDP}), size, and GDP decile dummies. H^{Lan}, H^{GDP}, and H^{City} are the Herfindahl indexes based on language, GDP, and city diversity of Guba message board posts. Size decile dummies and GDP decile dummies are based on sorts using average total market capitalization and the GDP per capita of the home province of each stock. The t-stats are in parentheses. Standard errors are clustered by province. The sample period is from 2008 to 2012.

Dependent Variable: Log(Turn)			
	(1)	(2)	(3)
Log(H^{Lan1})	-0.592		
	(-2.57)		
Log(H^{Lan2})		-0.523	
		(-2.31)	
Log(H^{Lan3})			-0.924
			(-3.12)
Log(H^{City})	-0.367	-0.372	-0.343
	(-9.23)	(-9.78)	(-6.97)
Log(H^{GDP})	0.010	0.035	-0.110
	(0.08)	(0.27)	(-0.57)
Size Dum	YES	YES	YES
GDP Dum	YES	YES	YES
Adj. R^2	0.270	0.274	0.246
# Obs	1,715	1,715	1,715

Table 11: HDI, Segregation, and Turnover

This table reports OLS regression results of turnover on Herfindahl-based linguistic diversity, and city share. For each stock, variables are averaged across the sample period. The dependent variable is log of mean quarterly turnover. Herfindahl-based linguistic diversity measure (HDI) is defined as 1 minus the language Herfindahl index in each province. Language Herfindahl index is measured by the fraction of population speaking each language by aggregating language speakers from all city. We assume residents of each city speak all languages in their respective cities. CS is the median number of languages spoken in cities of each province divide by LD. CS.High is a dummy variable which equals one if CS is greater than 0.4, and zero otherwise. CS.Low is a dummy variable which equals one if CS is lower than or equal to 0.4, and zero otherwise. Size, GDP, and VOL decile dummies are based on firm's average MV, home province GDP per capita, and volatility over the sample period. The sample includes Tier 1˜Tier 4 provinces and omit firms with MV in the lowest decile. Panel A reports results from 1998 to 2012, and Panel B from 2008 to 2012. T-stats are in parentheses. Standard errors are clustered by province.

Panel A: 1998-2012			
	(1)	(2)	(3)
Log(1+HDI)	0.106	-0.694	
	(1.11)	(-2.44)	
CS		-0.334	
		(-2.21)	
Log(1+HDI)*CS		1.024	
		(2.98)	
Log(1+HDI)*CS.High			0.159
			(1.63)
Log(1+HDI)*CS.Low			0.040
			(0.44)
Size Dum	YES	YES	YES
GDP Dum	YES	YES	YES
VOL Dum	YES	YES	YES
Adj. R^2	0.250	0.253	0.251
# Obs.	1,772	1,772	1,772
Panel B: 2008-2012			
	(1)	(2)	(3)
Log(1+HDI)	0.079	-0.531	
	(1.02)	(-2.55)	
CS		-0.274	
		(-2.80)	
Log(1+HDI)*CS		0.713	
		(2.37)	
Log(1+HDI)*CS.High			0.114
			(1.50)
Log(1+HDI)*CS.Low			0.034
			(0.45)
Size Dum	YES	YES	YES
GDP Dum	YES	YES	YES
VOL Dum	YES	YES	YES
Adj. R^2	0.415	0.416	0.415
# Obs.	1,717	1,717	1,717

Table 12: Linguistic Diversity with Sub-Languages

This table reports linguistic diversity measures with sub-languages. 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. HDI is the Herfindahl-based linguistic diversity measure, defined as 1 minus the language Herfindahl index in each province. Province language Herfindahl is measured by the fraction of population speaking each language by aggregating language speakers from all cities. HDI-SUB1, HDI-SUB2, and HDI-SUB3 denote the HDI measures based on level 1, 2, and 3 sub-languages, respectively. Panel A reports the linguistic diversity measures by province. The last two rows show the mean and standard deviations. Panel B reports the correlation coefficients.

Panel A

Province	LD	LD-SUB1	LD-SUB2	LD-SUB3	HDI	HDI-SUB1	HDI-SUB2	HDI-SUB3
Shanghai	1	1	1	1	0.00	0.00	0.00	0.00
Beijing	1	2	2	2	0.00	0.50	0.50	0.50
Tianjin	1	2	2	4	0.00	0.50	0.50	0.75
Zhejiang	3	11	16	16	0.41	0.88	0.92	0.92
Jiangsu	2	4	7	7	0.47	0.66	0.81	0.81
Guangdong	3	11	14	14	0.64	0.86	0.89	0.89
Shandong	1	3	8	11	0.00	0.66	0.85	0.88
Inner Mongolia	2	5	7	7	0.50	0.76	0.81	0.81
Liaoning	1	3	5	5	0.00	0.54	0.74	0.74
Fujian	4	12	14	14	0.52	0.87	0.89	0.89
Jilin	1	1	2	5	0.00	0.00	0.42	0.64
Hebei	2	5	8	14	0.44	0.68	0.85	0.91
Heilongjiang	1	3	5	7	0.00	0.23	0.66	0.80
Shanxi	2	8	10	12	0.45	0.85	0.87	0.90
Xinjiang	1	3	3	3	0.00	0.59	0.59	0.59
Hubei	2	3	6	6	0.32	0.60	0.81	0.81
Henan	2	3	8	9	0.24	0.30	0.75	0.75
Chongqing	1	1	1	1	0.00	0.00	0.00	0.00
Shaanxi	2	7	8	8	0.23	0.59	0.75	0.75
Ningxia	1	2	4	4	0.00	0.47	0.69	0.69
Hunan	4	12	18	18	0.72	0.85	0.89	0.89
Hainan	1	1	5	5	0.00	0.00	0.62	0.62
Qinghai	1	1	2	2	0.00	0.00	0.50	0.50
Sichuan	1	1	3	5	0.00	0.00	0.52	0.64
Jiangxi	5	12	12	12	0.69	0.91	0.91	0.91
Guangxi	6	12	15	15	0.80	0.87	0.89	0.89
Anhui	4	11	16	16	0.49	0.77	0.87	0.87
Yunnan	1	1	4	7	0.00	0.00	0.60	0.69
Gansu	1	3	5	5	0.00	0.53	0.70	0.70
Guizhou	1	1	5	5	0.00	0.00	0.67	0.67
Mean	1.97	4.83	7.20	8.00	0.23	0.48	0.68	0.71
Stdev	1.38	4.15	5.00	4.97	0.28	0.34	0.24	0.23

Table 12—*Continued*

Panel B								
	LD	LD-SUB1	LD-SUB2	LD-SUB3	HDI	HDI-SUB1	HDI-SUB2	HDI-SUB3
LD	1							
LD-SUB1	0.92	1						
LD-SUB2	0.86	0.94	1					
LD-SUB3	0.80	0.88	0.96	1				
HDI	0.91	0.89	0.86	0.82	1			
HDI-SUB1	0.70	0.83	0.77	0.73	0.78	1		
HDI-SUB2	0.59	0.68	0.77	0.78	0.67	0.77	1	
HDI-SUB3	0.51	0.61	0.69	0.75	0.59	0.70	0.96	1

Table 13: Linguistic Diversity and Turnover

This table reports OLS regression results of turnover on linguistic diversity. For each stock, variables are averaged across the sample period. The dependent variable is log of mean quarterly turnover over the sample period. Definitions of the measures for linguistic diversity are as follows: 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. HDI is the Herfindahl-based linguistic diversity measure (HDI), defined as 1 minus the language Herfindahl index in each province. Province language Herfindahl is measured by the fraction of population speaking each language by aggregating language speakers from all cities. HDI-SUB1, HDI-SUB2, and HDI-SUB3 denote the HDI measures based on level 1, 2, and 3 sub-languages. Size, GDP, and VOL decile dummies are based on firm's average market capitalization, home province GDP per capita, and volatility over the sample period. The sample includes Tier 1˜Tier 4 provinces and omit firms with size in the lowest decile. Panel A reports results from 1998 to 2012, and Panel B from 2008 to 2012. T-stats are in parentheses. Standard errors are clustered by province.

Panel A: 1998-2012

Dependent Variable: Log(Turn)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Log(LD)	Log(LD-	Log(LD-	Log(LD-	Log(1+	Log(1+	Log(1+	Log(1+
		SUB1)	SUB2)	SUB3)	HDI)	HDI-SUB1)	HDI-SUB2)	HDI-SUB3)
Diversity	0.065	0.041	0.065	0.065	0.106	0.236	0.288	0.245
	(1.89)	(1.57)	(2.18)	(1.88)	(1.11)	(2.81)	(3.17)	(2.34)
Size Dum	YES	YES	YES	YES	YES	YES	YES	YES
GDP Dum	YES	YES	YES	YES	YES	YES	YES	YES
VOL Dum	YES	YES	YES	YES	YES	YES	YES	YES
Adj. R^2	0.252	0.252	0.254	0.254	0.250	0.256	0.256	0.255
# Obs.	1,772	1,772	1,772	1,772	1,772	1,772	1,772	1,772

Panel B: 2008-2012

Dependent Variable: Log(Turn)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Log(LD)	Log(LD-	Log(LD-	Log(LD-	Log(1+	Log(1+	Log(1+	Log(1+
		SUB1)	SUB2)	SUB3)	HDI)	HDI-SUB1)	HDI-SUB2)	HDI-SUB3)
Diversity	0.053	0.042	0.058	0.055	0.079	0.238	0.321	0.294
	(1.93)	(1.57)	(2.23)	(2.00)	(1.02)	(2.84)	(4.27)	(3.61)
Size Dum	YES	YES	YES	YES	YES	YES	YES	YES
GDP Dum	YES	YES	YES	YES	YES	YES	YES	YES
VOL Dum	YES	YES	YES	YES	YES	YES	YES	YES
Adj. R^2	0.415	0.416	0.417	0.417	0.415	0.419	0.421	0.420
# Obs.	1,717	1,717	1,717	1,717	1,717	1,717	1,717	1,717

Table 14: Linguistic Diversity and Local Trade

This table reports OLS regression results of local turnover on linguistic diversity using Shanghai Stock Exchange (SSE) data. Every month for each stock, local turnover is measured by dividing volume from the firm's home province by its number of outstanding tradable shares. Log(Turn^{L}) is then log of average firm local turnover over the sample period. LD (Log(LD)) is the (log) number of languages spoken in each province. Size, GDP, and VOL decile dummies are based on firm's average market capitalization, home province GDP per capita, and volatility over the sample period. The sample includes Tier 1˜Tier 4 provinces with at least 15 firms listed on SSE and omit firms with size in the lowest decile. Panel A reports summary statistics for the key variables. Panel B reports the regression results. T-stats are in parentheses. Standard errors are clustered by province. The sample period is from April 2001 to August 2002.

Panel A		
	Mean	Stdev
Log(Turn^{L})	-5.422	0.934
Log(LD)	0.339	0.478
LD	1.595	0.902

Panel B		
Dependent Variable: Log(Turn^{L})		
	(1)	(2)
Log(LD)	0.852	
	(8.85)	
LD		0.379
		(7.11)
Size Dum	YES	YES
GDP Dum	YES	YES
VOL Dum	YES	YES
Adj. R^2	0.237	0.237
# Obs.	464	464

Table 15: Linguistic Diversity and Diversity of Opinions in Message Board Posts

This table reports OLS regression results of diversity of opinions measured by message board posts on linguistic diversity. Eight provinces are selected (four provinces with LD¿1, and four provinces with LD=1). A random sample of 30 firms is selected from each province, and we download the most recent 20 messages from Guba Eastmoney message boards. Firms listed on exchanges' secondary boards are dropped. The opinions in each post is coded by two graduate students independently with -2, -1, 0, 1, or 2, denoting Strong Sell, Sell, Neutral, Buy, and Strong Buy. STDEV1, STDEV2, and STDEVM are the firm-level standard deviations of the two students' codings and their average. Log(LD) is the log number of languages spoken in a firm's home province. Size, GDP, and VOL decile dummies are based on firm's average market capitalization, home province GDP per capita, and volatility from 2008 to 2012. Panel A reports summary statistics and Panel B reports regression results. T-stats are in parentheses. Standard errors are clustered by province.

Panel A							
	Mean	Stdev	Min	25%	50%	75%	Max
Log(LD)	0.62	0.63	0.00	0.00	0.55	1.17	1.39
STDEV1	0.75	0.13	0.308	0.671	0.768	0.838	1.026
Log(STDEV1)	-0.31	0.19	-1.18	-0.40	-0.26	-0.18	0.03
STDEV2	0.90	0.25	0.22	0.74	0.90	1.04	2.82
Log(STDEV2)	-0.15	0.27	-1.50	-0.30	-0.10	0.04	1.04
STDEV	0.76	0.16	0.28	0.65	0.77	0.87	1.55
Log(STDEV)	-0.30	0.23	-1.29	-0.42	-0.26	-0.14	0.44

Panel B			
	(1)	(2)	(3)
	Log(STDEV1)	Log(STDEV2)	Log(STDEVM)
Log(LD)	0.028	0.038	0.040
	(2.60)	(1.17)	(1.59)
Size Dum	YES	YES	YES
GDP Dum	YES	YES	YES
VOL Dum	YES	YES	YES
Adj. R^2	0.125	0.113	0.098
# Obs.	201	201	201

Table 16: 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., P(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 17: Linguistic Diversity and Diversity of Opinions in Message Board Posts—Full Sample

This table reports summary statistics and linguistic diversity regression with the full message sample. Naïve Bayes rule is applied to generate predicted buy/sell signal for each message. The full sample consists of 10 pages of messages per firm from eight provinces: Beijing, Fujian, Guangdong, Hunan, Shandong, Shanghai, Sichuan, and Zhejiang. A training sample of 20 messages from a random sample of 30 firms from each province is generated to measure the probabilities associated with each key word. Panel A reports the summary statistics. LD is the number of languages spoken, Num Msg. is the number of messages, STDEV is average firm-level pooled standard deviation of the generated buy/sell signal, Ratio2˜Ratio-2 are the ratios of each signal in the sample. TURN is average firm turnover and CORR is the correlation between firm-level TURN and STDEV. Panel B reports OLS results of divergence of opinion on linguistic diversity, measured by Log(LD) and LD. The dependent variable in Column (1) and (2) is log of STDEV, pooled standard deviation of the generated buy/sell signal. The dependent variable in Column (3) and (4) is log of mean daily standard deviation of the generated signal. Size, GDP, and VOL decile dummies are based on firms' average market capitalization, home province GDP per capita, and volatility from 2008 to 2012.

Panel A

Province	LD	Num Msg.	STDEV	Ratio 2	Ratio 1	Ratio 0	Ratio -1	Ratio -2	TURN	CORR
Beijing	1	115,765	1.08	0.10	0.10	0.38	0.32	0.10	1.56	0.13
Fujian	4	53,387	1.11	0.11	0.12	0.35	0.32	0.11	1.94	0.39
Guangdong	3	201,792	1.11	0.11	0.11	0.36	0.31	0.11	1.80	0.02
Hunan	4	42,900	1.07	0.09	0.12	0.35	0.33	0.11	1.93	0.18
Shandong	1	85,930	1.09	0.10	0.11	0.36	0.32	0.11	1.84	0.16
Shanghai	1	109,152	1.08	0.11	0.09	0.39	0.32	0.09	1.42	0.16
Sichuan	1	55,683	1.10	0.10	0.14	0.34	0.31	0.11	1.82	0.12
Zhejiang	3	132,198	1.09	0.10	0.14	0.34	0.31	0.11	2.13	0.20

Panel B

Dependent Variable: Log(STDEV)

	(1)	(2)	(3)	(4)
Log(LD)	0.0045		0.0112	
	(2.70)		(4.33)	
LD		0.0021		0.0062
		(2.70)		(4.33)
Size Dum	YES	YES	YES	YES
GDP Dum	YES	YES	YES	YES
VOL Dum	YES	YES	YES	YES
Adj. R^2	0.061	0.061	0.051	0.051
# Obs.	1,166	1,166	1,173	1,173

Table 18: Summary Statistics for Private Firms

This table reports summary statistics of the private firm data. The sample period is from 1999 to 2005. LD is the number of languages spoken in a firm's home province. NUM is the number of private firms. NUM.NEW is the number of new private firms. RATIO.EMP is the fraction of employees in each province employed by private firms. RATIO.ASSET is the fraction of asset in each province from private firms. Panel A reports the time series mean over the sample period for each province. Panel B shows the pooled summary statistics.

Panel A

Province	LD	NUM	NUM.NEW	RATIO.EMP	RATIO.ASSET
Shanghai	1	11,239	553	91.26%	78.97%
Beijing	1	4,991	231	92.01%	85.01%
Tianjin	1	5,528	268	98.02%	95.24%
Shandong	1	15,591	1355	96.39%	89.31%
Liaoning	1	6,914	464	97.54%	93.41%
Jilin	1	2,343	158	95.61%	90.59%
Heilongjiang	1	2,322	178	96.03%	89.70%
Xinjiang	1	1,043	80	92.48%	80.41%
Chongqing	1	2,024	118	95.12%	84.79%
Ningxia	1	415	36	85.98%	76.96%
Hainan	1	483	22	96.45%	86.20%
Qinghai	1	303	18	89.46%	71.70%
Sichuan	1	4,714	364	92.87%	82.67%
Jiangsu	2	24,462	1594	98.16%	92.93%
Inner Mongolia	2	1,238	139	91.41%	80.34%
Hebei	2	7,307	434	94.71%	87.02%
Shanxi	2	2,589	141	96.67%	88.39%
Hubei	2	5,830	415	96.15%	88.18%
Henan	2	8,902	482	97.04%	89.90%
Shaanxi	2	2,173	86	98.08%	92.64%
Zhejiang	3	24,435	1581	97.38%	94.12%
Guangdong	3	24,058	1382	97.58%	89.53%
Fujian	4	7,845	546	98.26%	91.68%
Hunan	4	4,989	440	95.29%	85.15%

Panel B

	Mean	Stdev	Min	25%	50%	75%	Max
LD	1.808	0.937	1	1	1	2	4
Log(LD)	0.409	0.485	0.00	0.00	0.00	0.693	1.386
NUM	7,156	8,174	267	2,026	5,022	8,748	40,372
NUM.NEW	403	592	10	96	210	477	4458
Log(NUM)	8.25	1.23	5.59	7.61	8.52	9.08	10.61
RATIO.EMP	94.99	3.38	84.2	93.57	96.06	97.48	99.12
RATIO.ASSET	86.87	6.33	63.44	82.69	88.84	91.56	95.98

Table 19: Linguistic Diversity and Entrepreneurship

This table reports OLS regression results of entrepreneurial activity on linguistic diversity, and city share. Log(NUM) is the log number of private firms. Log(NUM.NEW) is the log number of new private firms. RATIO.EMP is the fraction of employees in each province employed by private firms. RATIO.ASSET is the fraction of asset in each province from private firms. Log(LD) is the log number of languages spoken in the firm's home province. CS is the median of number of languages spoken in cities of each province divide by LD. CS.High is a dummy variable which equals one if CS is greater than 0.4, and zero otherwise. CS.Low is a dummy variable which equals one if CS is lower than or equal to 0.4, and zero otherwise. GDP decile dummies are based on a firm's home province GDP per capita. Year dummies are included. The sample includes Tier 1˜Tier 4 provinces. The sample period is from 1999 to 2005. T-stats are in parentheses and standard errors are clustered by province.

Dep. Variable	Log(NUM)		Log(NUM.NEW)		RATIO.EMP		RATIO.ASSET	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log(LD)	0.797	1.176	0.811	1.770	2.452	2.834	1.519	0.604
	(2.10)	(0.96)	(2.36)	(2.23)	(1.66)	(4.57)	(0.69)	(0.09)
CS		2.092		3.307		1.864		-3.239
		(1.77)		(2.42)		(9.76)		(0.23)
Log(LD)*CS		1.842		2.021		1.531		-2.038
		(1.34)		(1.82)		(7.41)		(-0.18)
GDP Dum	YES	YES	YES	YES	YES	YES	YES	YES
Year Dum	YES	YES	YES	YES	YES	YES	YES	YES
Adj. R^2	0.632	0.645	0.629	0.661	0.276	0.268	0.263	0.254
# Obs	168	168	168	168	168	168	168	168

Table 20: Summary Statistics for International Sample

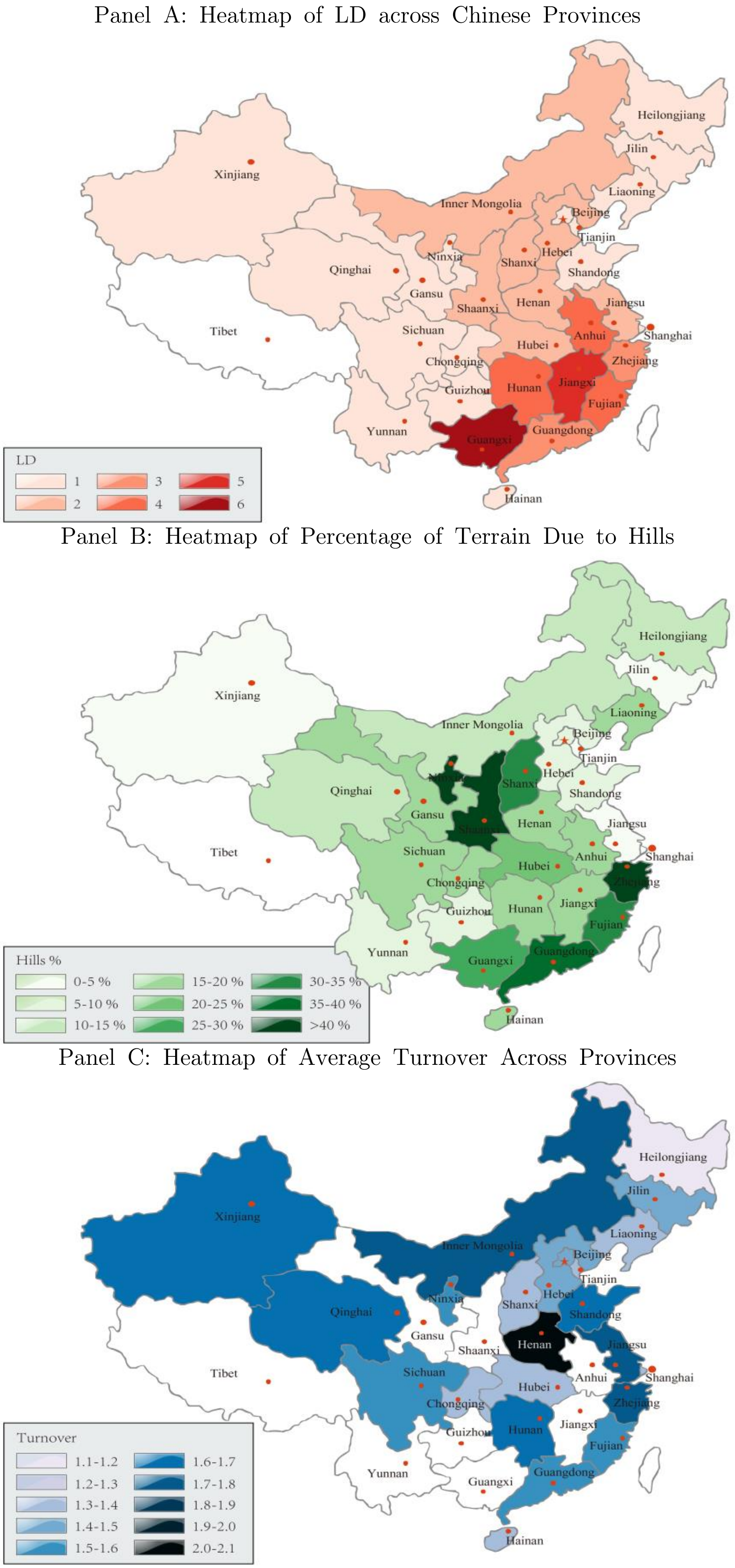
This table reports summary statistics of the international variables. IHDI, defined as one minus the language Herfindahl index, is from the UNESCO report based on the fraction of each language's speaker in a country's population. Turn is the time-series mean of the median monthly turnover from Hong and Yu (2009). MktCap is the size of a country's stock market capitalization by the end of 1999, in billions USD. GDPPC is the GDP per capita in 1999, in USD. AntiDir and JudEff are the anti-director index and judicial efficiency index from La Porta, Lopez-de-Silanes, Shleifer, and Vishny (1998). The sample consists of 41 countries.

	Min	25%	50%	75%	Max	Mean	Stdev
IHDI	0.00	0.14	0.38	0.66	0.93	0.40	0.29
Turn	0.00	0.01	0.02	0.03	0.23	0.03	0.04
Log(Turn)	-6.91	-4.34	-4.02	-3.44	-1.47	-4.04	0.98
GDPPC	287.92	2,021.68	9,554.44	21,715.10	38,290.67	12,804.93	11,817.37
Log (GDPPC)	5.66	7.61	9.16	9.99	10.55	8.73	1.46
MktCap	1.01	23.39	78.94	424.02	14,500.00	713.43	2,334.24
Log(Mkt Cap)	0.01	3.15	4.37	6.05	9.58	4.37	2.23
AntiDir	1	2	3	4	5	3.15	1.30
JudEff	2.5	6	7.25	10	10	7.55	2.09

Table 21: International Linguistic Diversity and Turnover

This table reports the OLS regression results of international turnover on linguistic diversity. The sample consists of 41 countries as described in Table 17. The dependent variable is the log of average median monthly turnover (Turn). IHDI, defined as one minus the language Herfindahl index, is from the UNESCO report based on the fraction of each language's speaker in a country's population. Other control variables: anti-director index (AntiDir), judicial efficiency index (JudEff), GDPPC and MktCap dummies denoting the decile (quintile) assignment of GDP per capita and stock market capitalization, Log(GDPPC) and Log(MktCap) are the logarithms of GDPPC and MktCap. T-stats are in parentheses.

	(1)	(2)	(3)	(4)	(5)
IHDI	0.897	0.887	0.921	0.953	0.908
	(1.61)	(1.62)	(1.77)	(1.58)	(1.89)
AntiDir		-0.142	-0.072	0.114	-0.034
		(-1.46)	(-0.73)	(0.91)	(-0.32)
JudEff			-0.179	-0.205	-0.228
			(-2.08)	(-2.18)	(-2.85)
Log(GDPPC)					0.514
					(2.82)
Log(MktCap)					0.081
					(0.85)
GDPPC Dum (10)	NO	NO	NO	YES	NO
MktCap Dum (10)	NO	NO	NO	YES	NO
GDPPC Dum (5)	YES	YES	YES	NO	NO
MktCap Dum (5)	YES	YES	YES	NO	NO
Adj. R^2	0.546	0.577	0.631	0.699	0.458
# Obs	41	41	41	41	41


Panel A: Heatmap of LD across Chinese Provinces

Panel B: Heatmap of Percentage of Terrain Due to Hills

Panel C: Heatmap of Average Turnover Across Provinces

Figure 1 Hills, LD, and Turnover. This figure plots the LD (Panel A), percentage of hill areas (Panel B), and average quarterly turnover (Panel C) for each province in China. Tibet is excluded from all three graphs. Gansu, Guangxi, Guizhou, Jiangxi, and Yunnan are excluded from Panel C. White area denotes the provinces that are excluded from the graphs.