# Back to ABCs: Clustering Alphabetically, Rather than Semantically, Enhances Vocabulary Learning

**Jingqi Yu (jy45@umail.iu.edu)**
Department of Psychological and Brain Sciences, 1101 East 10th Street, Bloomington, IN 47405 USA

**Veronica X. Yan (veronicy@usc.edu)**
Department of Psychology, University of Southern California, 3620 South McClintock Ave. Los Angeles, CA 90089 USA

**Elizabeth Ligon Bjork (elbjork@psych.ucla.edu)**
Department of Psychology, University of California, Los Angeles, 502 Portola Plaza, Los Angeles, CA 90095 USA

**Robert A. Bjork (rabjork@psych.ucla.edu)**
Department of Psychology, University of California, Los Angeles, 502 Portola Plaza, Los Angeles, CA 90095 USA

**Abstract**

Optimizing the study of vocabulary words for high-stakes tests such as the SAT or GRE prep can be problematic, given that many words are semantically, orthographically, or phonologically confusable. Companies marketing test preparation programs make multiple recommendations, such as clustering words on some basis, but little research has been carried out to examine what that basis should be. Across two experiments, we compare the efficacy of different types of clustering—categorical, alphabetical, and confusable--for the learning of semantically related words (Experiment 1) and confusable words (Experiment 2). We demonstrate that, in contrast to most learners' intuitions, an alphabetical sequence yields superior learning.

**Keywords:** memory, vocabulary learning, optimal sequencing, semantic clustering, alphabetical clustering

## Introduction

Vocabulary learning is a crucial component of learning languages, not only because knowing some minimum number of words is needed for a basic level of communication in a new language, but also because increasing one's vocabulary in one's own language can be critical in the context of high-stakes testing of various types. Applicants to undergraduate or graduate programs, for example, often spend months preparing for standardized examinations, such as the ACT, SAT, or GRE, and memorization plays an essential role in such preparation. Such vocabulary learning can be especially daunting for international applicants from non-English speaking backgrounds. To meet this demand from anxious test-takers, a huge test-preparation industry has sprung up, with each different business promising a different set of "secrets" to crack the SAT and GRE codes for a price.

Many of these organizations make recommendations for how very large sets of new words may be learned, including (a) using mnemonics (e.g., word imagery), (b) grouping words by their category membership, and (c) by Greek/Latin roots. Test preparation programs—and hence, their students—often strongly promote one over the other, yet there has been only minimal research testing the relative efficacy of such methods.

The best way(s) to learn new vocabulary, therefore, is still an open question, and what method yields the best learning outcomes may in part depend on the characteristics of the to-be-learned words. When it comes to learning new vocabulary, for example, in preparation for the GRE or SAT exams, there are two large sources of difficulty. First, these tests require individuals to learn and distinguish between many semantically related words (e.g., personality traits: *mendacious–callow*). Second, there is a need to distinguish between confusable but semantically distinct words, such as words that are similar-looking and/or similar-sounding (e.g., *decry–descry*).

While little to no research has been conducted on optimal sequencing for the learning of confusable pairs, there has been some research on semantic clustering, but evidence in support of semantic clustering, however, has been mixed. In the present studies, we specifically examine two popular methods—clustering by semantic category or by confusability, and alphabetically-clustered—and compare them against a random sequence. We examine the alphabetically-clustered sequence for a practical reason: lists of words are often organized alphabetically, and for this reason, there are many who may attempt to learn words in that order, out of convenience. Indeed—at least anecdotally—studying words alphabetically appears to be common among Chinese students preparing for the SAT and GRE exams. This alphabetical organization is not necessarily a conscious, explicit strategy, but simply a byproduct of how reading typically proceeds (i.e., start on page 1 and work your way through to the end). Studying words alphabetically does offer some structure,

but unlike semantic clustering or clustering of confusable words, clustering by initial letter appears somewhat arbitrary. However, given that it is a strategy that is widely used, and little is known as to whether this strategy is truly beneficial for vocabulary learning, we examine it in the present studies.

### Semantic Clustering

"Semantic clustering" refers to the practice of grouping vocabulary words into different categories based on their meanings (Tinkham, 1993). Such clustering is believed to be effective for several reasons, including that presenting words in semantic clusters allows for intra-category and inter-relational reinforcement (e.g., Seal, 1991), makes the meanings of words clearer by enabling learners to notice fine-grained distinctions between words (e.g., Gairns & Redman, 1986), better reflects semantic networks in the "mental lexicon" (e.g., Aitchison, 2002), and draws attention to the semantics which may lead to deeper levels of mental processing (Erten & Tekin, 2008). On the other hand, Schneider, Healy, and Bourne (1998) found that while clustering words semantically aided initial learning, it appeared to hinder relearning a week later.

What constitutes a "semantic cluster" has, however, differed greatly across researchers. Possible constructs include, but are not limited to, near synonyms (e.g., *man, fellow,* and *guy*; (Hippner-Page, 2000), topic-related items (e.g., *crime*: *smuggling*, *jury*, and *court*; Papasanasiou, 2009), and exemplars that fall under a super-ordinate category (e.g., *fruit*: *apple*, *pear*, and *peach*) (Waring, 1997). There is some, albeit limited, evidence supporting the facilitating effects of semantic clustering (e.g., Finkbeiner & Nicol, 2003), but these conclusions have largely been drawn from examination of acquisition during training, rather than on the basis of long-term memory tests. Considerable research, however, has demonstrated that the manipulations that boost performance during training do not always boost learning (see, e.g., Soderstrom & Bjork, 2015). Furthermore, conclusions are hard to draw because the random and semantically clustered conditions compare learning of different sets of words, making it unclear whether it is the sequence that enhances learning or whether a semantically related set of words is simply easier to learn.

## The Present Studies

We conducted two studies to examine the optimal sequence of vocabulary learning. Although previous studies have used separate lists for different conditions (e.g., Tinkham, 1993, 1997), we created one single set of GRE words for each experiment, ensuring that only sequencing would differ across conditions. Moreover, to test long-term memory, rather than just performance during acquisition, we employed final criterion tests that were delayed by at least 24 hours.

In Experiment 1, we examined the optimal sequencing for learning semantically related words by comparing the efficacy of alphabetical, categorical (i.e., semantic clusters), and random sequences.

In Experiment 2, we examined the optimal sequencing for learning confusable words by comparing the efficacy of alphabetical, paired (confusable clusters), and random sequences. Whether confusable clusters or random sequencing creates more difficulties is unclear. Because confusable pairs are similar-looking and/or similar-sounding, similarities in orthography and/or phonology could well interfere with performance during training, but then might also enhance retention performance. Alternatively, an alphabetical sequence resembles a hybrid of random and paired schedules to some degree, and thus might incorporate the best (or worst) of both worlds.

## Experiment 1

The purpose of Experiment 1 is to address the first challenge of vocabulary learning—the need to distinguish between semantically related words.

### Participants and Design

Participants were 152 undergraduates from the University of California, Los Angeles (UCLA) who participated in the experiment in exchange for course credit. Eight-two of them were native English-speakers, and 70 spoke English as a Second Language (ESL; $M_{\text{length of speaking English}} = 5.34$ years). ESL students were expected to be at a decent English level due to the University's requirements. Participants were randomly assigned to one of three sequencing conditions: categorical, alphabetical, and random.

### Materials

A pool of 36 GRE word-synonym pairs was selected from GREedge, a theme-based word list. This word list contained six words from each of six different categories: *Communication, Crime & Law, Nature, Personality, Thoughts & Ideas, and Time*. College-level participants were expected to know the meaning of each paired synonym. Table 1 shows an example of the "Communication" related words. Within each category, there was one word that began with one of six initial letters—*a, c, i, m, p, and s*—which allowed us to construct the six sets of words for the alphabetical condition. Table 2 shows an example of the words that begin with the letter *a*.

Table 1
Example of one of the six-word categorical sets: words relating to *"Communication"*

| Category | Initial | GRE Word | Synonym |
|---|---|---|---|
| Communication | *a* | ***acrimonious*** | *bitter* |
| | *c* | ***circumlocution*** | *rambling* |
| | *i* | ***importune*** | *beg* |
| | *m* | ***missive*** | *letter* |
| | *p* | ***prattle*** | *babble* |
| | *s* | ***sententious*** | *pithy* |

Table 2
Example of one of the six-word alphabetical sets: words beginning with the letter *a*

| Initial | Category | GRE Word | Synonym |
|---|---|---|---|
| a | *Communication* | ***acrimonious*** | *bitter* |
| | *Crime & Law* | ***abjure*** | *withdraw* |
| | *Nature* | ***arroyo*** | *environment* |
| | *Personality* | ***abstemious*** | *restrained* |
| | *Thoughts & Ideas* | ***apotheosis*** | *exaltation* |
| | *Time* | ***antediluvian*** | *ancient* |

## Procedure

Participants were told that their task was to learn 36 GRE words. Participants were presented one GRE word-synonym pair at a time, and, for each pair, they were first asked to generate the synonym (8 sec) before they were shown the correct answer (3 sec). Therefore, the study phase consisted of tests-with-feedback trials (for a discussion of the benefits of testing and feedback on long-term retention, see Roediger & Karpicke, 2006). The six pairs of a given set (e.g., a random set, a category set, or an alphabetical set) were always presented consecutively. Each set were presented three times, with the order of pairs randomized each time (hence, the alphabetical sequence was not strictly alphabetical, but grouped words starting with the same initial letter). Thus, any given pair was presented on an average spacing interval of 5-5. Between individuals, the order of the six sets was randomized.

After they completed studying all 36 pairs (3 times each), participants were asked to predict how many of the 36 pairs they would be able to answer correctly on the test the next day. They were then informed of the three different study sequences that had been used for different participants and asked to judge which one they believed would be most effective for learning vocabulary.

Twenty-four hours later, participants were emailed a link to take the final test. The test phase consisted of two portions: First, they were asked to complete a cued-recall task in which they were presented with the GRE word and asked to generate the synonym. Second, they were given a multiple-choice in which they were given the synonym and four options. The options, illustrated in Figure 1, were constructed such that one was the correct answer, one was a wrong answer from the same category, one was a wrong answer with the same initial letter, and one was a random lure selected from the other GRE words they had studied.

Finally, we collected information about participants' GRE preparation status and language fluency and number of languages spoken.

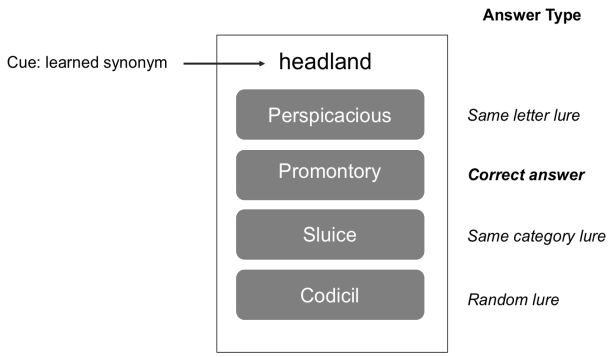


Figure 1: Example of a multiple-choice trial in Experiment 1. Participants were only shown contents inside the square. The rest is only for illustration purpose.

## Results and Discussion

Sixteen participants were removed from subsequent analysis for the following reasons: five looked up answers, seven reported serious technical issues with the experiment, and four people indicated both. Among the remaining 136 participants, 74 were native English-speakers and 62 were ESL students ($M_{\text{length of speaking English}}$ = 5.42 years).

**Acquisition** The increase in acquisition from the to the third time a given pair was presented during the study phase is shown in the left panel of Figure 2. By the end of the final trial, there was a trend for a difference between the three conditions, $F(2,133) = .2.24$, $MSE = .07$, $p = .11$. Post-hoc analyses revealed that the random condition ($M = .73$, $SD = .18$) was marginally worse than the alphabetical ($M = .79$, $SD = .18$, $p = .06$) and categorical conditions ($M = .79$, $SD = .18$, $p = .08$). No significant difference between the alphabetical and categorical conditions was obtained, $p = .96$.

**Final test** Performance on the final cued recall test, illustrated in the right panel of Figure 2, suggested some differences (although limited) between conditions, $F(2,133) = 3.105$, $MSE = .04$, $p = .048$, but the pattern was somewhat different from that of the acquisition pattern. Post-hoc pairwise comparisons revealed that participants in the alphabetical condition ($M = .50$, $SD = .03$) performed significantly better than those in the random condition ($M = .40$, $SE = .03$), $p = .02$), but neither was significantly different from the categorical condition ($M = .44$, $SD = .21$), $ps > .10$. No significant differences in performance on the multiple-choice test among conditions were observed, $F(2,133) = 1.764$, $p = .18$.

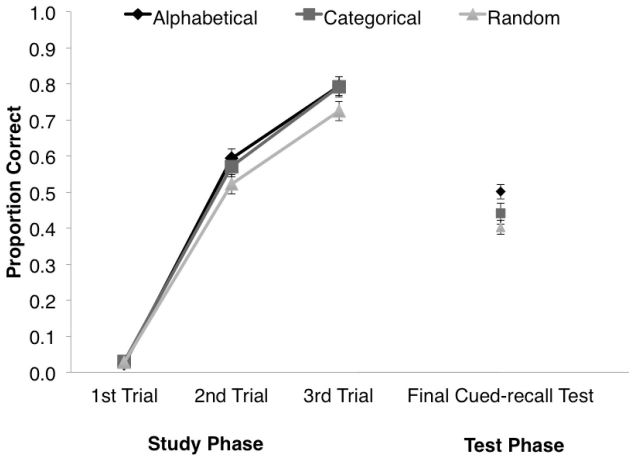


Figure 2: Study phase acquisition curves and final cued recall test in Experiment 1. Error bars represent one standard error of the mean.

**Metacognitive responses** With respect to the question "What sequence do you think is best for learning?" asked of participants at the end of the acquisition phase, 94 (70%) of the 134 participants responding reported believing that categorical clustering would lead to the best learning, 35 (26%) reported believing that a random sequence would be best, and only 5 (4%) reported believing that an alphabetical sequence would be best. These metacognitive responses did not differ by assigned condition, $\chi^2$ (4) =7.34, $p$ = .12.

Overall, while Experiment 1 neither show evidence in favor of nor against categorical sequencing, it surprisingly suggested some benefits associated with an alphabetical sequencing strategy. Participants' metacognitive beliefs, however, showed limited faith in this "new" strategy.

## Experiment 2

Experiment 2 aimed to address the second challenge in GRE word learning: students' ability to distinguish between similar-looking and/or similar-sounding words.

### Participants and Design

Participants were 112 undergraduates from UCLA who participated in the experiment in exchange for course credit. Sixty-five of those participants were native English-speakers, and 47 ESL students ($M_{\text{length of speaking English}}$ = 5.60 years). Despite being non-native English-speakers, these ELS students were expected to be at a decent English level due to the University's requirements. Participants were randomly assigned to one of three sequencing conditions: paired, alphabetical, and random.

### Materials

As in Experiment 1, we selected 36 GRE word- synonym pairs from several sites (e.g., Magoosh). However, each GRE word was also matched with another GRE word that was highly confusable. That is, 36 GRE words were broken down into 18 confusable pairs. The words in nine of these pairs shared the same initial letters (e.g., *augur-bode* and *auge*r- *drill*) and the remaining nine of these pairs did not (e.g., *astringent-bitter* and *stringent-strict*). Across all the GRE words, any initial letter was also made to appear at least three times (e.g., three words that began with 'a'). College-level participants were expected to know the meaning of each paired synonym.

In the random condition, one randomized sequence of the 36 words was created for each participant. In the paired condition, the confusable paired-GRE words were always presented consecutively (e.g., *augur-bode* followed by *auge*r- *drill*, or vice versa), and the order of pairs was randomized for each participant. In the alphabetical condition, the GRE words were presented in alphabetical order. Table 3 shows an example of two confusable pairs in alphabetical order: One same-initial pair (*veracious* and *voracious*) and one different-initial pair (*pabulum* and *vinculum*). As demonstrated below, Same-initial pairs by definition would still appear close to each other in alphabetical order (not necessarily consecutively), but different-initial pairs would for sure be shuffled.

Table 3
Example of two confusable pairs in alphabetical order: One same-initial pair and one different-initial pair.

| Initial | GRE Word | Synonym |
|---|---|---|
| p | ***pabulum*** | *sustenance* |
| … | … | … |
| v | ***veracious*** | *honest* |
| | ***vinculum*** | *bond* |
| | ***voracious*** | *greedy* |

### Procedure

The procedure of Experiment 2 was similar to that of Experiment 1 with three exceptions: (a) Instead of repeating in sets of six, all 36 pairs were presented before they were repeated, yielding an average spacing interval of 35-35; (b) the alphabetical order here was strictly alphabetical, so the order of words in each of the three cycles of 36 trials was the same; and (c) the multiple-choice test used the GRE word as the cue and presented four studied synonyms as the options (the correct answer, the synonym of the confusable GRE word, a synonym of a word sharing the same initial letter, and another random synonym).

### Results and Discussion

Sixteen participants were removed from subsequent analysis: five looked up answers, and 11 reported serious technical issues with the experiment. Among the remaining 96 participants, 57 were native English-speakers and 39 were ESL students ($M_{\text{length of speaking English}}$ = 5.76 years).

**Acquisition** The increase in acquisition from the first to the third time a given pair was presented during the study phase is shown in the left panel of Figure 3. Participants in different conditions exhibited a similar amount of accuracy boost by the end of the study phase, and the proportions of synonyms correctly recalled were not different, although performance in the alphabetical condition ($M$ = .43, $SD$ = .20) was numerically better than that in the random ($M$ = .39, $SD$ = .18) and paired ($M$ = .37, $SD$ = .18) conditions. Unlike in Experiment 1, the curves were nearly linear with no indication of deceleration in learning rates at the end of the 3rd presentation. One interpretation of this finding was that intrinsic confusion in confusable words made them naturally more difficult to learn than regular ones (indeed, accuracy rates were lower). Thus, three learning trials still left space for learning to improve at an accelerating rate before starting to slow down.

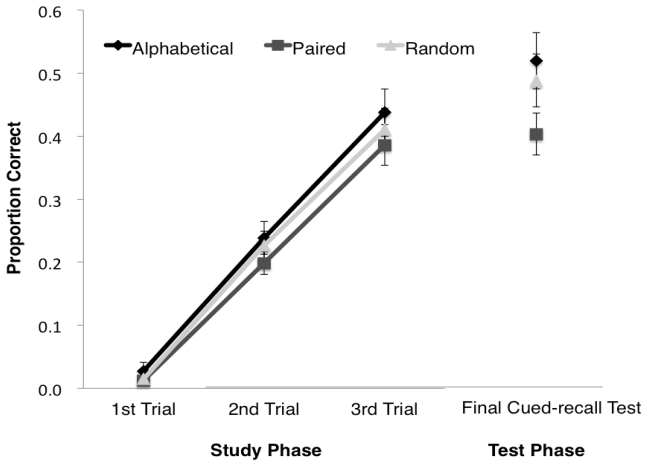


Figure 3: Study phase acquisition curves and final cued recall test in Experiment 2. Error bars represent one standard error of the mean.

**Final test** Performance on the final cued recall test is illustrated in the right panel of Figure 3. A 3 (sequencing condition) x 2 (pair type) mixed effects ANOVA performed on the cued-recall test performance, revealed a main effect of condition, $F(2,93) = 3.45$, $MSE = .10$, $p = .036$. Additionally, post-hoc pairwise comparisons revealed that the alphabetical condition ($M$ = .53, $SD$ = 25) yielded significantly better learning than the paired condition ($M$ = .39, $SD$ = .19, $p$ = .01). The random condition ($M$ = .44, $SD$ = .25) was not significantly different from either the paired condition ($p$ = .33) or the alphabetical condition ($p$ = .12).

A main effect of pair type was also observed, $F(1, 93) = 50.55$, $MSE = .01$, $p < .001$, with the GRE words from the pairs having different initial letters ($M$ = .50, $SD$ = .23) being learned significantly better than the words from pairs sharing the same initial letter ($M$ = .39, $SD$ = .24). There was no interaction between condition and pair type, $F(2,93) = 1.12$, $MSE = .01$, $p = .33$.

Figure 4 shows multiple-choice test performance by condition and pair type. A 3 (sequencing condition) x 2 (pair type) mixed effects ANOVA performed on the multiple-choice test performance revealed similar patterns to that obtained for the cued-recall test performance. Again, there was a trend-level effect of condition, $F(2,93) = 2.22$, $MSE = .10$, $p = .12$. Pairwise comparisons revealed a similar pattern as found with the cued-recall test: the alphabetical condition ($M$ = .77, $SD$ = .03) was marginally better than the paired ($M$ = .70, $SD$ = .03, $p$ = .06) and the random ($M$ = .70, $SD$ = .03, $p$ = .08) conditions.

Again, a significant effect of pair type, $F(1,93) = 53.03$, $MSE = .01$, $p < .001$, was observed, with the different-initial letter pairs ($M$ = .77, $SD$ = .18) learned better than the same-initial letter pairs ($M$ = .67, $SD$ = . 18). Finally, there was no condition x pair type interaction, $F(2,93) = 1.57$, $MSE = .01$, $p = .21$

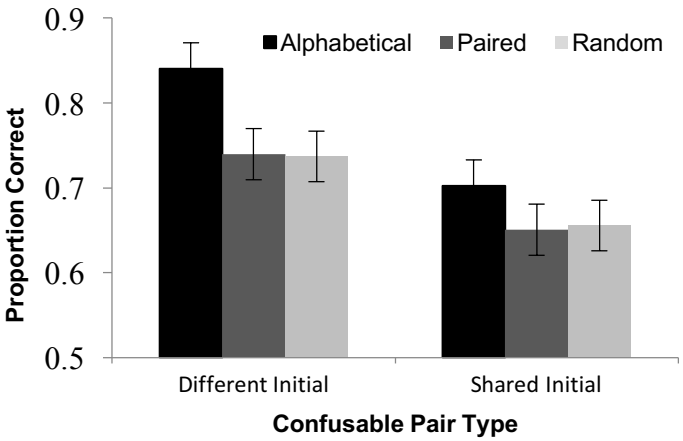


Figure 4: Multiple-choice test performance on Experiment 2 by condition and pair type. Error bars represent one standard error of the mean.

**Metacognitive Responses** When asked what sequence they believed was best for learning, 42 (44%) participants reported that they believed that pairing confusable words would lead to the best learning, 39 (41%) a random sequence would be best, and 15 (16%) believed that an alphabetical sequence would be best. These metacognitive responses did not differ by experienced condition, $\chi^2$ (4) = 6.30, $p$ = .18.

Overall, the findings of Experiment 2 demonstrate that when trying to learn confusable words, contrary to many people's belief (a majority of 84% thought that either random or paired order would be the best), an alphabetical order led to the greatest learning.

## General Discussion

The current study is one of the very few instances where the alphabetical order has been studied. Experiments 1 and 2 revealed some preliminary evidence on the merits of an

alphabetical word learning sequence. There was some suggestion that this strategy can be equally, if not more effective than strategies that are traditionally considered good (e.g., categorical or paired clustering).

It is worth pointing that while both being referred to as an "alphabetical sequence," the structures of the alphabetical order in Experiments 1 and 2 were not identical. In Experiment 1, the structure was not strictly alphabetical, but rather grouped words beginning with the same initial letter together. Participants went through an alphabetical cluster three times, so they were only exposed to one initial letter at any given time. Thus, it was more of an "alphabet-informed grouping." We speculate that one possibility is that this grouping offers an optimal level of support and difficulty: The support from the small degree of structure (i.e., shared initial letters) may ease extraneous cognitive processing load from the difficult word learning task; the otherwise-random nature of the words maintains a sense of difficulty to promote deeper processing. An alternative explanation could simply be that clustering alphabets gives learners another, redundant, cue to aid learning. The initial letter is another cue to the context in which words were learned. For example, the initials may have already narrowed down the list from 36 to six words, which may well require less cognitive effort to identify the correct answer, given that words in the same initial cluster are not too similar orthographically/phonologically. Therefore, while shared categories may be a contextual cue in categorical clustering, list position could be a contextual cue in an alphabetical sequencing. Each clustering method has its own advantage and both lead to respectable results.

In Experiment 2, the alphabetical sequence was truly alphabetical, with half same-initial and half different-initial confusable pairs. The advantage of this strategy was primarily reported in learning different-initial than same-initial pairs. Thus, the benefit of alphabetizing a confusable word-list was observed at a global (the entire list with multiple initials) rather than a local (same-initial pairs grouped to the same alphabetical cluster) level. Consequently, an alphabetical condition represents a hybrid of randomization of the easier pairs (i.e., different-initial pairs) and confusability- clustering of the more difficult pairs (i.e., same initial pairs), which may have incidentally created a degree of "desirable difficulty" (Bjork, 1994). Alternatively, because learners in Experiment 2 learned all 36 words before repeating, those in the alphabetical condition may have simply used list position as a contextual cue to aid learning. For example, they may have linked to-be-learned words to some known knowledge (i.e., an alphabetical list) to help memorize.

We have already demonstrated some caveats when alphabetizing to-be-learned words. As suggested above, there might be differences in the role of "alphabet-informed grouping vs. alphabetical order. It is therefore unclear how far the benefits of an alphabetical sequence would extend. In the present studies, we presented participants relatively difficult words in a language that they were familiar with Hence, knowledge about word etymology (e.g., Latin roots) or even passing familiarity with the to-be-learned words themselves (note the large jump in performance between the first and second trial of the study phase) may have supported learning. It is unclear how the optimal sequence might change for foreign language learning, where learners do not have this type of background knowledge.

As part of a critical factor in high-stakes tests, GRE word learning is a major concern of many students. The present studies extend the literature by suggesting the powerful potential of learning words in alphabetical order, a widely used, yet under-investigated alternative to clustering or random sequencing. Whether, however, the benefits of an alphabetical sequencing might generalize to vocabulary words that are less difficult and abstract than GRE words remains to be seen. In the meantime, it appears that generations of Chinese students who have been learning English vocabulary words grouped alphabetically may not have been engaging in what may seem, by some arguments, to be a misguided practice.

## References


Aitchison, J. (2002). *Words in the Mind: An Introduction to the Mental Lexicon (3rd Ed.).* Blackwell Publishers: Great Britain.

Bjork, R.A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Erten, I.H., & Tekin, M. (2008). Effects on vocabulary acquisition of presenting new words in semantic sets versus semantically-unrelated sets. *System, 36* (3), 407-422.

Finkbeiner, M., & Nicol, J. (2003). Categorical category effects in second language word learning. *Applied Psycholinguistics, 24*(3), 369-383.

Gairns, R., & Redman, S. (1986). *Working with words: A guide to teaching and learning vocabulary*. New York: Cambridge University Press.

Hippner-Page, T. (2000). *Semantic clustering versus thematic clustering of English vocabulary words for second language instruction: Which method is more effective?* Retrieved from ERIC database. (ED445550)

Seal, B. D. (1991). Vocabulary learning and teaching. In M. Celce–Murcia (Ed.), *Teaching English as a second or foreign language* (2nd ed., pp. 296–311). Boston: Heinle & Heinle.

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr (1998). Contextual interference effects in foreign language vocabulary acquisition and retention. In A. F. Healy & L. E. Bourne, Jr. (Eds.), *Foreign Language Learning: Psycholinguistic Studies on Training and Retention* (pp. 77-90). Mahwah, NJ: Erlbaum.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*, 176-199.

Tinkham, T. (1993). The effects of semantic clustering on the learning of second language vocabulary. *System, 21,* 371–380.

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research, 13*, 138–163.

Waring, R. (1997). The negative effects of learning words in semantic sets: A replication. *System, 25*, 261–274.