US-China Education Review A, April 2026, Vol. 16, No. 4, 242-249
doi: 10.17265/2161-623X/2026.04.009

An Empirical Study of the Effects of Multimodal Presentation on Korean-Speaking Children's Acquisition of Chinese Vocabulary—Evidence From Immediate and Delayed Testing

WANG Jiahao
Beijing Language and Culture University, Beijing, China
LIU Hongli
Beijing Normal University, Beijing, China

The effectiveness of multimodal resources in second-language vocabulary teaching remains a key issue, particularly regarding how different presentation modes influence learning outcomes. However, existing studies largely focus on short-term learning outcomes and assess vocabulary along a single dimension, obscuring how multimodal presentation affects vocabulary acquisition and retention. In this study, we examine the effects of three-modal combinations empirically—text, image, and audio—on Chinese vocabulary learning by Korean children. Immediate and delayed tests assessed three dimensions: meaning, word form, and grammatical collocations. Results showed that in the immediate test, the text-image group led in form/meaning, while text-image-audio excelled short-term in collocations. In the delayed test, the text-image group maintained stable performance and significantly outperformed the other two groups, whereas the text-image-audio group showed a marked decline. The findings indicate that the text-image integration more effectively supports long-term memory of Chinese vocabulary in children, whereas multimodal stacking may increase cognitive load and impair retention. This study provides pedagogical guidance for applying multimodal resources in children's L2 Chinese vocabulary instruction.

Keywords: multimodal presentation, Chinese vocabulary learning, Korean children, teaching Chinese as a second language

Introduction

Vocabulary proficiency is a key metric of successful L2 acquisition (Coady & Huckin, 1997). As the initial cognitive step, vocabulary presentation is central to L2 vocabulary research (Meara, 1980). Mayer (1997) noted that multimodal resources (visual, auditory, pictorial, linguistic) stimulate selective attention, organization, and integration. With technological advances, multimodal resources and language teaching increasingly converge. Multimodal vocabulary presentation refers to the deliberate instructional use of multiple semiotic systems—text, images, audio, video (Liu & Ouyang, 2019). A core question arises: What is multimodal instruction's actual impact, and which modality combination is optimal?

WANG Jiahao, College of International Chinese Language, Beijing Language and Culture University, Beijing, China.
LIU Hongli, College of International Chinese Language Education, Beijing Normal University, Beijing, China.

Scholarly consensus on multimodal presentation's effects remains divided. Some studies confirm its benefits; others indicate improper combinations may induce cognitive overload and hinder learning. Dual-coding theory (Paivio, 1986) posits two independent processing channels—verbal and imagery—which mutually activate when information is presented visually (e.g., pictures) and verbally (e.g., text), enhancing processing and long-term storage. This supports multimodality's positive potential. Mayer (1997) added that multimodal resources engage learners in active selection, organization, and integration. Empirical evidence corroborates this. Liu and Ouyang (2019) found adult English learners using text + image + audio outperformed text-only groups in immediate meaning recognition for 10 target words. Recent work extends to specific populations, such as ethnic minority children (Zhang & Xu, 2022), and AI (artificial intelligence) tutoring systems show promise in multimodal language learning (Liu et al., 2025).

However, not all research supports multimodality's benefits. Cognitive load theory (Sweller, van Merrienboer, & Paas, 1998) suggests improper modality combinations may increase extraneous load and impede learning. Lin and Yu (2017) similarly argued that poor instructional design elevates extraneous load, causing interference. Zhang (2000) found that pictures and animations could distract learners, making text-only presentation more effective. Multimodal effects may also vary by context and learner group (Moritz & Marinie, 2023). These discrepancies indicate that multimodal instruction's effectiveness is not absolute but moderated by modality combination, learner cognition, and instructional context.

The debate also involves defining "vocabulary learning" itself. Vocabulary knowledge is widely viewed as multidimensional. Nation (2001) delineated three core dimensions—form, meaning, and use—providing critical analytical vectors. Qian (2002) further emphasized depth: mastery of spelling, semantics, syntax, and collocation beyond mere size. Given this multidimensionality, assessment should transcend single dimensions (Schmitt, 2010). Yet existing multimodal research largely focuses on short-term meaning recognition, failing to clarify differential impacts on form and collocation, and lacking longitudinal, multidimensional retention studies.

These controversies underscore the need for refined empirical research. This study addresses current limitations through three innovations: (1) assessing meaning, form, and collocation comprehensively via immediate and delayed tests to reveal differential effects on acquisition and retention; (2) using the Beijing Language and Culture University Intelligent Chinese Teaching System (ICTS) to generate standardized, modality-controlled courseware, enhancing validity and replicability; (3) focusing on Korean children (ages 9-12, HSK 2-3), an understudied population compared to adult English learners.

Accordingly, this study addresses two research questions:

(1) How do text + image (A), text + audio (B), and text + image + audio (C) affect Chinese vocabulary learning in Korean children?

(2) Do these three modes differentially affect comprehension and memory of meaning, form, and grammatical collocations? If so, how?

Research Methodology

Experimental Design

This study employed an empirical research design to investigate the impact of three distinct multimodal presentation modes on vocabulary acquisition. Participants were allocated to one of three groups: Group A (text + images), Group B (text + audio), or Group C (text + images + audio). Vocabulary learning was assessed through

immediate test, conducted directly after the learning intervention, and delayed testing, administered one-week post-intervention. The evaluation encompassed three dimensions: semantic meaning, word form, and grammatical collocations.

Participants

A one-month empirical study was conducted using native Korean-speaking children recruited online as participants. The participants, all native Korean speakers aged 9-12, possessed an elementary Chinese proficiency level (HSK 2-3) and had been studying Chinese for approximately six months. A total of 45 participants completed all assessments and were randomly assigned to three groups: A, B, and C. The basic demographic information for the participants by group is presented as follows:

Table 1

Participant Allocation by Group

Grouping of Chinese multimodal lexical presentation	Number of people
Text + image (Group A)	15
Text + audio (Group B)	15
Text + image + audio (Group C)	15

All participants had no intellectual or learning disabilities. Prior to the experiment, they were not informed of the test details to avoid deliberate rehearsal. The experimental materials were produced using "ICTS", which provides standardized templates to facilitate the generation of multimodal instructional materials incorporating text, images, and audio. Participants had not previously used the system for learning and were only exposed to materials produced via the system.

Pilot Study

A one-month pilot study determined the optimal learning load. Two conditions were tested: five and 15 concrete nouns per set, presented in text + image format. The five-word condition produced a ceiling effect (excessive accuracy); the 15-word condition proved too difficult. Consequently, 10 words per set was selected for the formal experiment to maintain accuracy within an appropriate range.

Experimental Materials

Material preparation was based on "ICTS", ensuring consistency of multimodal instructional materials, including the Target Vocabulary Screening Test, the Immediate Test, and the Delayed Test. See the Section of Experimental Procedures for the specific usage requirements.

Experimental Procedures

The study was conducted in four stages:

(1) Target word screening: A pretest excluded known vocabulary, yielding 10 unfamiliar words (e.g., airplane, train, transportation) appropriate for children's proficiency.

(2) Vocabulary presentation: In an online classroom, Groups A, B, and C viewed text (word, pinyin, POS, meaning, example). Group A added an image; Group B added audio; Group C added both. Materials were ICTS-generated. Presentation lasted 10 minutes for self-guided study.

(3) Immediate testing: Three five-minute tasks: (a) Word form—circle the learned word among orthographic distractors; (b) word meaning—match three target words (plus two distractors) to definitions (Read, 2000); (c) Grammatical collocation—fill blanks requiring syntactic knowledge.

(4) Delayed testing: Administered one week later without prior notice. Content and scoring mirrored the immediate test; only word order changed.

Throughout the study, teachers only presented materials, avoiding intervention to ensure objectivity.

Experimental Results

Immediate Test

The immediate test was conducted online via Tencent Meeting. Participants used personal devices; the examiner screen-shared the test (PDF). Each section lasted five minutes; answers were written on prepared sheets, photographed, and uploaded. Data were independently double-entered, cross-checked, and analyzed in SPSS 26.0. A 3 (group: text + image, text + audio, text + image + audio) × 3 (task: form, meaning, collocation) repeated measures ANOVA assessed immediate scores. One-way ANOVA compared total scores across groups, with Bonferroni post-hoc correction. Figure 1 shows mean scores by group and dimension.

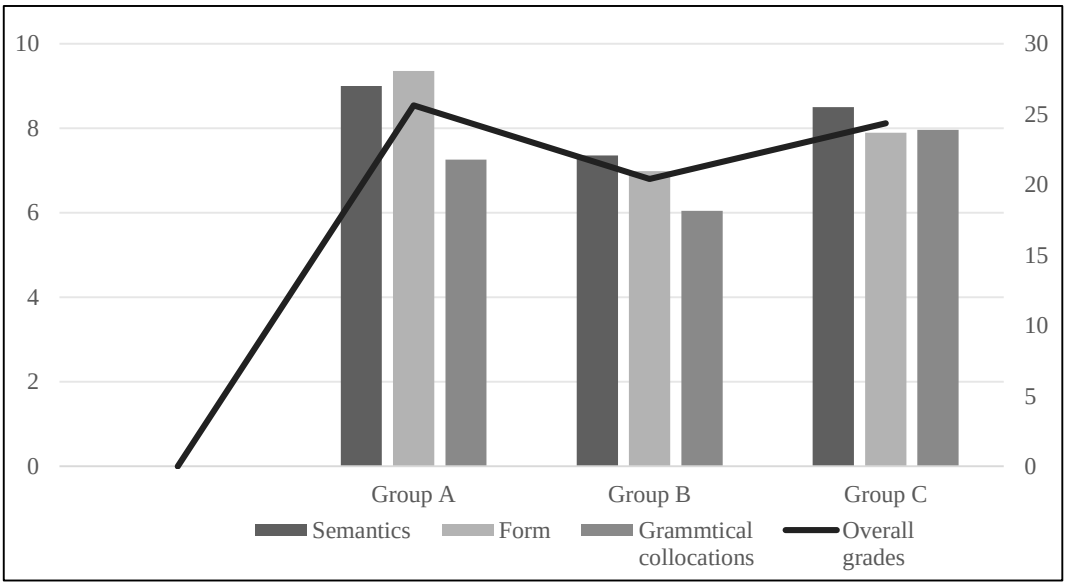

Figure 1. Comparison chart of the scores of each dimension and the overall score for the three groups of immediate tests.

Table 2

Descriptive Statistics of the Scores for Each Dimension in the Three Groups of Immediate Tests (M ± SD)

	Group A (N = 15)		Group B (N = 15)		Group C (N = 15)	
	M	SD	M	SD	M	
Word form	9.000	2.398	7.355	3.024	8.500	2.366
Meaning	9.355	1.245	6.987	2.359	7.890	2.106
Grammatical collocations	7.255	3.634	6.045	3.055	7.960	1.264

The overall immediate test scores ranked as A (25.61) > C (24.35) > B (20.39). An ANOVA revealed a significant main effect of group ($F(2,42) = 18.73$, $p < 0.001$, $\eta^2 = 0.47$). Post comparisons with Bonferroni correction showed that the immediate total scores for A and C were both significantly higher than B ($p < 0.001$), while A and C did not differ significantly ($p > 0.05$). This indicated that the text + image and text + image + audio conditions yielded superior immediate learning outcomes overall relative to text + audio alone.

According to the table, the ANOVA showed a significant three-way interaction among group, task type, and test time ($F(4,84) = 3.65$, $p < 0.05$, $\eta^2 = 0.15$). Simple effects analyses further showed that, in the immediate test for grammatical collocations, both A and C outperformed B ($p < 0.01$), while A and C did not differ ($p > 0.05$); in word form and meaning tasks, A and C also outperformed B ($p < 0.01$), with A slightly higher than C. These results indicate that the text + image pairing yields immediate advantages in word form and meaning learning, while the text + image + audio pairing yields grammar-collocation performance comparable to text + image, yet both outperform text + audio alone.

Delayed Test

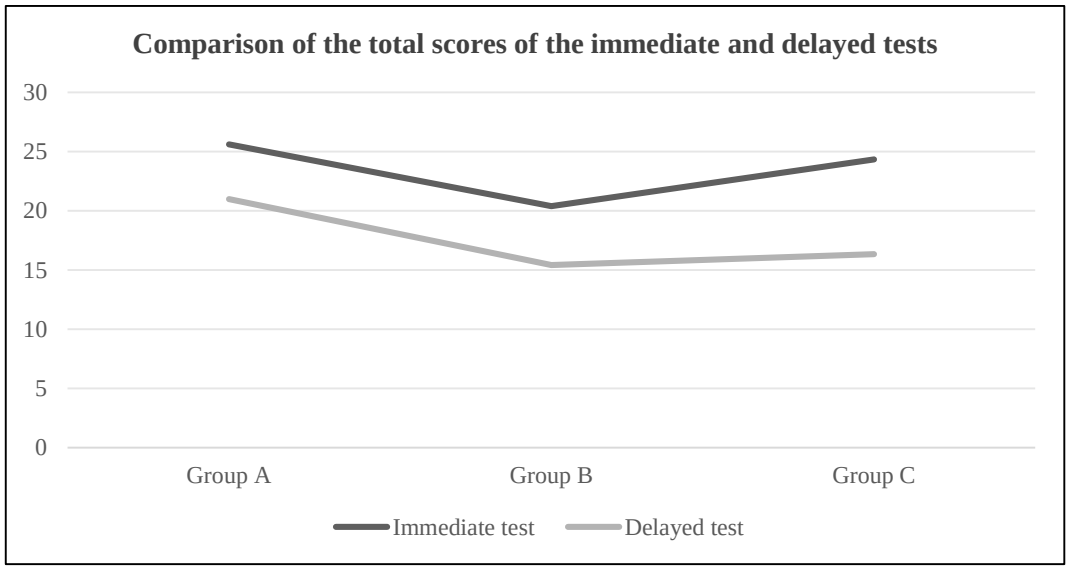

Figure 2. Comparison chart of the total scores of the three groups of immediate and delayed tests.

Table 3

Descriptive Statistics of the Scores for Each Dimension in the Three Groups of Delayed Tests (M ± SD)

	Group A (N = 15)		Group B (N = 15)		Group C (N = 15)	
	M	SD	M	SD	M	SD
Word form	7.650	2.568	6.074	1.025	6.215	2.335
Meaning	7.985	3.245	5.365	2.329	6.034	1.106
Grammatical collocations	5.346	1.632	3.984	1.054	4.096	3.274

Table 4

Descriptive Statistics of the Scores of the Three Groups Under Different Tasks and Testing Times (M ± SD)

Task type	Test time	Group A (text + image)	Group B (text + audio)	Group C (text + image + audio)
Word form	immediate	9.45 ±1.21	7.82 ±1.34	8.76 ±1.43
	delayed	8.12 ±1.56	6.21 ±1.89	6.84 ±1.67
Meaning	immediate	9.87 ±0.98	8.05 ±1.22	9.23 ±1.11
	delayed	8.65 ±1.34	6.54 ±1.78	7.21 ±1.45
Grammatical collocations	immediate	6.29 ±1.45	4.51 ±1.67	6.36 ±1.32
	delayed	5.21 ±1.67	3.67 ±1.89	4.28 ±1.56

Notes. Each group is scored out of 10 points, and the total score is 30 points (10 points for the word form, 10 points for the meaning, and 10 points for grammatical collocations).

In the delayed test, the total scores of the three groups declined relative to the immediate test, but Group A

remained ahead (20.98), significantly higher than Group B (15.42) and Group C (16.35), with Group C marginally higher than Group B. A repeated-measures ANOVA revealed a significant main effect of testing time ($F(1,42) = 89.34$, $p < 0.001$, $\eta^2 = 0.68$), with immediate performance surpassing delayed performance. The interaction between group and testing time was significant ($F(4,42) = 7.89$, $p < 0.01$, $\eta^2 = 0.27$). Simple effects showed that, in the delayed test, Group A scored significantly higher than Group B ($p < 0.001$) and Group C ($p < 0.01$), while there was no significant difference between Group C and Group B ($p > 0.05$). This indicates that the text + picture condition yields a clear advantage for memory retention, whereas the text + picture + audio condition, though comparable to Group A in the immediate test, loses its advantage after the delay, showing no significant difference from the text + audio group.

From the dimensional perspective, for word form, lexical meaning, and grammatical collocation tasks, the delayed test scores followed A > C > B. The interaction between task type and testing time was significant ($F(2,84) = 5.42$, $p < 0.01$, $\eta^2 = 0.11$). Further analyses showed that, in the word form and lexical meaning tasks, the declines in the delayed test were smaller (a decreases of 14.1% and 12.4%, respectively); in the grammatical collocation task, Group C exhibited the largest decline (32.7%), with Group A decreasing 17.2% and Group B 18.6%. The three-way interaction was significant ($F(4,84) = 3.65$, $p < 0.05$, $\eta^2 = 0.15$). Simple effects indicated that Group A maintained the most stable performance in word form and lexical meaning tasks; Group C, while performing well in the immediate test for grammatical collocation, showed the most pronounced decline in the delayed test; Group B performed weakest across all tasks, with notable declines in lexical meaning and grammatical collocation tasks.

Integrated results from immediate and delayed testing show that text + image (Group A) yields the strongest memory performance across all dimensions of vocabulary learning, with particularly stable outcomes in word form and word meaning tasks. Text + image + audio (Group C) performs comparably to Group A in the immediate test but exhibits pronounced memory decay in the delayed test, suggesting that information overload may impair long-term retention. Text + audio (Group B) shows the weakest overall performance, indicating that reliance on a single auditory channel provides limited benefits for children's learning of Chinese character vocabulary.

Discussion

The present study employed immediate and delayed tests to examine the performance of children with Korean as a native language background under three multimodal presentation modes—text + image, text + audio, and text + image + audio—in terms of meaning, word form, and grammatical collocations. The results indicate that, with respect to the word form and meaning, Group A (text plus image) performed best across both tests; with respect to grammatical collocation, Group C (text plus image plus audio) was optimal in the immediate test, but Group A outperformed others in the delayed test. The following discussion integrates theoretical and empirical perspectives.

Effects of Multimodal Presentation Modes

In the text + image mode, images reinforce text learning through visual representation, enhancing attention to phonology, orthography, semantics, and pragmatics. Co-encoded images cue lexical retrieval, aiding acquisition and retention. Multimodal instruction engages multiple senses, providing rich interaction that fosters internalization (Cai et al., 2021). Text + image aligns with dual coding theory: Learners link verbal and non-

verbal representations, improving processing, storage, and long-term retention (Li, 2022).

In text + audio mode, research on non-alphabetic visual-semantic access shows orthography, not phonology, and critically drives semantic retrieval. Audiovisual input provides more authentic, vivid exposure and richer multimodal context than audio alone, enhancing comprehension (Fu & Li, 2020). Yet for beginning Chinese learners, homophones and tonal semantics may limit audio's compensatory value.

In text + image + audio mode, multimodal input can improve learning, but simultaneous visual-auditory delivery risks overload. Studies show text + image + audio performs well immediately but declines sharply on delayed tests, suggesting excessive modalities increase cognitive load and impair retention (Li, 2022). Zhou (2020) confirms that multimodal overload heightens cognitive load, reducing decoding capacity.

Overall, text + image's stable advantage is explained by dual coding theory (Paivio, 1986): images and text engage distinct imaginal and verbal systems, dual encoding aiding processing and storage. For Korean children, Chinese characters as ideographs enable direct orthographic-semantic access, reinforced by images; audio may interfere via homophones and tonal complexity. Text + image + audio may aid short-term collocation learning, but cognitive load theory indicates excessive modalities increase extraneous load, causing delayed-test decline and impaired retention.

Differences in Word form, Meaning and Grammatical Collocations

For word form, both immediate and delayed scores ranked A > C > B. Visual-semantic processing of English largely supports phonological mediation (Wilson et al., 2011, p. 724), whereas Chinese character recognition favors direct access (Wang, 2011, p. 297). As a logographic script, Chinese lacks systematic grapheme-phoneme links, requiring non-phonological channels for semantic activation. Compared to audio, images better reinforce form-meaning connections.

For meaning, scores again followed A > C > B. This can be explained through embodied cognition and child development. Embodied cognition holds that perception, environment, body, and cognition are integrated (Feng & Zou, 2019). Ages 6-18 mark the abstract logical thinking stage, with growing generalization capacity (Xu, 1994). Participants aged 9-12 are early in this stage, with nascent abstract generalization. Images, perceptually concrete with salient features, help children extract "prototypes", and assimilate information through "prototype typification". Integrating lived experience, children build a "brain-body-environment" cognitive system, deepening lexical processing and enhancing outcomes.

For grammatical collocations, Group C performed best immediately, likely because dual visual-auditory channels captured attention, stimulated divergent thinking, and activated associative semantic networks, temporarily boosting performance.

Yet Group C's delayed scores fell below Group A's, reaffirming cognitive load theory: Information overload disperses processing depth. Though short-term associations are stimulated, deep encoding and long-term storage suffer.

Suggestions and Limitations

Based on the findings, the following recommendations are proposed:

(1) Prioritize the use of text-image combinations in Chinese character instruction to strengthen the association between images and character forms. Selected images should accurately represent lexical meaning and highlight distinctive features, thereby enhancing comprehensible input.

(2) Avoid the mechanistic superposition of multiple modalities. Instead, modalities should be judiciously

integrated according to learners' cognitive needs to reduce extraneous cognitive load and improve learning efficiency.

Furthermore, this study has certain limitations regarding sample size, control of variables, and assessment dimensions. Future research could expand the sample scope and incorporate neuroimaging techniques to further elucidate the underlying cognitive mechanisms of multimodal learning.

References

Cai, S., Jiao, X. Y., Yang, Y., Jiang, L. F., & Yu, S. Q. (2021). Practice of multimodal smart classroom in 5G environment. *Modern Distance Education Research, 33*(5), 103-112.

Coady, J., & Huckin, T. (1997). *Second language vocabulary acquisition: A rationale for pedagogy*. Cambridge: Cambridge University Press.

Feng, Y., & Zhou, R. (2019). The embodied cognition effect in second language verb processing under the spatial cueing paradigm. *Foreign Language Research, 206*(1), 71-78.

Fu, X., & Li, A. (2020). A review of research on audiovisual multimodal input in second language. *Journal of Yunnan Normal University (Teaching Chinese as a Foreign Language Edition), 18*(1), 7-16.

Li, W. (2022). An empirical study on the effects of multimodal vocabulary presentation on high school students' vocabulary memory. Master's thesis, Northwest Normal University.

Lin, C. C., & Yu, Y. C. (2017). Effects of presentation modes on mobile-assisted vocabulary learning and cognitive load. *Interactive Learning Environments, 25*(4), 528-542.

Liu, J., & Ouyang, J. (2019). Effects of vocabulary presentation modes and cognitive styles on second language vocabulary learning efficacy. *Journal of Shandong Normal University (Basic English Education), 21*(3), 19-27.

Liu, Z., Lin, G. Y., Tan, H. L., Zhang, H. Y., Lu, Y. F., Gao, X. X., … Chen, N. F. (2025). SingaKids: A multilingual multimodal dialogic tutor for language learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Vol. 6: Industry Track, pp. 1244-1253). Vienna, Austria.

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist, 32*, 1-19.

Meara, P. (1980). Vocabulary acquisition: A neglected aspect of language learning. *Language Teaching and Linguistics: Abstracts, 13*(4), 221-246.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Paivio, A. (1986). *Mental representations: A dual coding approach.* New York: Oxford University Press.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513-536.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Schmitt, N. (2015). *Researching vocabulary: A vocabulary research manual*. Beijing: Foreign Language Teaching and Research Press.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296.

Wang, K. (2011). An electrophysiological investigation of the role of orthography in accessing meaning of Chinese single-character words. *Neuroscience Letters, 487*(3), 297-301.

Wilson, L. B., Tregellas, J. R., Slason, E., Pasko, B. E., & Rojas, D. C. (2011). Implicit phonological priming during visual word recognition. *NeuroImage, 55*(2), 724-731.

Xu, Z. Y. (1994). The relationship between children's language and cognitive (thinking) development. *Acta Psychologica Sinica, 26*(4), 347-353.

Zhang, J., & Xu, H. (2022). Multimodal strategies for Chinese language teaching in Korean ethnic primary schools. *Sinogram Culture, 34*(10), 107-108 + 114.

Zhang, P. (2000). The effects of animation on information seeking performance on the World Wide Web: Securing attention or interfering with primary tasks? *Journal of the Association for Information Systems, 1*, 1-28.

Zhou, Z. (2020). Effects of input modality and practice frequency on interpreting quality: A study based on cognitive load theory. *Language Education, 8*(1), 8-14.