Chinese Lexicography and Stock Trading*

Cui Hu
Central University of Finance and Economics
P. R. China

Ben Li
Boston College
United States

This draft: April 9, 2019

Abstract

If the stock market is efficient, stock tickers per se should not matter. We find that Chinese stocks with lexicographically earlier tickers are traded more frequently. Quantitatively, all else held equal, a one standard deviation later lexicographic position is associated with a decrease in turnover by nearly ten percent. This lexicographic bias is heavier in the service sector and for obscure stocks. It is primarily driven by the initial (character) of stock tickers. Stocks switching to later lexicographic positions are penalized in trading frequency. Using alternative financial measures leads to the same findings.

JEL codes: G11, G12, G14.
Keywords: Efficient market, alphabetic bias, Chinese stock market, behavioral finance

1. Introduction

If the stock market is efficient, stock tickers should not matter since the information in them, if any, should have been reflected in stock prices. Theoretically, stock ticker is simply the index of an Arrow-Debrew security, whose role is no more than differentiating the security from other securities. Referring to a security with a different index should have little reason to change the existing market equilibrium or its underlying machinery. However, we find that stock tickers matter and — to be more specific — matter lexicographically. In data from the Chinese stock market, which has the third largest equity value in the world, we find that stocks with lexicographically earlier tickers are traded more frequently. Quantitatively, a one standard deviation later lexicographic position leads to a decrease in stock turnover by nearly ten percent.

We also find that the trading frequency difference related to lexicographic order, or the *lexicographic bias* as we call it, displays systematic patterns that illustrate the importance of company visibility in the stock market. First, the lexicographic bias is most (least) salient in the service (agricultural)

*✉ Li: ben.li@bc.edu, +1-617-552-4517; Boston College, 140 Commonwealth Ave, Chestnut Hill, MA 02467, USA.

sector, reflecting the differential needs for visibility across sectors. Second, the lexicographic bias is stronger for less visible companies, indicating that an earlier lexicographic position compensates for obscurity. Third, the lexicographic bias is concentrated in the initial (character) of stock tickers, which is consistent with the lexicographic ordering of stocks by investors. Lastly, switching to a later lexicographic position reduces a firm's stock turnover, an effect that attenuates over time.

To our knowledge, the literature most relevant to our findings is the study of alphabetic bias. Alphabetical bias is widely observed in the Anglosphere and throughout the Western world. The English alphabet, as a revised Latin alphabet, discriminates names in a straightforward way. This has been shown to result in alphabetically earlier "goods" conferring certain advantages. The bias is known to exist in various business where visibility is crucial, including academic citations (Ray and Robson, 2018; Einav and Yariv, 2006), charity donations (Meer and Rosen, 2011), and stock trading (Jacobs and Hillert, 2016; Itzkowitz, Itzkowitz, and Rothbort, 2016). Jacobs and Hillert (2016) and Itzkowitz et al. (2016) find that Western stocks with alphabetically earlier tickers are traded more. The alphabetic order is a lexicographic order where words are ranked using the orders of their initials within an alphabet. The purpose of our paper is to examine whether the lexicographic order in Chinese, a non-alphabetic language, also influences stock trading.

Our study makes three departures from the literature, contributing to a deeper understanding of the relationship between lexicographic order and information frictions in stock trading. First, Chinese lexicography does not have a strict ordering system. Western alphabets assign an automatic strict order to all words. Chinese, on the other hand, is a logographic language, the basic unit of which is a two-dimensional logogram known as *character*. The two-dimensional nature of these characters renders a strict ordering system impossible. Chinese lexicography stipulates that characters written with fewer strokes occur earlier in lexicographic order. It is common for words to have an equal count of strokes, and therefore there are a large number of lexicographic ties. Figure 1 presents two characters, *yuan* and *wu*, extracted from Chinese stock tickers. They have the same lexicographic value of 4 as they both are written with four strokes. In Chinese lexicography, there exists no method to decide which of the two characters should be lexicographically earlier. Such ties are common and thus substantially reduce the variations in lexicographic order. They keep lexicographic positions measured conservatively. To this end, our study contributes to the literature by examining a non-alphabetic language where strict ordering is absent.

Second, the official (default) identifiers of Chinese stocks are their numeric stock IDs, rather than their tickers written in Chinese. The Chinese tickers were introduced to supplement stock IDs as they can be cognitively linked to company names. In the media and in publications, the numeric ID and Chinese tickers are normally used together, to strike a balance between rigor in writing and convenience in reading. Taking advantage of this ID-ticker combining practice, we examine whether the lexicographic positions of the Chinese tickers remain to matter. Our results show that the Chinese tickers, despite not being official identifiers of stocks, still impact their trading. This is strong evidence of the lexicographic bias, which has been so far found only in markets where tickers are the official and

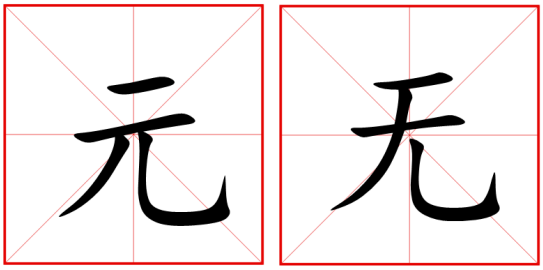


Figure 1: Chinese Characters with the Same Lexicographic Value

The two characters, *yuan* (left) and *wu* (right), have the same lexicographic value (stroke count) of 4.

default identifiers of stocks.

Third, the pervasive lexicographic ties mentioned earlier enable us to address endogenous choices of stock tickers. Listed companies have incentives to choose lexicographically earlier tickers to gain visibility in markets. This is especially true for marketing-savvy firms, which conceivably have better stock-market performances. Such endogeneity is left unaddressed in the existing studies and we fill the gap. Specifically, we control for lexicographic value fixed effect, such that in a given trading month, stock turnover is econometrically compared only among stocks with the same lexicographic value. That is, stocks with a lexicographic value, such as 4, have the same *lexicographic position* when they are traded together in a market. The variations in lexicographic position stem only from the different sets of stocks traded over time. For example, a variation in lexicographic position arises between i) when two stocks with a lexicographic value of 4 are traded along with stocks having a lexicographic value of 5, and ii) when the same two stocks with a lexicographic value of 4 are traded together with stocks having a lexicographic value of 6. In other words, adopting stock tickers with a lexicographic value of 4 is potentially linked to some marketing savvy of the companies, which however has been "controlled for" here. The marketing savvy, together with all other potential common characteristics of companies adopting tickers with a lexicographic value of 4, are held away from the variations in both stock turnover and lexicographic position that give us the statistical association between the two variables.

More broadly, our study is related to the literature on the cognitive biases associated with stock and fund names (Alter and Oppenheimer, 2006; Anderson and Larkin, 2019; Cooper, 2001; Cooper, Gulen, and Rau, 2005; Durham and Santhanakrishnan, 2016; Green and Jame, 2013; Itzkowitz and Itzkowitz, 2017; Krishnamurthy, Pelletier, and Warr, 2018). This literature finds that investors or more specifically their brains opt for stocks or funds whose names provide information that is somehow easier to digest, such as having simpler pronunciation or resembling known words. The literature has been limited to the Western world, and thus it is unclear whether the cognitive biases associated with names actually result from certain linguistic features of Western languages, such as phonetic spelling, Latin origins, and derivation based word formation. Languages are part of cultures and cultural traits are known to influence investment behaviors. We document the relevance of names in a completely different

cultural setup, which clearly signifies the cognitive nature of name related biases.

The rest of the paper is organized as follows. In Section 2, we describe Chinese lexicography and our identification strategy. In Section 3, we provide data sources and preliminary data patterns. Afterwards, we report our baseline findings and robustness checks in Section 4, experiment with different measures of lexicographic positions in Section 5, and examine firms that change their tickers in Section 6. We examine the relationship between other financial measures and lexicographic positions in Section 7. In Section 8, we conclude.

2. Background and Identification

2.1. The Chinese Lexicography of Stock Tickers

The basic unit of the Chinese language is a *character*, and every Chinese character consists of a specific number of strokes. The stroke count of a Chinese character is its lexicographic value.[1] In Chinese lexicography, characters with fewer strokes and thus lower lexicographic values are defined as occurring lexicographically earlier. The previous Figure 1 demonstrates two characters with the same stroke count of 4, where 4 is also their lexicographic values. The two terms, stroke count and lexicographic value, can be used interchangeably.[2] For consistency, we use the term lexicographic value (or *LV* for short) hereafter.

Just as in Figure 1, many characters share the same lexicographic value. In Chinese lexicography, characters with lower lexicographic values are used more frequently. In Figure 2, we plot *in blue (connected using hollow circles)* the distribution of lexicographic values of commonly used Chinese characters.[3] The lexicographic values range between 1 and 24, with a mean of 12.5 and a standard deviation of 7.7. Evidently, characters with low lexicographic values outnumber those with high lexicographic values.

Words, as linguistic units composed of more than one character, have the same lexicographic val-

[1] Chinese characters have to be written following a set of rules, which ensure that the stroke count of any Chinese character is unambiguous. The rules of writing were originally made to ensure that characters written by one person are legible to others. For a complete list of writing rules, see *Stroke Order National Standard (GF3002-1999 GB13000.1)*, a document issued by the Chinese government and made available at http://www.moe.edu.cn/s78/A19/yxs_left/moe_810/s230/201001/W020150902457900316281.pdf. In general, the writing rules specify five basic stroke types, such that every Chinese character can be deconstructed into strokes that fall into the five types in the same way regardless of the writer. Chinese learners are required to memorize those rules. Following the rules, Chinese learners can calculate every character's stroke count quickly in mind, which is useful when they look up characters in Chinese dictionaries.

[2] There are several ways to refine lexicographic values by accounting for the types of strokes used in characters with the same stroke count. We do not take these into consideration here because they fail to resolve many lexicographic ties, such as between the two characters presented in Figure 1.

[3] The distribution is based on the 3,500 Chinese characters included in the *List of commonly used characters* (*Xiandai Hanyu Changyong Zibiao*) published by the Ministry of Education of China in 1988, adjusted with modern Chinese vocabulary frequency calculated by Da (2004). In this study, we use the FFCell plugin in Microsoft Excel to calculate the stroke counts of Chinese characters. The plugin is available at www.ffcell.com.

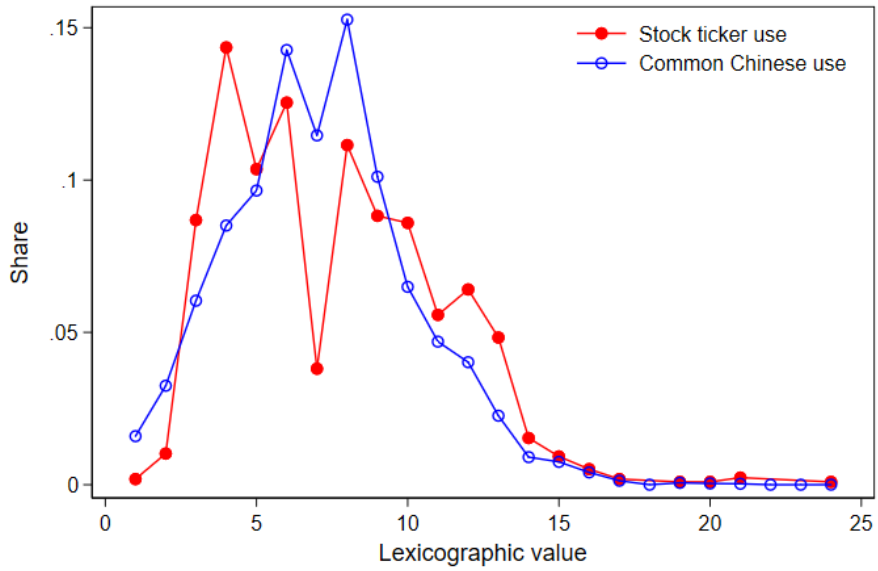


Figure 2: Distribution of Lexicographic Values in Stock Tickers and Common Chinese Use

ues as their initial characters. Therefore, all kinds of nouns, including the names of people, companies, and stocks, can be ranked lexicographically. Well-known use of lexicographic ordering in China includes the list of politburo members of the Chinese Communist Party and the list of congressmen in the National Congress of China. This convention in Chinese politics is used to avoid potential political implications of ordering the names of politicians in alternative or unconventional ways. Since the initial character in Chinese individual names comes from the surname, which is always listed before the given name, this convention is known as the "surname stroke order" (in Chinese, *an xingshi bihua paixu*). It is also used in other institutional arrangements for similar purposes, such as corporate boards, professional associations, and book series editors. In fact, the best example of the lexicographic ordering of words are reference books in Chinese, such as a handbook of electronic engineering, where voluminous technical terms have to be ordered in a way that any reader of the handbook who understands Chinese can find what she wants by browsing through the lexicographic values starting from 1.[4]

Every Chinese stock has a word as its ticker. The ticker is required to be related to the stock-issuing company's full name. For example, *yuan cheng gu fen* is the stock ticker of the Yuan Cheng Environmental Co. (*gu fen* means stocks in Chinese) In practice, stock tickers are oftentimes used as the short names of companies. Listed companies also have numeric IDs as their official (default) identifiers. For example, the Yuan Cheng Environmental Co. has 603388 as its numeric ID. Both stock tickers and numeric stock IDs are determined prior to initial public offerings. Numeric stock IDs are permanent,

[4]There are only two Chinese characters that have one stroke (i.e. the lexicographic value of 1).

while stock tickers can change from time to time, reflecting changes to company names, capital structure, business concentration, or business strategy. Stock IDs and tickers are normally used together in the media and in publications to strike a balance between rigor in writing and convenience in reading.

In the present study, we calculate the lexicographic values of all stocks that have ever been traded in mainland China. As an example, stocks traded in December 2015 have a lexicographic value distribution shown by the *red line (connected using solid circles)* in Figure 2. The stroke counts range between 1 and 24 just as in the common Chinese use mentioned earlier, with [3,6] being the densest range. There are no stock whose lexicographic values equal 18, 22, or 23. The lexicographic values range between 1 and 24, with a mean of 7.5 and a standard deviation of 3.5.

A comparison between the two distributions in Figure 2 demonstrates that, relative to the common Chinese use, listed companies tend to choose Chinese tickers with low lexicographic values. The tendency is most evident in the [3,4] range. Otherwise, the two distributions have similar patterns, suggesting that the character choices for stock tickers do not depart far from the common Chinese use.

2.2. Identification Strategy

Position vs. value. The variable of interest in our analysis is lexicographic *position* rather than lexicographic *value*. Intuitively, lexicographic bias arises as a relative phenomenon. That is, stock tickers with a lexicographic value of 4 could be lexicographically earlier or later, depending on the tickers of other stocks traded together in the market. The lexicographic value of 4 is generally to the left of the spectrum (recall Figure 2), though it remains possible that it would be lexicographically late if, for example, the stocks having a lexicographic value of 4 are traded only with stocks having a lexicographic value of 3. Intuitively, in this scenario, stocks with a lexicographic value of 3 may have visibility advantages compared to those with a lexicographic value of 4, which is the hypothesis that we will test later. This example is extreme but introduces a notion that is crucial for our identification strategy. That is, we are interested in the relative lexicographic situations of stocks when they are compared with each other. The lexicographic values have little relevance on their own.

The previous Figure 2 suggests another reason for focusing on lexicographic positions rather than lexicographic values. The dispersion of lexicographic values are highly skewed. In fact, quite a few high lexicographic values have few or even no corresponding stocks. This would engender econometric issues if lexicographic value were used as the variable of interest in regressions.

Both reasons above apply as well to the existing studies on alphabetic bias in Western stock markets. They constructed alphabetic positions of stocks traded in the market every month, rather than numeric value of letters (such as 1 for A and 26 for Z). We follow their procedure to construct lexicographic positions for Chinese stock tickers. Our construction also accounts for the unique elements in Chinese lexicography. Denote the lexicographic value of stock s by $LV_s \in [1,24]\backslash\{18,22,23\}$, where 18, 22, or 23 are the lexicographic values to which no stock ticker correspond. We rank all stocks traded

in the market in a given month t in ascending order of their LV values, to obtain their stock-month duplet level rank $R_{s,t} = 1, 2, 3, ..., 21$, where the maximum value of $R_{s,t}$ is 21 rather than 24 because of the "missing" 18, 22, or 23 values. Tie cases take equal values in $R_{s,t}$.

China has two stock exchanges, one in Shanghai and the other in Shenzhen. On each of the two exchanges, there are two types (A and B) of stocks traded. Type-A stocks are traded in the Chinese currency yuan while Type-B stocks are traded in foreign currencies. Combining the two stock exchanges and the two stock types, we have four *markets*. In a given time period, a given stock is traded only in one out of the four markets (i.e. on one exchange, in one currency). Denote a market by j, such that every stock s is by definition traded in only one j.[5]

For each market-month (jt) pair, we calculate the maximum $R_{s,t}$ and use it to construct a normalized *lexicographic position*. That is, the lexicographic position of stock s in its market (j) in month t equals

$$Lexi\_position_{s,t} = \frac{R_{s,t}}{R_{jt}^{\max}}, \tag{1}$$

where

$$R_{jt}^{\max} \equiv \max_{s \in jt}\{R_{s,t}\}. \tag{2}$$

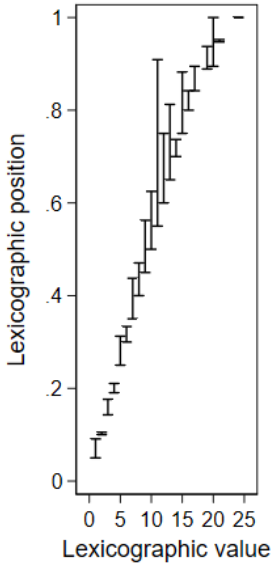
The normalization ensures that $Lexi\_position_{s,t}$ must be between 0 and 1 so that it is comparable across market-month (jt) duplets.[6] Notice that $R_{s,t}$ and its normalization $Lexi\_position_{s,t}$ of stock s depend not only on lexicographic value LV_s but also on the tickers of other stocks traded in the same market.

Two implications immediately follow from the above setting: (i) technically, a low lexicographic value (LV) is neither sufficient nor necessary for an early lexicographic position $Lexi\_position_{s,t}$, whereas (ii) statistically, a low lexicographic value is more likely to have an early lexicographic position. The statistical implication (ii) is due to the fact that a low lexicographic value faces less "competition" across the spectrum of lexicographic values (recall Figure 2). The statistical implication is evidenced by our sample, as shown in Figure 3, where the 10th and 90th percentile of lexicographic positions are provided for each lexicographic value (details of the sample will be provided later).

In Figure 3, the length of each bar represents the within-lexicographic-value variation. Across lexicographic values there emerges a pattern that the within-lexicographic-value variations of mid-range lexicographic values are larger than those of low and high lexicographic values. Intuitively, stocks with low (high) lexicographic values are more likely to have early (late) lexicographic positions. For example, it is unlikely for a lexicographic value of 4 to be ranked relatively late, while it is similarly unlikely for a lexicographic value of 24 to be ranked early. The lexicographic positions of those stocks with mid-range

[5] None of the Chinese listed companies issue stocks on both exchanges. But on one exchange, a company may issue both type-A and type-B stocks. When this happens, the two stocks are considered as two stocks because they have numeric tickers distinct from each other. For our research purposes, they should also be considered as two separate stocks because they have different peers in their markets.

[6] The upper bound 1 applies if a stock s is the lexicographically latest one in its market in month t.



Summary Statistics of Lexicographic Positions across Lexicographic Values*

Mean *Lexi_position*$_{st}$	
Mean:	0.563
Std.:	0.303
Min:	0.061
Max:	1.000

Median *Lexi_position*$_{st}$	
Mean:	0.548
Std.:	0.306
Min:	0.053
Max:	1.000

(* 21 unique values)

Figure 3: Lexicographic Position vs. Lexicographic Value

Chinese tickers with the same lexicographic value have different lexicographic positions, depending on what other stocks are being traded in the market at the same time. For each lexicographic value (from 1 to 24, with 18, 22, and 23 excluded), we find the mean and median lexicographic positions. Across the lexicographic *values*, we report mean, standard deviation, minimum and maximum of the mean and median lexicographic *positions* (see the right). We plot the 10th and 90th percentiles of lexicographic positions across lexicographic values (see the left). In our later regressions, because of the lexicographic value fixed effects, the variations in use are essentially those within lexicographic values (vertical bars in the figure).

lexicographic values are more dependent on other stocks traded at the same time. Our later regressions, as explained below, mainly use the variations within each lexicographic value, or graphically speaking, within each vertical bar shown in Figure 3.

In addition, for each lexicographic value, we calculate the mean and median of lexicographic positions. Their summary statistics are reported in the right part of Figure 3. The mean and median lexicographic positions across lexicographic values turn out to be close to each other, suggesting an overall symmetric distribution of lexicographic positions.

Regression. We hypothesize that stocks with lexicographically earlier tickers are traded more frequently. Intuitively, this phenomenon emerges so long as *some* investors with limited attention *sometimes* view stocks lexicographically when they decide what stocks to invest in. Our regression follows Chordia, Huh, and Subrahmanyam (2007) and Jacobs and Hillert (2016) in using observable firm char-

acteristics to predict future turnover:

$$\ln Turnover_{s,t+1} = \beta Lexi\_position_{s,t} + \bar{\delta}' X_{s,t} + \lambda_{LV} + \lambda_I + \lambda_Y + \epsilon_{s,t+1}. \tag{3}$$

where $Turnover_{s,t+1}$ is defined as the number of shares traded divided by the number of shares outstanding for stock s in month $t+1$ (we also use other trading frequency measures; see Section 7). β is the parameter of interest. $X_{s,t}$ is a vector of firm-level control variables. λ_I is an industry fixed effect, while λ_Y is a year fixed effect. $\epsilon_{s,t+1}$ is the error term, double-clustered by stock and month in estimation.

At the core of our identification strategy is the lexicographic value (LV) fixed effect λ_{LV} in regression (3). Recall that the range of lexicographic value LV is $[1,24]\backslash\{18,22,23\}$. For each $LV = l$ value in the range, we set up a dummy variable $D_l = 1$ (otherwise, $D_l = 0$). When D_l is present along with the $Lexi\_position_{s,t}$ in equation (1), it absorbs the non-relative variations in $Lexi\_position_{s,t}$. The coefficients of the dummy variables, namely λ_{LV}, capture the lexicographic value fixed effect. For each lexicographic value, λ_{LV} ensures that the variations used in our study come from relative lexicographic positions. It also addresses potential endogenous choices of tickers in favor of certain lexicographic values. For example, marketing-savvy firms might be more likely to choose lexicographic values 2 and 3, but the resulting turnover premium has been absorbed by $D_{l=2}$ and $D_{l=3}$.

Numeric ID order. A remarkable control variable in the $X_{s,t}$ of regression (3) is the numeric ID position of a stock s in month t. Since numeric IDs of stocks are their official (default) identifiers and numeric values are *a priori* orderable, we construct a numeric ID position following the same method we used to construct the lexicographic position. Denote the numeric ID of stock s by n_s, and its numeric ID rank in its market (j) in month t by $N_{s,t}$. $N_{s,t}$ is increasing in n_s. We can then construct the numeric ID position of stock s in market-month (jt) duplet as

$$Num\_position_{s,t} = \frac{N_{s,t}}{N_{jt}^{\max}}, \tag{4}$$

where

$$N_{jt}^{\max} \equiv \max_{s \in jt}\{N_{s,t}\}. \tag{5}$$

Although the numeric ID position of a stock also influences its visibility, we do not have a specific prior on how it influences the visibility. On the one hand, stocks with smaller numeric IDs are more likely to appear at the top of portfolios, generating a visibility advantage. On the other, the newest stocks, which automatically have larger numeric IDs, usually draw heavy attention in the market. We remain agnostic on the coefficient of $Num\_position_{s,t}$ in regression (3) and consider the variable simply as a covariate of stock turnover to hold constant.

3. Data

Sources. The main data source used in this study is the China Stock Market and Accounting Research Database (CSMAR), maintained by the GTA Information Technology Co., Ltd. The CSMAR database records all stock transactions beginning in the year 1990 when China (re)opened stock exchanges in Shanghai and Shenzhen.[7] It also provides a wide range of information on publicly listed firms in mainland China. All variables in this study, except Chinese tickers, are obtained from the CSMAR database. The CSMAR database refers to all stocks using their present tickers. That is, it links the transaction records of stocks previously under different tickers to their current tickers. This renders the study of ticker changes infeasible. To address the issue, we resort to the Windinfo database where historical tickers of stocks were all recorded. Since the numeric stock IDs are permanent, we merge historical tickers of stocks through corresponding stock IDs into the CSMAR data.

This study was initiated in the year 2018, when the 2017 data in the CSMAR database were not yet completed (some company-level information has delayed availability). We decided to end our sample coverage in December 2015 rather than December 2016, because there was a market crash in Chinese stock exchanges in January 2016. Also in January of that year, the China Securities Regulatory Commission introduced trading curbs, which had a tremendous impact on the turnover of stocks. Our working dataset is at the stock-month level, covering the time period starting in December 1990 and ending in December 2015.[8]

Patterns. We start with plotting stock turnover against lexicographic *value* to uncover basic correlations. Recall that lexicographic values are time-invariant characteristics of firms (except for companies that ever changed their tickers). We take the mean of turnover for each firm's stock across its trading months to obtain its average turnover. The average turnover is shown as the green line (connected by circles) in Figure 4. A Lowess smoothing is done to the average turnover across lexicographic values, as demonstrated by the red (dash) line in the figure. At the bottom of the figure we reproduce the distribution of lexicographic values across firms (recall the red line in Figure 2, now presented in the form of bars). An apparent downward pattern is revealed in the figure, alluding to the negative relationship between turnover and lexicographic value at the company level.

A natural question arises as to the magnitudes of the variations in use. We first calculate the standard deviation of lexicographic position *within* each lexicographic value and report them in the upper panel of Table 1. We next calculate and report weighted averages of the within-lexicographic-value standard deviations across lexicographic values. The weight schemes in use include simply (i.e. equally) weighted, number-of-firms weighted, and market-capital weighted. The average of the standard deviations ranges between 0.037 and 0.048. To gauge the magnitude of these means, we calculate

[7] There was a Shanghai Stock Exchange operated in China from 1904 through 1949. It was closed in the year 1949 when the Chinese Communist Party took power.

[8] Stocks issued by Chinese firms in foreign stock markets are not considered in this study.

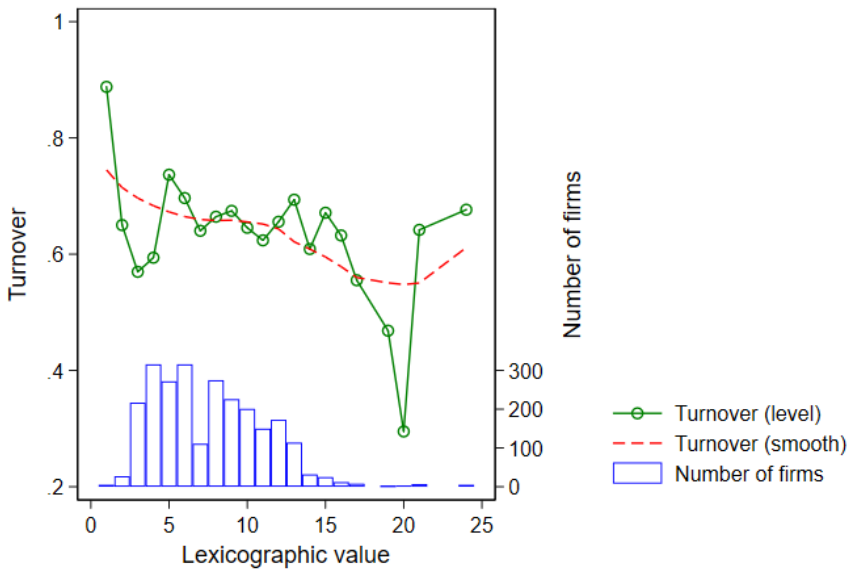


Figure 4: Turnovers and Lexicographic Values

In this figure, the average turnovers of stocks associated with different lexicographic values are plotted against lexicographic values. The solid (dashed) curve represents the actual averages (Lowess-smoothed). The bars at the bottom represent corresponding numbers of firms.

the cross (lexicographic value) averages of the within (lexicographic value) means. Mechanically, the simply weighted cross average of the within mean, 0.563, is the same 0.563 reported in the right part of Figure 3. Lastly, we calculate the corresponding coefficients of variation (i.e. standard deviation divided by mean) and report them at the bottom of Table 1. Overall, the standard deviation behaves stably in response to the use of different weights. There are enough within-lexicographic-value variations in lexicographic position (except when the lexicographic value equals 24 (the maximum), a case excluded in later regressions).

As noted earlier, the range of lexicographic position is between 0 and 1. On average, one standard deviation accounts for 7.5% to 12.2% of the mean value in lexicographic position.

Table 2 reports the summary statistics by lexicographic position. Specifically, we divide observations within each [industry,lexicographic value] duplet into early (upper-half) and late (lower-half) groups using the duplet-level lexicographic position mean, and compare turnover and other firm characteristics used in regression (3) between the two groups. The firm characteristics compared here are self-explanatory (a detailed description of the variables is provided in Table A1). It is clear that stocks in the upper half demonstrate advantages over those in the lower half, including higher turnovers, more positive returns, larger market capitalization, more R&D and advertisement expenditures, and a greater

number of related analysts. The only exception is that the leverage ratio does not significantly differ between the two groups.

Table 1: Magnitudes of Variations in Use

In Panel A, the standard deviation of lexicographic positions within each lexicographic value (LV) is reported. In Panel B, we report the averages, weighted across lexicographic values, of (i) the above standard deviations, (ii) the corresponding means, and (iii) the corresponding coefficients of variations (i.e. standard deviation divided by mean). Three different types of weights are used here: simple weights (all LV values receive equal weights), number-of-firms weights, and market-capital weights. Number of firms and market capital refer to their aggregate levels at the lexicographic value level.

Panel A: Within-LV standard deviattions of lexicographic positions

LV values*	Standard deviation of lexicographic position	LV values*	Standard deviation of lexicographic position
1	0.017	12	0.109
2	0.007	13	0.083
3	0.020	14	0.048
4	0.020	15	0.059
5	0.029	16	0.017
6	0.023	17	0.020
7	0.055	19	0.024
8	0.047	20	0.050
9	0.058	21	0.014
10	0.061	24	0.000
11	0.127	* 21 unique values	

Panel B: Cross-LV variations of within-LV moments

Cross-LV average of within-LV standard deviations		
	Simply weighted	0.042
	Number weighted	0.048
	Capital weighted	0.037
Cross-LV average of within-LV means		
	Simply weighted	0.563
	Number weighted	0.395
	Market capital weighted	0.328
Cross-LV average of coefficients of variation		
	Simply weighted	0.075
	Number weighted	0.122
	Market capital weighted	0.113

Table 2: Summary Statistics by Early and Late Lexicographic Positions

This table provides the summary statistics of the data in the following fashion. The median of lexicographic position (as defined in the text) is calculated for each [industry-LV] duplet. Using those medians, we divide stocks into early (upper-half) and late (lower-half) groups. For each group, the mean, median, and standard deviation of various firm characteristics are reported. For each firm characteristic, we conduct a comparision (t-test with sample size adjusted) between the two groups and report the differences (early minus later) and p-values in the last two columns.

Variable	Mean	Median	Std.	Mean	Median	Std.	Difference	p-value
	Early (upper-half)			Late (lower-half)			Early minus late	
Turnover	0.48	0.31	0.52	0.45	0.27	0.51	0.03	0.000
Positive return	0.53	1	0.50	0.51	1	0.50	0.02	0.000
Negative return	0.46	0	0.50	0.49	0	0.50	-0.02	0.000
Price	11.66	8.93	11.06	12.11	9.15	32.04	-0.45	0.000
Return volatility	0.03	0.03	0.09	0.04	0.02	0.48	-0.01	0.000
Market capitalization	1.04E+07	3.15E+06	5.79E+07	3.85E+06	2.07E+06	9.96E+06	6.55E+06	0.000
Sales	8.70E+09	1.36E+09	6.61E+10	2.19E+09	6.23E+08	1.57E+10	6.50E+09	0.000
Age	12.90	13	5.73	8.00	7	4.68	4.91	0.000
Leverage	2.37	1.88	15.38	2.31	1.80	8.57	0.06	0.127
Advertisement	3.57E+06	0	9.53E+07	1.69E+05	0	7.44E+06	3.40E+06	0.000
No. of analysts	27.69	3	52.56	8.33	0	26.56	19.36	0.000
R&D	1.08E+07	0	1.14E+08	6.66E+05	0	9.09E+06	1.02E+07	0.000
Book to market ratio	1.13	0.71	1.57	0.87	0.65	0.91	0.26	0.000

Table 3: Lexicographic Position and Firm Characteristics

This table shows coefficients and t-statistics (in parentheses) obtained from estimating the correlation between lexicographic position and firm characteristics. Dependent variables are either rank value $R_{s,t}$ or lexicographic position (labeled in column headings). Standard errors are double-clustered by firm and month. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

	(1) Rank value (L_{st})	(2) Lexicographic position	(3) Lexicographic position	(4) Rank value (L_{st})	(5) Lexicographic position
Estimation method	Panel regression			Fama/MacBeth regression	
Positive return	-0.012	-0.001	-0.001	0.020	0.003
	(-0.151)	(-0.316)	(-0.682)	(0.189)	(0.372)
Negative return	-0.034	-0.003	-0.001	0.029	0.014
	(-0.430)	(-0.717)	(-1.158)	(0.245)	(0.924)
Price	0.203*	0.010*	-0.001	0.260***	0.015***
	(1.944)	(1.797)	(-0.579)	(4.774)	(3.907)
Return volatility	-0.504***	-0.035***	-0.009*	-0.387	-0.120**
	(-3.633)	(-3.828)	(-1.826)	(-0.676)	(-2.552)
Market capitalization	-0.347***	-0.018***	-0.000	-0.148***	-0.002
	(-3.888)	(-3.816)	(-0.100)	(-3.187)	(-0.266)
Leverage	0.001	0.000	-0.000	0.018	0.000
	(0.567)	(0.346)	(-1.116)	(0.799)	(0.071)
Sales	-0.023	-0.003	-0.002**	-0.070***	-0.002
	(-0.515)	(-1.122)	(-2.309)	(-2.669)	(-0.462)
Age	0.200	0.006	-0.005***	0.129***	0.003
	(1.503)	(0.825)	(-3.083)	(2.946)	(1.078)
Book to market ratio	-0.029	0.008**	0.009***	0.091	-0.003
	(-0.612)	(1.971)	(4.509)	(1.256)	(-0.171)
Numeric ID position	0.000	-0.000**	-0.000***	-0.001	-0.003
	(0.217)	(-2.382)	(-4.235)	(-0.132)	(-0.985)
Advertisement	0.061	0.002	-0.001**	-0.056	-0.003
	(0.902)	(0.646)	(-2.179)	(-0.520)	(-0.616)
Missing advertisement dum.	0.971	0.038	-0.012*	-0.311	-0.022
	(1.003)	(0.780)	(-1.954)	(-0.210)	(-0.292)
No. of analysts	0.076	0.004*	0.000	0.376**	0.020***
	(1.592)	(1.757)	(1.010)	(2.560)	(2.612)
Missing No. of analysts dum.	0.057	0.010	0.007***	0.359**	0.023***
	(0.514)	(1.629)	(4.265)	(2.554)	(3.197)
R&D	-0.038	-0.001	0.001	-6.692**	-0.358**
	(-0.484)	(-0.296)	(1.522)	(-2.261)	(-2.264)
Missing R&D dummy	-1.243	-0.052	0.012	-111.908**	-5.992**
	(-0.923)	(-0.774)	(1.412)	(-2.252)	(-2.256)
Fixed effects	Industry, year	Industry, year	Industry, year, and LV	NA	NA
Observations	336,774	336,774	336,774	336,774	336,774
R-squared	0.050	0.070	0.963	0.183	0.189

An obvious caveat when one interprets the comparison in Table 2 is the absence of control vari-

ables. In Table 3, we regress both the previous lexicographic position $Lexi\_position_{s,t}$ and its non-normalized counterpart $R_{s,t}$ on firm characteristics. These firm characteristics will also serve as control variables in our later regressions. Variables related to advertisement, analysts and R&D have a large number of missing values. We treat their missing values as zero and add corresponding missing variable dummies to address the truncation. In the first three columns we use industry and year fixed effects (in column (3), the lexicographic value (LV) fixed effect as well).[9] In the last two columns, we use instead the estimation approach proposed by Fama and MacBeth (1973). Econometrically, the fixed effect regressions weight firm-month's equally, whereas the Fama and MacBeth (1973) regressions weight months equally.

In general, lexicographic position has only few robust associations with firm-level characteristics.[10] Among columns (1)-(3) that use the fixed-effect estimation, column (3) has the most stringent specification (with LV fixed effects included).[11] By comparing that column with column (5) that also uses lexicographic position, one can see that the only coefficient appearing to be statistically significant at the 1% level in both columns is the dummy variable related to missing analysts. Missing analysts is suggestive of having either no or a very small number of recorded analysts. That is, a later lexicographic position is associated with a higher probability of missing analysts-related information, indicating a visibility disadvantage.

We are now ready for the regression analysis.

4. Main Findings

Reverting back to regression (3) as specified earlier, our parameter of interest is β, namely the coefficient of $Lexi\_position_{s,t}$. Our baseline results are reported in Table 4. Across the five columns, differences in specifications lie in the use of control variables. We incrementally add control variables but reach similar findings across columns. The sample in use is consistent across columns. The only exception is column (5), where the addition of "return 12 months ago" shrinks the sample size by approximately nine percent, since the newest stocks might not have been traded for 12 months by the end of the sample. To retain a large sample size, our later regressions will use column (4) in this table as the benchmark specification.

The results reported in Table 4 demonstrate that stocks with an earlier lexicographic position have higher turnover. The coefficients of control variables have expected signs. We have to be cautious when interpreting the magnitudes of the coefficients. Given that the dependent variable is in logarithm, the $\hat{\beta}$ in Table 4 suggests a quantitatively substantial impact of $Lexi\_position_{s,t}$ on stock turnover. It

[9] The specification of column (3) in Table 3 is the most similar to our main specification later, except that the dependent variable in column (3) is lexicographic position.

[10] The correlation patterns shown in Table 3 are similar to what Jacobs and Hillert (2016) find (Table 1 in their paper).

[11] The R-squared substantially rises after λ_{LV} is included, though that is driven mechanically by the inclusion of fixed effects rather than driven by improvements in linear fitting.

should be remembered that only within-lexicographic-value (LV) variations are used in regressions. A one standard deviation change in $Lexi\_position_{s,t}$ is between 0.037 and 0.048 (recall Table 1) depending on what weights are used. Take the 0.042 (simply weighted mean) and our preferred specification in column (4). A one standard deviation increase in $Lexi\_position_{s,t}$ (i.e. moving lexicographically later) leads to a decrease in turnover by 9.6 percent ($-2.280 \times 0.042 = -0.096$).[12]

We next slice the data in three different ways and rerun regression (3). We apply the specification in column (4) of Table 4 to all regressions in Table 5. In Panel A, the full sample is divided into three different sectors, where we find the strongest impact of lexicographic position in the service sector, and the weakest in the agriculture sector. This is consistent with the literature, as well as with our expectation that the service sector is the one where visibility matters the most.[13] In comparison, the agriculture sector has the weakest (though still statistically significant) lexicographic bias. In Panel B of Table 5, we divide observations according to analyst coverage. The high analysts coverage group, which are less obscure companies, show little lexicographic bias.

5. Different Measures of Lexicographic Bias

Other characters. In previous tables, the variable of interest is the lexicographic position of the initial (first) character in stock tickers. However, the presence of lexicographic bias may not be limited to the first character. We now construct the lexicographic positions of the second and third characters in stock tickers following the procedure described in Section 2.2, since lexicographically speaking, the stroke counts of later characters can be used to address tie cases when needed. Although the Chinese ticker is not the primary stock identifier in our context and thus tie cases need not be addressed, we remain interested in the primary source of the lexicographic bias. In columns (1) to (3) of Table 6, we first add each of the three lexicographic positions separately, each corresponding to the first, second, or third character in stock tickers. Lexicographic value (LV) fixed effects are not used here, in order to pool cross-LV and within-LV variations as an experiment. The previous lexicographic bias is not found. In fact, column (3) displays a slightly positive coefficient of the third-character lexicographic position.

In the rest of Table 6, we add second- and third-character lexicographic positions to supplement the initial-character lexicographic position. Specifically, we start without adding second- and third-character lexicographic positions. In column (4), we simply reproduce the results from the first-character lexicographic position (equivalent to column (4) in Table 4) for comparison. Next, we add second-

[12] This magnitude is comparable with the coefficient in Jacobs and Hillert (2016) (their Table 2), where the continuous lexicographic position (also ranging between 0 and 1) has coefficients around -0.10 that suggests a magnitude around 10 percent as well. Unlike in their case, where the lexicographic position can change fully across the range 0 to 1, our $Lexi\_position_{s,t}$ has been conditional on lexicographic value fixed effects such that its effective range in our study should be gauged using the magnitudes in our Table 1.

[13] Einav and Yariv (2006) note that the Los Angeles Yellow Pages have more than 450 listed businesses with names containing a redundant initial A.

Table 4: Turnover and Lexicographic Position

This table reports the coefficients and t-statistics (in parentheses) obtained from the panel regression discussed in the text. Dependent variable is logarithmized turnover. The explanatory variable of interest is the lexicographic position (between 0 and 1) in the current month. All regressions include LV (as explained in the text), industry, and year fixed effects. Standard errors are double-clustered by firm and month. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

	(1)	(2)	(3)	(4)	(5)
Lexicographic position	-2.353***	-2.353***	-2.285***	-2.283***	-2.481***
	(-6.059)	(-6.064)	(-6.040)	(-6.049)	(-6.073)
Positive return	0.169***	0.169***	0.164***	0.163***	0.147***
	(5.756)	(5.749)	(5.559)	(5.526)	(4.912)
Negative return	0.023	0.023	0.021	0.020	0.014
	(0.653)	(0.651)	(0.580)	(0.558)	(0.394)
Price	0.179***	0.180***	0.187***	0.183***	0.140***
	(6.810)	(6.809)	(6.762)	(6.670)	(5.034)
Return volatility	2.222***	2.222***	2.207***	2.206***	9.045***
	(8.496)	(8.496)	(8.498)	(8.504)	(6.479)
Market capitalization	-0.181***	-0.181***	-0.166***	-0.162***	-0.151***
	(-11.809)	(-11.760)	(-9.999)	(-9.828)	(-9.007)
Leverage	0.001**	0.001**	0.001**	0.001**	0.001**
	(2.030)	(2.029)	(2.287)	(2.296)	(2.257)
Sales	-0.016**	-0.016**	-0.019**	-0.018**	-0.014*
	(-2.025)	(-2.019)	(-2.480)	(-2.409)	(-1.966)
Age	-0.096***	-0.096***	-0.094***	-0.092***	-0.023
	(-6.041)	(-6.020)	(-6.077)	(-5.991)	(-1.314)
Book to market ratio	-0.106***	-0.105***	-0.103***	-0.103***	-0.094***
	(-7.467)	(-7.426)	(-7.333)	(-7.297)	(-6.966)
Numeric ID position	0.000***	0.000***	0.000***	0.000***	0.000***
	(5.240)	(5.235)	(5.183)	(5.064)	(5.209)
Advertisement		-0.013	-0.007	-0.004	-0.008
		(-1.067)	(-0.540)	(-0.340)	(-0.605)
Missing advertisement dum.		-0.211	-0.112	-0.071	-0.114
		(-1.187)	(-0.643)	(-0.405)	(-0.646)
No. of analysts			-0.066***	-0.064***	-0.067***
			(-7.177)	(-7.053)	(-7.517)
Missing No. of analysts dum.			-0.274***	-0.270***	-0.256***
			(-13.230)	(-13.095)	(-13.196)
R&D				-0.054***	-0.046***
				(-3.803)	(-3.352)
Missing R&D dummy				-0.986***	-0.833***
				(-4.097)	(-3.560)
Return 12 months ago					0.028**
					(2.475)
Observations	336,774	336,774	336,774	336,774	308,395
R-squared	0.355	0.355	0.360	0.360	0.380

and third-character lexicographic positions separately. The coefficient of the second-character lexicographic position is negative and statistically significant, with a smaller magnitude compared with the coefficient of the first-character lexicographic position. The coefficient of the third-character lexicographic position is statistically insignificant.

Table 5: Turnover and Lexicographic Position, Various Subsamples

This table reports the coefficients and t-statistics (in parentheses) obtained from various subsamples. The observations in Table 4 (full sample) are now differently grouped in the three panels; otherwise the regressions here follow the specifications used in column (4), Table 4. In Panel A, observatitons are grouped into three sectors (agriculture, manufacturing, and service) according to their firm-level industry codes. In Panel B, observations are divided into visible and less visible groups using the median of analyst coverage. In both panels, dependent variable is logarithmized turnover. Coefficients of control variables are not reported to save space. All regressions include LV, industry, and year fixed effects. Standard errors are double-clustered by firm and month. Statistical significance at the 10% and 1% level are indicated by * and ***, respectively.

	Panel A: By sector		
	Agriculture	Manufacturing	Service
Lexicographic position	-3.147*	-1.556***	-2.698***
	(-1.893)	(-3.166)	(-5.171)
Observations	15,932	183,241	137,600
R-squared	0.411	0.362	0.341
	Panel B: By analyst coverage		
	Visible	Less visible	
Lexicographic position	-0.991	-2.203***	
	(-0.848)	(-6.711)	
Observations	119,745	217,029	
R-squared	0.398	0.406	
By aggregation	-2.425***	-2.211***	
	(-4.659)	(-7.028)	
Observations	149,084	187,689	
R-squared	0.379	0.407	

Table 6: With Other Characters Considered

This table reports the coefficients and t-statistics (in parentheses) when other characters in stock tickers are considered. The regressions here follow the specifications used in column (4), Table 4, except introducing the additional position variables. For comparison, the coefficient in column (4) here is highlighted. It is the same as the coefficient in column (4) of Table 4 since the two columns have the same specificatiton and regressors. Coefficients of control variables are not reported to save space. Both columns include industry and year fixed effects. Standard errors are double-clustered by firm and month. Statistical significance at the 5% and 1% level are indicated by ** and ***, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Lexi. position (initial character)	-0.026			-2.283***	-2.317***	-2.281***
	(-0.644)			(-6.049)	(-6.516)	(-6.157)
Lexi. position (second character)		0.008			-1.194***	
		(0.213)			(-3.963)	
Lexi. position (third character)			0.099**			-0.019
			(2.365)			(-0.377)
Industry and year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Corresponding LV fixed effects	No	No	No	Yes	Yes	Yes
Observations	336,774	336,774	336,774	336,774	336,774	336,774
R-squared	0.355	0.355	0.356	0.360	0.363	0.363

The results in Table 6 have two implications. First, the lexicographic bias also exists, but with a small magnitude, in later characters of stock tickers. We do not place emphasis on these characters in this study, because stock tickers have different character lengths (up to four characters) which complicates the comparison. Second, the lexicographic bias is mainly driven by within-lexicographic-value (LV) variations. When within-LV variations and cross-LV variations are pooled, lexicographic bias does not appear (recall column (1) in Table 6). Notice that this does not necessarily imply that cross-LV variations do not generate lexicographic bias. Rather, econometrically, the omitted LV fixed effects are correlated with the regressors in unknown ways and thus lead to estimates biased towards unknown directions.

Cross-LV variations. We next examine the relationship between stock turnover and lexicographic value (LV) cross-sectionally. The sample used in previous tables cannot be directly used here, because lexicographic value is more aggregated than the firm level. Notice that lexicographic position is a firm-month level characteristic and our sample is at the firm-month level. If we directly used the firm-month level data and replaced the firm-month level lexicographic position with firm-level lexicographic value, the standard errors of coefficients would be downward biased (correspondingly, the significance level would be upward biased). We turn to the following experiment to investigate how lexicographic values impact stock turnover. We take the averages of all variables at the LV-industry-year level, and use the averaged variables to run the regression.[14] Our experiment uses the log values of the variable means. That is, we artificially create "averaged firms" with lexicographic values of 1, 2, 3, and so on. The results are reported in Table 7, where the set of control variables is the same as in earlier tables but excludes those that are "unaverageable."[15] The results show evidence of a lexicographic bias. This lexicographic bias, as we intended, comes solely from distinct lexicographic values. The coefficients of the control variables have similar signs as in Table 4.

6. Lexicographic Bias and Ticker Change

We now turn to ticker changes. Remember that the variations in lexicographic position featured in previous sections are due to peer effects, namely how the lexicographic position *relative to other stocks in the same market* makes a difference. Here we continue this focus. The ticker change effect that interests us now is how switching to a later lexicographic position affects a stock's turnover. Consider a stock that switches to a ticker with a greater lexicographic value. This will bring the stock lexicographically later, and meanwhile bring the rest of stocks in the same market lexicographically earlier. The lexico-

[14] This is different from a between-estimation as opposed to the previous within-estimation (fixed-effect estimation), because many variables in the regression need to be in logarithmic terms and the mean of a logged variable X is unequal to the log value of the mean of variable X (Jensen's inequality).

[15] Remember that the control variables related to advertisement, analysts and R&D have a large number of missing values. Our solution in the previous tables was to treat missing values as zero and add corresponding missing variable dummies. That solution does not work here when all variables are averaged.

Table 7: Averaged Firms

This table reports the coefficients and t-statistics (in parentheses) obtained from a variant of the panel regression in Table 4. For each market-year pair, we average firms with the same lexicographic value (LV) into one firm. Dependent variable is the logarithm of the average-firm turnover. The explanatory variable of interest is the lexicographic value. All other variables are averaged in the same fashion. Standard errors are double clustered at the LV-industry pair and month. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Lexicographic value (LV)	-0.012**	-0.008**
	(-2.181)	(-2.107)
Positive return		0.597***
		(3.415)
Negative return		0.384**
		(2.160)
Price		0.159*
		(1.955)
Return volatility		0.392***
		(8.561)
Market capitalization		-0.041
		(-0.772)
Leverage		0.008**
		(2.438)
Sales		-0.024
		(-0.632)
Age		0.025
		(0.292)
Book to market ratio		-0.139***
		(-3.217)
Numeric ID position		-0.000
		(-1.406)
Observations	16,604	13,351
R-squared	0.595	0.622

graphic change caused by this stock is voluntary while the resulting change for other stocks in the same market is involuntary. Between the two types of changes, voluntary changes are arguably endogenous, because the decision to change tickers is associated with other factors that may affect stock performances (e.g. company name changes as a result of ownership changes). In comparison, involuntary changes are largely exogenous because they occur for stocks who do not change their tickers.

Our identification strategy is as follows. We first construct a dummy variable $DLLMn_{s,t}$ that equals 1 if the current month t is the n-th month after stock s switches to a lexicographically later ticker. $DLLMn$ is an abbreviation of "dummy of n months after becoming lexicographically later." Then we add this dummy variable to the previous regression (3). Notice that the current lexicographic position, namely $Lexi\_position_{s,t}$, remains in the regression such that the lexicographic bias we analyzed earlier is still captured by the coefficient of the variable. The newly introduced dummy variable $DLLMn_{s,t}$ aims to capture the additional penalty — if any — caused by moving to an later lexicographical position (either voluntarily or involuntarily).

The design of the above identification strategy aims to separate two lexicographic effects from each other. When a stock's lexicographic position moves later (either voluntarily or involuntarily), it will receive less visibility as opposed to its peers in the same market. This is the lexicographic effect captured by $Lexi\_position_{s,t}$ just as in earlier sections. But remember that this stock, compared with its own past, is hypothesized to lose visibility, and this penalty, if existing, will be captured by the newly introduced $DLLMn_{s,t}$. Voluntary lexicographic changes are not excluded here. Rather, we add another control variable $DECHG_s$ (an abbreviation of "dummy of ever change") to denote if stock s ever changes its ticker lexicographically later.

The results are reported in Panel A of Table 8. Evidently, there is a visibility penalty associated with taking a later lexicographic position. This effect attenuates over time (disappears in the fourth month after the change). The dummy variable that denotes whether a firm has ever changed its ticker is statistically insignificant. These regressions otherwise follow the previous benchmark specification (Column (4), Table 4). The coefficient of the current lexicographic position remains as before.

Since lexicographic position is relative, the coefficients of $DLLMn_{s,t}$, $n = 1,2,3,4$, captures the turnover decrease of stocks whose lexicographic positions become later, relative to both (a) their own past and (b) stocks whose lexicographic positions become earlier. We next exclude group (b) by dropping stocks whose lexicographic positions become earlier. That is, we include only firm-month pairs where firms change to a later position. Notice that the sample size shrinks only by 8.4 percent (from 336,774 to 308,414), indicating that only a small portion of firms move lexicographically earlier. The results are reported in Panel B of Table 8. As expected, the coefficients of $DLLMn_{s,t}$, $n = 1,2,3,4$, shrink in both magnitude and significance level since the contrast is toned down when group (b) is excluded. Though, the findings from Panel A still hold here in Panel B.

7. Other Financial Measures

We next experiment with other financial measures and specifications. The results are reported in Table 9, where the regressions follow column (4) in Table 4 except the differences noted in the column headings. We start with the same stock turnover as before but this time use its level rather than logarithm. Then, we experiment with the illiquidity measure (proposed by Amihud (2002)), log share volume, log dollar volume, fraction of zero return days, and log high-low spread (proposed by Corwin and Schultz (2012)). Then, we rerun our benchmark specification without stocks having bottom 5% turnover to address possible concerns over outliers. Last, we rerun the regression using the Fama-MacBeth method (as in Table 3). Evidently, stocks with an earlier lexicographic position have higher turnovers, higher share volumes, higher dollar volumes, lower illiquidity, greater spreads, and more trading activities overall.

Table 8: Lexicographic Position Changes Over Time

This table reports the coefficients and t-statistics (in parentheses) when ticker changes are considered. Dependent variable is logarithmized turnover. Lexicographic position at the bottom of each panel is as defined in the text. In Panel A, the full sample is used and the variables/coefficients of interest are reported. In Panel B, we use the subsample of firm-month pairs that are after firms change their tickers lexicographically later. Otherwise, the regressions in both panels follow the specifications used in column (4), Table 4. Coefficients of control variables are not reported to save space. All regressions include LV, industry, and year fixed effects. Standard errors are double-clustered by firm and month. Statistical significance at the 10%, 5% and 1% level is indicated by *,** and ***, respectively.

Panel A: Full sample			
DLLM1	-0.071***	-0.077***	-0.080***
(dummy: 1st month after taking a later lexicographic position)	(-4.531)	(-4.683)	(-4.749)
DLLM2	-0.067***	-0.073***	-0.077***
(dummy: 2nd month after taking a later lexicographic position)	(-4.208)	(-4.317)	(-4.346)
DLLM3		-0.038*	-0.042*
(dummy: 3rd month after taking a later lexicographic position)		(-1.784)	(-1.871)
DLLM4			-0.031
(dummy: 4th month after taking a later lexicographic position)			(-1.298)
DVCHG	0.091	0.094	0.097
(dummy: firm voluntarily changes to a later lexicographic position)	(0.886)	(0.917)	(0.946)
Lexicographic position	-2.317***	-2.340***	-2.352***
	(-6.245)	(-6.306)	(-6.324)
Observations	336,774	336,774	336,774
R-squared	0.362	0.362	0.362
Panel B: Subsample (firm-month pairs after taking lexicographically later positions)			
DLLM1	-0.030*	-0.030*	-0.031*
(dummy: 1st month after taking a later lexicographic position)	(-1.960)	(-1.908)	(-1.878)
DLLM2	-0.032**	-0.032*	-0.033*
(dummy: 2nd month after taking a later lexicographic position)	(-2.005)	(-1.914)	(-1.853)
DLLM3		-0.004	-0.004
(dummy: 3rd month after taking a later lexicographic position)		(-0.171)	(-0.190)
DLLM4			-0.004
(dummy: 4th month after taking a later lexicographic position)			(-0.188)
DECHG	0.094	0.094	0.095
(dummy: firm voluntarily changes to a later lexicographic position)	(0.925)	(0.928)	(0.932)
Lexicographic position	-2.238***	-2.240***	-2.242***
	(-5.272)	(-5.269)	(-5.261)
Observations	308,414	308,414	308,414
R-squared	0.361	0.361	0.361

8. Concluding Remarks

One might assume that stock tickers should not matter. This efficient-market view is countered by our findings. We find that stocks occurring earlier in lexicographic order are traded more frequently in China. Unlike alphabetic biases in the Western world, Chinese lexicography does not construct a strict ordering system. Stocks in China have numeric IDs as their primary identifiers. Also, our identification strategy was designed to address endogenous ticker choices. Nevertheless, lexicographic bias exists in the Chinese stock market, and the effect is persistent. In this study, we assess its magnitude, primary source, varying severity across sectors and visibility groups, changes resulting from ticker changes and

Table 9: Other Financial Measures and Specifications

This table reports the coefficients and t-statistics (in parentheses). Dependent variables are noted in the headings of columns. Otherwise the regressions follow the specifications used in column (4), Table 4. Coefficients of control variables are not reported to save space. All regressions include LV, sector, and year fixed effects. Standard errors are double-clustered by firm and month. Statistical significance at the 5% and 1% level is indicated by ** and ***, respectively.

	(1) Turnover level	(2) Amihud illiquidity	(3) Share volume (log)	(4) Value volume (log)
Lexicographic position	-2.281***	3.291***	-3.202***	-3.888***
	(-6.046)	(8.305)	(-8.386)	(-6.278)
Observations	336,774	332,087	332,087	332,087
R-squared	0.360	0.734	0.701	0.716
	(5) Corwin & Schultz spread	(6) Fraction of zero return days	(7) Turnover (bottom 5% excluded)	(8) Fama/MacBeth regression
Lexicographic position	0.012**	0.062***	-2.182***	-0.154***
	(2.288)	(5.388)	(-5.939)	(-3.660)
Observations	332,087	332,087	319,655	336,774
R-squared	0.395	0.139	0.364	0.415

their attenuation over time.

The Chinese stock market, because of its tremendous size and lack of transparency, is worth examining under the scope of informational frictions in financial investments. Abnormalities found in Western markets may not hold in China because cultural and regulative institutions of the West are intertwined with its stock market performances. In this case, the Western alphabetic bias finds its counterpart in China, despite the distinct languages and lexicographies involved. Possible avenues for future research include examining how Chinese linguistics, which is rich in name stereotypes, fluency differences, verbal ambiguities, and cognitive complexities, relate to financial market performances.

References

Alter, Adam L. and Daniel M. Oppenheimer (2006), "Predicting short-term stock fluctuations by using processing fluency." *Proceedings of the National Academy of Sciences of the United States of America*, 103, 9369–9372.

Amihud, Yakov (2002), "Illiquidity and stock returns: cross-section and time-series effects." *Journal of Financial Markets*, 5, 31–56.

Anderson, Alyssa G. and Yelena Larkin (2019), "Does noninformative text affect investor behavior?" *Financial Management*, 48, 257–289.

Chordia, Tarun, Sahn-Wook Huh, and Avanidhar Subrahmanyam (2007), "The cross-section of expected trading activity." *Review of financial studies*, 20, 709–740.

Cooper, Michael J. (2001), "A Rose.com by Any Other Name." *Journal of Finance*, 56, 2371–2388.

Cooper, Michael J., Huseyin Gulen, and P. Raghavendra Rau (2005), "Changing Names with Style: Mutual Fund Name Changes and Their Effects on Fund Flows." *Journal of Finance*, 60, 2825–2858.

Corwin, Shane A. and Paul Schultz (2012), "A simple way to estimate bid-ask spreads from daily high and low prices." *Journal of Finance*, 67, 719–760.

Da, Jun (2004), "Xiandai hanyu changyong zibiao: Individual character frequencies based on statistics from the modern chinese corpus." Technical report, http://lingua.mtsu.edu/chinese-computing/statistics/char/listchangyong.php.

Durham, Greg and Mukunthan Santhanakrishnan (2016), "Ticker fluency, sentiment, and asset valuation." *The Quarterly Review of Economics and Finance*, 61, 89–96.

Einav, Liran and Leeat Yariv (2006), "What's in a Surname? The Effects of Surname Initials on Academic Success." *Journal of Economic Perspectives*, 20, 175–187.

Fama, Eugene and James D MacBeth (1973), "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy*, 81, 607–36.

Green, T. Clifton and Russell Jame (2013), "Company name fluency, investor recognition, and firm value." *Journal of Financial Economics*, 109, 813–834.

Itzkowitz, Jennifer and Jesse Itzkowitz (2017), "Name-Based Behavioral Biases: Are Expert Investors Immune?" *Journal of Behavioral Finance*, 18, 180–188.

Itzkowitz, Jennifer, Jesse Itzkowitz, and Scott Rothbort (2016), "ABCs of Trading: Behavioral Biases affect Stock Turnover and Value." *Review of Finance*, 20, 663–692.

Jacobs, Heiko and Alexander Hillert (2016), "Alphabetic Bias, Investor Recognition, and Trading Behavior." *Review of Finance*, 20, 693–723.

Krishnamurthy, Srinivasan, Denis Pelletier, and Richard S. Warr (2018), "Inflation and equity mutual fund flows." *Journal of Financial Markets*, 37, 52–69.

Meer, Jonathan and Harvey S. Rosen (2011), "The ABCs of charitable solicitation." *Journal of Public Economics*, 95, 363–371.

Ray, Debraj and Arthur Robson (2018), "Certified random: A new order for co-authorship." *American Economic Review*, 108, 489–520.

Table A1
Description of Variables

Variable	Description
Turnover	Number of shares traded divided by the number of shares outstanding.
Positive return	A dummy variable that equals 1 if stock return in the previous month was positive and 0 otherwise.
Negative return	A dummy variable that equals 1 if stock return in the previous month was negative and 0 otherwise.
Price	Nominal share price at the end of each month. We add one to the price before taking logarithm.
Return volatility	We calculate standard deviation of daily returns over each month. Return volatility equals the logarithm of the standard deviation plus one.
Market capitalization	The logarithm of market capitalization at the end of month.
Sales	The logarithm of total sale by the end of year.
Age	The logarithm of number of years since birth plus one.
Leverage	Book debt scaled by equity.
Advertisement	The logarithm of advertisement expenditure by the end of year. We have a separate dummy variable for missing advertisement expenditure.
No. of analysts	CSMAR provides names of analysts who provide reports for a stock by the end of each year. We count the names for each stock and define number of analyst as the logarithm of the count plus one. We have a separate dummy variable for missing analyst information.
R&D	The logarithm of Research and Development expenditure plus one. We have a separate dummy variable for missing R&D expenditure.
Book to market ratio	Total asset divided by market capitalization. Total asset and market capitalization are both calculated at the end of year. If these values are missing at the end of year, we use the value at the end of September, June, or March instead (that is, use the end-of-June value if the end-of-September value is unavailable, and use the end-of-March value if the end-of-June value is unavailable).