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

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

This Draft: July 1, 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 fosters 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. Provinces with more linguistic diversity have more diverse opinions as measured by greater trading volume of local stocks and disagreement in stock message boards. Linguistic diversity is uncorrelated with province GDP per capita but is 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 Fudan University, 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 ethnic, 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 varieties 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.

Empirically, most of the evidence for this premise thus far has been confined to field and laboratory experiments. 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)). Given the relevance of this premise, research interest into why diversity might be good for productivity extends into other social sciences. For instance, an emerging consensus from experiments in the fields of psychology and education points to exposure to multiple cultures and languages training the brain to pay better attention, have better recall and be more creative (Bialystok and Martin (2004), Maddux and Galinsky (2009), Maddux, Adam, and Galinsky (2010), and Kovacs and Mehler (2009)).

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. In short, 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. The persistence of local languages is recognized by the Chinese themselves in the saying, "Nothing feels lonelier than two Shanghainese speaking Mandarin to each other." 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 (1988), 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 and a Herfindahl-based measure of the fraction of the population in the province speaking each language. 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, using data from the Thematic Database for Human-Earth System. Our measure based on the number of languages lines up the best with the terrain of China and also seems the most robust in the analysis below, though our Herfindahl-based measure is also correlated and yield similar results. So we use as our baseline linguistic diversity measure the number of languages spoken in the province and use the Herfindahl-based measure in our robustness checks. One reason is that the estimates of the fraction of the population actually speaking each language is very noisy whereas the number of languages spoken is measured precisely. Perhaps more importantly, both are uncorrelated with economic development measures like

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

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 has persistent influences on economic behavior even today. Their measures like ours, is a "deeper-rooted" measure 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 a large literature in behavioral finance which establishes that both retail and institutional investors over-weigh stocks with headquarters located near them and trade more when they have divergent opinions about the prospects of the local stocks (Odean (1999), Barber and Odean (2001),Grinblatt and Keloharju (2001)). In short, investors trade local stocks because they have different opinions for the future prospects of those stocks (Varian (1989), Harris and Raviv (1993) and Kandel and Pearson (1995)). 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. 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".

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 log of the number of languages spoken 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. We find an economically meaningful and statistically significant effect with t-statistics of around 2 where we have conservatively clustered standard errors by province. A one standard deviation increase in the log of the number of languages spoken in the province where the firm is headquartered is associated with an increase in the stocks log share turnover that is around 7% of the standard deviation of the left-hand side variable.

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. 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 measure a stock's investor base, given that we do not have data on who owns the shares, 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 (1955) 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.

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 firms 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.

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.

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 can be juxtaposed with a growing body of work on the importance of trust in economic exchange. Guiso, Sapienza, and Zingales (2009) argue that a lack of trust due to differing cultures leads to less trade. Our findings might be interpreted as surprising in light of theirs to the extent that linguistically rich regions might have lower levels of trust. Another literature is the impact of the structure of language on economic behavior. 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 paper proceeds as follows. We describe our datasets in Section 2. We present the main results in Section 3. We discuss and attempt to discriminate among alternative explanations in Section 4. We present the international sample results in Section 5. We conclude in Section 6.

2. Linguistic Diversity, Geography and GDP

We hand collect the languages spoken in the provinces of China from S.A. Wurm and Lee (1988). 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 languages spoken in each province, defined as LanNum, which is our main measure of linguistic diversity across Chinese provinces. The results are sorted by the log GDP per capita.

The province language is simply the union of the number of languages spoken in the various cities in a 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.

We can see the results of Table 1 in Figure 1, a map of the Chinese provinces along with the number of languages spoken in a circle and the GDP per capita. The distribution of these provinces along the coast will help us below 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 Table 1. Notice 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 LDI that is simply one minus the Herfindahl-based on share of the 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 LDI measure and our LanNum measure are highly correlated at .91.

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.

In Table 2 column (1), we regress 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 .107. Notice that the coefficient on Log (GDP) is statistically insignificant. The coefficient is -3.718 but only has a t-statistic of -.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 do 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 if 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 LanNum 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 LDI measure as the left-hand side variable rather than LanNum and we find that LanNum is more explained by terrain than LDI. The R^2's of the regressions for LanNum are larger than for LDI. Also, given that LDI depends on certain assumptions we make, we use LanNum as our preferred measure of diversity and resort to LDI as a robustness check. It turns out that the results do not differ too much in any event as we show in the robustness check section.

3. Baseline Result: 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. In Section 3.3 we show that results are robust to alternative sample cuts.

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. We also report the LanNum in each province again for convenience. Next to LanNum, 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 .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 LanNum.

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:

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

where the dependent variable, Log(TURN) is log of mean quarterly turnover over the sample period, LanNum is the number of languages spoken in a firms home province, and Log(LanNum) is the log of this variable. The control variables $\mathbf{X}$ 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 .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 is 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 LanNum 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 .052 with a t-statistic of 1.98 in column (1) and .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.

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 LanNum specification, we even get t-statistics now close to 2. As such, we conclude from Tables 4 and 5 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

But 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 both Table 1 and 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 .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 .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:

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

where we interact Log(LanNum) 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

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

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 6. Panel A has the results for the full sample. In column (1), we report the results for the Log(LanNum) specifications. First, the interaction coefficient of interest in column (1) is .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(LanNum) with CS, we are essentially re-weighting the Log(LanNum), whereby provinces with high CS effectively get a higher diversity score. For instance, a province with two languages but with a CS score of .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 .4 to get enough provinces into the low CS group. Here, the coefficient

for high CS provinces is .129 with a t-statistic of 2.98. The coefficient in front of low CS provinces is .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 LanNum 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.[2]

[2]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

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 in quarter is calculated as follows:

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

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 in quarter 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, $Prob(Guan)_{Hubei} = Prob(Gan)_{Hubei} = 0.5$. Finally, we drop the firm-quarter language Herfindahl measure if there are posts with non-identifiable IP addresses.

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:

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

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 is the number of message board posts for stock i, and n_i^j is the number of posts originated from Tier j cities for stock i.

The idea is that we estimate the effect of a stock's language Herfindahl on turnover while controlling for it's 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 should be 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 other two city and GDP Herfindahls. So we should be able to adequately address all remaining concerns on omitted stock characteristics driving our results.

Table 7 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 omitting 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 average number of posts per quarter and the average 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 8, 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

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

where $Log(H_i^{Lan})$ is the log of average language Herfindahl, and $LanNum_i$ is the 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 -0.107 with a t-statistic of -7.68. We further consider two alternative measures for number of languages. First, we take simply the number of languages, $LanNum_i$. In this case, the coefficient is -0.047 with a t-statistic of -6.72. Second, we use a dummy variable, $LanDum$, which equals 1 if the firm's headquarter province has more than one language, and zero otherwise. The coefficient is -0.090 with a t-statistic of -8.77. The results suggest that language is a strong instrument for a stock's language Herfindahl.

In Table 9, 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 three first-stage regressions, and the other covariates are listed in the captions of Table 9. All three specifications of our instrumental variables give consistent estimates. The implied economic effect, for instance from the first specification, for log Herfindahl of language on log turnover is .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 Measure 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 LDI that is simply one minus the Herfindahl-based on share of the 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 10, we then re-run our earlier baseline regressions using as the right-hand side variable Log(1+LDI). 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 LanNum. But the conclusions we draw are similar.

Finally, we have also performed a battery of robustness tests for our baseline regressions (Table 4), identification strategy 1 with city share (Table 6), and identification strategy 2 with language Herfindahls (Table 9). 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.

4. Small 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 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 11. The first is simply the 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. A second 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 third 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 12, while controlling for GDP per capita of the province. We find economically and statistically significant effects for our first two measures Log(NUM) or log of the number of private enterprises in a province regressed on the linguistic diversity of that province. In column (1), the coefficient in front of Log(LanNum) is .797 with a t-statistic of 2.1. We see again that the effects in columns (2) and (3) are somewhat larger in high CS provinces than low CS.

In columns (4)-(6), the dependent variable is RATIO.EMP. The coefficient in front of Log(LanNum) is 2.452 with a t-statistic of 1.66 in the univariate specification. The interaction of Log(LanNum) with CS yields a statistically significant coefficient of 1.531 with a t-statistic of 7.41 in column (5). When we decompose the baseline coefficient into a CS.High and CS.Low, we find again that the coefficient is somewhat larger in high CS provinces.

In columns (7)-(9), notice that we get a positive coefficient for the univariate specification but it is statistically insignificant. And neither of the other other specifications in columns (8) and (9) are 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 dominates 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.

5. 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 Linguistic Diversity Index (LDI) published by the United Nations Educational, Scientific and Cultural Organization (UNESCO). The LDI 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[3]. To be consistent with our variable definitions in Section 3, we use the language Herfindahl measure (H^{Lan}) instead of LDI for all following analyses. 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 Banks 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 13. 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 14 reports the regression results. Our benchmark specification (1) is as follows:

$$Log(Turn_i) = \alpha + \beta_1 H_i^{Lan} + \gamma' \mathbf{X} + \epsilon_i \tag{9}$$

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

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 H_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 $MktCap$. The coefficients on H_i^{Lan} are consistently economically significant with t-stats ranging from -1.58 to -1.89. Overall, the results in Table14 are consistent with the firm-level results in China that higher linguistic diversity leads to higher stock turnover.

6. 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.

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*.

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.

Barber, B. M., and T. Odean, 2001, "Boys Will Be Boys: Gender, Overconfidence, And Common Stock Investment," *The Quarterly Journal of Economics*, 116(1), 261–292.

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*.

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

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

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, 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.

Herfindahl, O., 1955, "Comment on Rosenbluths measures of concentration, in George J. Stigler, ed.," .

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

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, 65566560.

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 RomeLearn 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.

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

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.

S.A. Wurm, T. B., and M. W. Lee, 1988, *Language Atlas of China*. Longman Group, Hong Kong.

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 and the number of languages in each province (LanNum) in each province are obtained from the Language Atlas of China (1988). Linguistic diversity indexe (LDI) 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 in each city, assuming residents of each city speak all languages in their cities. H%, M%, and W% are the fraction of hills, mountains and waters 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)	LanNum	LDI	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	Minyu, 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	Minyu, 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	NA	NA	NA
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

Table 1—*Continued*

Province	Languages	Log(GDP)	LanNum	LDI	H%	M%	W%
Hunan	Guan, Gan, Kejia, Xiang	9.67	4	0.72	0.17	0.44	0.02
Hainan	Minyu	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, Minyu, 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%). Log(GDP) is the log of province GDP per capita. In Panel A the dependent variable is the number of languages spoken in each province (LanNum). The dependent variables in Panel B is linguistic diversity index (LDI), defined as 1 minus the language Herfindahl index in each province. Province language Herfindahl is measured by the fraction of population speaking by aggregating language speakers in each city. We assume residents of each city speak all languages in their respective cities. The sample excludes provinces that are adjacent to Mt. Everest: Chongqing, Sichuan, and Yunan. T-statistics are in parentheses.

Panel A			
Dependent Variable: LanNum			
	(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: LDI			
	(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. LanNum is the number of languages spoken in each province. 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 LanNum. 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	LanNum	CS	Turn	MV	VOL
Beijing	1	1	1.46	23.14	0.13
Shanghai	1	1	1.30	5.58	0.14
Tianjin	1	1	1.30	5.21	0.14
Jilin	1	1	1.44	2.31	0.19
Liaoning	1	1	1.30	2.46	0.15
Shandong	1	1	1.64	2.58	0.14
Chongqing	1	1	1.37	1.84	0.14
Heilongjiang	1	1	1.19	2.19	0.19
Xinjiang	1	1	1.63	3.37	0.15
Hainan	1	1	1.32	2.35	0.16
Ningxia	1	1	1.50	1.62	0.15
Qinghai	1	1	1.66	3.98	0.18
Sichuan	1	1	1.59	2.67	0.15
Jiangsu	2	0.50	1.77	2.48	0.13
Hebei	2	0.50	1.45	2.89	0.15
Neimenggu	2	0.50	1.70	3.43	0.14
Henan	2	0.50	2.05	2.88	0.14
Hubei	2	0.50	1.38	2.08	0.14
Shanxi	2	0.50	1.35	8.83	0.14
Shaanxi	2	0.50	1.54	2.65	0.14
Guangdong	3	0.33	1.58	4.56	0.14
Zhejiang	3	0.33	1.72	2.18	0.13
Fujian	4	0.25	1.58	4.11	0.15
Hunan	4	0.50	1.64	2.84	0.14

Panel B

	Mean	Stdev	Min	25%	50%	75%	Max
LanNum	1.94	0.99	1.00	1.00	2.00	3.00	4.00
Log(LanNum)	0.53	0.51	0.00	0.00	0.69	1.10	1.39
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. LanNum is the 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				
Dependent Variable: Log(Turn)				
	(1)	(2)	(3)	(4)
Log(LanNum)	0.060	0.065		
	(1.81)	(1.89)		
LanNum			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(LanNum)	0.052	0.053		
	(1.98)	(1.93)		
LanNum			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. For each stock, variables are averaged across the sample period. The dependent variable is log of mean quarterly turnover over the sample period. LanNum is the number of languages spoken in a firm's home province. Log(Pop) is the log population 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				
Dependent Variable: Log(Turn)				
	(1)	(2)	(3)	(4)
Log(LanNum)	0.053	0.057		
	(1.64)	(1.64)		
LanNum			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(LanNum)	0.058	0.057		
	(2.26)	(2.13)		
LanNum			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, 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. LanNum is the 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 LanNum. 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(LanNum)	-0.048			
	(-0.54)			
LanNum			-0.049	
			(-1.16)	
CS	-0.017		-0.272	
	(-0.14)		(-1.70)	
Log(LanNum)*CS	0.330			
	(2.03)			
LanNum*CS			0.177	
			(2.17)	
Log(LanNum)*CS.High		0.129		
		(2.98)		
Log(LanNum)*CS.Low		0.056		
		(2.05)		
LanNum*CS.High				0.066
				(3.38)
LanNum*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 6—*Continued*

Panel B: 2008-2012				
Dependent Variable: Log(Turn)				
	(1)	(2)	(3)	(4)
Log(LanNum)	-0.027			
	(-0.34)			
LanNum			-0.027	
			(-0.73)	
CS	-0.033		-0.182	
	(-0.39)		(-1.42)	
Log(LanNum)*CS	0.199			
	(1.48)			
LanNum*CS			0.102	
			(1.54)	
Log(LanNum)*CS.High		0.097		
		(3.19)		
Log(LanNum)*CS.Low		0.047		
		(2.15)		
LanNum*CS.High				0.050
				(3.50)
LanNum*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 7: 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. LanNum 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 per quarter. # 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	LanNum	Turn		H^{Lan}		H^{GDP}		H^{City}		# Posts	# Firms
		Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev		
Beijing	1.00	5.73	3.47	0.27	0.03	0.32	0.03	0.03	0.01	490,035	155
Shanghai	1.00	5.29	2.97	0.26	0.02	0.33	0.04	0.04	0.02	605,232	169
Tianjin	1.00	5.38	2.44	0.27	0.03	0.33	0.03	0.03	0.01	121,287	31
Jiangsu	2.00	6.84	3.09	0.27	0.03	0.32	0.03	0.03	0.01	613,631	191
Guangdong	3.00	6.61	3.39	0.24	0.03	0.33	0.03	0.03	0.01	1,349,423	288
Zhejiang	3.00	7.26	3.42	0.25	0.02	0.32	0.04	0.03	0.01	628,601	198
Jilin	1.00	6.34	3.16	0.27	0.04	0.31	0.03	0.03	0.01	92,706	41
Liaoning	1.00	6.44	3.49	0.27	0.03	0.31	0.03	0.03	0.01	268,858	63
Shandong	1.00	6.80	2.95	0.28	0.04	0.31	0.03	0.03	0.01	513,957	132
Hebei	2.00	5.96	2.68	0.26	0.02	0.31	0.03	0.03	0.01	223,352	47
Neimenggu	2.00	6.75	3.11	0.26	0.02	0.30	0.02	0.03	0.01	30,380	23
Fujian	4.00	7.09	2.87	0.25	0.02	0.31	0.03	0.03	0.01	343,760	81
Chongqing	1.00	5.93	2.34	0.29	0.03	0.30	0.04	0.03	0.01	128,539	34
Heilongjiang	1.00	5.80	2.56	0.27	0.03	0.31	0.03	0.03	0.01	126,213	33
Xinjiang	1.00	6.47	2.62	0.27	0.03	0.30	0.02	0.03	0.01	38,350	35
Henan	2.00	7.11	3.24	0.25	0.03	0.30	0.03	0.03	0.02	295,135	58
Hubei	2.00	6.27	2.64	0.26	0.02	0.30	0.03	0.03	0.01	296,090	70
Shanxi	2.00	5.56	2.50	0.25	0.02	0.30	0.03	0.03	0.01	89,377	29
Hainan	1.00	7.52	3.47	0.26	0.02	0.31	0.02	0.02	0.01	41,153	26

Table 7—*Continued*

Ningxia	1.00	6.58	1.94	0.29	0.03	0.29	0.02	0.02	0.01	15,572	12
Qinghai	1.00	6.97	3.48	0.26	0.02	0.31	0.03	0.02	0.00	7,867	11
Sichuan	1.00	6.63	2.86	0.28	0.03	0.30	0.03	0.03	0.03	322,742	89
Shaanxi	2.00	6.35	2.64	0.26	0.02	0.30	0.03	0.03	0.01	193,183	32
Hunan	4.00	6.96	2.75	0.25	0.03	0.29	0.02	0.03	0.01	247,116	61

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
LanNum	1.99	1.03	1.00	1.00	2.00	3.00	6.00
Log(LanNum)	0.55	0.52	0.00	0.00	0.69	1.10	1.79

Table 8: 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 $\text{Log}(\text{H}^{\text{Lan}})$. The primary independent variables of interest are (1) Log(LanNum): log number of languages of the firm's home province, (2) LanNum: number of languages of the firm's home province, and (3) LanDum: language dummy variable which equals one if the firm's home province speaks more than one language, and zero otherwise. Other control variables are $\text{Log}(\text{H}^{\text{City}})$, $\text{Log}(\text{H}^{\text{GDP}})$, 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: $\text{Log}(\text{H}^{\text{Lan}})$			
	(1)	(2)	(3)
Log(LanNum)	-0.107		
	(-7.68)		
LanNum		-0.047	
		(-6.72)	
LanDum			-0.090
			(-8.77)
$\text{Log}(\text{H}^{\text{City}})$	0.077	0.077	0.077
	(4.12)	(4.13)	(4.06)
$\text{Log}(\text{H}^{\text{GDP}})$	-0.375	-0.374	-0.380
	(-6.55)	(-6.54)	(-6.52)
Size Dum	YES	YES	YES
GDP Dum	YES	YES	YES
Adj. R^2	0.233	0.229	0.208
# Obs	1,926	1,926	1,926

Table 9: 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(LanNum): log number of languages of the firm's home province, (2) LanNum: number of languages of the firm's home province, and (3) LanDum: language dummy variable which equals one if the firm's home province speaks more than one language, and zero otherwise. $Log(H^{Lan1})$, $Log(H^{Lan2})$, and $Log(H^{Lan3})$ are language Herfindahl index instrumented by (1), (2), and (3) above, respectively. 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(Turn)			
	(1)	(2)	(3)
$Log(H^{Lan1})$	-0.659		
	(-2.57)		
$Log(H^{Lan2})$		-0.637	
		(-2.58)	
$Log(H^{Lan3})$			-0.746
			(-2.30)
$Log(H^{City})$	-0.326	-0.328	-0.320
	(-8.07)	(-8.02)	(-7.23)
$Log(H^{GDP})$	-0.118	-0.110	-0.151
	(-0.98)	(-0.95)	(-0.82)
Size Dum	YES	YES	YES
GDP Dum	YES	YES	YES
Adj. R^2	0.262	0.263	0.256
# Obs	1,926	1,926	1,926

Table 10: LDI, Segregation, and Turnover

This table reports OLS regression results of turnover on log(1+LDI), and city share. For each stock, variables are averaged across the sample period. The dependent variable is log of mean quarterly turnover. Linguistic diversity index (LDI) 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 in each 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 LanNum. 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+LDI)	0.106	-0.694	
	(1.11)	(-2.44)	
CS		-0.334	
		(-2.21)	
Log(1+LDI)*CS		1.024	
		(2.98)	
Log(1+LDI)*CS.High			0.159
			(1.63)
Log(1+LDI)*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+LDI)	0.079	-0.531	
	(1.02)	(-2.55)	
CS		-0.274	
		(-2.80)	
Log(1+LDI)*CS		0.713	
		(2.37)	
Log(1+LDI)*CS.High			0.114
			(1.50)
Log(1+LDI)*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 11: Summary Statistics for Private Firms

This table reports summary statistics of the private firm data. The sample period is from 1999 to 2005. NUM is the number of 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	LanNum	CS	NUM	RATIO.EMP	RATIO.ASSET
Beijing	1	1	4,991	92.01%	85.01%
Shanghai	1	1	11,239	91.26%	78.97%
Tianjin	1	1	5,528	98.02%	95.24%
Jilin	1	1	2,343	95.61%	90.59%
Liaoning	1	1	6,914	97.54%	93.41%
Shandong	1	1	15,591	96.39%	89.31%
Chongqing	1	1	2,024	95.12%	84.79%
Heilongjiang	1	1	2,322	96.03%	89.70%
Xinjiang	1	1	1,043	92.48%	80.41%
Hainan	1	1	483	96.45%	86.20%
Ningxia	1	1	415	85.98%	76.96%
Qinghai	1	1	303	89.46%	71.70%
Sichuan	1	1	4,714	92.87%	82.67%
Jiangsu	2	0.50	24,462	98.16%	92.93%
Hebei	2	0.50	7,307	94.71%	87.02%
Neimenggu	2	0.50	1,238	91.41%	80.34%
Henan	2	0.50	8,902	97.04%	89.90%
Hubei	2	0.50	5,830	96.15%	88.18%
Shanxi	2	0.50	2,589	96.67%	88.39%
Shaanxi	2	0.50	2,173	98.08%	92.64%
Guangdong	3	0.33	24,058	97.58%	89.53%
Zhejiang	3	0.33	24,435	97.38%	94.12%
Fujian	4	0.25	7,845	98.26%	91.68%
Hunan	4	0.50	4,989	95.29%	85.15%

Panel B

	Mean	Stdev	Min	25%	50%	75%	Max
LanNum	1.808	0.937	1	1	1	2	4
Log(LanNum)	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
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 12: Linguistic Diversity and Entrepreneurship

This table reports OLS regression results of entrepreneurial activity on linguistic diversity, and city share. The dependent variable is the log number of private firms, Log(NUM), in each province-year. Log(LanNum) 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 LanNum. 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)			RATIO.EMP			RATIO.ASSET		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Log(LanNum)	0.797	1.176		2.452	2.834		1.519	0.604	
	(2.10)	(0.96)		(1.66)	(4.57)		(0.69)	(0.09)	
CS		2.092			1.864			-3.239	
		(1.77)			(9.76)			(0.23)	
Log(LanNum)*CS		1.842			1.531			-2.038	
		(1.34)			(7.41)			(-0.18)	
Log(LanNum)*CS.High			0.924			2.646			1.537
			(1.70)			(1.22)			(0.48)
Log(LanNum)*CS.Low			0.622			2.184			1.494
			(2.65)			(1.06)			(0.88)
GDP Dum	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year Dum	YES	YES	YES	YES	YES	YES	YES	YES	YES
Adj. R^2	0.632	0.645	0.633	0.276	0.268	0.273	0.263	0.254	0.258
# Obs	168	168	168	168	168	168	168	168	168

Table 13: Summary Statistics for International Sample

This table reports summary statistics of the international variables. H^{Lan} is defined as one minus the Language Diversity Index (LDI) of a country from the UNESCO report. LDI is the summed squared proportion 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
H^{Lan}	0.07	0.34	0.62	0.86	1.00	0.60	0.30
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 14: 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 13. The dependent variable is the log of average median monthly turnover (Turn). The measure for linguistic diversity (H^{Lan}) is one minus the LDI index from the UNESCO report. LDI is the summed squared proportion 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)
H^{Lan}	-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

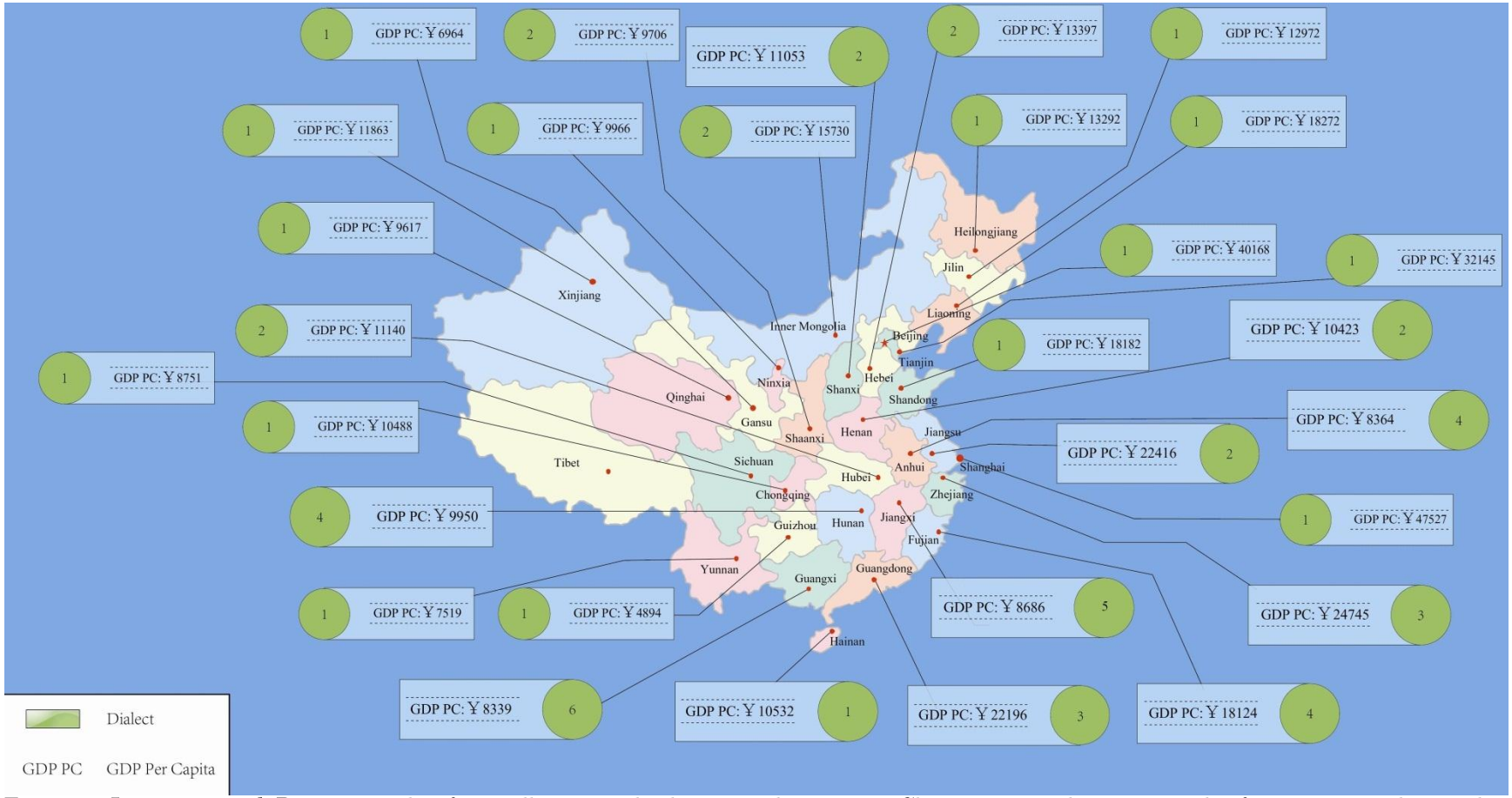

Figure 1. Languages and Provinces. This figure illustrates the language diversity in China. For each province, the figure reports the number of languages spoken (in green circle), and the GDP per capita (GDP PC) in Yuan.