Behav Res (2018) 50:1763–1777
DOI 10.3758/s13428-017-0944-0

MELD-SCH: A megastudy of lexical decision in simplified Chinese

Yiu-Kei Tsang[1,2] · Jian Huang[3] · Ming Lui[1,2] · Mingfeng Xue[3] · Yin-Wah Fiona Chan[4] · Suiping Wang[3] · Hsuan-Chih Chen[5]

Published online: 4 August 2017

Abstract Here we report on MELD-SCH (MEgastudy of Lexical Decision in Simplified CHinese), a dataset that contains the lexical decision data of 1,020 one-character, 10,022 two-character, 949 three-character, and 587 four-character simplified Chinese words obtained from 504 native Chinese users. It also includes a number of word-level and character-level variables. Analyses showed that the reliability of the dataset is satisfactory, as indicated by split-half correlations and comparisons with other datasets. Item-based regression showed that both word-level and character-level variables contributed significantly to the reaction times and error rates of lexical decision. Moreover, we discovered a U-shape relationship between word-length and reaction times, which has not been reported in Chinese before. MELD-SCH can facilitate research in Chinese word recognition by providing high quality normative data and information of different linguistic variables. It also encourages researchers to extend their empirical findings, which are mostly based on one-character and two-character words, to words of different lengths.

Keywords Megastudy · Chinese · Word recognition · Lexical decision

Electronic supplementary material The online version of this article (doi:10.3758/s13428-017-0944-0) contains supplementary material, which is available to authorized users.

✉ Yiu-Kei Tsang
yktsang@hkbu.edu.hk

[1] Department of Education Studies, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong

[2] Centre for Learning Sciences, Hong Kong Baptist University, Kowloon, Hong Kong

[3] School of Psychology, South China Normal University, Guangzhou, People's Republic of China

[4] The Psychometrics Centre, University of Cambridge, Cambridge, UK

[5] Department of Psychology, The Chinese University of Hong Kong, Shatin, Hong Kong

Introduction

Language is core to human cognition. It is the primary medium of human thinking and communication. Given its significance, psychologists have made tremendous effort to understand the mechanisms behind language processing. Most previous studies have adopted the experimental approach, in which researchers manipulated one or two variables of interest and controlled other variables as much as they could. Although the experimental approach has generated a lot of important findings about language processing, its limitations are also obvious. First, because many lexical variables are correlated, it is difficult to select stimuli that differ only on one or two dimensions but are matched in all other aspects. Uncontrolled variables may undermine the validity of the conclusions. Second, researchers may end up having only a small, restricted set of stimuli that fulfil their selection criteria, which limits the generalizability of their results. Third, continuous lexical variables (e.g., word frequency) are often dichotomized in experiments (e.g., high vs. low frequency), which obscure any non-linear effects and may lead to spurious statistical results (Maxwell & Delaney, 1993).

To solve these problems, researchers began to employ a new approach called "megastudy" to investigate language processing, and, specifically, word recognition (Balota, Yap, Hutchison, & Cortese, 2012). In a typical megastudy, researchers obtain lexical decision or naming data of thousands or tens of thousands of words. The large amount of stimuli guarantees better generalizability, and make the examination

of the effects of many continuous lexical variables simultaneously with regression-based analyses possible. These allow megastudies to provide complementary data to experiments. Given its huge potential, the megastudy approach has developed rapidly in the last decade. Since the seminal work of the (American) English Lexicon Project (ELP) by Balota et al. (2007), megastudies of word recognition are now available in British English (Keuleers, Lacey, Rastle, & Brysbaert, 2012), Dutch (Keuleers, Diependaele, & Brysbaert, 2010), French (Ferrand et al., 2010), and Malay (Yap, Rickard Liow, Jalil, & Faizal, 2010). Recently, it has also been extended to simplified Chinese (one-character words: Liu, Shu, & Li, 2007; Sze, Rickard Liow, & Yap, 2014) and traditional Chinese (one-character and two-character words: Chang, Hsu, Tsai, Chen, & Lee, 2016; Lee, Hsu, Chang, Chen, & Chao, 2015; Tse et al., in press).

In this article, we report the newly developed MELD-SCH (MEgastudy of Lexical Decision-Simplified CHinese), a dataset that contains various linguistic variables and the lexical decision data of 12,578 simplified Chinese words of different word lengths (from one-character to four-character) obtained from 504 native Mandarin Chinese participants. We first introduce the unique features of written Chinese that make it an interesting case to compare with alphabetic languages. Second, we review existing megastudies in Chinese and introduce two unanswered questions. Third, we describe the methods in obtaining the data in MELD-SCH. In the Result section, we compare MELD-SCH with previous megastudies in Chinese to establish its reliability. We also present item-based regression analyses that aim to answer the two unanswered questions. Finally, we discuss how MELD-SCH can contribute to the understanding of Chinese language processing.

Features of written Chinese

Adopting the logographic system, written word recognition in Chinese may differ from the more widely-studied alphabetic languages in significant aspects (Hoosain, 1991; Tsang & Chen, 2012). Specifically, there are a number of unique features in Chinese script. For example, characters, rather than letters, are the building blocks of Chinese words. While the different letters in alphabetic languages have a similar level of visual complexity, Chinese characters can be as simple as "乙", or as complex as "癵". Nevertheless, no matter how complex the character is, all its visual information is packed within the same box-shaped area, which results in a relatively high information density, as compared to alphabetic scripts. In addition, over 70% of the Chinese characters are phonograms (Lee et al., 2015). These characters are constructed by combining sublexical (sub-character) units called "semantic radicals" and "phonetic radicals." As their names imply, the semantic and phonetic radicals provide cues to the meanings and pronunciations of the whole-characters, respectively. For example, the semantic radical "火" (/*huǒ*/, fire) and the phonetic radical "考" (/*kǎo*/, to examine) combine to form the character "烤" (/*kǎo*/, to roast). Although the cues provided are unreliable (e.g., the phonetic radicals provide valid cues in at most 50% of the characters in Mandarin Chinese; Hsiao & Shillcock, 2006), native Chinese users often rely on the semantic and phonetic radicals when reading unknown or newly learned characters. For example, Shu and Anderson (1997) showed that third and fifth graders were able to infer the meaning of the unfamiliar character "眺" (/tiào/, "to look at") based on the semantic radical "目" (/mù/, eye). Similarly, they often pronounced the unfamiliar character "琼" (/*qióng*/, "jade") incorrectly as/*jīng*/, based on the pronunciation of the phonetic radical "京" (/jīng/, the capital of a country). A similar functional distinction of sublexical units is not found in most alphabetic languages (Tsang, Wu, Ng, & Chen, 2017).

In most cases, each Chinese character corresponds to both a syllable and a morpheme. Therefore, Chinese is also regarded as a morphosyllabic language. While word recognition in alphabetic languages is generally assumed to rely heavily on phonological encoding, it has been under debate whether the same assumption is applicable to Chinese, or instead is a direct mapping between the orthography and semantics the key (Chen & Shu, 2001; Perfetti & Tan, 1998; Sze, Yap, & Rickard Liow, 2015)? Phonology may play a less important role in Chinese word recognition because the form-sound correspondence is highly opaque. First, as aforementioned, phonetic radicals only provide unreliable cues to character pronunciations. Second, even when the phonetic radials are reliable, their pronunciations are learned by rote. Therefore, the retrieval of pronunciations in Chinese through phonetic radicals is much more resource-demanding than applying the letter-sound rules in alphabetic languages. Third, there are many homophonic characters in Chinese (around 1,300 possible Mandarin syllable-plus-tone combinations and over 5,000 characters). Accessing meanings through character pronunciations is thus highly error-prone. Indeed, developmental research has shown that phonological awareness plays a less important role in reading acquisition in Chinese than in English or Korean (McBride-Chang et al., 2005).

At the same time, many characters are associated with more than one meaning, which results in an ambiguous character-morpheme mapping. For example, the character "花" (/*huā*/) can mean "flower" or "to spend." The exact meanings of the characters can be identified only by considering the context. For example, the "花" in "花瓶" (/*huāpíng*/, vase) means "flower," while that in "花费" (/*huā*fèi/, expense) means "to spend." Although a similar phenomenon also exists in English and other languages (e.g., "-er" in "teacher" and "taller"), it is much more prominent in Chinese (about 50% of all characters have two or more distinctive meanings; Liu et al., 2007). This leads to a high degree of context dependence in Chinese,

which has significant impact on word recognition. In a series of experiments, Tsang and Chen (2010, 2013a, b) showed that it takes about 200 ms to resolve the character-morpheme ambiguity, such that the appropriate meanings can be selected for integration and comprehension. In other words, ambiguous characters may make word recognition more difficult.

Megastudies in Chinese word recognition

Naming aloud and lexical decision (i.e., participants decide whether the presented stimuli are existing lexical items) are two popular tasks for investigating word recognition. Although the two tasks differ in response modality, it has been shown that they rely on similar word recognition and lexical retrieval processes (Carreiras, Mechelli, Estévez, & Price, 2007). There are several existing megastudies in Chinese word recognition that adopt these tasks. To the best of our knowledge, the first published one was a character-naming dataset developed by Liu et al. (2007), who collected the naming data of 2,423 one-character words in simplified Chinese from 480 native Mandarin speakers. Their primary goal was to examine the effects of 15 variables, including word frequency, number of strokes, homophone density, number of meanings, etc., on naming latency. Two sets of multiple regression analyses were run. In the first analysis, the 15 variables were entered directly into the regression model. Ten variables had significant effects on the naming latency, among which "age of acquisition" (i.e., the age when a typical Chinese user has learned the meaning and pronunciation of a character) and "cumulative frequency" (i.e., the sum of frequencies of all words that contain the character as a constituent morpheme) had the strongest impact. Together, the ten variables alone accounted for 55.8% of the variance.

In the second analysis, the variables were grouped into five factors by principal component analysis before entering into the regression model. This helped avoid multi-collinearity and improve the clarity of interpretation. Results showed that frequency, visual complexity, and semantic factors all contributed to naming latency, but the effect of phonology was non-significant. Specifically, the phonology component consisted of homophone density (i.e., number of characters that share the same syllable) and phonological frequency. Their combined effect was non-significant in accounting for naming latency in the second analysis. This stands in sharp contrast to the findings in alphabetic languages, in which various phonological factors (e.g., phonological neighbors, syllable frequency, etc.) had a strong role to play in determining word naming latency (Perea & Carreiras, 1998; Yap & Balota, 2009).

Sze et al. (2014) developed the Chinese Lexicon Project (CLP), which includes lexical decision (character decision) data for 2,500 one-character words in simplified Chinese from 35 highly proficient native Mandarin speakers. The pseudo-characters used in the lexical decision task were created by randomly switching the semantic radicals of the real characters to form structurally legal but non-existing combinations. Based on the data in CLP, Sze et al. (2015) showed that orthographic (e.g., number of strokes) and semantic variables (e.g., number of meanings) both contributed strongly to lexical decision latency, but the effects of most phonological variables (e.g., the sum of frequencies of all homophonic characters) were not significant. The only phonological variable that had a significant impact was feedforward consistency (i.e., whether characters that share the same phonetic radicals have the same pronunciation). Responses were faster for consistent characters. Overall, both the orthographic and semantic factors could account for more unique variance in the lexical decision latency than the phonological factors. The lexical decision results are thus consistent with the naming data by Liu et al. (2007) in suggesting that Chinese readers may rely more on orthographic and semantic information than on phonological information in Chinese word recognition.

There are also megastudies in traditional Chinese. For example, Chang et al. (2016) collected the naming data of 3,314 phonograms. As in Sze et al. (2014), they showed that while feedforward consistency significantly affected naming latency, homophone density had no effects. At the same time, the effects of orthographic and semantic variables (e.g., frequency, number of meanings, etc.) were significant. Similar results were obtained in a lexical decision megastudy of 3,423 one-character words developed by the same group (Lee et al., 2015). The authors were interested in separating the effects of feedforward consistency and feedback consistency (i.e., whether homophonic characters share the same phonetic radicals). For example, most characters that are pronounced as/mǎ/contained the phonetic radical "马" (e.g., "玛", "蚂", and "码") and they are thus having high feedback consistency. On the other hand, characters that contained the phonetic radical "马" can also take other pronunciations (e.g., "冯"/féng/) and the feedforward consistency is low. When both consistency measures were entered into a regression model, only the effect of feedback consistency remained significant, indicating that feedback consistency is a more important phonological consistency measure than feedforward consistency. Again, homophone density had no significant effect, while word frequency, number of strokes, and number of meanings continued to exert strong impact on lexical decision latency.

These megastudies include only one-character Chinese words, which is understandable because characters are the building units of Chinese words. In addition, given that the number of characters is more restricted (around 5,000 commonly used characters), many linguistic variables can be defined more easily at the character level. On the other hand, Tse et al. (in press) recently extended the CLP (we refer to it as CLP-Tse) by including lexical decision data of 25,286 two-character traditional Chinese words and pseudo-words

(constructed by randomly recombining the characters in words to form orthographically legal but non-existing combinations) obtained from 594 native Chinese users in Hong Kong. This could be a huge improvement because one-character words account for less than 10% of all Chinese word types, while two-character words account for around 60%. Besides, two-character words are arguably the most popular materials in experimental studies of Chinese language processing.[1] CLP-Tse provided valuable complementary data to previous experiments.

Based on their megastudy data, Tse et al. (in press) conducted several virtual experiments and successfully replicated some controversial findings of Chinese word recognition. For example, the interaction between character frequency and semantic transparency observed in Peng, Liu, and Wang (1999) was replicated, such that character frequency had a facilitative effect on transparent words, but its effect on opaque words was non-significant. Similarly, the authors replicated the findings of Leong and Cheng (2003), who showed that two-character words that contained phonologically inconsistent second characters (i.e., the characters have different pronunciations in different words; e.g., "躲藏-to hide" is pronounced as/*duǒcáng*/, but "西藏-Tibet" was pronounced as/*Xīzàng*/) were responded to faster than words that contained phonologically consistent characters.

Tse et al. (in press) also compared several Chinese word frequency measures and showed that the subtitle corpus developed by Cai and Brysbaert (2010) outperformed other larger corpora. In particular, the best frequency measure, namely the context diversity (i.e., the number of movies and television shows a word appears in), accounts for 34.72% and 17.36% of the variance of lexical decision latency and accuracy, respectively. Moreover, Tse et al. collected semantic transparency data by asking participants to rate the strength of the relationship between the meaning of each character and the whole word. They showed that these semantic transparency ratings could explain extra variance of lexical decision performance after the effects of other more well-established factors had been accounted for (e.g., number of strokes, frequency of each character, and whole words).

Goal of our megastudy

Tse et al. (in press) have extended the megastudy approach to two-character words in traditional Chinese. However, the same character can be quite dissimilar visually in different Chinese scripts. For example, the character for "head" is "头" in simplified Chinese but "頭" in traditional Chinese. Readers familiar with one script would sometimes find it difficult to recognize characters in the other script. Therefore, megastudies in traditional Chinese may not be directly applicable to users of simplified Chinese. Given that simplified Chinese is used by the majority of Chinese users (mainland China, Singapore, and many L2 learners), it is desirable to develop a megastudy for simplified Chinese words.

We also intend to use this megastudy to investigate two questions that remain unresolved in previous experiments and megastudies. First, although most Chinese words consist of one or two characters, about 20% of the words are three-character or four-character in length. Yet, the effect of word length has rarely been examined directly in previous experiments. And none of the existing megastudies in Chinese have incorporated three-character and four-character words. Indeed, researchers have generally manipulated the number of strokes instead of word length to examine the effect of visual complexity in Chinese word recognition. However, our understanding of Chinese word recognition would be incomplete without considering the longer words. To obtain empirical data regarding the word length effect in Chinese and to facilitate future research on the issue, we used words of one-character to four-character long in this megastudy.

Longer words usually have more visual features and should require more time for recognition. This received preliminary support from an eye-tracking study that examined the size of perceptual span in Chinese reading (Inhoff & Liu, 1998). Although the word length effect was not their primary focus, the study did provide data about the reading times for one-character, two-character, and three-/four-character words. During natural reading (i.e., no restriction on parafoveal preview), the gaze durations for one-character and two-character words were around 200 ms, while those for three-/four-character words were around 400 ms. However, the Chinese perceptual span size is about four characters and Chinese readers can obtain semantic information in the parafovea (Yan, Richter, Shu, & Kliegl, 2009). Therefore, the apparent word length effect could also be explained by a more semantic parafoveal preview for shorter words, which reduced the processing time when the words were actually fixated. Indeed, Su and Samuels (2010) only observed a non-significant trend of slower responses to four-character words as compared to two-character or three-character words during isolated word recognition. In this megastudy, the word length effect was tested in the lexical decision task. Regression analyses also allowed us to isolate the impact of word length after the effects of other variables (e.g., word frequency and number of strokes) were accounted for.

Second, it has been under debate whether Chinese words are recognized holistically or through combining the constituent characters (Tsang & Chen, 2012). Given the prevalence

[1] An informal survey on PsycINFO on 9 June 2017 with the searching keywords "Chinese one-character words" and "Chinese single-character words" returned nine and 39 entries, respectively, as compared to 103 entries for "Chinese two-character words." "Chinese three-character words" and "Chinese four-character words" returned six and nine entries, respectively. When the keywords were "Chinese character recognition" and "Chinese word recognition," the numbers of entries were 703 and 961, respectively.

of homophones and meaning ambiguity at the character level (as reviewed above), Packard (1999) argued that Chinese word recognition should be holistic and bypass the constituent characters because a time-consuming and resource-demanding disambiguation process is needed before individual characters can be integrated for word recognition. Such a process is too inefficient to explain the high performance in daily language task. Consistent with Packard's proposal, Bai, Yan, Liversedge, Zang, and Rayner (2008) showed that inserting spaces between characters disrupted the perceptual unity of Chinese words and slowed down reading. This effect could not be attributed to reading a less familiar spacing text because inserting spaces between words led to similar reading times to the normal unspaced text.

On the other hand, developmental study showed that a better understanding of character meanings (i.e., higher morphological awareness) is predictive of better reading performance in Chinese (Tong, McBride-Chang, Shu, & Wong, 2009). Furthermore, priming experiments indicated that character-sharing between primes and targets facilitated target word recognition and the effect could not be explained by lexical level factors (Tsang & Chen, 2013a, 2013b). In CLP-Tse, character frequency and semantic transparency also explained unique variance of the lexical decision latency (Tse et al., in press). These findings are more consistent with the proposal that characters play an important role during Chinese word recognition. To clarify these contradictory claims, we include and analyze the effects of different character-level variables in this megastudy.

Method

Participants

Five hundred and four undergraduate students from different universities in Guangzhou were gathered at South China Normal University to participate in the experiment. All of them were native Chinese users and reported no language-related disorders. They completed a questionnaire when they were recruited, which inquired demographic data (age, gender, and university), language background (number of hours in Chinese reading per week, Chinese subject score in the National Higher Education Entrance Examination, self-rated Chinese language proficiency on a 10-point Likert scale, and the number of foreign languages known). In addition, they completed a vocabulary knowledge test and a dictation test. Table 1 shows these participant characteristics. As shown by the performance in the language-related tests, our participants have a medium to high level of Chinese proficiency. We did not restrict our recruitment to only highly proficient Chinese users (as in Sze et al., 2014), so that our data will be more representative to the average undergraduate students. A larger variation in language proficiency also allows the exploration of individual differences in language processing in future research (Yap, Balota, Sibley, & Ratcliff, 2012).

Materials

All words in MELD-SCH were selected from the SUBTLEX-CH frequency corpus (Cai & Brysbaert, 2010). First, we randomly chose 20,000 words of one-character to four-character length from the corpus. Because the corpus was based on movie and TV show subtitles, there were a large number of proper names for persons and places. Our second step was to remove the proper names, except those that had become lexicalized and productive due to frequent use. For example, "孔子" (/kǒngzǐ/, Confucius) and "日本" (/rìběn/, Japan) were retained because they were so frequently used that their constituent characters had taken on new meanings and could form new words such as "孔教" (/kǒngjiào/, Confucianism) and "日元" (/rìyuán/, Japanese Yen). Third, because Sze et al. (2014) had already constructed the CLP, an excellent dataset of lexical decision for 2,500 one-character simplified Chinese words, we intentionally replaced some one-character words with those that had not been included in the CLP. Nevertheless, 381 one-character words were shared with Sze et al. (2014), which allowed us to compare MELD-SCH with CLP.

After these steps, there were 1,020 one-character words, 10,022 two-character words, 949 three-character words, and 587 four-character words.[2] Because the number of Chinese characters is limited, character repetition is unavoidable in constructing this megastudy dataset. An equal number of pseudowords for each word length was prepared for the lexical decision task. The one-character pseudowords were constructed by adding/deleting strokes from real characters or by combining radicals to form non-existing characters. Some of these one-character pseudowords were listed on the CJK Unicode. We confirmed their pseudoword status by ensuring that they did not have corresponding entries in the Baidu online dictionary (http://dict.baidu.com/), the Chinese character database: With word formation (http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/), and the SUBTLEX-CH frequency corpus (Cai & Brysbaert, 2010). Two-character, three-character, and four-character pseudowords were constructed by

[2] In Chinese, the boundary between words and phrases is often unclear. For example, "吃飯-to eat" literally means "eat rice" and is technically a phrase. Yet, it is treated as a word in most Chinese word dictionaries and CLP-Tse (Tse et al., in press). In this study, we considered all materials listed on the SUBTLEX-CH frequency corpus (Cai & Brysbaert, 2010) as words, which were segmented from movie subtitles by computer software.

Table 1 Background characteristics of the participants

	All (N = 504)	Male (N = 228)	Female (N = 276)
Age, y	19.56 (1.36)	19.78 (1.28)	19.43 (1.38)
Hours of Chinese reading per week	5.21 (5.50)	5.39 (5.58)	5.07 (5.44)
Chinese subject score in the National Higher Education Entrance Examination (max score = 150)	114.37 (9.24)	112.11 (9.32)	116.27 (8.76)
Self-rated Chinese proficiency level (1–10; 10 = highly proficient)	6.32 (1.20)	6.51 (1.08)	6.16 (1.27)
Number of foreign languages known	1.16 (0.45)	1.15 (0.48)	1.16 (0.42)
Accuracy of vocabulary knowledge test (%)	70.40 (12.29)	70.23 (11.77)	70.55 (12.73)
Accuracy of dictation test (%)	83.16 (9.57)	82.75 (8.28)	83.47 (10.53)

Note: Standard deviations are shown in parentheses

randomly recombining the constituent characters of the word counterparts with character position preserved. Again, their pseudoword status was checked against Baidu online dictionary and SUBTLEX-CH.

MELD-SCH can be found in the Online Supplementary Material. It is available both in xlsx and pdf format. Besides the behavioral data of the lexical decision task and word-level lexical variables (e.g., word frequency and word length), we also added a number of character-level variables to address our research question of whether characters play a role in Chinese word recognition. Because there are different numbers of character-level variables for words of different lengths (i.e., one set of variables for one-character words, two sets of variables for two-character words, and so on), the empty cells in the Supplementary Materials do not imply missing data. For pseudowords, only the behavioral data of the lexical decision task is available. Table 2 summarizes the variables listed in MELD-SCH. While the meanings of lexical-level

Table 2 Variables listed in MELD-SCH

Variable name	Description
id	Unique numeric id for each item
word	Words/pseudowords in simplified Chinese
lexicality	Numeric code to indicate words (1) and pseudowords (2)
N	Number of data point contributing to the mean reaction times
RT	Mean reaction time across participants
RTSD	Standard deviation of reaction time across participants
zRT	Mean standardized reaction time across participants
zRTSD	Standard deviation of standardized reaction time across participants
ERR	Mean error rate across participants
length	Word length in number of characters
tstroke	Total number of strokes of the whole-word
wfreq	Raw word frequency count based on SUBTLEX-CH
wfreq/mil	Word frequency (per million count) based on SUBTLEX-CH
cd	Contextual diversity based on SUBTLEX-CH
C1stroke1 to C4stroke	Number of strokes of constituent characters 1 to 4
C1nwf to C4nwf	Total number of words formed by constituent characters 1 to 4
C1cf to C4cf	Cumulative frequency of constituent characters 1 to 4
C1nom to C4nom	Number of meanings of constituent characters 1 to 4
C1nop to C4nop	Number of pronunciations of constituent characters 1 to 4
C1pr1 to C4pr5	Pronunciations (max: 5) of constituent characters 1 to 4

variables are straight-forward, some character-level variables are constructed for this megastudy. The meanings of these variables are elaborated in the following:

Number of words formed (nwf) refers to the number of words formed by each character, regardless of the character position. Position was not taken into account because there is evidence that Chinese readers easily confused words containing characters in transposed positions (e.g., "帶領-to lead" and "領帶-tie"), which indicates that character position is not important in the representations of Chinese words and Chinese readers probably need a "checking back procedure" to ultimately discriminate between these words (Taft, Zhu, & Peng, 1999). This variable was constructed by counting the number of words sharing the same character in SUBTLEX-CH (Cai & Brysbaert, 2010), after removing the entries of proper names for persons and places.

Cumulative frequency (cf) of a character refers to the sum of frequencies of all words that contain this character. Again, character position is not considered. This variable was based on the raw character frequency count in SUBTLEX-CH (Cai & Brysbaert, 2010).

Number of meanings (nom) refers to the number of meanings a character has. It was estimated by counting the number of separate entries for each character in the Baidu online dictionary. We present the raw counts instead of categorizing this continuous variable, as in Liu et al. (2007).

Number of pronunciations (nop) refers to the number of pronunciations a character has. Those that share syllables but differ in tones were also considered as different pronunciations. For example, "予" can be pronounced as/*yú*/ or/*yǔ*/, so its nop is two. The pronunciations were obtained from the Baidu online dictionary.

Procedure

The stimulus list contains 12,578 simplified Chinese words of various lengths and 12,578 corresponding pseudowords. The 25,156 items were divided into 12 lists. For list 1, there were 2,100 items (1,050 words and 1,050 pseudowords). For all other lists, there were 2,096 items (1,048 words and 1,048 pseudowords). Words/pseudowords of different lengths were roughly equally distributed across lists. Within each list, the items were further divided into ten blocks of equal size. The blocks were constructed to minimize character repetition within a block. But character repetition was unavoidable across the entire experimental list. Blocks within list, and items within blocks, were presented randomly to the participants. In between blocks, participants were allowed to take a break. Participants were randomly assigned to the 12 lists, such that there were 42 participants in each list.

All participants arrived at the computer laboratory in South China Normal University for the experimental sessions. Stimulus presentation and response recording was controlled by DMDX (Forster & Forster, 2003). Each trial began with a fixation cross presented for 300 ms, followed by a blank screen for 200 ms. Then, the target stimulus appeared and stayed on the screen until the participant responded. Participants were instructed to judge whether the stimulus was a real Chinese word or not. The instruction was shown on the computer screen and the participants could ask for clarifications. Both speed and accuracy were emphasized in the instruction. Responses were collected with keyboards, with the "M" and "Z" keys for words and pseudowords, respectively. When a response was recorded, or if the participant did not respond within 3,000 ms after stimulus onset, the next trial started after a blank screen of 200 ms. Forty practice trials with items unused in the main experimental lists were prepared to ensure good understanding of the instruction. All participants completed the lexical decision experiment in about 90 min (including reading instruction, practice, break, and the main experiment).

Results and discussion

Although we intended to recruit participants with diverse language proficiency levels, 16 participants were discarded because of excessive errors (>25%), which might indicate inattentiveness during the experiment. Extreme response times (i.e., faster than 200 ms or slower than 2,500 ms) were also discarded. Then, we calculated the mean and standard deviation (SD) of the reaction time (RT) for each participant. Extreme RTs (±2.5 SD from the individual's mean) were excluded from further analyses. After excluding the outliers, the RTs of each participant were standardized by z-transformation based on the individual's mean and SD. The mean and SD of each item were averaged based on both the raw RT and the standardized RT (zRT) of the participants. We reported the raw RTs for easy comparison with other studies, but the subsequent analyses were conducted with the zRTs because this had the advantages of reducing noises due to individual differences and at the same time preserving item-based variability (Tse et al., in press).

Following previous megastudies (e.g., Ferrand et al., 2010; Tse et al., in press), we calculated the split-half correlation of all items (i.e., randomly assigning participants into half A and half B, and then calculating the item-level correlation between the two halves). We also applied the Spearman-Brown formula $(2 \times r)/(1 + r)$ to correct for differences in the number of observations in the two halves. The corrected split-half correlation for RT, zRT, and error rate (ERR) were .83, .85, and .87,

respectively, which were comparable to other megastudies (Ferrand et al., 2010; Keuleers et al., 2010; Tse et al., in press). Sze et al. (2014) and Tse et al. (in press) removed items with error rates above 30%. Similarly, we removed items with error rates above 25%. For the remaining items, the behavioral data and the information of word-level and character-level variables are displayed in Table 3. In the Online Supplementary Material, the data for all items are included such that researchers can also identify items with high error rates.

Comparisons with other megastudies in Chinese

Sze et al. (2014) developed the CLP, a lexical decision megastudy of 2,500 one-character words in simplified Chinese. The mean RT and ERR across all words in CLP were 601.70 ms and 4.77%, respectively. In comparison, the mean RT and ERR of all one-character words in MELD-SCH were 790.40 ms and 7.24%, respectively. This could be partly related to the fact that we deliberately chose some low frequency words that were not included in the CLP. However, when we restricted our comparison to the 381 words shared between the CLP and MELD-SCH (15 of which had error rates >25% in either CLP or MELD-SCH and were not included in the analyses), performance was still better in CLP (mean RT and ERR = 588.82 ms and 3.06%) than MELD-SCH (mean RT and ERR = 762.27 ms and 5.19%; $ts(365) = 47.23$ and 8.31, both $ps < .001$). We believe that several factors might be responsible for the differences. First, language proficiency of the participants might play a role. In Sze et al. (2014), all participants were highly proficient Chinese users, while we recruited participants from more diverse language background (e.g., different universities and major subjects, a broad range of performance in the Chinese subject in National Higher Education Entrance Examination). Second, our experimental session was slightly longer (about 2,100 items vs. 1,700 or 1,600 items). Third, mixing Chinese words with different lengths in MELD-SCH might have the unexpected effect of making the lexical decision task more difficult.

More importantly, the data in CLP and MELD-SCH were well correlated. For the shared items, Pearson's correlation coefficients in zRTs and ERRs were .68 and .52, respectively,

Table 3 Lexical decision performance, word-level, and character level variables

	One-character words (N = 779)	Two-character words (N =9,495)	Three-character words (N = 903)	Four-character words (N = 578)
RT (ms)	790.45 (107.00)	749.89 (91.30)	758.67 (95.14)	814.53 (94.19)
zRT	−0.16 (0.35)	−0.30 (0.30)	−0.26 (0.31)	−0.08 (0.31)
ERR (%)	7.24 (6.63)	4.20 (5.14)	3.67 (4.90)	3.45 (4.21)
cd	541.82 (1,145.91)	303.03 (732.58)	81.66 (264.81)	38.79 (102.26)
C1stroke	9.95 (3.24)	8.48 (3.26)	7.45 (3.26)	6.95 (3.58)
C2stroke	-	8.23 (3.29)	7.81 (3.11)	7.53 (3.31)
C3stroke	-	-	7.76 (3.26)	7.09 (3.45)
C4stroke	-	-	-	8.11 (3.39)
C1nwf	29.72 (55.48)	86.87 (105.92)	127.47 (133.34)	179.45 (217.11)
C2nwf	-	110.30 (128.08)	130.74 (130.26)	154.83 (179.96)
C3nwf	-	-	172.87 (162.91)	180.77 (210.23)
C4nwf	-	-	-	116.90 (133.54)
C1cf	8,475.97 (39,322.64)	31,703.73 (87,933.11)	41,399.37 (93,376.88)	86,735.99 (190,203.35)
C2cf	-	34,905.33 (81,119.89)	42,144.42 (108,455.69)	64,020.00 (152,825.44)
C3cf	-	-	43,921.94 (84,416.19)	102,517.76 (215,552.23)
C4cf	-	-	-	44,172.12 (94,962.56)
C1nom	3.49 (2.45)	5.74 (3.76)	6.60 (4.00)	6.35(3.92)
C2nom	-	5.86 (3.72)	6.40 (3.93)	6.07 (3.57)
C3nom	-	-	6.11 (3.56)	6.32 (4.11)
C4nom	-	-	-	6.04 (3.85)
C1nop	1.20 (0.51)	1.27 (0.57)	1.28 (0.54)	1.28 (0.58)
C2nop	-	1.26 (0.54)	1.26 (0.56)	1.22 (0.53)
C3nop	-	-	1.28 (0.53)	1.26 (0.52)
C4nop	-	-	-	1.26 (0.57)

Note: Standard deviations are shown in parentheses. Please see Table 2 for an explanation for the abbreviations

which could be considered strong associations according to Cohen's guideline (1992). In other words, despite the apparent differences in absolute RT and ERR in the two megastudies, there were indeed considerable similarities in the relative performance across different words. This provides additional evidence for the reliability of MELD-SCH.

We also compared our data with Tse et al. (in press), who recently extended the CLP to include the lexical decision data of two-character words in traditional Chinese. The mean RT and ERR in CLP-Tse were 656.76 ms and 10.25%, respectively. In comparison, the mean RT and ERR of all two-character words in MELD-SCH were 749.90 ms and 4.20%, respectively. Then, we restricted our analyses to the words that were shared between CLP-Tse and MELD-SCH. We excluded words with high error rates (>25% in either CLP-Tse or MELD-SCH) and words that did not have a one-to-one correspondence between simplified and traditional Chinese. For example, both the traditional Chinese words "事蹟" and "事跡" become "事迹" during simplification and were thus excluded from the following analyses. Thus 8,165 two-character words remained. For this shared set of words, the participants in CLP-Tse responded faster than those in MELD-SCH (624.43 ms vs. 738.14 ms, $t(8,164) = 154.10$, $p < .001$), but they also made more errors (4.58% vs. 3.52%, $t(8164) = 17.90$, $p < .001$). The differences might be related to differences in language proficiency for the China participants in MELD-SCH compared to the Hong Kong ones in CLP-Tse. Specifically, the Hong Kong participants might know fewer words (as indicated by the higher error rate in CLP-Tse), which allowed them to receive less interference from orthographically similar words but also made them mistake more words as nonwords. Alternatively, this might reflect differences between processing simplified and traditional Chinese scripts. These possibilities could be examined in future studies by including standardized tests of Chinese proficiency in simplified and traditional Chinese to control for participants' language proficiency level.

There were moderate to strong correlations in zRT ($r = .64$) and ERR ($r = .36$) between CLP-Tse and MELD-SCH. Although the correlations are numerically not very strong, they were interesting because they indicated that despite the variations in surface orthographic form between the traditional and simplified scripts, and participants' language background (Mandarin vs. Cantonese mother tongue), there is still remarkable sharing in the core processes underlying simplified and traditional Chinese word recognition. These core processes may engage semantics more heavily than the orthographic and phonological forms (Liu et al., 2007), which were varied between CLP-Tse and MELD-SCH (e.g., simplified Chinese words usually have fewer strokes, and the participants in CLP-Tse speak Cantonese, while those in MELD-SCH speak Mandarin).

Item-based regression analyses

Table 4 displays the inter-correlations among the behavioral measures and the word-level and character-level variables of interest. All variables, except the number of pronunciations, were significantly correlated with zRT and ERR. We conducted item-based regression analyses to investigate how different word-level and character-level properties affected Chinese word recognition as measured by lexical decision performance. Word-level factors included word length and word frequency as measured by the logarithm of context diversity in SUBTLEX-CH (Cai & Brysbaert, 2010), which was shown to be the most effective word frequency measure in CLP-Tse (Tse et al., in press). Moreover, as shown in Table 3, there is an obvious U-shape relationship between word length and RT/zRT. Therefore, we also included the quadratic term of word length to capture this pattern.

Because the number of character-level factors increases as word length increases, we averaged the character-level factors at different character positions and entered the averaged values in our regression models. A similar averaging procedure has been adopted in a megastudy that investigated how spelling-to-sound consistency of onsets and rimes affected nonword naming in English, such that the consistencies of onsets and rimes were averaged (Mousikou, Sadat, Lucas, & Rastle, 2017). The character-level factors included the number of words formed, cumulative character frequency, number of meanings, number of strokes, and number of pronunciations. The first three variables were highly positively skewed so their logarithmic values were used.[3] As shown in Table 4, the number of words formed was also highly correlated with cumulative character frequency. To avoid a multi-collinearity problem, we conducted a separate regression model with cumulative frequency as the dependent variable and number of words formed as the predictor. We then calculated the residual of cumulative frequency after accounting for number of words formed. This residual term of cumulative frequency was used in the final item-based regression models for zRT and ERR.

Table 5 displays the results of the regression models, with zRT and ERR as the dependent variables. For zRT, the final

[3] We averaged the log-transformed values across characters, which was different from log-transforming the averaged values. We believe that the former procedure might provide a fairer treatment for words that contain characters with extreme values. For example, the context diversity counts of the first and second characters for "一贯" (/*yīguàn*/, consistent) are 604,056 and 486, and the averaged log-transformed and log-transformed averaged values are 5.48 and 4.23, respectively. The context diversity counts for "人情" (/*rénqíng*/, humanity) are 373,292 and 69,279, and the averaged log-transformed and log-transformed averaged values are 5.34 and 5.20, respectively. As shown in this example, the two words appear to have similar log-transformed averaged context diversity, but this is due to the high value of the first character in "一贯." In contrast, the averaged log-transformed values could better capture the difference between the two words. Having said this, better treatments probably exist. Therefore, the raw values for each character are included in MELD-SCH to allow researchers to test different treatments.

Table 4 Inter-correlations among the behavioral measures and word-level and character-level variables used in the regression models

		1	2	3	4	5	6	7	8
1	zRT								
2	ERR	.64***							
3	log_cd	−.60***	−.41***						
4	length	.07***	−.11***	−.24***					
5	avg(stroke)	.14***	.05***	−.06***	−.18***				
6	avg(log_nwf)	−.24***	−.18***	.17***	.33***	−.49***			
7	avg(log_cf)	−.29***	−.20***	.34***	.28***	−.45***	.86***		
8	avg(log_nom)	−.13***	−.09***	.13***	.18***	−.33***	.64***	.59***	
9	avg(nop)	−.004	−.006	.06***	.02*	−.06***	.13***	.16***	.35***

* $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed). Please see Table 2 for an explanation for the abbreviations

model that included all variables accounted for 40.1% of the variance, which was statistically significant (R^2 = .401, $F(8,11746) = 982.0$, $p < .001$). For ERR, the final model accounted for 22.4% of the variance, which was also significant (R^2 = .224, $F(8,11746) = 424.6$, $p < .001$). Moreover, the effects of individual variables were mostly reasonable and as expected. Consistent with previous megastudies (Sze et al., 2014; Tse et al., in press; Yap & Balota, 2009), the effect of word frequency (measured as context diversity) was strong and facilitative to both zRT and ERR. The effects of number of meanings (significant in both zRT and ERR models) and number of pronunciations (only significant in the zRT model) were also straightforward: Both variables had inhibitory effects on lexical decision performance, suggesting that semantic and phonological ambiguities took time to be resolved for successful word recognition. It is interesting to note that number of pronunciations did not correlate with zRT (Table 4), but had a significant effect in the regression model. This might be an example of "suppression" in regression analysis, which refers to the phenomenon that a predictor of small zero-order correlation with the dependent variable turns out to be important when other predictors are considered simultaneously. It occurs when the other predictors in the regression model "clean up" part of the variances in zRT that are errors between number of pronunciations and zRT. The significance of number of pronunciations only in the regression model also illustrates the usefulness of megastudies in testing the effects of multiple variables simultaneously.

Table 5 Results of item-level regression analyses

	zRT		ERR	
	β	*t*	*β*	*t*
log_cd	−.619	−74.00***	−.485	−50.93***
length	−.548	−15.14***	−.579	−14.06***
quadratic term of length	.516	14.63***	.364	9.07***
avg(stroke)	.056	6.76***	−.033	−3.54***
avg(log_nwf)	−.078	−6.73***	−.022	−1.65+
residual of avg (log_cf)	.063	7.73***	.076	8.23***
avg(log_nom)	.047	4.74***	.037	3.27**
avg(nop)	.028	3.60***	.012	1.32

+ $p < .1$, * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed). Please see Table 2 for an explanation for the abbreviations

Somewhat unexpectedly, word length had a facilitative effect on both zRT and ERR. This might reflect the difficulty in recognizing one-character words (it was just slightly faster than four-character words and had the highest error rates; see Table 3). Together with the finding that a larger average number of strokes led to slower but more accurate lexical decision responses, the word length effect observed suggests that visually more complex words do not necessarily increase the processing demand in Chinese word recognition. In addition, the quadratic length effect was also significant, which confirmed the U-shaped relationship between word length and response latency as shown in Table 3. To the best of our knowledge, although such a quadratic pattern has been observed in word recognition in alphabetic languages (Ferrand et al., 2010; New, Ferrand, Pallier, & Brysbaert, 2006; Yap & Balota, 2009), it is the first report of a similar pattern in Chinese.

Alternatively, it could be argued that one-character words were particularly difficult to process in the lexical decision experiment because the corresponding pseudowords were constructed by removing or adding strokes, or by recombining radicals, while multi-character pseudowords were constructed by recombining real characters to form non-existing combinations. This made the difference between one-character words and pseudowords more subtle than the multi-character ones. We believe that this explanation was unlikely. First, if the

construction of pseudowords was the primary reason for the U-shaped pattern, we would expect the same difficulty to reject one-character pseudowords. However, the averaged response times were 808.26 ms, 928.36 ms, 912.45 ms, and 986.62 ms for one-character to four-character pseudowords, respectively. The faster response to one-character pseudowords was inconsistent with this argument. Second, we conducted an extra item-based regression analysis with multi-character words only, such that the pseudowords were constructed under the same principle. Besides, given that the word frequency of two-character words was much higher than that of three-character and four-character words (Table 3), we selected a subgroup of words with matched frequency to ensure that the quadratic-length effect could not be attributed to extreme differences in word frequency. Specifically, there were 578 words for each word length and the average log(context diversity) values were 1.12, 1.08, and 1.07 for two-character, three-character, and four-character words, respectively. The quadratic-length effect on zRT was still significant ($\beta = 1.85$, $t = 8.38$, $p < .01$).[4]

In the regression analyses, the most interesting finding involved the total number of words formed and cumulative character frequency. As shown in Table 4, both variables correlated negatively with zRT and ERR. The facilitative effects suggested that when a character was used more often (forming more words and had higher cumulative frequency), its representation could be activated more easily to pass on activation to the corresponding word-level representations. However, when both variables were entered simultaneously into the regression models, only the effects of number of words formed remained facilitative (marginally significant in the ERR model). The (residual) effects of cumulative frequency reversed and became inhibitory.

Although this finding might look surprising initially, it became more sensible when we looked more closely at the words that produced good and bad word recognition performance. The results of the regression analyses indicated that fast and accurate responses occurred when a word contained characters that could form many words, but each of these words was used infrequently, leading to a low cumulative character frequency. This reflects the expected facilitation by number of words formed. In contrast, responses were slow and error-prone for words that contained characters with low number of words formed and high cumulative frequency. This would only be possible when at least one of the words that shared the character had a very high frequency of usage, such that it got activated easily and interfered with the recognition of the actual target word. In other words, the inhibitory effect by cumulative frequency might be driven by the presence of high frequency orthographically similar words.[5] This proposal can be tested in future studies by directly examining how high frequency orthographically similar words influence Chinese word recognition.

General discussion

Chinese is one of the most widely used languages in the world. It has the largest pool of native users and an increasing number of foreign language learners. The development of more effective strategies in Chinese language learning is an important issue in education. Moreover, the unique features in Chinese that are not found in alphabetic languages also make it an appealing test case for psycholinguists who are interested in discovering universal aspects of language processing. Therefore, it is both practically and theoretically meaningful to investigate Chinese language processing. In the last two decades, much progress has been made with the experimental approach. Recently, researchers have begun to conduct megastudies to study Chinese language. Most of these megastudies only contain one-character words (e.g., Liu et al., 2007; Sze et al., 2014), although the newly developed CLP-Tse (Tse et al., in press) contains two-character words in traditional Chinese.

In this article, we report on MELD-SCH, a new megastudy of simplified Chinese word recognition. MELD-SCH distinguishes itself from existing megastudies in two ways. First, it contains the lexical decision data of over 12,000 Chinese words of different lengths. The proportion of one-character, two-character, three-character, and four-character words in MELD-SCH is similar to the proportion in the entire Chinese language, which makes it more representative. It also allows the investigation of word length in Chinese, which has not received adequate attention in previous studies. Second, MELD-SCH contains the data of different character-level variables, which reflects the continuing interest in testing whether the constituent characters are activated during Chinese word recognition (Packard, 1999; Tsang & Chen, 2012). These

[4] The β for the quadratic-length effect is larger than 1 because of its collinearity with word length.

[5] This proposal may seem to suggest an interaction between number of words formed and cumulative character frequency. Actually, their effects are assumed to be additive. When number of words formed increases, the character-level processing can be completed faster, thereby facilitating lexical decision performance. When cumulative character frequency increases, there is a higher chance that at least one of the orthographically similar words is of high frequency, which can become activated easily to inhibit the activation of target word. Therefore, the overall effect is facilitation for words that are high in number of words formed but low in cumulative frequency, and inhibition for words that are low in number of words formed but high in cumulative frequency. For words that are high or low on both dimensions, the two effects add up and cancel out each other, leading to intermediate lexical decision performance.

variables are orthographic (e.g., number of strokes), phonological (e.g., number of pronunciations), or semantic (e.g., number of meanings) in nature, thereby allowing a more in-depth understanding about which aspects of the constituent characters could potentially influence processing.

Following previous megastudies (e.g., Ferrand et al., 2010; Tse et al., in press), we established the reliability of MELD-SCH by calculating the split-half correlation with Spearman-Brown correction. The corrected split-half correlations for RT, zRT, and ERR were satisfactory (all above .80) and comparable to previous megastudies. We further ensured the reliability of MELD-SCH by correlating our lexical decision data of one-character words with CLP (Sze et al., 2014) and two-character words with CLP-Tse (Tse et al., in press). Moderate to strong correlations (.36 to .68) were found between the zRT and ERR in MELD-SCH and these two datasets, despite the many procedural differences among studies (e.g., location of testing, participants' language proficiency level, language background, etc.). Based on these reliability analyses, we are confident that MELD-SCH provides reliable and high quality lexical decision data in simplified Chinese.

We also demonstrated how MELD-SCH could be used to investigate Chinese language processing. Specifically, we conducted item-based regression analyses to examine the effects of word length and character-level variables in Chinese word recognition. Regarding word length, we discovered a significant U-shaped relationship with lexical decision performance after the effects of other lexical-level and character-level variables had been accounted for. The recognition of one-character and four-character words was slower than two-character and three-character words. Although the range of word length in Chinese is much more restricted than in alphabetic languages, this U-shape pattern is consistent with the findings in English and French (Ferrand et al., 2010; New et al., 2006). Ferrand et al. (2010) suggested that short words are orthographically similar to many other words, producing more interference and difficulty in recognition even though they are perceptually simpler. In contrast, longer words are orthographically more distinctive, which can compensate the perceptual disadvantage. In other words, increasing word length has two opposing effects, namely the facilitation due to higher distinctiveness and the inhibition due to higher visual complexity (i.e., having more visual units for integration). The actual outcome depends on the balance between facilitation and inhibition.

Similar mechanisms might be at work in Chinese. Increasing word length can improve the clarity and distinctiveness of short Chinese words. For example, the one-character words "爱" (/*ài*/, love) and "受" (/*shòu*/, to receive) are visually similar, but two-character words that contain them, like "爱人" (/*àirén*/, lover) and "受伤" (/*shòushāng*/, be injured), are visually more distinctive. This outweighs the cost of increased perceptual complexity and leads to an overall facilitation. In contrast, longer words are orthographically quite distinctive already, so the gain is small compared to the cost of increased perceptual complexity. This results in slower recognition of four-character than three-character words. To test this proposal directly, future research would need to establish more robust operational definitions of orthographic similarity and orthographic neighborhood (e.g., orthographic Levenshtein distance; Yarkoni, Balota, & Yap, 2008) in Chinese. MELD-SCH will contribute to this process by providing the benchmark behavioral data and information of other linguistic variables against which the effectiveness of the new measures can be verified.

Regarding the role of characters in word recognition, the regression analyses clearly indicated that characters are activated during Chinese word recognition. Moreover, the characters do not simply act as an orthographic code for accessing the representations of words. Their phonological and semantic properties are also activated, as shown by the significant number of pronunciations and number of meanings effects. These effects are particularly interesting because character-level ambiguity was the primary reason why Packard (1999) proposed that Chinese words should be recognized holistically. Indeed, given that these effects were both inhibitory in the regression analyses, Packard was correct that ambiguous character meanings and pronunciations would slow down word recognition. However, this does not lead Chinese readers to process words holistically. It is possible that Chinese words are automatically and obligatorily decomposed into constituent characters, just like opaque or pseudo-complex words (e.g., department and corner) are decomposed into the constituent morphemes in English and other alphabetic scripts (Rastle, Davis, & New, 2004). Although decomposition may impose a cost for subsequent integration, it affords more economical storage in the mental lexicon. As shown in Table 3, it is common for a character to form more than 100 words, which results in a high degree of storage redundancy if all words have to be stored separately. Therefore, after balancing the needs of storage and processing, it may be more efficient overall to store the characters individually and combine them online during word recognition.

At the same time, there are words where combining characters would lead to entirely incorrect word meanings. For example, "派对" (/*pàiduì*/, party) is a loanword, such that its constituent characters (literally "to give correct") are irrelevant to the whole-word meanings. In CLP-Tse, Tse et al. (in press) showed that these words were recognized more slowly and less accurately than words with semantically transparent constituent

characters. In a priming experiment, Tsang and Chen (2014) showed that the character meanings for these words were activated at an early stage of word recognition, which was then corrected at a later stage, probably by relying on holistic word representations for these special cases. These results indicate that the benefits of more economical storage might even outweigh the processing disadvantages, such that "decomposition and integration" continues to be the default mechanism even when it will fail sometimes. The constraints of such decomposition and integration process should be investigated in future studies, perhaps by examining how pseudowords that contained interpretable combinations of constituent characters (e.g., "家力", which literally means "family force") are processed differently from the non-interpretable ones (e.g., "家鱗, which literally means "family fish scale").

The effects observed in the regression analyses impose significant constraints on the development of Chinese word recognition models. Besides testing important theoretical questions, MELD-SCH also contributes to the understanding of Chinese language processing in different ways. Balota et al. (2012) and Tse et al. (in press) have made a good summary on how megastudies like MELD-SCH could facilitate empirical research, such as providing benchmark data for testing the effects of new linguistic variables, supporting material selection in experimental studies, performing virtual experiments to replicate previous findings, and so on. We want to discuss two further possible contributions. First, it provides the data necessary for simulating Chinese word processing. Although computer simulation has the advantages of high clarity and transparency as compared to traditional "box-and-arrow" models, it has only been used occasionally to study Chinese language processing (e.g., visual word recognition: Yang, McCandliss, Shu, & Zevin, 2009; spoken word recognition: Shuai & Malins, 2017; reading: Rayner, Li, & Pollatsek, 2007). One major obstacle for performing computer simulation in Chinese is the absence of large-scale corpus data of various linguistic variables and benchmarking behavioral data, both of which are available in MELD-SCH. We also hope that linguists and psychologists would find MELD-SCH useful in testing new linguistic variables, thereby further enriching the database and allowing more complex simulations.

Second, MELD-SCH can be considered as normative data of simplified Chinese word recognition by native Chinese university students, which can be used to construct Chinese proficiency tests. This approach was demonstrated by Brysbaert (2013), who selected French words and nonwords of different difficulty levels based on the data in the French Lexicon Project (Ferrand et al., 2010). The materials chosen were then used to construct the LEXTALE-FR, a French language proficiency test. This test has good reliability and validity. Most items also have good psychometric properties as revealed by item response theory analyses. This test can be completed in 5 min and can be used to assess French proficiency level for both native and second language users. It is possible to apply the same approach to construct a similar test in Chinese. Moreover, given the large number of words in megastudies, test developers can select multiple sets of words and nonwords with similar difficulty levels and construct parallel tests, which allows multiple assessments of language proficiency for the same participants without repeating the same set of items.

To summarize, we believe MELD-SCH can make unique contributions to research in Chinese language processing. With the normative data provided in MELD-SCH, future studies can ensure the generalizability of findings to three-character and four-character words and examine in more details how characters influence Chinese word recognition.

Acknowledgments This research was supported by a 3-way Partnership Research Grant from the Faculty of Social Science, Hong Kong Baptist University (SOSC/14-15/GRFIAS-1), Guangzhou Philosophy and Social Science Foundation (14Q15), Natural Science Foundation of China (31500880), Natural Science Foundation of Guangdong Province (2014A030311016), and Key Institute of Humanities and Social Sciences (16JJD880025). We thank two anonymous reviewers for their comments on an earlier version of this manuscript.

References

Bai, X., Yan, G., Liversedge, S. P., Zang, C., & Rayner, K. (2008). Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 1277–1287. doi:10.1037/0096-1523.34.5.1277

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445–459. doi:10.3758/BF03193014

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition* (pp. 90–115). Hove: Psychology Press.

Brysbaert, M. (2013). Lextale_FR A fast, free, and efficient test to measure language proficiency in French. *Psychologica Belgica, 53,* 23–37. doi:10.5334/pb-53-1-23

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE, 5,* e10729. doi:10.1371/journal.pone.0010729

Carreiras, M., Mechelli, A., Estévez, A., & Price, C. J. (2007). Brain activation for lexical decision and reading aloud: Two sides of the same coin? *Journal of Cognitive Neuroscience, 19,* 433–444. doi:10.1162/jocn.2007.19.3.433

Chang, Y.-N., Hsu, C.-H., Tsai, J.-L., Chen, C.-L., & Lee, C.-Y. (2016). A psycholinguistic database for traditional Chinese character naming.

Behavior Research Methods, 48, 112–122. doi:10.3758/s13428-014-0559-7

Chen, H.-C., & Shu, H. (2001). Lexical activation during the recognition of Chinese characters: Evidence against early phonological activation. *Psychonomic Bulletin & Review, 8,* 511–518. doi:10.3758/BF03196186

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. doi:10.1037/0033-2909.112.1.155

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., … Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42,* 488–496. doi:10.3758/BRM.42.2.488

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35,* 116–124. doi:10.3758/BF03195503

Hoosain, R. (1991). *Psycholinguistic implications for linguistic relativity: A case study of Chinese*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hsiao, J. H. W., & Shillcock, R. (2006). Analysis of a Chinese phonetic compound database: Implications for orthographic processing. *Journal of Psycholinguistic Research, 35,* 405–426. doi:10.1007/s10936-006-9022-y

Inhoff, A. W., & Liu, W. (1998). The perceptual span and oculomotor activity during the reading of Chinese sentences. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 20–34. doi:10.1037//0096-1523.24.1.20

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology, 1,* 174. doi:10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44,* 287–304. doi:10.3758/s13428-011-0118-4

Lee, C.-Y., Hsu, C.-H., Chang, Y.-N., Chen, W.-F., & Chao, P.-C. (2015). The feedback consistency effect in Chinese character recognition: Evidence from a psycholinguistic norm. *Language and Linguistics, 16,* 535–554. doi:10.1177/1606822X15583238

Leong, C.-K., & Cheng, P.-W. (2003). Consistency effects on lexical decision and naming of two-character Chinese words. *Reading and Writing, 16,* 455–474. doi:10.1023/A:1024243507278

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods, 39,* 192–198. doi:10.3758/BF03193147

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113,* 181–190. doi:10.1037/0033-2909.113.1.181

McBride-Chang, C., Cho, J. R., Liu, H., Wagner, R. K., Shu, H., Zhou, A., … Muse, A. (2005). Changing models across cultures: Associations of phonological awareness and morphological structure awareness with vocabulary and word recognition in second graders from Beijing, Hong Kong, Korea, and the United States. *Journal of Experimental Child Psychology, 92,* 140–160. doi:10.1037/0022-0663.97.1.81

Mousikou, P., Sadat, J., Lucas, R., & Rastle, K. (2017). Moving beyond the monosyllable in models of skilled reading: Mega-study of disyllabic nonword reading. *Journal of Memory and Language, 93,* 169–192. doi:10.1016/j.jml.2016.09.003

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13,* 45–52. doi:10.3758/BF03193811

Packard, J. L. (1999). Lexical access in Chinese speech comprehension and production. *Brain and Language, 68,* 89–94. doi:10.1006/brln.1999.2102

Peng, D. L., Liu, Y., & Wang, C. (1999). How is access representation organized? The relation of polymorphemic words and their morphemes in Chinese. In J. Wang, A. W. Inhoff, & H.-C. Chen (Eds.), *Reading Chinese script: A cognitive analysis* (pp. 65–89). NJ: Lawrence Erlbaum Associates.

Perea, M., & Carreiras, M. (1998). Effects of syllable frequency and syllable neighborhood frequency in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 134–144. doi:10.1037/0096-1523.24.1.134

Perfetti, C. A., & Tan, L. H. (1998). The time course of graphic, phonological, and semantic activation in Chinese character identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 101–118. doi:10.1037/0278-7393.24.1.101

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review, 11,* 1090–1098. doi:10.3758/BF03196742

Rayner, K., Li, X., & Pollatsek, A. (2007). Extending the E-Z reader model of eye movement control to Chinese readers. *Cognitive Science, 31,* 1021–1033. doi:10.1080/03640210701703824

Shu, H., & Anderson, R. C. (1997). Role of radical awareness in the character and word acquisition of Chinese children. *Reading Research Quarterly, 32,* 78–89. doi:10.1598/RRQ.32.1.5

Shuai, L., & Malins, J. G. (2017). Encoding lexical tones in jTRACE: A simulation of monosyllabic spoken word recognition in Mandarin Chinese. *Behavior Research Methods, 49,* 230–241. doi:10.3758/s13428-015-0690-0

Su, Y. F., & Samuels, S. J. (2010). Developmental changes in character-complexity and word-length effects when reading Chinese script. *Reading and Writing, 23,* 1085–1108. doi:10.1007/s11145-009-9197-3

Sze, W. P., Rickard Liow, S. J. R., & Yap, M. J. (2014). The Chinese Lexicon Project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods, 46,* 263–273. doi:10.3758/s13428-013-0355-9

Sze, W. P., Yap, M. J., & Rickard Liow, S. J. R. (2015). The role of lexical variables in the visual recognition of Chinese characters: A megastudy analysis. *The Quarterly Journal of Experimental Psychology, 68,* 1541–1570. doi:10.1080/17470218.2014.985234

Taft, M., Zhu, X., & Peng, D. (1999). Positional specificity of radicals in Chinese character recognition. *Journal of Memory and Language, 40,* 498–519. doi:10.1006/jmla.1998.2625

Tong, X., McBride-Chang, C., Shu, H., & Wong, A. M. (2009). Morphological awareness, orthographic knowledge, and spelling errors: Keys to understanding early Chinese literacy acquisition. *Scientific Studies of Reading, 13,* 426–452. doi:10.1080/10888430903162910

Tsang, Y.-K., & Chen, H.-C. (2010). Morphemic ambiguity resolution in Chinese: Activation of the subordinate meaning with a prior dominant-biased context. *Psychonomic Bulletin & Review, 17,* 875–881. doi:10.3758/PBR.17.6.875

Tsang, Y.-K., & Chen, H.-C. (2012). Eye movement control in reading: Logographic Chinese versus alphabetic scripts. *PsyCh Journal, 1,* 128–142. doi:10.1002/pchj.10

Tsang, Y. K., & Chen, H. C. (2013a). Early morphological processing is sensitive to morphemic meanings: Evidence from processing ambiguous morphemes. *Journal of Memory and Language, 68,* 223–239. doi:10.1016/j.jml.2012.11.003

Tsang, Y. K., & Chen, H. C. (2013b). Morpho-semantic processing in word recognition: Evidence from balanced and biased ambiguous morphemes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1990–2001. doi:10.1037/a0033701

Tsang, Y.-K., & Chen, H.-C. (2014). Activation of morphemic meanings in processing opaque words. *Psychonomic Bulletin & Review, 21,* 1281–1286. doi:10.3758/s13423-014-0589-2

Tsang, Y. K., Wu, Y., Ng, H. T. Y., & Chen, H. C. (2017). Semantic activation of phonetic radicals in Chinese. *Language, Cognition and Neuroscience, 32,* 102–116. doi:10.1080/23273798.2016.1246744

Tse, C.-S., Yap, M. J., Chan, Y.-L., Sze, W. P., Shaoul, C., & Lin, D. (in press). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*. doi:10.3758/s13428-016-0810-5

Yan, M., Richter, E. M., Shu, H., & Kliegl, R. (2009). Readers of Chinese extract semantic information from parafoveal words. *Psychonomic Bulletin & Review, 16,* 561–566. doi:10.3758/PBR.16.3.561

Yang, J., McCandliss, B. D., Shu, H., & Zevin, J. D. (2009). Simulating language-specific and language-general effects in a statistical learning model of Chinese reading. *Journal of Memory and Language, 61,* 238–257. doi:10.1016/j.jml.2009.05.001

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60,* 502–529. doi:10.1016/j.jml.2009.02.001

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance, 38,* 53–79. doi:10.1037/a0024177

Yap, M. J., Rickard Liow, S. J., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods, 42,* 992–1003. doi:10.3758/BRM.42.4.992

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15,* 971–979. doi:10.3758/PBR.15.5.971