Behavior Research Methods (2020) 52:440–445
https://doi.org/10.3758/s13428-019-01240-2

Normative data for Chinese-English paired associates

Kit W. Cho[1] · Chi-Shing Tse[2] · Yuen-Lai Chan[2]

Published online: 8 April 2019

Abstract
Paired-associate learning is one of the most commonly used paradigms to study human memory. In many of these studies, participants are typically told to learn foreign language–English translations, such as Swahili–English or Lithuanian–English pairs. One limitation of these currently available foreign language–English translation norms is that their foreign languages are based on the alphabetic writing system, thereby preventing researchers from generalizing their findings to languages based on logographic writing systems. In the present study we collected normative data for 160 Chinese–English word pairs. Participants completed three study–test cycles, followed by metacognitive judgments on their learning experience. For each pair, we report recall performance, recall latency, ease of learning, and judgments of learning. A simultaneous multiple regression analysis with frequency (of both the English word and the Chinese character), word length (English), and number of strokes (Chinese) as predictors revealed that a greater number of strokes (or higher visual complexity) for the Chinese characters predicted lower target recall.

Keywords Paired associates · Chinese–English translations · Cued recall · Episodic memory

Paired-associate learning is one of the most widely used paradigms to study human memory. In many of these studies, researchers have their participants learn foreign language translations because it mimics part of what students are required to do in the classroom: namely, to learn new facts, vocabulary words, or a foreign language. Currently, the most widely used foreign language–English translation norms are Swahili–English (Nelson & Dunlosky, 1994), in part because many people, especially undergraduates attending a university or college in the United States, do not speak Swahili. For many years, the Swahili–English norms were the only foreign language–English translation corpus available; thus, Grimaldi, Pyc, and Rawson (2010) noted that its popularity could potentially functionally limit the sample size of a researcher's participant pool, due to excluding participants from a future experiment for having completed another study that used the same materials. Consequently, Grimaldi et al. gathered normative data for another set of foreign language (Lithuanian)–English translations. Similar to the rationale of using Swahili as the foreign language for students to learn, Lithuanian was chosen because it is a language that is unfamiliar to most undergraduates.

However, the main limitation of the currently existing translation norms is that both foreign languages (i.e., Swahili and Lithuanian) are based on an alphabetic writing system. Researchers who are studying paired-associate learning, such as the testing or retrieval-practice effect (e.g., Coppens, Verkoeijen, & Rikers, 2011; Kang, 2010; Wartenweiler, 2011), may be interested in generalizing their results to a logographic (or nonalphabetic) writing system, which relies more on spatial features. For example, the retrieval-practice or testing effect is the finding that self-testing enhances memory over other commonly used study strategies (see Roediger, & Karpicke, 2006, for a review). One popular explanation of this effect, as applied to paired-associate learning, is that participants learning weakly associatively related pairs (e.g., *mother–child*) activate mediators (e.g., *father*) that serve as links between the cue and the target (see Carpenter, 2011; but see Cho, Neely, Brennan, Vitrano, & Crocco, 2017). Mediators can thus be semantically related to the cue and the target or, when participants have to learn a foreign language, can be orthographically and phonologically related to the cue and semantically related to the target (e.g., for the Swahili–English pair *wingu–cloud*, the mediator could be *wing*; see Pyc & Rawson, 2010). Although this mediator effectiveness hypothesis can be used to account for the testing effect when participants have to learn an alphabetic writing system, it is unclear how it can account for the testing effect when observed using a logographic writing

✉ Kit W. Cho
ChoK@uhd.edu

[1] Department of Social Sciences, University of Houston–Downtown, Houston, TX, USA

[2] Department of Educational Psychology, Chinese University of Hong Kong, Hong Kong, Hong Kong

system (e.g., Cho & Powers, 2019; Kang, 2010), which raises the intriguing possibility of using nonalphabetic mediators.

Moreover, some unique features are associated with a logographic language, in particular Chinese, that make them particularly appealing for researchers to use. First, many (about 90%) of Chinese characters are made up of radicals, or subcharacters that can provide a hint to the meaning or pronunciation of the whole character (Hoosain, 1991). As an example, the radical " " translates to "wood." This radical is found in other characters for which the meaning of the character is related to wood: 林 (forest), 杪 (twig), and 枝 (branches). Thus, radicals can function as both free morphemes (words or characters that stand alone) and bound morphemes (words or characters that cannot stand alone and must be attached to free morphemes—e.g., the suffix "-er" in English, which is commonly used to denote a person who performs an action as part of their occupation). This unique radical system will be especially of interest to researchers who are interested in conceptual or categorical learning (Cho & Powers, 2019; Goldstone, 1994; Lau, Alger, & Fishbein, 2011) of educationally relevant materials. For example, researchers can assess whether exposing participants to characters with the same radical can facilitate learning of other characters sharing that same radical. Indeed, among native Chinese speakers, the ability to identify and understand radicals is positively correlated with performance on literacy measures (Tong, McBride-Chang, Shu, & Wong, 2009; Wu et al., 2009). Another unique characteristic of Chinese characters is that visually complex characters (i.e., high stroke counts) occupy the same amount of space as non-visually-complex characters (e.g., 墙 vs. 力), thereby allowing researchers to select characters with either a high or low information density. Research has shown that RTs in a lexical decision task increase as the complexity of the characters increases (e.g., Tse & Yap, 2018; Tse et al., 2017). Thus, the purpose of the present study was to gather normative data for Chinese–English pairs, which will be of high interest to those wishing to generalize their research to a logographic language, which has many unique characteristics that could further inform psychological theories.

To our knowledge, only one other study (Nishimoto, Ueda, Miyawaki, Une, & Takahashi, 2010) reports normative data for nonalphabetic writing–English pairs. Nishimoto et al.'s stimuli (what they referred to as *droodles*) are outlines of incomplete pictures of objects or scenes. Normative cued-recall data were gathered by first having participants list a label or phrase for a picture (e.g., minced and steamed fish and a mechanical pencil looked at from below) and then measuring how quickly another group of participants could learn these labels. However, the major limitation of their norms is that each correct response was a phrase and therefore was convoluted to learn and score. Another limitation is that the droodle norms consist of only 98 items, which might restrict the number of observations per condition for researchers who would use them in a within-subjects experiment.

Method

Materials and design

We collected data for 160 Chinese–English nouns.[1] The overall lexical characteristics (i.e., frequency, number of strokes, and word length) of the pairs are presented in Table 1. The frequency for each English word and Chinese character is presented in the supplementary file. The Chinese characters were based on the simplified writing system. We selected familiar words (i.e., concepts) along with Chinese characters from the Chinese Lexicon Project corpus (Sze, Rickard Liow, & Yap, 2014) that varied in their number of strokes, as we thought that might influence the difficulty of learning the pairs. The English translation for each Chinese character was taken from the Chinese–English dictionary CC-CEDICT (MDBG, 2018). For counterbalancing purposes, the 160 pairs were randomly divided into four sets of 40, and each participant studied only one of the four sets. Five items were used as a primacy buffer for both the study and test phases. Each pair was studied and tested three times, across three separate study–test cycles. We collected target recall accuracy and reaction times for each test trial, yielding three sets of target recall accuracies and reaction times for each pair for each participant. In addition, we collected two metacognitive judgments: ease of learning (perceived difficulty of learning each pair) and judgment of learning (how likely participants thought they would be to remember each pair if they were tested on it a week later). The experiment was administered using the E-Prime 2.0 software (Psychology Software Tools, 2002; Pittsburgh, PA).

Participants

One hundred seventy-six undergraduates from the University of Houston–Downtown were recruited from the Psychology Participant Pool and took part in exchange for research credit. The data from one participant were excluded because he/she was not attentive during the task, and the data from three participants were excluded due to low accuracy (see the Results section below). Forty-three participants were assigned to each of the four counterbalancing lists. The mean age was 21.59 years (SD = 5.42), and 74% of the participants were female.[2] Both the solicitation statement and the consent form for the study, which participants signed, explicitly noted that participants were eligible to participate in the study only if

[1] In addition to providing the stimuli in text format (Excel file), we also provide each character as an image file (.png format). We recommend that researchers who are presenting the stimuli in a computer program use the image file, to ensure that each character will display properly.

[2] Due to experimenter error, demographic data were not solicited from 42 participants.

Table 1 Lexical characteristics of English–Chinese pairs: Mean (*SD*)

	Log Frequency[a]	Word Length[b]	Number of Strokes[c]
English	10.00 (1.14)	5.16 (1.50)	–
Chinese	3.80 (0.58)	–	8.14 (3.22)

[a] English: English Lexicon Project (Balota et al., 2007); Chinese: SUBTLEX-CH (Cai & Brysbaert, 2010). [b] English Lexicon Project. [c] CC-CEDICT (MDBG, 2018)

they had no or limited knowledge of logographic languages such as Chinese, Japanese, and Korean.

Procedure

Our procedure was identical to that used in Grimaldi et al. (2010). At the beginning of the study, participants were told that there would be three identical study–test cycles and two judgment tasks. They were told that they would be given a Chinese character and that they would have to retrieve its English translation. They were also told that the instructions for the judgment tasks would be given to them later. In each study phase, each of the 40 pairs was presented for 10,000 ms and was proceeded by a 100-ms blank screen. The Chinese character appeared on the left-hand side of the screen, with its English translation appearing on the right-hand side. After each study phase, the participants immediately proceeded to the test phase. In the test phase, each pair was proceeded by a 100-ms blank screen, followed by a screen with the Chinese character and to its right was an equal sign that was followed by a question mark, indicating that the participant should retrieve its English translation. Participant had up to 12,000 ms to retrieve a response. If a participant did not proffer a response within that time frame, the computer automatically advanced to the next trial. If a participant provided a response before 12,000 ms had passed, the participant pressed the Enter key to advance to the next trial. Reaction time was measured from the onset of the cue to the time that a participant typed a letter on the keyboard. Each Chinese characters were presented as an image (.png file), and its English translation was presented in size 36, Times New Roman, white-colored font on a black background. Participants completed the study–test cycles three times.

Following the third (last) test phase, participants proceeded to the metacognitive judgments. For the ease-of-learning task, participants were told that they would be presented with the pairs they had just studied and that they would have to indicate, using a number from 0 (*denoting very hard*) to 100 (*very easy*), how difficult it had been for them to learn that pair. A continuous scale with five anchors (0, 25, 50, 75, 100) remained on the screen, to encourage participants to use the whole range of the distribution. After they had rated all pairs for ease of learning, participants were given instructions for the judgment-of-learning task. They were told that they would now have to rate each pair on the likelihood that, if tested one week later, they would be able to recall the English translation for the Chinese character, on a scale from 0% (*will not recall*) to 100% (*likely to recall*). As with the ease-of-learning judgments, a continuous scale with the same five anchors was provided. For both metacognitive judgment tasks, we presented participants with a blank box in which to input their response, and they had an unlimited amount of time to respond. For each participant, all pairs were presented in a new, randomized order in each of the three study and test phases, yielding a total of six different presentation orders (three for the study phases and three for the test phases).

Results

The data we report in this study are averaged across participants for each item for each of the three test trials. Below, we report the overall data for recall, latency, ease of learning, and judgments of learning. The data for each individual item are presented in a supplemental file (https://osf.io/4mzxr/), listed in descending order by the accuracy of the first recall trial. For all dependent variables, we also report the split-half reliability by first randomly assigning participants within each counterbalancing list to one of two groups (determined randomly using a random-number generator in Excel), to yield two groups of 86 participants. We then computed a correlation coefficient between the two groups for each dependent variable. Minor spelling errors, which could consist of a missing letter (e.g., *summe* for *summer*), a typo (e.g., *unbrella* for *umbrella*), or an abbreviation for the word (e.g., *grandma* for *grandmother*), were scored as correct. (These corrections accounted for 0.19% of all trials.) Three participants' data were removed because their performance on the third test was more than two *SD*s (.25) less than the overall mean performance (.67) of all participants on the third test.

Recall

Overall, accuracy increased across the three test trials: Trial 1 = .16 (*SD* = .13), Trial 2 = .44 (*SD* = .19), Trial 3 = .65 (*SD* = .16), $F(2, 318) = 1{,}960.61$, $p < .001$, $\eta_p^2 = .925$. The split-half reliabilities of Trials 1–3 were all strong: *r*s = .714, .749, and .722, respectively, all *p*s < .001. The relative difficulty of the items remained the same across trials, as indicated by the strong correlations between Trial 1 and Trial 2, $r = .820$, and between Trial 2 and Trial 3, $r = .886$, both *p*s < .001. To further investigate the lexical characteristics that predicted the difficulty of learning these pairs, we computed a simultaneous multiple regression for each of the three test trials, using the following as predictors: frequency (of both the English word and Chinese character), word length (English), and number of

strokes (Chinese). The results are presented in the top section of Table 2. Altogether, these factors accounted for 13%, 18%, and 16% of the variance of participants' target recall in Trials 1–3, respectively, all Fs > 5.8, ps < .001. However, as can be seen in the table, the number of strokes was the only reliable predictor of participants' target recall performance—targets cued by Chinese characters with fewer strokes were more likely to be remembered.

Reaction time (RT)

The overall latency, which was conditionalized on correct responses, decreased across the three test trials: Trial 1 = 5,766 ms (SD = 1,524 ms), Trial 2 = 4,541 ms (SD = 980 ms), and Trial 3 = 3,932 ms (SD = 832 ms), $F(2, 312)$ = 256.57, p < .001, η_p^2 = .622. The stability of these RTs was significant, both from Trial 1 to Trial 2 and from Trial 2 to Trial 3, r = .459 and r = .818, respectively, ps < .001. The split-half reliabilities of Trials 1–3 were rs = .264, .484, and .693, respectively, all ps < .003. The reliabilities, in particular, of the first and second trials were not as strong as in the case of recall. We computed the same multiple regression analyses for the RT data as we had for the recall data. These data are presented in the second section of Table 2. They are similar to those for the recall data, in that the number of strokes was likewise a reliable predictor of RT, with RTs being longer for characters with more strokes. In addition, the length of the target word also predicted RTs. Overall, the factors accounted for 16%, 30%, and 40% of the total variance for Trials 1–3, respectively, Fs > 7.20, ps < .001.

Metacognitive judgments

Ease of learning The overall ease of learning was 50 (SD = 16), indicating that participants thought that the pairs were moderately difficult to learn. The split-half reliability was strong, .747, p < .001. The correlations between ease-of-learning judgments and actual recall were significant (ps < .001) and robust for all three test trials, rs = .823, .918, and .873 for Trials 1–3, respectively. The same aforementioned multiple analysis revealed that number of strokes was the only significant predictor (see the third section of Table 2), with more strokes predicting reduced ease of learning. Overall, all

Table 2 Lexical characteristics of pairs predicting: Accuracy and reaction times on Test Trials 1–3, ease-of-learning judgments, and judgments of learning

Dependent Variable	Predictors	Test Trial								
		1			2			3		
		β	t	p	β	t	p	β	t	p
Target recall	Frequency									
	English	– .168	– 1.648	.10	– .153	– 1.541	.13	– .115	– 1.141	.26
	Chinese	.182	1.878	.06	.097	1.030	.31	.067	0.701	.48
	Word length	.062	0.761	.45	.050	0.624	.53	.031	0.388	.70
	Number of strokes	– .300	– 4.000	< .001	– .411	– 5.431	< .001	– .391	– 5.091	< .001
Reaction time	Frequency									
	English	.092	0.915	.36	.062	0.675	.50	– .076	– 0.892	.37
	Chinese	– .123	– 1.282	.20	– .084	– 0.963	.34	– .020	– 0.243	.81
	Word length	.226	2.793	.01	.365	4.986	< .001	.475	6.970	< .001
	Number of strokes	.289	3.749	< .001	.362	5.189	< .001	.291	4.485	< .001
Ease of learning^	Frequency									
	English	– .147	– 1.503	.13	–	–		–	–	–
	Chinese	.090	0.967	.34	–	–		–	–	–
	Word length	.045	0.573	.57	–	–		–	–	–
	Number of strokes	– .438	– 5.871	< .001	–	–		–	–	–
Judgment of learning^	Frequency									
	English	– .048	– 0.480	.63	–	–		–	–	–
	Chinese	.086	0.909	.36	–	–		–	–	–
	Word length	.077	0.975	.33	–	–		–	–	–
	Number of strokes	– .405	– 5.352	<.001	–	–		–	–	–

The reaction time data are conditionalized on correct responses. ^Participants provided their responses to these queries after having completed all three study–test cycles.

predictors together accounted for 20% of the variance, $F(4, 155) = 9.659$, $p < .001$.

Judgments of learning The overall judgment of learning score was 46 ($SD = 16$). These data were nearly identical to the ease-of-learning data, in that (1) the split-half reliability was strong, $r = .802$, $p < .001$; (2) the correlations between judgments of learning and actual recall were significant (ps < .001) and robust across all three test trials, rs = .813, .921, and .866 for Trials 1–3, respectively; and (3) the multiple regression analysis, which accounted for 18% of the total variance, $F(4, 155) = 8.349$, $p < .001$, showed that more strokes predicted lower judgments of learning.

Discussion

Although there are two sets of foreign language–English translation norms (Grimaldi et al., 2010; Nelson & Dunlosky, 1994), both foreign languages use an alphabetic writing system. The present set of norms will allow researchers to generalize their findings to a logographic foreign language. Furthermore, our stimuli impart greater practical or educational relevance to a researcher's study by having participants learn one of the most widely spoken languages in the world (Eberhard, Simons, & Fennig, 2015). On that same note, researchers who are considering using these data for a memory study should take precautions to ensure that their participants do not have a familiarity with any logographic language. The variability in the recall accuracy of our word pairs (range and standard deviation on Trial 1 and Trial 3 of recall: .00–.67 and .13, and .30–.98 and .16, respectively) allows researchers to tailor their learning materials to match their desired level of difficulty, an issue that has arisen in previous research using Chinese–English translations (Kang, 2010). Participants' metacognitive judgments showed that they were cognizant of how well they had learned the material, even on their very first test. One limitation of our metacognitive judgment data is that we did not solicit these responses prior to participants taking the first test. We therefore caution against the use of our data to show that participants have insight into how well they can predict remembering a Chinese character and its English translation on a future memory test.

The results of the multiple regression analysis show that of the four lexical characteristics (i.e., word and character frequencies, word length, and number of strokes) used to predict target recall, number of strokes (i.e., the visual complexity of the Chinese character) was the only reliable predictor. This information can help guide future researchers in selecting additional Chinese–English words tailored to their desired level of difficulty that were not normed in the present study.

The stability was high for both our recall and RTs across trials, replicating findings from the Swahili–English (Nelson & Dunlosky, 1994) and Lithuanian–English (Grimaldi et al., 2010) norms. However, unlike those norms that reported a moderately sized positive correlation between English word frequency and target recall on Trial 1, we did not observe such a pattern here. Rather, the only significant predictor was the number of strokes. It could be that in a logographic writing system, the visual complexity of the stimuli (i.e., the number of strokes, in the present study) overwhelms any effects of word frequency. That said, reconciling the difference between these discrepant findings was not the main purpose of the present study, and thus our finding can serve as a fruitful direction for future research.

Within the past decade, interest in using Chinese stimuli in psycholinguistic research has grown, as evidenced by the development of Chinese characters databases with lexical and semantic variables (e.g., Chang, Hsu, Tsai, Chen, & Lee, 2016; Liu, Shu, & Li, 2007; Sze et al., 2014). Findings that lexical variables such as word frequency, which is negatively correlated with naming and lexical decision RTs in alphabetic languages (e.g., English: see Balota, 1994; Seidenberg, 1995; Spanish: see Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2011), have also been observed with Chinese characters (e.g., Chen, Wang, Wang, & Peng, 2004; Liu, Zhang, & Shu, 2006; Tse & Yap, 2018; Tse et al., 2017). Thus, we believe that the creation of our Chinese–English translation norms is timely.

Author note We thank Silvia Caamano, Jessica Cantu, Erin Chaniago, Jennifer A. Lara, Nikolas Morgan, Monica Orellana, and Elizabeth Ruiz-Harris for their assistance in collecting data for the study.

References

Balota, D. (1994). Visual word recognition: The journey from features to meaning. In M. A. Gernsbacher (Ed.), Handbook of psycholinguistics (pp. 303–356). San Diego, CA: Academic Press

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

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5*, e10729:1–8. https://doi.org/10.1371/journal.pone.0010729

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1547–1552. https://doi.org/10.1037/a0024140

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. https://doi.org/10.3758/s13428-014-0559-7

Cho, K. W., Neely, J. H., Brennan, M. K., Vitrano, D., & Crocco, S. (2017). Does testing increase spontaneous mediation in learning semantically related paired associates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1768–1778. https://doi.org/10.1037/xlm0000414.

Cho, K. W., & Powers, A. (2019). Testing enhances both memorization and conceptual learning of categorical materials. *Journal of Applied Research in Memory and Cognition*. Advance online publication. https://doi.org/10.1016/j.jarmac.2019.01.003

Coppens, L. C., Verkoeijen, P. P., & Rikers, R. M. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology*, *23*, 351–357. https://doi.org/10.1080/20445911.2011.507188

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica*, *32*, 133–143.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2015). Ethnologue: Languages of the world (18th ed.). Dallas, TX: SIL International. Retrieved from http://www.ethnologue.com

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*, 125–157. https://doi.org/10.1016/0010-0277(94)90065-5

Grimaldi, P. J., Pyc, M. A., & Rawson, K. A. (2010). Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for Lithuanian–English paired associates. *Behavior Research Methods*, *42*, 634–642. https://doi.org/10.3758/BRM.42.3.634

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

Kang, S. H. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition*, *38*, 1009–1017. https://doi.org/10.3758/MC.38.8.1009

Lau, H., Alger, S. E., & Fishbein, W. (2011). Relational memory: A daytime nap facilitates the abstraction of general concepts. *PLoS ONE*, *6*, e27139. https://doi.org/10.1371/journal.pone.0027139

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

MDBG. (2018). CC-CEDICT [Machine-readable dictionary]. Retrieved October 5, 2018, from https://www.mdbg.net/chinese/dictionary?page=radicals

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory*, *2*, 325–335. https://doi.org/10.1080/09658219408258951

Nishimoto, T., Ueda, T., Miyawaki, K., Une, Y., & Takahashi, M. (2010). A normative set of 98 pairs of nonsensical pictures (droodles). *Behavior Research Methods*, *42*, 685–691. https://doi.org/10.3758/BRM.42.3.685

Psychology Software Tools, Inc. (2002). E-Prime 2.0 (Software). Retrieved from www.pstnet.com

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335–335. https://doi.org/10.1126/science.1191465

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Seidenberg, M. S. (1995). Visual word recognition: An overview. In P. Eimas & J. L. Miller (Eds.), Handbook of perception and cognition: Language (pp. 137–179). New York, NY: Academic Press.

Sze, W. P., Rickard Liow, S. J., & 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. https://doi.org/10.3758/s13428-013-0355-9

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. https://doi.org/10.1080/10888430903162910

Tse, C.-S., & Yap, M. J. (2018). The role of lexical variables in the visual recognition of two-character Chinese compound words: A megastudy analysis. *Quarterly Journal of Experimental Psychology*, *71*, 2022–2038.

Tse, C.-S., Yap, M. J., Chan, Y. L., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*, *49*, 1503–1519. https://doi.org/10.3758/s13428-016-0810-5

Wartenweiler, D. (2011). Testing effect for visual-symbolic material: Enhancing the learning of Filipino children of low socio-economic status in the public school system. *International Journal of Research and Review*, *6*, 74–93.

Wu, X., Anderson, R. C., Li, W., Wu, X., Li, H., Zhang, J., . . . Gaffney, J. S. (2009). Morphological awareness and Chinese children's literacy development: An intervention study. *Scientific Studies of Reading*, *13*, 26–52. https://doi.org/10.1080/10888430802631734