The Role of Online Assignments and Feedback to Chinese Language Learning Efficiency

by
Xiwen Lu

A Dissertation
Submitted to the Faculty
of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy
in
Learning Sciences & Technologies
August 2023

APPROVED:

Prof. Neil Heffernan, Advisor, Worcester Polytechnic Institute
Prof. Stacy Shaw, Worcester Polytechnic Institute
Prof. Adam Sales, Worcester Polytechnic Institute
Prof. Shijuan Liu, Indiana University of Pennsylvania

Abstract

The projects in this dissertation focus on the efficiency of feedback on Chinese as a foreign language (CFL) assignments using an intelligent tutoring system (ITS), with a particular interest in understanding how to provide feedback effectively to students to promote CFL learning. The first three projects compared Chinese learning efficiency between the ITS and traditional handwriting methods. The initial project assessed online Chinese word recognition challenges. The second and third projects delved deeper into Chinese word recognition with ITS, examining the effectiveness of ITS versus handwriting. These projects, comprising five experiments, collectively confirmed that recognizing Chinese words through ITS is more efficient than handwriting. The fourth and fifth projects examined the efficiency of feedback timing in CFL learning with ITS. The fourth project compared immediate and delayed feedback, demonstrating that immediate feedback outperformed delayed feedback in conceptual knowledge learning. The fifth project focused on immediate feedback timing, revealing that end-of-question feedback was more effective than end-of-assignment feedback. Based on the previous projects, the final project focused on immediate feedback through semi-open-ended questions in ITS for CFL. A teacher survey and in-class experiment revealed that providing only correct responses led to the highest learning gains, requiring less time for feedback and enhancing student learning judgment. This dissertation offers insight into effective teaching practices in foreign language education and ITS interventions. It contributes to developing efficient approaches to CFL instruction.

Acknowledgments

This dissertation represents the culmination of my seven-year journey at Worcester Polytechnic Institute as a part-time PhD student in learning sciences and technology. I want to express my sincerest gratitude to my committee members: Neil Heffernan, Adam Sales, Stacy Shaw, and Shijuan Liu. Their involvement and dedication played a pivotal role in my journey. I am immensely grateful for the invaluable knowledge and wisdom I have gained from them.

Particularly, I want to express my heartfelt appreciation to my advisor, Neil Heffernan, who embraced my unique research background and generously supported my endeavors. His openness and trust granted me the opportunity to pursue my passion.

I would also like to extend my heartfelt thanks to my colleagues, friends, and students who contributed to the success of my research. Their commitment to my experiments and support during my busiest times were invaluable.

Last, this achievement would not have been possible without the unwavering support of my family. Above all, I thank my husband, Zhilong, for his endless support. When I hesitated to start my PhD, Zhilong steadfastly supported me. He willingly took on additional family responsibilities during overwhelming times, granting me more time to focus on my research. I also thank my children, Alex and Claire, for always forgiving me for being busy and uplifting my spirits. Moreover, I thank my parents for supporting me unwaveringly in pursuing my dreams. I am overjoyed to share my graduation celebration with all of you. Thank you all for being an integral part of my journey.

Table of Contents

Chapter 1. Introduction

Many speakers of non-logographic languages believe that the Chinese language is difficult to learn. As a logographic language, Chinese uses characters, which are often only vaguely related to the meaning or pronunciation of the words. In order to successfully read Chinese, students studying the language must memorize all the three aspects of each word: visual (orthography), pronunciation (phonology), and meaning (semantics), and practice combinations of these aspects for reading and writing (Shen, 2004).

Orthographic elements have been shown to increase the burden of retrieval and retention (Chinese Language Committee, 2009). Handwritten character practice has been isolated as the most time-consuming activity for Chinese as foreign language (CFL) learners (Walker, 1989), significantly slowing the learning process and preventing students from engaging in meaningful communication, especially in the earliest stages of learning (De Francis, 1984; Allen 2008). Since Chinese characters are not directly related to pronunciation or meaning to improve Chinese reading and writing ability, students historically were tasked with copying characters by hand repetitively in order to then remember and output them. Technological development has brought new opportunities for Chinese learners. With the popularization of computers and the internet, as well as the help of the pinyin system, students can easily input the pinyin on keyboards and then select the correct words to complete writing tasks. This kind of writing is commonly called "digital writing (电写）"(Xie, 2014). Digital writing will likely become a weapon that subverts the way the Chinese language is learned as a second language.

As a CFL instructor by profession and observing students arrive at my class with passion for learning, yet flinching in the face of the Chinese language, I established my career trajectory to improve students' learning efficiency. I sought opportunities to allow my students to believe

that CFL learning is more similar to a pleasant stroll in the park rather than a strenuous climb. This goal motivated me to undertake an exploration of the learning sciences and technology field. I am deeply convinced that technology can change and improve the efficiency of CFL.

As I received more exposure to different aspects of learning sciences and technology, I became interested in the learning procedure and how various kinds of student interaction affect the learning results, as well as the psychological reasons behind these learning behaviors. The first thing I have noticed throughout my research relates to the timing of feedback. The timely feedback function of online learning platforms provides great convenience for the teaching and learning of CFL. The effectiveness of feedback timing on online platforms has been extensively studied in the learning science field (Butler & Woodward, 2018; van der Kleij et al., 2015). Some studies support immediate feedback, while others support delayed feedback (Attali & van der Kleij, 2017; Corral et al., 2021, Lefevre & Cox, 2017; Mullet et al., 2014; Sinha & Glass, 2015; van der Kleij et al., 2012). While most of these studies are based on STEM learning, there is limited research conducted on foreign language learning. In addition, the impact of learning science and technology on language learning, especially Chinese, is also a field full of unknowns and challenges. Therefore, substantial research is needed on CFL online learning, and it is critical to provide effective instructional support that promotes students' CFL online learning that occurs outside of the classroom setting.

Appropriate online assignments and feedback have the potential to enhance CFL learning efficiency, both in and outside of classrooms. Informed by Stephen Krashen's theory of second language acquisition (Krashen, 1988) and Long's Interaction Hypothesis (Long, 1996), I investigate how to improve the learning efficiency of CFL learning through online assignments.

My goal is to help students overcome the difficulty of learning Chinese and make the learning process more efficient.

Specifically, I present six projects designed to advance the efficiency of Chinese learning.

The first project explores the difficulties of learning word recognition online. Because Chinese words are not phonetic, Mandarin Chinese learners must construct six-way mental connections in order to learn new words, linking characters, meanings, and sounds. Very little research has focused on the difficulties inherent to each specific component involved in this process, especially within digital learning environments. The manuscript *Understanding the Complexities of Chinese Word Acquisition Within an Online Learning Platform* (Lu, Ostrow, & Heffernan, 2019a) examines Chinese word acquisition within ASSISTments, an online learning platform commonly known for mathematics education. Students were randomly assigned to one of three conditions in which researchers manipulated a learning assignment to exclude one of three bi-directional connections thought to be required for Chinese language acquisition (i.e., sound-meaning and meaning-sound). Researchers then examined whether students' performance differed significantly when the learning assignment lacked sound-character, character-meaning, or meaning-sound connection pairs, and whether certain problem types were more challenging for students than others. Assessment of problems by component type (i.e., characters, meanings, and sounds) showed that students exhibited higher accuracy with fewer attempts and a lesser need for system feedback when sounds were used for the prompt. However, analysis revealed no significant differences in word acquisition by condition, as evidenced by next-day post-test scores or pre- to post-test gain scores.

The second and third projects further examine online CFL word recognition and the effects of supplemental handwriting practice. In these two projects, a series of four experiments

have been conducted to investigate the efficiency of online Chinese word recognition both with and without handwriting. The second project *Save Your Strokes: Chinese Handwriting Practice Makes for Ineffective Use of Instructional Time in Second Language Classrooms* (Lu, Ostrow, & Heffernan, 2019b) explores the efficiency of Chinese handwriting practice and computer-mediated typing practice on word recognition. Handwriting practice is the most time-consuming activity for learners of CFL. CFL instructors commonly allocate at least one-third of their course time to handwriting practice, despite the fact that it prevents students from engaging in meaningful communication, especially in the earliest stages of learning. The amount of time students spend in a college course is relatively fixed, so the present study sought to understand the best use of students' time if their primary goals are word acquisition and communication. Although a significant amount of literature has supported the correlation between Chinese reading and handwriting, the principle of "read and type more, handwrite less" has gained more traction in the CFL field. This project aimed at testing the effectiveness of online versus handwriting practices regarding word recognition. We examined word acquisition and recognition while manipulating the two conditions of No-Handwriting (NH) practice and With-Handwriting (WH) practice, and post-test point (1 (immediate), 2 (one day delay), and 3 (one week delay)). Two-way repeated-measures ANOVAs revealed significant differences between conditions and test points in online and on-paper measures of word recognition and handwriting, respectively. The second project serves as a replication of our pilot work. Prior to data collection and analysis, this study was accepted as a pre-registered publication in AERA Open. The pilot work is included in the third project as the first experiment.

Building from the findings of the second project (Lu et al., 2019b), I present the third project manuscript as part of the dissertation research. The manuscript *Chinese Word*

Recognition Practices in Novice Language Learners: Typing Proven a More Effective Practice as Compared to Handwriting (submitted as a book chapter for intended publication by the Publishers in the work *Transforming L2 Hanzi Teaching and Learning in the Age of Digital Writing: Theory, Research, and Pedagogy 电写时代汉字教学的理论与实践*) conducted three experiments to examine CFL word recognition in an online learning environment and the effects of supplemental handwriting practices. The three experiments included conditions with-or-without handwriting in an allotted amount of time. The first experiment consisted of three conditions: no-handwriting (NH) and with-handwriting (WH), and the second experiment was a replication of the first experiment which conducted at a different university to test the generalizability of the results. The third experiment further explored the results by adding another condition to the comparison: 70% of NH practices, which reduced the handwriting time by 30% and shortened participants practicing time to 70%, in only online platforms. The same to the second project, the results of the three experiments in the third project revealed a significant difference between NH and WH conditions in the online posttest results. Participants in the study performed better on the word recognition tasks when their practice time was spent entirely on digital writing.

Whereas the previous two studies demonstrated the efficiency of online Chinese word recognition practice, my fourth and fifth projects explore the efficiency of feedback timing as it relates to online text-based Chinese learning practice. Researchers have held different views on the effects of feedback timing for decades. A closer reading of the timing of feedback literature that favored delayed feedback revealed that this conclusion may have been reached prematurely. Results might have been affected by the time interval between feedback and a posttest. The fourth project differs from previous feedback timing studies in three distinct ways: First, this

study addressed the limitations of previous studies by holding a time interval between the feedback (either immediate or delayed) and the posttest constant. Second, this study included various types of knowledge and investigated the interaction between feedback timing and different knowledge types. Third, most studies that investigate the comparative effectiveness of immediate and delayed feedback on written assignments were conducted in the STEM fields, whereas few studies can be found in the second language learning field. The fourth project *Immediate Versus Delayed Feedback on Learning: Do People's Instincts Really Conflict with Reality?* (Lu, Sales, & Heffernan, 2021) explores the efficiency of immediate and delayed feedback on CFL written assignments. It reveals that the immediate feedback condition significantly outperformed the delayed feedback condition on conceptual knowledge learning, however, no difference between the two conditions was found on situational knowledge learning. Results of this study contradicted the findings that claim a significant delay-retention effect, and supported the effectiveness of immediate feedback when learning conceptual knowledge such as grammar.

A closer reading of the research related to immediate feedback, however, reveals that the definition of "immediate feedback" is inconsistent. Findings from the STEM literature were not well supported by other fields. As a result, clarification is needed in order to assess which type of immediate feedback leads to improved performance in a computer-assisted learning environment. Research related to the effects of immediate feedback outside of STEM classes is meaningful in order to better understand whether the findings can be generalized. The fifth project *Immediate Feedback Timing on Second Language Text-Based Assignments: How Immediate Should Immediate Feedback Be?* (submitted to *Computers & Education Open* for review) investigated the effects of immediate feedback timing in online language learning

exercises. Three conditions were examined: no feedback, end-of-question feedback, and end-of-assignment feedback. A Planned Contrasts test revealed that with a pretest functioning as the covariate, the end-of-question feedback condition received significantly higher grades in the posttest, compared to the end-of-assignment feedback condition, and students' learning improved significantly while taking assignments in the end-of-question feedback condition. Students with lower pretest scores used more attempts, although their learning progress was not significantly better as compared to students with higher prior knowledge. The findings of this project provide insights into the use of immediate feedback to improve learning as part of foreign language classroom instruction.

Finally, I propose to extend the two projects by investigating the effectiveness of immediate feedback types toward semi-open-ended short answer questions in Chinese language learning, followed by the status of work and a proposed timeline.

References

Allen, J. R. (2008). Why Learning to Write Chinese Is a Waste of Time: A Modest Proposal. *Foreign Language Annals*. 41(2).

Attali, Y., & van der Kleij, F. (2017). Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers & Education, 110,* 154-169. https://doi.org/10.1016/j.compedu.2017.03.012

Butler, A. C., & Woodward, N. R. (2018). Toward consilience in the use of task-level feedback to promote learning. *Psychology of Learning and Motivation, 69*. https://doi.org/10.1016/bs.plm.2018.09.001

Chinese Language Committee. (2009). *Modern Chinese common character list (Hanyu Tongyong Zibiao).* Beijing, China: Commercial Press.

Corral, D., Carpenter, S. K., & Clingan-Siverly, S. (2021). The effects of immediate versus delayed feedback on complex concept learning. *Quarterly Journal of Experimental Psychology, 74*(4), 786–799. https://doi.org/10.1177/1747021820977739

De Francis, J. (1984). *The Chinese Language: Fact and Fantasy*. Honolulu: University of Hawaii Press.

Krashen, S. D. (1988). *Second Language Acquisition and Second Language Learning*. Prentice-Hall International.

Lefevre, D., & Cox, B. (2017). Delayed instructional feedback may be more effective, but is this contrary to learners' preferences? *British Journal of Educational Technology, 48*(6), 1357–1367. https://doi.org/10.1111/bjet.12495

Long, M. (1996). The role of the linguistic environment in second language acquisition. In: Ritchie, W., Bhatia, T. (Eds.), *Handbook of second language acquisition* (pp. 413–468). San Diego, CA: Academic Press.

Lu, X., Ostrow, K. S., & Heffernan, N. T. (2019). Save Your Strokes: Chinese Handwriting Practice Makes for Ineffective Use of Instructional Time in Second Language Classrooms. *AERA Open.* https://doi.org/10.1177/2332858419890326

Lu, X., Ostrow, K. S., Heffernan, N. T. (2019) Understanding the Complexities of Chinese Word Acquisition Within an Online Learning Platform. *Proceedings of the 11th International Conference on Computer Supported Education, Greek, 1,* 321-329.

Lu, X., Ostrow, K. S., Yang, Q., & Heffernan, N. T. (submitted for review). Chinese Word Recognition Practices in Novice Language Learners: Typing Proven a More Effective Practice as Compared to Handwriting. In Chu, C., Coss, M., & Zhang, P. N. (Eds.), *Transforming L2 Hanzi Teaching & Learning in the Age of Digital Writing: Theory and Pedagogy (《电写时代汉字教学的理论与实践》).* Routledge, UK.

Lu, X., Sales, A., & Heffernan, N. T. (2021). Immediate Versus Delayed Feedback on Online Learning: Do people's Instincts Really Conflict with Reality? *Journal of Higher Education Theory and Practice, 21*(16). https://doi.org/10.33423/jhetp.v21i16.4925

Lu, X., Wang, W., Motz, B., Ye, W., Heffernan, N. T. (submitted for review). *How Immediate Should Immediate Feedback Be? Immediate Feedback Timing and its Effects on Learning Performance of Written Homework Assignments.*

Mullet, H. G., Butler, A. C., Verdin, B., von Borries, R., & Marsh, E. J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback

immediately. *Journal of Applied Research in Memory and Cognition, 3*(3), 222–229. https://doi.org/10.1016/j.jarmac.2014.05.001

Shen, H. H. (2004). Level of cognitive processing: Effects on character learning among non-native learners of Chinese as a foreign language. *Language and Education, 18,* 167–182.

Sinha, N., & Glass, A. L. (2015). Delayed, but not immediate, feedback after multiple-choice questions increases performance on a subsequent short-answer, but not multiple-choice, exam: evidence for the dual-process theory of memory. *The Journal of General Psychology, 142*(2), 118–134. https://doi.org/10.1080/00221309.2015.1024600

van der Kleij, F. M., Eggen, T. J.H.M., Timmers, C. F., & Veldkamp, B. P. (2012) Effects of feedback in a computer-based assessment for learning. *Computers & Education, 58* (1), 263-272. https://doi.org/10.1016/j.compedu.2011.07.020

van der Kleij, F.M., Feskens, R.C.W., & Eggen, T.J.H.M. (2015) Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*, 475-511. https://doi.org/10.3102/0034654314564881

Walker, G. (1989). Intensive Chinese curriculum: The EASLI model. *Journal of the Chinese Language Teachers Association, 24*(2), 43-84.

Xie, T. (2014). 中文教学与时俱进 [Chinese teaching in modern society]. In Zhou, X. (Ed.), *Chinese language in the world, 3,* 69–74. Guangzhou, China: Zhongshan University Press. Retrieved from http://web.csulb.edu/~txie/papers/Chinese_teaching_in_modern_society.pdf

Chapter 2: Understanding the Complexities of Chinese Word Acquisition Within an Online Learning Platform

This chapter presents the following manuscript:

Lu, X., Ostrow, K. S., Heffernan, N. T. (2019) Understanding the Complexities of Chinese Word Acquisition Within an Online Learning Platform. *Proceedings of The 11th International Conference on Computer Supported Education*, *Greek, 1,* 321-329.

Abstract

Because Chinese reading and writing systems are not phonetic, The Mandarin Chinese learners must build six-way connections in order to learn a word, linking characters, meanings, and sounds. Little research has focused on the difficulties inherent to each specific component involved in Chinese language acquisition, especially within the digital world. The present work studied Chinese word acquisition within the ASSISTments online learning platform. Students were randomly assigned to one of three conditions examining the loss of each connection pair within Chinese language acquisition (i.e., sound-meaning and meaning-sound). The research questions were: do students' performance differ significantly when word assignments lack sound-character, character-meaning, or meaning-sound connection pairs? Which type of questions is the most difficult one for word acquisition? Assessment of problems by their component type revealed support for the relative ease of problems that provided sounds, with students showing higher accuracy, fewer attempts, and less need for system feedback when sounds were included. However, analysis revealed no significant differences in pre-to-posttest gains by condition. Implications and suggestions for future work are discussed.

Key Words: Chinese (Mandarin), foreign/second language learning/acquisition, strategies, computer-assisted language learning (CALL)

Introduction

Mandarin Chinese is one of the most difficult languages for a native English speaker to learn. In 1982, the Foreign Service Institute (FSI) of the U.S. Department of State published a ranking comparing the approximate amount of time required for native English-speaking students to achieve General Professional Proficiency in Speaking and General Professional Proficiency in Reading in various foreign languages. The report lists Chinese as one of the five most difficult languages to learn, requiring 2,200 class hours to achieve speaking and reading proficiency, whereas French and Spanish require less than 600 hours (Liskin-Gasparro, J., 1982, Wolff, D., 1989). Chinese takes substantially longer to master than traditional European languages currently being taught in American public schools (e.g., French, Spanish, German, etc.) due to its lack of common vocabulary roots, its novel tonal and writing systems, and its distinctly different syntactic structure.

As many other researchers, De Francis, J. (1984) has pointed out that "the most difficult and time-consuming aspects of learning Chinese are character recognition and handwriting". Because Chinese reading and writing systems are not phonetic, learners must build six-way connections in order to learn a word. To learn a word, students must learn a specific character and successfully link the character to the proper sound and meaning. These connections must go both ways for successful use of the word in conversation, writing, and reading. On the other hand, learners of traditional European languages usually only need to build two-way connections to be able to read or produce a word because of the language's phonetic nature. Native speakers of these traditionally phonetic languages struggle when learning Mandarin because they cannot "spell" Chinese characters and, often, there is no obvious link between a character and its sound.

Even native Chinese speakers may come up blank when called upon to write the character for a relatively common word due to the lack of intuitive connections.

These connections must go both ways for successful use of the word in conversation, writing, and reading. On the other hand, learners of traditional European languages usually only need to build two-way connections to be able to read or produce a word because of the language's phonetic nature. Native speakers of these traditionally phonetic languages struggle when learning Mandarin because they cannot "spell" Chinese characters and, often, there is no obvious link between a character and its sound. Even native Chinese speakers may come up blank when called upon to write the character for a relatively common word due to the lack of intuitive connections.

Warschauer, M. & Healey, D. (1998) pointed out that the development of information technology has provided foreign language instructors and learners with new possibilities for practicing language acquisition. For instance, the extremely successful language-learning app Duolingo (2017) offers 27 gamified, self-paced courses for native English speakers to learn a new language (NPR/TED Staff, 2014). The app simultaneously uses learners' responses as a verification process to translate sites and articles on the Internet into foreign languages. However, Duolingo has not branched into Mandarin until 2017 (Hagiwara, M. 2017), likely due to its time consuming and difficult nature. Alternatively, the app ChineseSkill follows a gamified format similar to that of Duolingo, but focuses strictly on Mandarin (ChineseSkill Co., Ltd., 2017). These applications broaden the reach of the Chinese language to those who may have otherwise been intimidated by the learning curve.

Research has shown that reading and writing Chinese characters are two separate information acquisition processes with different influencing factors (Jiang, X., 2007). The use of

Pinyin, a Romanization system for Mandarin, helps to link these processes by transforming characters into phonetic words. Through Pinyin, online learning platforms and applications allow learners to type a phonetic version of Chinese characters without having to write the characters by hand. Zhu, Z., Liu, L., Ding, G. & Peng, D. (2009) indicated that Pinyin as a digital input method can strengthen character recognition through the consolidation of pronunciation capability. While tablets and other touch devices may allow for character drawing, Pinyin bridges the availability of Chinese learning acquisition to broader digital environments. Of course, learners must still memorize characters for the sake of recognition, and in order to connect the character and its Pinyin equivalent.

Despite Chinese language acquisition requiring characters, meanings, sounds, and often, Pinyin, little research has been done on the difficulty inherent to each of these specific components. Even less work has focused on how Chinese language acquisition has adapted to the digital world. As a Chinese language instructor at a major institution in New England, the first author observed that students typically begin word memorization by practicing connections between sound and meaning, considering the character/meaning connection as a secondary task. Tan, L. & Perfetti, C. A. (1999) suggested that phonology is an obligatory component of word identification in Chinese reading. Perfetti, C. A. & Liu. Y. (2006) supported this idea, stating that phonology is automatically activated in reading words, regardless of whether activation occurs before or after the moment of lexical access and regardless of whether it is instrumental in retrieving the word's meaning. Essentially, Chinese characters activate pronunciation, even when the reader's goal is to determine the character's meaning.

Based on past research and considering the six connections required for Chinese word acquisition, it is easy to speculate that connections between meaning and character are more

difficult because sound is also accessed, even if unintentionally. It becomes difficult to discern if and how these components can be teased apart in language acquisition, and whether providing particular types of connections more frequently than others has the potential to produce more robust learning. As such, our hypotheses were that students' performance are different when word assignments lack sound-meaning, meaning-character, or character-sound connection pairs, and questions providing sound are the easiest one for word acquisition when looking into the word assignment procedure.

The present study manipulated three levels of conditions examining the removal of a pair of connections involved in language acquisition: sound focused condition, meaning focused condition, and character focused condition. Students' correct rates were dependent variables. When measuring the difficulties of question types during assignments, questions providing sounds, meanings, and characters were compared, and students' learning gains on these question types, assignment accuracy, hint counts, attempts counts, and answer requests during practices were measured. This study sought to answer the following research questions:

1. Do students' performance on the platform, as measured at posttest, differ significantly when word assignments lack sound-meaning, meaning-character, or character-sound connection pairs?

2. Which type of questions is the most difficult one for word acquisition? Are word practices that lack sound connections more difficult? That is, do students require significantly more attempts or feedback when completing these types of problems?

Methods

Participants

Participants included 60 students enrolled in an Intermediate Level Chinese class at a major university in New England, conducted during the Fall 2016 semester. Students were participating in the course for credit, and had been enrolled at the intermediate level through a placement exam or following experience in a preliminary Chinese course. A total of 60 students were assigned the pretest and assignment as a single problem set on Day 1. Of these students, two did not have access to video content and were removed from the study. Additionally, three students failed to complete any problems and were not assigned to a condition. The remaining 55 students were randomly assigned to one of three experimental conditions, a sound focused practice, a meaning focused practice, or a character focused practice. Student-level randomization did not result in a particularly normal distribution across conditions, but attrition between conditions was not significantly different. Three students failed to complete the Day 1 assignment, and 51 students were assigned the posttest on Day 2. Of the 50 students completing the posttest, 46 had first completed the Day 1 assignment.

Setting

All study materials were delivered within a graded classwork assignment, and all subjects participated during their regular class period. The ASSISTments platform was a novel tool for language acquisition to all students in the study. The present work was the first of its kind to be conducted within ASSISTments, an online learning platform typically used for middle school mathematics. The platform is offered as a free service of WPI, with the goal of providing students with instructional ASSISTance while offering teachers reports for formative assessMENT, establishing its moniker (Heffernan, N. and Heffernan, C., 2014). The system also allows researchers to conduct randomized controlled trials within a library of premade content,

providing access to a large subject pool of 50,000 student users. Additionally, the platform can also be used to develop material and research in other domains, and is growing with regard to Physics, Chemistry, and Language.

All of the data produced by the ASSISTments platform are anonymized. In the form that researchers fill out before running research, they agree to not de-anonymize the data. Current study was overseeing by the ASSISTments' IRB approval, and signing agreements from participants were waived.

Materials

The first author, a lecturer of the Intermediate Level Chinese course but not the active teacher of study participants, worked collaboratively with the sitting lecturer to select ten novel words from the textbook to assign as learning targets. The ten words can be found at Lu, X. (2017). It was expected that the students would not have prior experience with the chosen words. A pretest was used to assess participants' knowledge of these words prior to beginning the classwork assignment. Students that knew any or all words were still required to complete the assignment, but their data was not included in any analysis examining word acquisition.

The classwork assignment consisted of sixty possible questions. For each of the ten new words, questions were established corresponding to each of the six connections between language acquisition components of sound, meaning, and character. Each word included questions prompting students to provide solutions transitioning from sound to meaning, sound to character, meaning to sound, meaning to character, character to sound, and character to meaning. Questions with a sound component included a brief YouTube video providing the audio. All questions are available at Lu, X. (2017) for additional reference.

Procedures

The experimental assignment spanned two class days and students were allowed to work at their own pace. On the first day, students created ASSISTments accounts and were provided an explanation of how to proceed with their assignment. At the start of the assignment, students completed a question assessing their ability to access YouTube videos to verify that students would receive the sound component of problems or hints.

If students could not access video content, they were routed into an alternative assignment and excluded from the study. Participants were then randomly assigned to one of three conditions examining the removal of a pair of connections involved in language acquisition. Each condition's assignment began with a ten-question pretest assessing knowledge of each novel word. Each problem followed the format: "Please write down the English meaning of the word "孩子, hai2zi0". If you don't know this word, please enter the word "no"."

Following the pretest, participants in Condition 1 received four problem types related to sound-based connections for each target word (e.g., sound to meaning, sound to character, meaning to sound, character to sound), totaling 40 problems. Participants in Condition 2 received 40 problems related to meaning-based connections, and participants in Condition 3 received 40 problems related to character-based connections. For each problem, if students were unable to provide the correct answer, they could ask for a single hint. Each hint provided the student with the language component missing from the problem. For instance, if the problem asked the student to convert a character to it's meaning, the hint would provide the character's sound through a YouTube video. If the student was still unable to reach the solution after realizing all three components of the word, they were able to access the correct answer in order to move on to the next problem.

On the second day, students began class by logging into ASSISTments and taking a posttest with the 10 problems. Students were not overtly aware that their Day 1 assignment was part of an experiment, and therefore the interleaving was used to make the posttest more comprehensive of their overall classroom learning. Students were told that their score on the Day 2 assignment would count as a daily quiz grade.

Analyses

Data included the pretest and posttest results, as well as students' performance on the first day's assignments including accuracy, response time, attempt count, hint and answer requests. All data was logged by the ASSISTments learning platform, and analyzed by SPSS.

The analysis of learning gains focused on the pretest, assignment, and posttest correct rate comparison using ANNOA. Descriptive statistics of learning gains across conditions were calculated and listed in Table 1.

The analysis of learning difficulty focused on students' performances during the first day assignments, including accuracy, hint count, attempts count, and answer requests. For each dependent variable, ANOVAs were conducted to examine differences across problem types and across assigned conditions. Examining averages for these measures controlled for the number of problems students experienced within each condition. Descriptive statistics were calculated and listed in Table 2.

Table 1

Means and SDs of learning gains exhibited across groups

	n	Pretest	Assignment	Posttest	Gain (Pre-Post)
Group 1 – Sound Focused Practice	10	0.32 (0.15)	0.69 (0.11)	0.78 (0.17)	0.43 (0.25)
Group 2 – Meaning Focused Practice	22	0.41 (0.20)	0.80 (0.08)	0.81 (0.13)	0.40 (0.27)
Group 3 – Character Focused Practice	14	0.42 (0.26)	0.76 (0.16)	0.83 (0.19)	0.47 (0.25)

Note. Mean (*SD*).

Table 2

Descriptive statistics and ANOVA results within and between groups

Problem Type	Accuracy	Hint Count	Attempt Count	Answer Requests
Providing Sound	$F(2, 52) = 8.29^{**}$	$F(2, 52) = 20.91^{***}$	$F(2, 52) = 6.32^{**}$	--
Group 1 – Sound Focused Practice	0.76 (0.14)	0.08 (0.05)	1.38 (0.28)	0.00 (0.00)
Group 2 – Meaning Focused Practice	0.90 (0.10)	0.00 (0.00)	1.14 (0.16)	0.00 (0.00)
Group 3 – Character Focused Practice	0.90 (0.11)	0.01 (0.05)	1.16 (0.20)	0.00 (0.00)
Total	0.87 (0.12)	0.02 (0.05)	1.20 (0.23)	0.00 (0.00)
Providing Meaning	$F(2, 52) = 57.80^{***}$	$F(2, 52) = 31.60^{***}$	$F(2, 52) = 24.56^{***}$	$F(2, 52) = 17.16^{***}$
Group 1 – Sound Focused Practice	0.32 (0.25)	1.02 (0.58)	3.53 (1.62)	0.39 (0.31)
Group 2 – Meaning Focused Practice	0.65 (0.14)	0.52 (0.33)	2.18 (0.85)	0.21 (0.17)
Group 3 – Character Focused Practice	0.95 (0.08)	0.00 (0.00)	1.08 (0.13)	0.00 (0.00)
Total	0.67 (0.28)	0.47 (0.52)	2.14 (1.32)	0.19 (0.24)
Providing Character	$F(2, 53) = 48.98^{***}$	$F(2, 53) = 20.33^{***}$	$F(2, 53) = 18.03^{***}$	$F(2, 53) = 14.75^{***}$
Group 1 – Sound Focused Practice	0.93 (0.08)	0.08 (0.13)	1.19 (0.27)	0.02 (0.04)
Group 2 – Meaning Focused Practice	0.98 (0.04)	0.00 (0.00)	1.02 (0.04)	0.00 (0.00)
Group 3 – Character Focused Practice	0.63 (0.19)	0.44 (0.38)	2.43 (1.36)	0.17 (0.18)
Total	0.85 (0.20)	0.17 (0.30)	1.54 (1.02)	0.06 (0.13)

Note. Mean (SD). *** $p < .001$, ** $p < .01$

Results

The hypotheses of this study were that student performance differ significantly when word assignments lack sound-meaning, meaning-character, or character-sound connection pairs, and questions providing sound are the easiest one for word acquisition when looking into the word practice procedure.

Students' learning gains, assignment accuracy, hint count, attempts count, and answer requests during practices were measured.

Learning Gains

Average target word scores and standard deviations across conditions were presented in Table 1 for students' performance on the pretest, assignment, and posttest. In order to assess whether student performance, as measured at posttest, differed significantly when word

assignments lacked sound-meaning, meaning-character, or character-sound component pairs, the performance of the 46 students completing both Day 1 and Day 2 assignments were analyzed. An analysis of variance (ANOVA) revealed that conditions were not significantly different at pretest, $F(2, 53) = 0.87$, $p > .05$, eta squared= .03, despite a lower average for those assigned to receive sound focused assignment. Average scores on the 40 problem word acquisition assignment were significantly different across conditions, $F(2, 53) = 3.38$, $p < .05$, eta squared = .11, driven by a significant difference between sound focused assignment and meaning focused assignment. Students receiving meaning focused assignment solved assignment problems with significantly more accuracy in post hoc tests, $p = .04$, 95%CI[-.21, -.01], Cohen's d=-.26. Despite significant differences amongst assignment scores, no significant differences were observed at posttest, $F(2, 43) = 0.29$, $p > .05$, eta squared = .01. Additionally, pre-to-post test gains were not significantly different between conditions, $F(2, 42) = 0.31$, $p > .05$, Partial eta squared=.01.

Learning Difficulty

In order to assess whether word assignments that lack sound connections were more difficult, the Day 1 assignment performance of the 56 students assigned to conditions was analyzed. In order to examine the effect of providing-sound questions, the data was resorted to consider question type. Problem types included those providing sounds, those providing meanings, and those providing characters as prompts. Using this reorganization, students were asked 10 or 20 problems of each type, depending on their assigned condition. Dependent variables used to explore difficulty included students' average accuracy on problems in the assignment, the number of hints requested during the assignment, the number of attempts required, and the number of answers requested. For each dependent variable, ANOVAs were

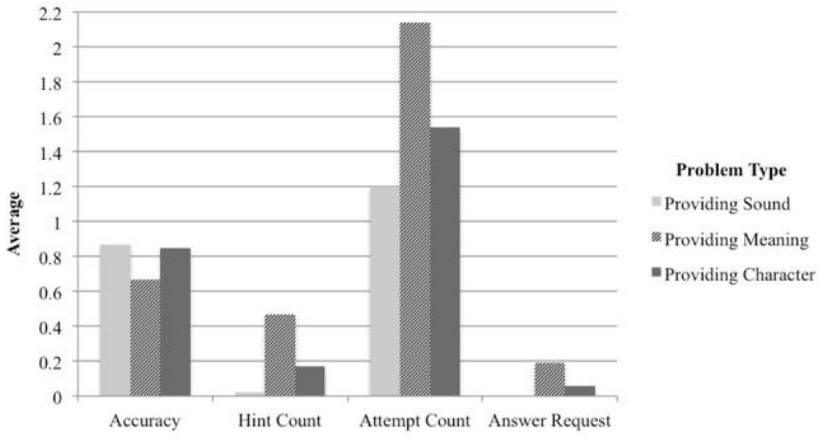
conducted to examine differences across problem types and across assigned conditions. Examining averages for these measures controlled for the number of problems students experienced within each condition. Means and standard deviations for all metrics within and between conditions were presented in Table 2. Problem type averages were also presented visually in Figure 1.

Accuracy

Significant differences were observed between students' accuracy on differing problem types, $F(2,163) = 14.49$, $p < 0.001$, eta squared=.15. Specifically, significant differences were observed between problems providing sounds and those providing meanings, $p < .001$, Cohen's $d = .93$, as well as between problems providing meanings and those providing characters, $p < .001$, Cohen's $d = .74$. Problems that provided meanings resulted in the lowest accuracy ($M = 0.67$, $SD = 0.28$). In contrast, problems that provided sounds and those that provided characters resulted in higher accuracy, ($M = 0.67$, $SD = 0.28$ and $M = 0.67$, $SD = 0.28$, respectively). Significant differences were also observed between experimental conditions with regard to problems providing sounds, meanings, and characters, as shown in Table 2, all at $p < .01$.

Figure 1

Average accuracy, hint count, attempt count, and answer request across question types



Hint Count

Significant differences were also observed between the number of hints requested by students on differing problem types, $F(2,163) = 23.58$, $p < 0.001$, eta squared = .22. Specifically, significant differences were again observed between problems providing sounds and those providing meanings, $p < .001$, Cohen's $d = 1.21$, as well as between problems providing meanings and those providing characters, $p < .001$, Cohen's $d = 0.71$. Students required the most hints on problems that provided meanings ($M = 0.47$, $SD = 0.52$), and the least hints on problems that provided sounds ($M = 0.02$, $SD = 0.05$). Students required a moderate amount of hints on problems that provided characters ($M = 0.17$, $SD = 0.30$). Significant differences were also observed between experimental conditions with regard to problems providing sounds, meanings, and characters, as shown in Table 2, all at $p < .001$.

Attempt Count

Significant differences were also observed between the number of attempts students made on differing problem types, $F(2,163) = 13.11$, $p < 0.001$, eta squared = .14. Specifically, significant differences were again observed between problems providing sounds and those providing meanings, $p < .001$, Cohen's $d = .99$, as well as between problems providing meanings and those providing characters, $p = .002$, Cohen's $d = 0.51$. Students made the most attempts on problems that provided meanings ($M = 2.14$, $SD = 1.32$), and the least attempts on problems that provided sounds ($M = 1.20$, $SD = 0.23$). Students made a moderate number of attempts on problems that provided characters ($M = 1.54$, $SD = 1.02$). Significant differences were also observed between experimental conditions with regard to problems providing meanings at $p < .001$, as shown in Table 2.

Answer Requests

Significant differences were also observed between the number of answer requests students made on differing problem types, F (2,163) = 20.54, $p < 0.001$, eta squared = .20. Interestingly, regardless of condition, students did not request answers at all when confronted with problems that provided sounds ($M = 0.00$, $SD = 0.00$, all conditions). Significant differences were observed between problems providing sounds and those providing meanings, $p < .001$, as well as between problems providing meanings and those providing characters, $p < .001$, Cohen's $d = .67$. Students requested answers most frequently when working on problems that provided meanings ($M = 0.19$, $SD = 0.24$), but fewer when working on problems that provided characters ($M = 0.06$, $SD = 0.13$). Significant differences were also observed across problem types between conditions with assignments focused on meanings and with assignments focused on characters, as shown in Table 2, at $p < .01$.

Discussion

Past research has confirmed that Mandarin Chinese is one of the most difficult languages for native English speakers to acquire, requiring almost four times as much class time to reach the same level of speaking and reading proficiency as French or Spanish (Liskin-Gasparro, J., 1982, Wolff, D., 1989). Because Chinese reading and writing systems are not phonetic, learners must build six-way connections between sounds, meaning, and characters in order to learn words. However, little work has focused on the relative importance of each pair of connections. Online learning platforms and applications have allowed learners to broach language acquisition in new and unique ways, while allowing researchers the flexibility to investigate the complexities of word acquisition. The present work sought to examine the components of Chinese word acquisition (sounds, meanings, and characters) within an online learning platform in which language research was novel.

Specifically, the present work first sought to understand if removing a pair of component connections (i.e., sound to meaning and meaning to sound) would significantly alter students' learning as measured by a delayed posttest. An assessment of average target word posttest scores for 46 participants revealed no significant differences, despite significantly different assignment performance across conditions. Following 40-word assignments lacking a component pair by condition, gain scores from pre-to-post test also revealed no significant condition differences. Thus, despite the emphasis that past research has placed on sound components within Chinese word acquisition, the present work offered no evidence that the removal of sound disproportionately hindered word acquisition.

Additionally, based on past research (Perfetti, C. A. & Liu. Y., 2006; Perfetti, C. A., Zhang, S. and Berent, I., 1992; Tan, L. & Perfetti, C. A., 1999) the present work sought to examine whether word assignments lacking sound connections would be more difficult to students, considering their average accuracy and need for system provided feedback elements within a classwork assignment. Using data resorted by problem type, analysis of 52 students revealed significant differences in within-assignment difficulty as measured by students' average accuracy, hint usage, attempt counts, and answers requested across assignment problems. For each dependent variable, significant differences were observed between problems providing sounds and those providing meanings, as well as between problems providing meanings and those providing characters.

Problems that provided sounds were the most likely to be solved accurately, while requiring the fewest hints and attempts, and never causing students to request answers. In contrast, problems that provided meanings were the most likely to be solved inaccurately, while requiring the most hints and attempts, and causing students to request answers most frequently

on average. Problems that provided characters resulted in moderate performance across metrics of difficulty. These results suggested that problems providing meanings were most difficult within assignment (meaning to sound, meaning to character) while problems providing sounds were least difficult within assignment (sound to meaning, sound to character). While this work did not directly replicate the Universal Phonological Principle described by Perfetti, C. A., Zhang, S. and Berent, I. (1992), it reflected the principle from an alternative perspective. Based on these findings, teachers can expect that practices providing sounds will be easier for students, while those providing meanings will be more difficult. When designing assignments, it may be better to start with practices that provide sounds, and spend additional time practicing meaning connections as secondary instruction.

The present work presented several limitations. First, despite student-level random assignment conducted by the ASSISTments platform, distribution across the three experimental conditions was not well balanced. As the unbalanced distribution came by chance, a larger sample size may have resolved this issue. A larger sample size may have also revealed greater differences in learning gains between conditions. Given the observation that problems providing meanings posed greater difficulty for students, while problems providing sounds were met with the most ease, future work should consider how altering practice strictly by component (i.e., removing sound prompting problems by removing sound to meaning and sound to character problems) rather than by a component connection pair (i.e., removing sound to meaning and meaning to sound) alters students' learning and retention. This would likely result in significantly different learning gains, but was not employed in the present work for the sake of fair learning within an authentic classwork assignment.

Further, the present work was based on only four word practices per target word. While this resulted in 40 problems spanning the ten target words, future work should examine how adding additional practice instances might ultimately enhance learning. Learning gains may be stronger through additional practice, or the current work may have caught a ceiling effect, which future investigation could reveal or confirm. Additionally, future work should consider long-term retention, as learning gains spanning longer terms may function differently than gains observed with the brief delay of a single day.

Future work should also further examine why problem types guided by acquisition component (sound, meaning, or character) pose different levels of difficulty to students. What is it that makes a particular type of problem more difficult to answer? Do problems providing meaning strain recall? How can teachers and learning platforms better assist students with these types of problems, and, is the added difficulty inherently beneficial for later word retention?

Conclusions

Little work has focused on the relative importance and difficulty of the connections between sound, meaning, and character components required for successful Chinese language acquisition. The present work teased apart these components within the context of an online learning platform, finding that problems had significantly different difficulty levels by component type, but that the removal of particular connections between components did not significantly impact learning as measured at posttest. Results suggest that problems providing meanings are most difficult within practice (meaning to sound, meaning to character) while problems providing sounds are least difficult within practice (sound to meaning, sound to character), ultimately suggesting that teachers of Chinese as a foreign language, and those building online learning content for Mandarin, should begin practices with sound connections and spend extra time on

meaning connections later in practice. While subtle, this finding has the potential to enhance the way Chinese language is taught in foreign language classrooms and in online learning environments, reducing students' difficulty and, perhaps, enhancing their motivation to continue in the pursuit of language acquisition.

Acknowledgements

We acknowledge funding from multiple NSF grants (ACI-1440753, DRL-1252297, DRL-1109483, DRL-1316736 & DRL-1031398), the U.S. Department of Education (IES R305A120125 & R305C100024 and GAANN), the ONR, and the Gates Foundation.

References

ChineseSkill Co., Ltd. 2017. *ChineseSkill - Learn Chinese/Mandarin Language for free.* Retrieved from http://www.chinese-skill.com/cs.html.

De Francis, J. 1984. *The Chinese Language: Fact and Fantasy*. Honolulu, HI: University of Hawaii Press.

Duolingo. 2017. *Language Courses for English Speakers.* Retrieved from https://www.duolingo.com/courses.

Hagiwara, M. 2017. *Duolingo now supports Chinese, but it probably won't help you become fluent.* Retrieved from https://www.theverge.com/2017/11/16/16598626/duolingo-chinese-app-language-learning-simplified-vocabulary-chatbots

Heffernan, N.T. & Heffernan, C.L. 2014. The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for

Minimally Invasive Research on Human Learning and Teaching. *International Journal of Artificial Intelligence in Education, 24*(4), 470-497. doi:10.1007/s40593-014-0024-x

Jiang, X. 2007. An Experimental Study on the Effect of the Method of "Teaching the Learner to Recognize Characters More Than Writing". *Chinese Teaching in the World*, *2007*(2): 91-97.

Liskin-Gasparro, J. 1982. *ETS oral proficiency test manual.* Princeton, NJ: Educational Testing Service.

Lu, X. 2017. Experiment data available at: https://sites.google.com/view/xiwenlu/data-document?authuser=1

NPR/TED Staff. 2014. Translating the Web with Millions: Luis Von Ahn Answers Your Questions. *TED Radio Hour*. Retrieved from:

http://www.npr.org/2014/06/10/319071368/translating-the-web-with-millions-luis-von-ahn-answers-your-questions.

Wolff, D. 1989. Teaching language in context. Proficiency-oriented instruction: Omaggio, Alice C., Boston: Heinle and Heinle Publishers, Inc., 1986, 479 pp. *System*, 17(2), 286-288. doi:10.1016/0346-251X(89)90047-X

Perfetti, C. A. & Liu. Y. 2006. Reading Chinese characters: Orthography, phonology, meaning, and the lexical constituency model. In P. Li, L. H., Tan, E. Bates, & O. J. L. Tzeng (Eds.), *The handbook of East Asian Psycholinguistics, Volume 1: Chinese*, 225-236. New York: Cambridge University Press. Retrieved from http://www.pitt.edu/~perfetti/PDF/Read%20Chinese%20char,%20ortho%20and%20phono%20(chapt)-%20Liu.pdf

Perfetti, C. A., Zhang, S., & Berent, I. 1992. Reading in English and Chinese: Evidence for a "universal" phonological principle. In R. Frost & L. Katz (Eds.), Advances in psychology, Vol. 94. *Orthography, phonology, morphology, and meaning,* 227-248. Oxford, England: North-Holland. http://dx.doi.org/10.1016/S0166-4115(08)62798-3

Tan, L. & Perfetti, C. A. 1999. Phonological activation in visual identification of Chinese two-character words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 25, 2 (Mar. 1999), 382-393. http://dx.doi.org/10.1037/0278-7393.25.2.382

Warschauer, M., & Healey, D. 1998. Computers and language learning: An overview. *Language Teaching*, 31(2), 57-71. https://doi.org/10.1017/S0261444800012970

Zhu, Z., Liu, L., Ding, G. & Peng, D. 2009. The Influence of Pinyin Typewriting Experience on Orthographic and Phonological Processing of Chinese Characters. *Acta Psychologica Sinica*. 41(09), 785-792. http://dx.doi.org/10.3724/SP.J.1041.2009.00785

Chapter 3: Save Your Strokes: Chinese Handwriting Practice Makes for Ineffective Use of Instructional Time in Second Language Classrooms

This chapter presents the following manuscript:

Lu, X., Ostrow, K. S., & Heffernan, N. T. (2019). Save Your Strokes: Chinese Handwriting Practice Makes for Ineffective Use of Instructional Time in Second Language Classrooms. *AERA Open.* https://doi.org/10.1177/2332858419890326

Abstract

Handwriting practice is the most time consuming activity for learners of Chinese as a foreign language (CFL). CFL instructors commonly allocate at least one third of their course time to handwriting practice even though it prevents students from engaging in meaningful communication, especially in the earliest stages of learning. As the amount of time students spend in a college course is relatively fixed, the present study sought to understand the best use of students' time if their primary goals are word acquisition and communication. This work replicates a pilot study examining CFL word recognition in an online learning environment (ASSISTments) and the effects of supplemental handwriting practice. We examined word acquisition and recognition while manipulating condition (No-Handwriting (NH) practice and With-Handwriting (WH) practice), and Posttest test point (1 (immediate), 2 (one day delay), and 3 (one week delay)). Two-way repeated measures ANOVAs revealed significant main effects for both condition and test point in online and on paper measures of word recognition and handwriting, respectively. Potential implications for CFL instruction and directions for future work are discussed.

Keywords: Chinese as a foreign language (CFL) instruction, word acquisition, word recognition, handwriting, randomized controlled trial.

Background

As a logographic language, Chinese has numerous features that set it apart from Western languages. Chinese reading and writing systems use characters that are formed with radicals and strokes that are often only vaguely related to their meaning or pronunciation. As such, it can be difficult for learners to extract accurate meaning or pronunciation from characters alone. Characters are not equivalent to words in the Chinese language. Chinese words are formed by one or more characters, and most words are disyllabic in nature, or formed by two characters. Handwriting refers to the action of producing these characters or words by hand from memory.

Chinese is also a tonal language. Pinyin, the standard Romanized system for transliteration, is often used to help learners understand pronunciation and to connect a word's sound to its meaning. In order to successfully read Chinese, students must learn to rapidly combine the three aspects of each word: visual (orthography), pronunciation (phonology), and meaning (semantics) (Shen, 2004). Research has shown that Chinese courses exhibit greater difficulty conforming to the ACTFL (American Council on the Teaching of Foreign Languages) proficiency guidelines (ACTFL, 2012), and that learners taking Chinese as a foreign language (CFL) often find it difficult due to its orthographic nature, which increases the burden of retrieval and retention (Chinese Language Committee, 2009).

De Francis (1984) and Allen (2008) both isolated handwriting practice as the most time consuming activity for CFL learners, noting that it significantly slows the learning process and prevents students from engaging in meaningful communication, especially in the earliest stages of learning. Hand copying characters by following stroke order is one of the most commonly used practices for writing and word recognition. Many CFL teachers believe that this type of mechanical repetition helps students solidify word recognition skills. However, the need to write characters

and words by hand has rapidly declined in nearly all Chinese social settings (Allen, 2008; Xie, 2014). Still, support for handwriting practice has been found in Chinese-language literacy contexts (as a native language) in both reading-development research (e.g., Huang & Hanley, 1994; Leck, Weekes & Chen, 1995; Tan, Spinks, Eden, Perfetti & Siok, 2005; Chan, Ho, Tsang, Lee & Chung, 2006; Packard et al., 2006) and neuroimaging studies of normal adult subjects (e.g., Siok, Perfetti, Jin & Tan, 2004). Tan et al. (2005) indicated that for Chinese children, the ability to read Chinese words is more strongly related to writing skills than phonological awareness, as compared to the results of children learning alphabetic native languages. Tan, Xu, Chang & Siok (2013) also indicated that primary-school children's reading development may be negatively impacted by transitioning handwriting practice to Pinyin or typed practice. Such theories have enhanced CFL teachers' beliefs that handwriting should be required to help reading development. But these hypotheses were based on native-language learners rather than second-language learners; should the effects of writing share the same mechanism?

Although different CFL programs maintain different requirements and pedagogical goals for learning (Everson, 2011), several researchers have suggested a close relationship between Chinese word recognition and handwriting (e.g., Ke, 1998; Cao et al., 2012; Cao et al., 2013; Guan et al., 2011). For instance, Guan et al. (2011) compared the effects of three types of online writing tutors: handwriting, reading only, and Pinyin typing. They suggested that both writing skill and phonological awareness, as proposed by Tan et al. (2005), may play roles in CFL reading, noting practical implications for the integration of handwriting and Pinyin typing in promoting reading Chinese in second-language contexts. Xu et al. (2013) compared the effectiveness of different approaches on CFL learners' orthographic knowledge of Chinese and found that writing and animation helped improve form recognition, that reading led to stronger recall of meaning and

sound, and that writing promoted character reproduction from memory. However, their results were based on a sample of foreign-language learners with orthographic knowledge and general understanding of stroke-order rules – students who were regularly assigned the task of writing words from memory on homework and quizzes. Unfortunately, when foreign-language learners *begin* learning Chinese, they lack this knowledge and understanding. Hsiung et al. (2017) also found that writing exercises helped students to memorize the orthography and output of Chinese characters. However, their experiment was conducted on CFL learners who were studying abroad in Taiwan; learners who had not only mastered orthographic knowledge, but who had spent much of their study time focused on Chinese language learning. This suggests that their results were not particularly generalizable.

Guan et al.'s (2011) hypothesis that writing helps reading in Chinese might be true for native-language learners; there is no doubt that language teachers and learners alike should continue to value the tradition and art of handwriting. However, as researchers and educators have come to realize the unbalanced input and output of handwriting at novice levels, the principle of "listening and speaking first" has gained traction, proposing non-synchronized character/word recognition and production (Jiang, 2007; Cui, 1999). This principle makes sense for CFL teachers, who often share the primary concerns of: 1) finding the most efficient ways to reach high-proficiency levels; and 2) helping lower-level students communicate as quickly as possible in order to maintain interest. Real-world communication typically requires skill in speaking and reading. As such, it makes sense to reduce the amount of time that lower-level students are expected to spend on handwriting practice.

To better understand the current state of Novice- and Intermediate-Level CFL teaching, we surveyed 27 secondary- and college-level instructors in the United States. Responses were

representative of 17 college instructors and 10 secondary-level instructors at 26 schools. Of those polled, 70% of instructors required their students to be able to write all learned Chinese words by hand. This statistic increased to 88% when only considering college level instructors. Many CFL instructors believe that handwriting is the most reliable approach for novel word acquisition, recognition, and retention, and feel it should be undertaken from the start. When asked, "When assigning homework to students, what percentage of time do you expect students to spend on writing by hand? (Handwriting-required practices include: copying characters/words, essay writing, answering questions based on text, translating English to Chinese, etc.)," instructors at the college level who required students to be able to write all words by hand expected students to spend an average of 44% of their homework time on handwriting. Even instructors without strict handwriting requirements expected students to spend an average of 30% of their homework time on new word memorization techniques. However, these instructors are allocating at least one third of their course time on practice that may not actually be helping their students learn and retain novel words.

Non-native beginners typically use few orthographic strategies in their approaches to character learning, while more advanced students tend to rely more heavily on orthographic knowledge (Shen, 2005). The time commitment required to build such a knowledge base adds significantly to a student's work load (Shen, 2005). As such, we believe that the most valuable focus for CFL instructors is not whether handwriting should be included, but rather, isolating the most efficient strategies for teaching CFL and understanding how those strategies may differ for novices.

Before the popularity of technology, "reading before writing" was a rather contrived concept. Warschauer and Healey (1998) stated that the development of information technology

has provided foreign-language instructors and learners with new possibilities. Computerized Chinese instruction began more than a decade ago and instructors are frequently turning to computer-based tools in the digital age (Xie, 2014). Language learning principles have been developed to suit the context of computerized language instruction and researchers believe that computerization has made proficiency-based Chinese instruction more efficient and more aligned to the guidelines of foreign-language learning in the 21st century (Xie, 2014). Zhu, Liu, Ding & Peng (2009) noted that Pinyin typewriting can be beneficial for both the phonological and orthographic processing of Chinese characters. Zhu et al. (2016) even suggested that CFL beginners should rely on the Pinyin-input method found in word-processors (e.g., Microsoft Word) rather than practicing conventional handwriting techniques, as the former medium led to better performance in essay-writing tasks. Whereas students would traditionally combine orthography, phonology, and semantics (Shen, 2004), Zhu et al.'s (2016) work suggested that reducing some of this burden through the affordances of technology led to more efficient and effective outcomes (Appendix A explains how the Pinyin input method can be used to type in Chinese). In the present study, online practices were used to teach novel target words by providing one or two formats of each word (orthography, phonology, or semantics) and asking students to supply the third. For instance, when prompted to supply the orthographic aspect of a word, students entered their response using Pinyin rather than hand-copying characters. Similarly, when prompted to supply the phonological aspect of a word, they entered Pinyin with tone marks.

The present study specifically considered recognition of disyllabic, or two-character, words. Word recognition is the ability of a reader to recognize written words correctly and effortlessly. Everson (1998) defined Chinese word recognition as "deriving both the phonetic codes (or pronunciation) as well as lexical meaning from printed Chinese characters." In the present study,

we measured “isolated word recognition" or a reader's ability to recognize words individually without contextual help. It should be noted that in Chinese, word recognition is different from character recognition. From a psychological perspective, two-character Chinese words require that readers assemble characters, increasing processing complexity above that necessary for single characters (Tan & Perfetti, 1999). Previous work has suggested that rapid word recognition is the main component of fluent reading. When considering communication, words (rather than characters) are the basic unit of a sentence; the vocabulary list of most CFL textbooks are built on the foundation of words.

A pilot version of this work (Lu, Ostrow & Heffernan, In Preparation) revealed that handwriting practice was an ineffective use of instruction time for CFL learners; participants scored significantly lower on online portions of word recognition posttests after spending 30% of their practice time on character/word hand-copying exercises. These significant differences were apparent immediately following word acquisition practice sessions and upon repeated testing one day and even one week later. Worse, results of handwriting posttests did not reveal significant gains for those who had spent time focused on hand copying. As such, pilot results lent credibility to the proposition that CFL instructors should not sacrifice time in the early stages of learning on handwriting when word acquisition and communication are of primary concern.

The present study serves as a replication of our pilot work. We considered word recognition while manipulating two independent variables: condition and test point. Condition included two levels: No-Handwriting (NH) practice and With-Handwriting (WH) practice. Test point indicated posttest time point and included three levels: 1 (immediate), 2 (one day delay), and 3 (one week delay). Given that the amount of time students spend in a college course is relatively fixed, we sought to understand the best use of students’ time if the primary goal is word acquisition and

communication. We sought to confirm that students perform better on word-recognition tasks when conventional amounts of handwriting practice are eliminated to make more time for online word-acquisition practice. Thus, we hypothesized that, as observed in our pilot work, students would score significantly higher on word recognition posttests when they were not subjected to hand copying exercises.

Methods

Participants

Participants included 60 students enrolled in an Intermediate Chinese class in the fall 2018 semester at a university in the northeastern United States. The average age of participants was 19.45 years (*SD* = 1.25 years), with a distribution of 25 males and 35 females, 13 freshmen, 29 sophomores, 9 juniors, 8 seniors and 1 graduate student. Students participated in the course for credit and were enrolled via a placement exam or following experience in a preliminary Chinese course after displaying "Novice-High" ACTFL Levels of language proficiency. The participants had prior knowledge of Pinyin and were required to be able to type and read all new words, but were not required to hand copy words from memory on homework or quizzes.

On the first day of the experiment, 52 of the 60 sampled students participated in the Pretest, word practice session, and Posttest 1; the remaining eight students missed class. On the second day, 51 of the remaining 52 participants participated in Posttest 2. On the eighth day, 49 of the remaining 51 participants participated in Posttest 3. Analyses were conducted using *treated* rather than *intent-to-treat* methodologies, therefore taking these smaller samples into account.

Setting

This study was conducted using ASSISTments, an online learning platform that provides students with immediate feedback and teachers and researchers with robust student-level data

(Heffernan & Heffernan, 2014). While the system is more commonly used for mathematics education, content from other domains including statistics, physics, chemistry, electronics, biology, history, English, and now Chinese has been created and researched using its research infrastructure, *E-TRIALS*. Previous work on this platform has explored the intricacies of CFL, including approaches to word recognition (Lu, Ostrow & Heffernan, 2016; Lu, Ostrow & Heffernan, 2018) and the benefits of various feedback mediums (Lu, Xiong & Heffernan, 2017). All participants took daily in-class quizzes using their laptops and had completed at least two in-class quizzes on ASSISTments before the study began, establishing class-wide familiarity with the system. The Pretest, word practice session, and Posttests were delivered using ASSISTments during class time. Supplemental handwriting practice was conducted on paper. The pilot version of this work followed the same structure.

Materials

Two sets of five Chinese words were curated by the first author of this work. These sets, "Word Set X" (with words 1-5) and "Word Set Y" (with words 6-10) are shown in Table 1. All ten target words were two-character words selected from the Second-Class Vocabulary listed in the Outline of Graded Vocabulary for HSK (HSK Department, Chinese government, 1992). To increase the likelihood that these words were novel to study participants, we verified that none could be found in preliminary course textbooks.

For the word practice session, the first author developed two 85-second introductory videos for Word Set X and Word Set Y, available for reference at Lu (2018). These videos introduced the new words by reading each word aloud twice in Chinese, reading its English meaning aloud twice, and then reading the word twice aloud again in Chinese. Chinese characters, along with their Pinyin and English meanings, were shown on screen while each word was read aloud. To create

the practice session, the first author then developed seven question types for each word, as shown in Table 2. After watching the appropriate introductory video, participants cycled through these seven question types two to three times during a 13-minute, timed word-practice session. When students were randomly assigned to practice handwriting for a particular word set, they had to hand copy each word three times using proper stroke order on paper at the beginning of the practice round. A sample of the hand-copy practice sheet is provided in Figure 1.

The Pretest and the online portion of all Posttests contained the same ten problems (one for each word) presented randomly to each participant at each test point (see Appendix B). Each problem contained two sub-tasks focused on word recognition, requiring participants to enter the meaning or pronunciation of each word using Pinyin after viewing the Chinese characters for the word. All test points included online word recognition tasks and on paper handwriting tasks. Participants were expected to answer online portions using ASSISTments (following the format of the Pretest) and produce characters from memory on a provided paper worksheet, as prompted by the word's Pinyin and English meaning. The order of the ten target words on the handwriting worksheet was randomized for each test point. A Posttest worksheet sample is included as Appendix C.

Design

The present study was comprised of two conditions that all participants experienced using a crossover design: No-Handwriting (NH) and With-Handwriting (WH). The only difference between these conditions was that students in the WH condition began each round of practice with a handwriting exercise on paper that took approximately 30% of their practice time (see Figure 1), while those in the NH condition spent all of their practice time in ASSISTments cycling through word acquisition practices (see Table 2). In a counterbalanced fashion, participants received Word

Set X or Word Set Y in their randomly assigned condition, followed by an immediate posttest, before moving to the remaining set and the remaining condition, as shown in Figure 2. This approach was used to control for both word order and condition order. Minimal washout was thought to be required based on word novelty.

Within the NH condition, participants practiced the ten target words using ASSISTments. Thirteen minute practice sessions for each word set were designed using two rounds: the first round contained question types one to four, while the second contained question types five to seven (see Table 2). Rounds were assigned randomly to balance the distribution of question types. If participants were able to finish the second round within the allotted time, they were offered another iteration of the first- and second-rounds. To help participants become familiar with the novel words at the start of the practice session, question types one and two were presented in a linear fashion for each word. Subsequent rounds featured fully randomized question types. While practicing within ASSISTments, participants were able to access hints (orthographic, phonological, or semantic, as shown in Table 2) and could ultimately access the correct answer in order to move on to the next question. Hints always provided the missing facet in the word acquisition task; for instance, if participants were given a sound and asked to choose the orthography, the hint gave the meaning of the word. All materials are available at Lu (2018) for further reference.

Within the WH condition, students began each 13 minute practice session by completing a hand copying worksheet (see Figure 1). A prompt in ASSISTments (see Figure 1) directed students to complete and hand in this worksheet before moving on to the online content already described in the NH condition. As such, approximately on third of learning time in the WH condition was lost to handwriting practice in order to allow students to practice hand copying characters, mimicking the course structure followed by most CFL instructors.

Procedure

Blocking Participants and Randomization

To increase the chance that participants were randomly assigned to groups with equal variance, we blocked and randomized students based on prior knowledge, producing groups that were sufficiently homogenous. We used students' average grades from daily class quizzes prior to participation in the study as a measure of prior knowledge. To block and randomize participants, we rank ordered these scores, paired the two highest-performing participants and all subsequent pairs, and then randomly assigned each pair member to a condition progression (NH → WH or WH → NH). The same method was then used to divide condition progression groups into four sub-groups in order to counterbalance the influence of word set order (as shown in Figure 2).

Experimental Process

Prior to the experiment, the first author explained the procedure to participants and reminded them to take advantage of each question type and to use the hint function as necessary to keep moving forward. Participants then took an online Pretest assessing their knowledge of the ten target words. Participants then began the first round of their randomly assigned practice content. They were given 13 minutes for the first practice session, which began with an 85-second video introducing their assigned word set. After watching the video, participants were expected to work through the practice questions at their own pace. If assigned to the WH condition, participants were expected to first submit a handwriting practice worksheet before moving on to their online content. A Posttest (1a) on the assigned word set was provided immediately following the 13 minute practice session. This process was then repeated for a second session of practice in which participants experienced the alternate word set and condition. They again had 13 minutes, including an introductory video, to proceed through handwriting (if applicable) and online practice

as assigned. A Posttest (1b) on the new word set was provided immediately following this second practice session. Posttest 2 covering both word sets was given on day three of the experiment during a regularly scheduled course meeting. Posttest 3 was then given during scheduled course time eight days later. This experimental design is depicted in Figure 2.

Training and Delivery

The first author was in charge of enacting the procedure and monitored the whole experiment together with two teaching assistants. To ensure that the experiment ran smoothly, the first author, who also conducted the pilot study, trained two teaching assistants on the procedure using the flowchart shown in Figure 2. All 52 participants that attended day one engaged in the Pretest, the word practice session, and the immediate Posttests (1 a & b). A stopwatch was used to ensure that all participants received the same practice time (13 minutes) in each session, but Posttests were self-paced and participants could take as much time as they needed.

Scoring Protocol

All online Pretest and Posttest scores were sourced from log data collected by ASSISTments. Data were anonymized and are available at Lu (2018) for further reference. Participants' average posttest scores were calculated by adding the number of questions answered correctly and dividing that sum by the total number of questions on each test. Partial credit scores were generated for answers with otherwise accurate Pinyin using the wrong tones. The first author scored the handwriting portions of each Posttest (blindly) by calculating the number of characters written accurately divided by the total number of characters on each test.

Results

Our null hypothesis was that there would be no difference between NH and WH conditions on either online or on paper portions of word recognition and handwriting Posttests. Further, we

predicted that scores would decrease with each test point, as observed in our pilot work, denoting forgetting. Tables 3 and 4 present mean scores and standard deviations for all test points by condition, with Table 4 providing practice session metrics and a comparison to our pilot work.

As anticipated, most participants got zeroes on both online (M = 0.05, SD = 0.11) and on paper (M = 0.04, SD = 0.10) portions of the Pretest. As we counterbalanced practice orders and all students participated in both conditions, we performed two paired t-tests to determine whether there were within group differences at Pretest by condition. No significant differences were observed within groups in the online portions of the pretest, with students performing approximately the same in the NH condition (M = 0.05, SD = 0.09) and the WH Condition (M = 0.06, SD = 0.12), $t(51) = -1.0$, $p = .322$. Similarly, no significant differences were observed in the on paper portion of the pretest, with students performing approximately the same in the NH condition (M = 0.04, SD = 0.08) and the WH Condition (M = 0.05, SD = 0.12) , $t(51) = -1.42$, $p = .162$. Thus, we concluded that our groups were largely homogeneous and that we could proceed with planned Posttest analyses.

We then conducted a two-way repeated measures ANOVA to examine the main effects of condition (NH, WH) and posttest test point (1, 2, 3), and to examine the interaction effects of these independent variables on the online portion of posttest scores. There was a significant main effect of condition, $F(1, 47) = 6.49$, $p = .014$, with a moderate effect size ($\eta_p^2 = 0.12$) indicating both a statistically and practically significant difference between online word acquisition practice (M = 0.47, SE = 0.04) and with-handwriting practice (M = 0.41, SE = 0.04). We explored this main effect further using post hoc paired t tests. Figure 3a shows the scores of each condition by test point. As hypothesized, students exhibited a clear downward trend over time, representing forgetting. This graph also depicts relatively stable reliable differences between conditions. There

was a significant difference observed between conditions on Posttest 1, with students in the NH condition (M = 0.67, SD = 0.25) outperforming those in the WH condition (M = 0.56, SD = 0.28), $t(51) = 3.05$, $p = .004$, 95% CI [0.04, 0.18], Cohen's $d = 0.41$. There was also a marginally significant difference observed between conditions on Posttest 2, with students in the NH condition (M = 0.43, SD = 0.29) slightly outperforming those in the WH condition (M = 0.39, SD = 0.29), $t(50) = 1.95$, $p = .058$, 95% CI [-0.001, 0.09], Cohen's $d = 0.14$. However, significant differences were not observed between conditions by Posttest 3, with students scoring approximately the same in both the NH condition (M = 0.32, SD = 0.27) and the WH condition (M = 0.29, SD = 0.26), $t(48) = 0.87$, $p = .391$. These results suggest that the NH condition produced better word acquisition on average than the WH condition, especially in early measures of learning, reaffirming that time spent on handwriting instruction in CFL classes may be misplaced.

There was also a significant main effect of test point $F(2, 94) = 80.30$, $p<.001$, with an impressive effect size ($\eta_p^2 = 0.63$), indicating we could reject the null hypothesis that there was no change across test points. Results revealed evidence of learning immediately following practice sessions (Posttest 1) and were suggestive of forgetting as anticipated. Pairwise comparisons revealed significant differences between Posttest 1 (M = 0.61, SE = 0.03) and Posttest 2 (M = 0.40, SE = 0.04), $p < .001$, Posttest 1 and Posttest 3 (M = 0.31, SE = 0.04), $p < .001$, and Posttest 2 and Posttest 3, $p < .001$.

A second two-way repeated measures ANOVA was conducted to examine the main effects of condition (NH, WH) and test point (1, 2, 3), as well as the interaction effects of these independent variables, with regard to students' scores on the on paper portions of each posttest. There was a marginally significant main effect of condition, with students in the WH condition (M = 0.39, SE = 0.05) outperforming those in the NH condition (M = 0.31, SE = 0.04), $F(1, 36) = 3.91$,

$p = .056$, $\eta_p^2 = 0.10$. We explored this main effect further using post hoc paired *t* tests and observed a significant difference between conditions on Posttest 1 and Posttest 2. At Posttest 1, students in the WH condition ($M = 0.45$, $SD = 0.34$) outperformed those in the NH condition ($M = 0.33$, $SD = 0.26$), $t(51) = -2.92$, $p = .005$, 95% CI [-0.20, -0.04], Cohen's $d = 0.40$. This difference remained at Posttest 2, with students in the WH condition ($M = 0.35$, $SD = 0.29$) outperforming those in the NH condition ($M = 0.29$, $SD = 0.26$), $t(43) = -2.13$, $p = 0.039$, 95% CI [-0.12, 0.003], Cohen's $d = 0.22$. However, significant differences were no longer observed between conditions by Posttest 3, with students in the WH condition ($M = 0.31$, $SD = 0.28$) performing approximately the same as those in the NH condition ($M = 0.28$, $SD = 0.24$), $t(40) = -0.68$, $p = .503$. Figure 3b shows scores by condition and test point. These results suggest that while the WH condition led to better hand copying skill, differences were nonexistent at one week, suggesting no lasting impact of limited handwriting practice.

On paper portions of Posttests also exhibited a significant main effect of test point, $F(2, 72) = 15.33$, $p < .001$, with a large effect size ($\eta_p^2 = 0.30$), indicating we could reject the null hypothesis that there were no changes across test points. Pairwise comparisons revealed significant differences between Posttest 1 ($M = 0.40$, $SE = 0.05$) and Posttest 2 ($M = 0.33$, $SE = 0.04$), $p < .001$, and between Posttest 1 and Posttest 3 ($M = 0.30$, $SE = 0.04$), $p < .001$. There was no significant difference observed between Posttest 2 and Posttest 3, $p > .05$.

We also examined word practice session data from ASSISTments to help inform our analysis. Data revealed that each participant saw an average of 31.86 problems ($SD = 17.58$), made an average of 1.29 attempts per problem ($SD = 0.28$), and used an average of 0.14 hints per problem ($SD = 0.11$). Average time spent per problem was 27.84 seconds ($SD = 35.01$ seconds), while median time spent was 9.62 seconds ($SD = 7.85$ seconds). While assigned to the WH condition,

all participants completed the first round of handwriting practice within the allotted time. Twenty-one participants were then able to start a second round of handwriting practice, and three participants were able to start a third round. When considering completed handwriting problems, participants spent an average of 4.51 minutes (270.54 seconds, *SD* = 126.62 seconds) on handwriting practice, or 34.69% of overall practice time.

Discussion

We hypothesized that students would perform better on word recognition tasks if they were able to spend more of their practice time on word acquisition activities instead of allocating approximately 30% of their practice time to handwriting practice. Findings suggested a significant difference between practice with and without handwriting when considering immediate word recognition outcomes, favoring the removal of handwriting exercises ($p = .004$), with a marginally significant lasting impact three days later ($p = .058$). Although these results were not as robust as those observed in our pilot study, they trended in the same direction and reaffirmed that replacing handwriting with additional word acquisition training may lead to stronger word recognition in the short term. This gain was lost one week later, denoting a natural forgetting curve. These findings suggest that handwriting is an ineffective use of practice time; students scored significantly lower on word recognition tasks after spending 34.69% of their practice time on handwriting exercises, mimicking the structure of a traditional CFL course. This aligned with the findings of our pilot work in which students scored significantly lower on word recognition tasks after spending 30.38% of their practice time on handwriting exercises.

When considering students' performance on handwriting tasks, findings suggested a significant difference between practice with and without handwriting when considering immediate handwriting outcomes, favoring the inclusion of handwriting practice exercises ($p = .005$), with a

marginally significant lasting impact three days later (p = .039). This gain was lost one week later, again denoting a natural forgetting curve. These findings differed from those observed in our pilot work (interestingly, our pilot work did not reveal significant differences on handwriting outcomes), but they did not deviate from our expectations. The format of the handwriting portion of each Posttest asked students to write a word's characters as prompted by its Pinyin and meaning. The purpose of this exercise was to measure how well students could write the target words from memory, not to measure word recognition. It is logical that students who practiced handwriting were more likely to successfully write each word. However, it is worth reminding readers that proficiency in handwriting is not necessarily helpful for beginning CFL learners hoping to communicate and efficiently build their vocabulary. Thus, we prioritized gains in word recognition in the present study as suggesting greater promise for CFL learners. Plus, observed gains in both word recognition and handwriting were lost after one week, suggesting that spending 30% of course time on handwriting practice may be ineffective, but that simply filling that time with additional word acquisition practice may not be a viable solution to support long term retention.

Still, the results of both our pilot work and the present study indicated that spending 30-35% of practice time on handwriting hinders students' immediate word recognition. While these results do not suggest that handwriting should be removed from CFL curricula all together, they speak to the efficient use of students' learning time. As the amount of time students spend in a college course is relatively fixed, our results support that the best use of students' time, if their primary goals are communication and vocabulary growth, is on word acquisition tasks rather than handwriting practice.

Our pilot work and the present study both took place at the same university in two class sessions of the same course over a two year period. Both studies consisted of the same ten target

words. The two classes requirements for handwriting were the same, and neither class required students to be able to write characters from memory in everyday homework or exams. Both classes frequently utilized computers, were familiar with typing characters and words, and had been exposed to the learning platform used for study implementation, ASSISTments. Pretests for both the pilot study and the present work confirmed that students were not familiar with the target words. Learning curves, forgetting, and observed differences between conditions and across test points were largely comparable. Learning habits within practice sessions were also largely comparable, as shown in Table 4, but participants in the present study spent considerably more time per problem (M = 27.84, SD = 35.01, seconds) than those in the pilot study (M = 16.92, SD = 62.21). This would have led to fewer repetitions of target words experienced within allotted practice time in the present study, which may explain deflated scores on Posttest 2 in comparison to our pilot work.

One major limitation of this work was our measure of "long-term" results. It is possible that word recognition may change over longer periods of time and future work should explore longer-term outcomes by extending the duration of practice and by considering Posttests with greater delay. It is also important to note that handwriting practice is not the only activity that may facilitate performance on handwriting Posttests; word familiarity may also influence performance. Participants who did not practice handwriting may have been able to write out the target words based on familiarity from word acquisition practice. It is easy to imagine that characters with fewer strokes would be easier to remember and write out. Unfortunately, the present study did not control for stroke number across target words or consider confounding character difficulty.

Further, findings from the present study suggested that students scored significantly higher on word recognition Posttests when they were not subjected to handwriting practices. However, both our pilot work and the present study took place in a Chinese course that had adopted a

computer-assisted learning approach, one that did not focus on strengthening handwriting practices in its everyday structure. It is possible that this approach may have weakened students' performance on handwriting tasks and iterations of this work should be considered in more traditional CFL settings in which 30% of course time is regularly spent on handwriting practice. Future work could also extend beyond word recognition to examine the efficacy of handwriting on reading outcomes, in order to determine the optimal length of handwriting practice required to balance reading gains and resulting handwriting skill.

The present study improved upon our pilot work by pretesting students' handwriting ability, thereby enhancing the validity of the experiment. It also raised supplemental questions regarding the importance of "efficiency" as a criteria in CFL learning. While many studies have supported handwriting in Chinese instruction (e.g., Hsiung et al. 2017), most have failed to consider the efficient use of students' time and some have even failed to rule out learning time as a major confounding factor. As such, the present work fills a critical gap in CFL literature while presenting results that challenge the standard of practice in CFL instruction. Essentially, CFL students should save their strokes: handwriting practice is an ineffective use of instructional time.

Acknowledgement

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors of this work have been funded in part by NSF grants (e.g., 1440753, 1252297, 1109483, 1316736, 1535428, 1724889, 1636782, 1759229, 1822830, 1031398, 1903304, 1931523, 1940236, 1917713), the US Department of Education Institute for Education Sciences (e.g., R305A170137, R305A170243, R305A180401, R305A120125, & R305C100024), the Graduate Assistance in Areas of National Need program (e.g., P200A180088 & P200A150306), EIR and the Office of Naval Research (N00014-18-1-

2768), and Schmidt Futures. Views presented herein are those of the authors and are not necessarily endorsed by these funding agencies.

References

ACTFL. (2012). *ACTFL proficiency guidelines* [Electronic version]. Retrieved November 26th, 2017 from https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012.

Allen, J. R. (2008). Why learning to write Chinese is a waste of time: A modest proposal. *Foreign Language Annals, 41*(2): 237-251. doi: 10.1111/j.1944-9720.2008.tb03291.x.

Cao, F., Rickles, B., Vu, M., Zhu, Z., Chan, H., Harris, L., Stafura, J., Xu, Y. & Perfetti, C. A. (2013). Early-stage visual-orthographic processes predict long-term retention of word form and meaning: A visual encoding training study. *Journal of Neurolinguistics, 26*(4): 440–461. doi: 10.1016/j.jneuroling.2013.01.003.

Cao, F., Vu, M., Chan, D. H. L., Lawrence, J. M., Harris, L. N., Guan, Q., Xu, Y. & Perfetti, C. A. (2012). Writing affects the brain network of reading in Chinese: A functional magnetic resonance imaging study. *Human Brain Mapping*, *34*(7): 1670–1684. doi: 10.1002/hbm.22017.

Chan, D. W., Ho, C. S.-H., Tsang, S.-M., Lee, S.-H. & Chung, K. K. H. (2006). Exploring the reading-writing connection in Chinese children with dyslexia in Hong Kong. *Reading and Writing, 19*(6): 543–561. doi: 10.1007/s11145-006-9008-z.

Chinese Language Committee. (2009). *Modern Chinese common character list (Hanyu Tongyong Zibiao).* Beijing, China: Commercial Press.

Cui, Y. (1999) About the reforms in patterns of basic Chinese teaching. *Chinese Teaching in the World, 1999*(1): 3-8.

De Francis, J. (1984). *The Chinese Language: Fact and Fantasy*. Honolulu, HI: University of Hawaii Press.

Everson, M. E. (1998). Word recognition among learners of Chinese as a foreign language: Investigating the relationship between naming and knowing. *Modern Language Journal, 82*(2): 194-204. doi: 10.2307/329208.

Everson, M. E. (2011). Best practices in teaching logographic and non-Roman writing systems to L2 learners. *Annual Review of Applied Linguistics, 31*: 249–274. doi: 10.1017/S0267190511000171.

Guan, C. Q., Liu, Y., Chan, D. H. L., Ye, F. & Perfetti, C. A. (2011). Writing strengthens orthography and alphabetic-coding strengthens phonology in learning to read Chinese. *Journal of Educational Psychology, 103*(3): 509-522. doi: 10.1037/a0023730.

Heffernan, N.T. & Heffernan, C.L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education, 24*(4): 470-497. doi: 10.1007/s40593-014-0024-x.

Hsiung, H., Chang, Y., Chen, H. & Sung, Y. (2017). Effect of stroke-order learning and handwriting exercises on recognizing and writing Chinese characters by Chinese as a foreign language learners. *Computers in Human Behavior*, *74*: 303-310. doi: 10.1016/j.chb.2017.04.022.

HSK Department, Chinese government. (1992). *The Handbook Outlining the Graded Vocabulary for HSK (《汉语水平（词汇）等级大纲》).* Beijing, China: Beijing Language Institute Press.

Huang H. & Hanley, R. (1994). Phonological awareness and visual skills in learning to read Chinese and English. *Cognition, 54*(1): 73–98. doi: 10.1016/0010-0277(94)00641-W.

Jiang, X. (2007). An experimental study on the effect of the method of "teaching the learner to recognize characters more than writing." *Chinese Teaching in the World*, *2007*(2): 91-97.

Ke, C. (1998). Effects of strategies on the learning of Chinese characters among foreign language students. *Journal of the Chinese Language Teachers Association, 33*(2): 93–112. doi: 10.1111/j.1944-9720.1998.tb01335.x.

Lu, X. (2018). Data Documentation for "Save Your Strokes: Chinese Handwriting Practice Makes for Ineffective Use of Instructional Time in Second Language Contexts." Retrieved from: http://tiny.cc/LuOstrowHeffernan2019Data.

Lu, X., Ostrow, K. S. & Heffernan, N. T. (2016). *Improving students' character recognition ability using on-line practice*. Paper presented at ACTFL Annual Convention and World Languages Expo, Boston, MA.

Lu, X., Ostrow, K. S. & Heffernan, N. T. (In Preparation) *Comparing instructional policies for Chinese as a second language: Is handwriting worthwhile?*

Lu, X., Ostrow, K. S. & Heffernan, N. T. (2018). *Understanding the complexities of Chinese word acquisition within an online Learning Platform*. Paper presented at the National Chinese Language Conference. Salt Lake City, UT.

Lu, X., Xiong, X. & Heffernan, N. T. (2017). *Experimenting choices of video and text feedback in authentic foreign language assignments at scale*. Fourth Annual ACM Conference on Learning at Scale. Boston, MA.

Packard, J. L., Chen, X., Li, W., Wu, X., Gaffney, J. S., Li, H. & Anderson, R. C. (2006). Explicit instruction in orthographic structure and word morphology helps Chinese children learn to write characters. *Reading and Writing, 19*(5): 457–487. doi: 10.1007/s11145-006-9003-4.

Shen, H. H. (2004). Level of cognitive processing: Effects on character learning among non-native learners of Chinese as a foreign language. *Language and Education, 18*(2): 167–182. doi: 10.1080/09500780408666873.

Shen, H. H. (2005). An investigation of Chinese-character learning strategies among non-native speakers of Chinese. *System, 33*(1): 49-68. doi: 10.1016/j.system.2004.11.001.

Siok, W. T., Perfetti, C. A., Jin Z., Tan, L. H. (2004). Biological abnormality of impaired reading is constrained by culture. *Nature, 431*(7004): 71-76. doi: 10.1038/nature02865.

Tan L. H., Perfetti. C. A. (1999). Phonological activation in visual identification of Chinese two–character words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2): 382–393.

Tan, L. H., Spinks, J. A., Eden, G. F., Perfetti, C. A. & Siok, W. T. (2005). Reading depends on writing, in Chinese. *Proceedings of the National Academy of Sciences of the United States of America, 102*(24): 8781–8785. doi: 10.1073/pnas.0503523102.

Tan L. H., Xu M., Chang C. Q. & Siok W. T. (2013). China's language input system in the digital age affects children's reading development. *Proceedings of the National Academy of Sciences of the United States of America, 110*(3): 1119-1123. doi: 10.1073/pnas.1213586110.

Warschauer, M. & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching*, 31(2): 57-71. doi: 10.1017/S0261444800012970.

Xie, T. (2014). 中文教学与时俱进. In X. Zhou (Eds.), *Chinese Language in the World, 3:* 69-74. Guangzhou, China: Zhongshan University Press. Retrieved from: http://web.csulb.edu/~txie/papers/Chinese_teaching_in_modern_society.pdf

Xu, Y., Chang, L., Zhang, J. & Perfetti, C. (2013). Reading, writing, and animation in character learning in Chinese as a foreign language. *Foreign Language Annals, 46*(3): 423-444. doi: 10.1111/flan.12040.

Zhu, Y., Shum, S. M., Tse. S. B. & Liu, J. J. (2016). Word-processor or pencil-and-paper? A comparison of students' writing in Chinese as a foreign language. *Computer Assisted Language Learning, 29*(3): 596-617. doi: 10.1080/09588221.2014.1000932.

Zhu, Z., Liu, L., Ding, G. & Peng, D. (2009). The influence of Pinyin typewriting experience on orthographic and phonological processing of Chinese characters. *Acta Psychologica Sinica*, 41(09): 785-792. doi: 10.3724/SP.J.1041.2009.00785.

Appendix A

Entering Chinese Characters on a Computer

All computers come with built in input method editors. First time users may need to set Chinese as their input method, allowing them to type Chinese using Pinyin. Pinyin is the standard Romanized system for transliterating Chinese and borrows from the English alphabet. As such, users can type Pinyin using a standard QWERTY keyboard. To generate a word, users enter Pinyin based on how the word sounds and a list pops up displaying characters or multi-character words that match the supplied Pinyin (as shown below).

Many Chinese words and characters sound similar but look different, so the user is prompted to choose the intended character or word from a list. The image above shows what the user sees when typing “Shanghai.” The first listed choice provides the characters for writing the city name Shanghai.

Appendix B

Pretest and Posttests

1) Problem #PRABEJGC "PRABEJGC - pre03"

A) Write down the English meaning of Chinese word "标准". If you do not remember, please feel free to answer "dnk".

B) Type out Pinyin of the Chinese word: 标准

Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out “ni3hao3”. If you do not remember, please feel free to answer "dnk".

2) Problem #PRABEJGD "PRABEJGD - pre04"

A) Write down the English meaning of Chinese word "原因". If you do not remember, please feel free to answer "dnk".

B) Type out Pinyin of Chinese word 原因:

Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out “ni3hao3”. If you do not remember, please feel free to answer "dnk".

3) Problem #PRABEJGF "PRABEJGF - pre01"

A) Write down the English meaning of Chinese word "稳定". If you do not remember, please feel free to answer "dnk".

B) Type out Pinyin of the Chinese word: 稳定/穩定

Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out “ni3hao3”. If you do not remember, please feel free to answer "dnk".

4) Problem #PRABEJGJ "PRABEJGJ - pre02"

A) Write down the English meaning of Chinese word "仔细". If you do not remember, please feel free to answer "dnk".

B) Type out Pinyin of the Chinese word: 仔细/仔細

Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out “ni3hao3”. If you do not remember, please feel free to answer "dnk".

5) Problem #PRABEJGN "PRABEJGN - pre05"

A) Write down the English meaning of Chinese word "其实". If you do not remember, please feel free to answer "dnk".

B) Type out Pinyin of the Chinese word: 其实/其實

Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out “ni3hao3”. If you do not remember, please feel free to answer "dnk".

6) Problem #PRABEJGE "PRABEJGE - pre06"

A) Write down the English meaning of Chinese word "距离". If you do not remember, please feel free to answer "dnk".

B) Type out Pinyin of the Chinese word: 距离/距離

Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out “ni3hao3”. If you do not remember, please feel free to answer "dnk".

7) Problem #PRABEJGG "PRABEJGG - pre07"

A) Write down the English meaning of Chinese word "后悔". If you do not remember, please feel free to answer "dnk".

B) Type out Pinyin of the Chinese word: 后悔/後悔

Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out “ni3hao3”. If you do not remember, please feel free to answer "dnk".

8) Problem #PRABEJGH "PRABEJGH - pre08"

A) Write down the English meaning of Chinese word "熟悉". If you do not remember, please feel free to answer "dnk".

B) Type out Pinyin of the Chinese word: 熟悉

Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out “ni3hao3”. If you do not remember, please feel free to answer "dnk".

9) Problem #PRABEJGK "PRABEJGK - pre09"

A) Write down the English meaning of Chinese word "主动". If you do not remember, please feel free to answer "dnk".

B) Type out Pinyin of the Chinese word: 主动/主動

Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out “ni3hao3”. If you do not remember, please feel free to answer "dnk".

10) Problem #PRABEJGM "PRABEJGM - pre10"

A) Write down the English meaning of Chinese word "于是". If you do not remember, please feel free to answer "dnk".

B) Type out Pinyin of the Chinese word: 于是/於是

Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out “ni3hao3”. If you do not remember, please feel free to answer "dnk".

11) Problem #PRABGMJW "PRABGMJW - Please ask for ap..."

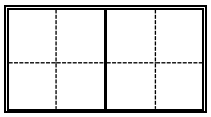
Please ask for a test sheet from teacher, and try your best to write down characters based on meaning and pinyin. Feel free to leave any of them blank. Once completed, please click "completed" on ASSISTments.

Appendix C

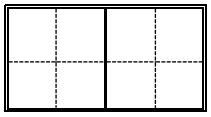
Posttest Worksheet Sample

Please try to write down characters for the following words:

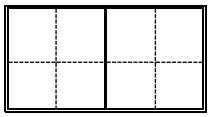
1. Meaning: be familiar with; know well; be acquainted with
   Pinyin: shúxī



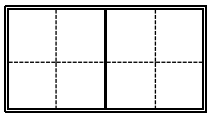
2. Meaning: hence; as a result; and then
   Pinyin: yúshì



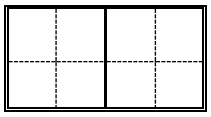
3. Meaning: Standard, criterion
   Pinyin: biāozhǔn



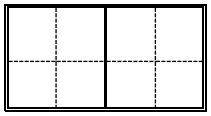
4. Meaning: Carefully; attentive
   Pinyin: zǐxì



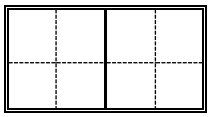
5. Meaning: Distance; range
   Pinyin: jùlí



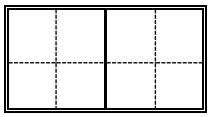
6. Meaning: reason; cause
   Pinyin: yuányīn



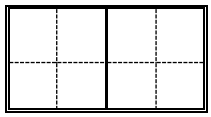
7. Meaning: Regret
   Pinyin: hòuhuǐ



8. Meaning: Initiative
   Pinyin: zhǔdòng



9. Meaning: actually, in fact, as a matter of fact
   Pinyin: qíshí



10. Meaning: stable
    Pinyin: wěndìng

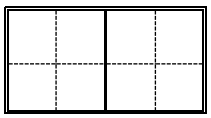


Table 1

Target words, Pinyin Pronunciations, and English Meanings

Target Word	Pinyin	English Meaning
Word Set X		
稳定	wěndìng	stable
仔细	zǐxì	carefully; attentive
标准	biāozhǔn	standard; criterion
原因	yuányīn	reason; cause
其实	qíshí	actually; in fact; as a matter of fact
Word Set Y		
距离	jùlí	distance; range
后悔	hòuhuǐ	regret
熟悉	shúxī	be familiar with; know well
主动	zhǔdòng	initiative
于是	yúshì	hence; as a result; and then

Table 2

Question, Answer, and Hint Types and Correct Responses for the word 稳定

Question Type	**Answer Type**	**Hint Type**	**Correct Response**
1. Choose the correct meaning of the word.	Multiple Choice	wěndìng	稳定 → stable
2. Choose the correct Pinyin of the word.	Multiple Choice	stable	稳定 → wen3ding4
3. Listen and choose the word that matches the sound.*	Multiple Choice	stable	wen3ding4 → 稳定
4. Type out the English meaning of the word.	Entry	wěndìng	稳定 → stable
5. Choose the word that matches the meaning.	Multiple Choice	wěndìng	stable → 稳定
6. Type out Pinyin of the Chinese word. Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out "ni3hao3".	Entry	stable	稳定 → wen3ding4
7. Type out words based on meaning.	Entry	wěndìng	stable → 稳定

*Audio was provided through a YouTube video showing the question, with the word read aloud twice. Students could replay the video if needed.

Table 3

Sample Sizes, Pairwise Comparisons, Means and (Standard Deviations) for Pretest and Posttest Test Points by Condition for Online and On Paper Scores

	Online						On Paper					
	NH		WH				NH		WH			
Test Point	*N*	*M (SD)*	*N*	*M (SD)*	*t*	*p*	*N*	*M (SD)*	*N*	*M (SD)*	*t*	*p*
Pretest	52	0.05 (0.09)	52	0.06 (0.12)	-1.00	.322	52	0.04 (0.08)	52	0.05 (0.12)	-1.42	.162
Posttest 1	52	0.67 (0.25)	52	0.56 (0.28)	3.05	.004**	52	0.33 (0.26)	52	0.45 (0.34)	-2.92	.005**
Posttest 2	51	0.43 (0.29)	51	0.39 (0.29)	1.95	.058	44	0.29 (0.26)	44	0.35 (0.29)	-2.13	.039*
Posttest 3	49	0.32 (0.27)	49	0.29 (0.26)	0.87	.391	41	0.28 (0.24)	41	0.31 (0.28)	-0.68	.503

Notes. NH = No Handwriting. WH = With Handwriting.

* $p < .05$.

** $p < .01$.

Table 4

Mean and (Standard Deviation) for Pretest, Posttest Test Points, and Practice Session Metrics across Pilot and Present Studies by Condition for Online and On Paper Scores

			Pilot Study	Present Study
Pretest	Online	NH	0.04 (0.07)	0.05 (0.09)
		WH	0.05 (0.08)	0.06 (0.12)
	On Paper	NH	N/A	0.04 (0.08)
		WH		0.05 (0.12)
Posttest 1	Online	NH	0.76 (0.22)	0.67 (0.25)
		WH	0.64 (0.27)	0.56 (0.28)
	On Paper	NH	0.09 (0.10)	0.33 (0.26)
		WH	0.13 (0.13)	0.45 (0.34)
Posttest 2	Online	NH	0.50 (0.28)	0.43 (0.29)
		WH	0.41 (0.30)	0.39 (0.29)
	On Paper	NH	0.08 (0.09)	0.29 (0.26)
		WH	0.10 (0.12)	0.35 (0.29)
Posttest 3	Online	NH	0.39 (0.27)	0.32 (0.27)
		WH	0.32 (0.24)	0.29 (0.26)
	On Paper	NH	0.11 (0.10)	0.28 (0.24)
		WH	0.11 (0.12)	0.31 (0.28)
Problems Seen			37.46 (12.12)	31.86 (17.58)
Attempts			1.27 (0.22)	1.29 (0.28)
Hints			0.15 (0.11)	0.14 (0.11)
Time per Online Problem (sec)			16.92 (62.21)	27.84 (35.01)
			Median = 7.11	*Median = 9.62*
Time per Hand-Copy Practice (sec)			236.98 (95.82)	270.54 (126.62)
			30.38% of overall practice time	*34.69% of overall practice time*

Note. NH = No Handwriting. WH = With Handwriting.

Figure 1

Hand-copy practice prompt and worksheet example

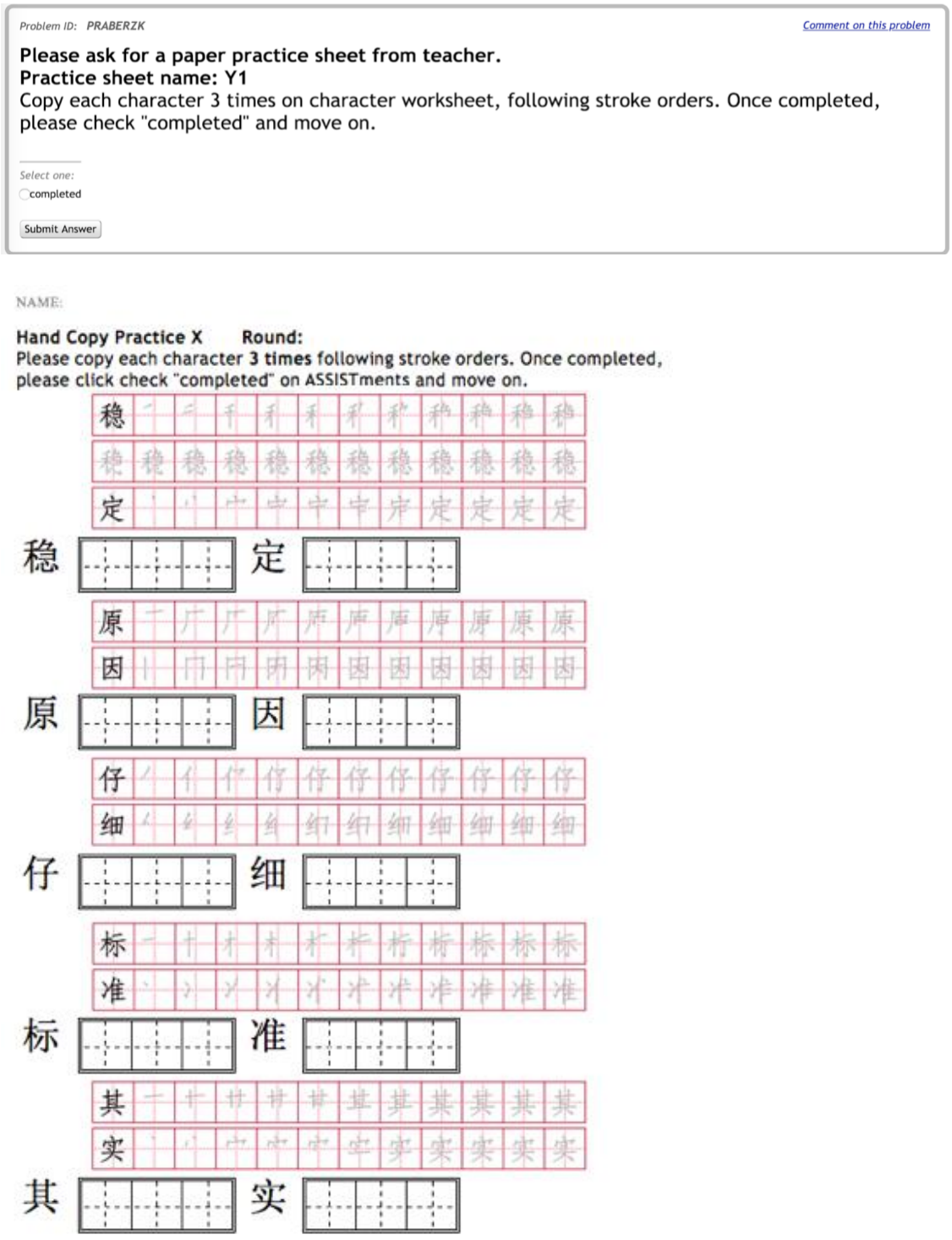


Figure 2

Experimental design flowchart

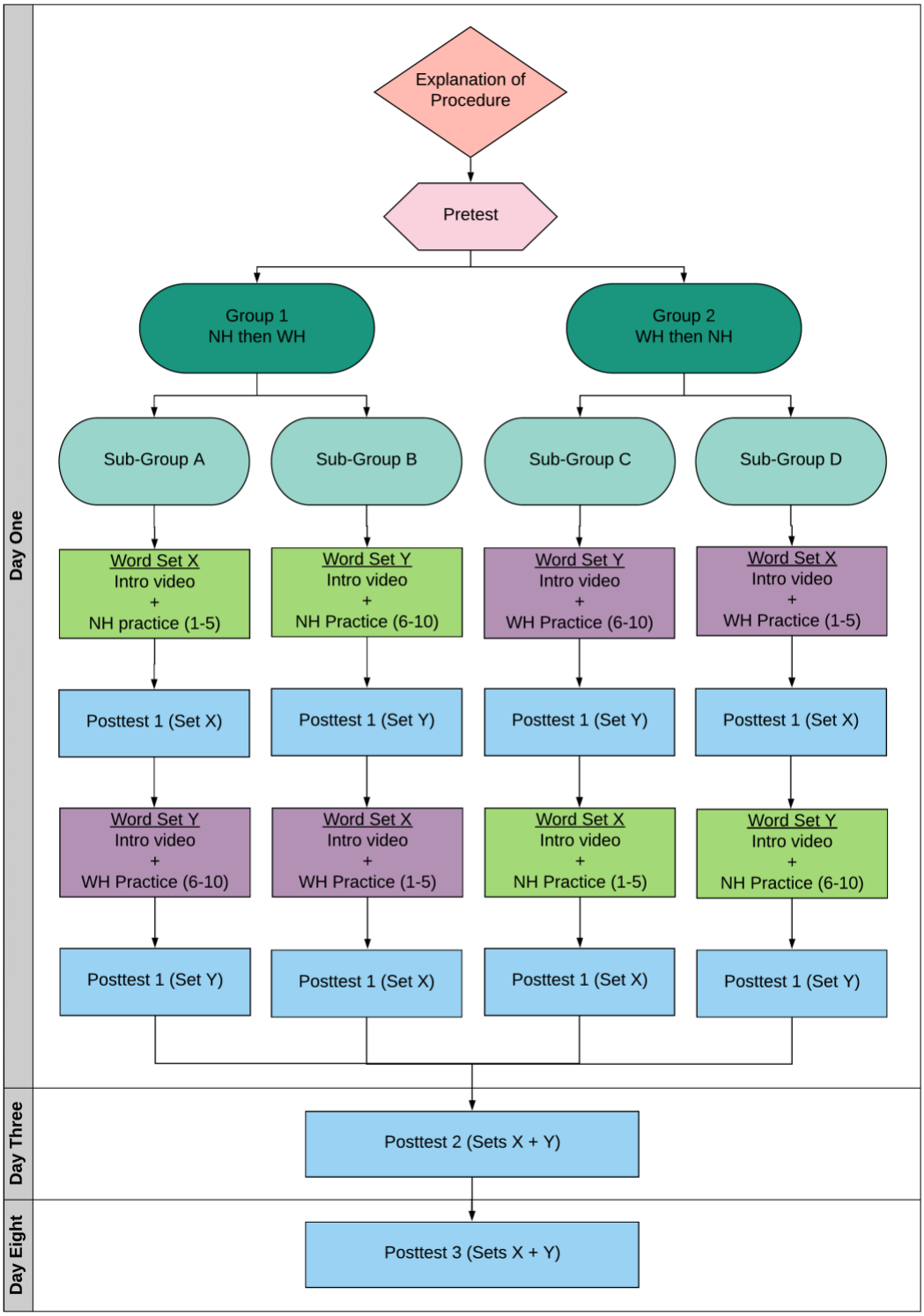
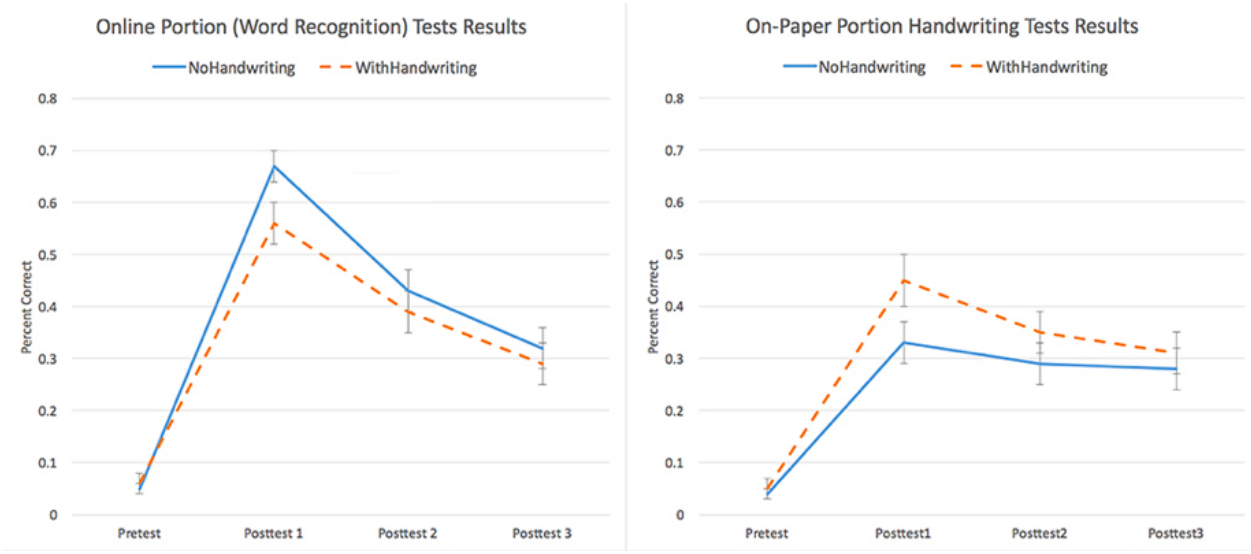


Figure 3

Percent Correct on Pretest and each Posttest Test Point by Condition for a) Online Word Recognition Tasks (left) and b) On Paper Handwriting Tasks (right)



Chapter 4: Save Your Strokes: Further Studies on the Efficiency of Learning Chinese Words Without Hand-Writing

This chapter presents the manuscript accepted by *Transforming L2 Hanzi Teaching & Learning in the Age of Digital Writing: Theory and Pedagogy (《电写时代汉字教学的理论与实践》), Routledge, UK* for publication as a book chapter.

Lu, X., Ostrow, K. S., Yang, Q., & Heffernan, N. T.

Abstract

This study aimed to test the effectiveness of no-hand-writing (NH) and with-hand-writing (WH) practices for word recognition among lower-level Chinese as foreign language learners. It included four experiments in an online learning environment. Each experiment had NH or WH conditions for an allotted time. The online post-test results for Experiment 1 revealed significant differences between the NH and WH conditions. Experiments 2 and 3 were replications of Experiment 1 to test the findings' generalizability. They both confirmed Experiment 1's results. Experiment 4 further explored the results by adding another condition to the comparison: 70% of NH practice, which eliminated 30% of hand-writing time and shortened the participants' practicing time to 70% (online only). Participants still performed better on the word recognition tasks when their practice time did not include hand-writing tasks, and no differences were found between the 70% NH condition and the WH conditions. The practical implications for teaching and learning CFL word recognition are discussed.

Keywords: Chinese word recognition, Chinese as a foreign language (CFL), online learning, hand-writing, second language acquisition, classroom-based experiment

Introduction

Achieving functional literacy in L2 Chinese is a well-documented challenge (Allen, 2008; Everson, 2009). According to a survey by Lu et al. (2019), the majority of Chinese as foreign language (CFL) instructors still try to improve students' word recognition through traditional means of assigning hand-writing tasks both in and out of class. Some instructors (70.37%) self-reported "strict hand-writing requirements", that is, the requirement that students to be able to handwrite all the words they have taught them (rather than, say, a subset). Lu and colleagues' survey also found that even instructors without strict hand-writing requirements during class time (22.22%) still expected lower-level students to spend an average of 30% of their homework time hand-writing, though unlike instructors with strict requirements, these instructors only required students to be able to handwrite a portion of their target vocabulary in an effort to alleviate some of the burden of hand-writing. With the emerging interest in pedagogies focused on e-writing in Chinese, recent research has questioned the efficiency of hand-writing in CFL learning. Among the factors contributing to CFL learning, word recognition plays a fundamental role in reading and comprehension (Grabe, 1991), and some scholars have insisted that hand-writing is a necessary precursor to reading ability (e.g., Guan et al., 2011). However, most studies in the CFL field have focused on character recognition instead of word recognition, and even fewer studies have focused on the efficiency of CFL learning when discussing these issues.

This chapter summarizes a series of four experiments measuring the efficiency of word recognition practice with or without hand-writing practice for lower-level learners. Experiment 1 examined word recognition while manipulating the conditions (no-hand-writing [NH] practice and with-hand-writing [WH] practice) and post-test test points (1 [immediate], 2 [one-day delay],

and 3 [one-week delay]). The NH condition outperformed the WH condition on all the three post-tests. Experiment 2 (Lu et al., 2019), which was published separately, served as a replication of the first experiment in the same institution with the same learning material. The hypothesis was that students' word recognition performance would improve if they spent more time practicing word acquisition activities online rather than allocating some of their practice time to hand-writing practice. The results again revealed a significant difference in short-term learning outcomes between the NH and WH conditions, and suggested that removing hand-writing exercises led to better word recognition performance. The first two experiments revealed a significant difference in learning outcomes between the NH and WH conditions, particularly in the short-term. Namely, the results strongly suggested that lower-level students using online questions to practice new words without hand-writing recognized words better than students practicing hand-writing for the same amount of time. Experiment 3 sought to further develop this line of research by replicating the experiment used for Experiments 1 and 2 in a different institution, and Experiment 4 further advanced the research by adding an additional condition to investigate the effectiveness of the time spent on hand-writing practice. The primary goal of this collective body of experiments was to explore whether teachers should ask students to spend time on hand-writing in the early stages of learning practice when word recognition and communication are the ultimate goals. In other words, if it is assumed that a primary goal of learning is word recognition and communication for lower-level (i.e., beginner) learners, is hand-writing actually necessary or facilitative of these learning outcomes?

Literature Review

Chinese Word Recognition Differs from Character Recognition

A word is defined as “a speech sound or series of speech sounds that symbolizes and communicates a meaning usually without being divisible into smaller units capable of independent use” (Brown, 2020). In Chinese, a word can be formed by one, two, three, or more characters/syllables. Statistics have shown that two-character words account for more than 60% of all Chinese words (Su, 2001), and 75.57% of CFL target words are two-character words (Hanban, 2010). Word recognition is “the ability to accurately identify printed words” (Hayes & Flanigan, 2014). Word recognition should be effortless and automatic to ensure reading comprehension (Garnett, 2011). Previous studies have suggested that word recognition is central to reading proficiency (e.g., Koda, 1996). As such, the vocabulary in most CFL textbooks and curricula is built upon the foundation of words, not characters. Although controversy exists, the majority of studies have concurred that Chinese reading is word based (Li & McBride-Chang, 2014). Li et al. (2014) conducted an eye movement study to explore reading in Chinese from a psychological perspective, finding that Chinese reading is similar to that of alphabetic languages and is word-based rather than character-based. A meta-analysis of fMRI studies by Zhao et al. (2017) further supported this finding.

Chinese word and character reading play different roles in Chinese reading comprehension (Pan, et al., 2021; Yang, et al., 2022). The fundamental process of word recognition differs significantly from character recognition. Characters function as lexical morphemes in most Chinese words, hence recognizing that single characters do not equal the word-level combination of sound and writing system. For example, The Chinese character "水" (shuǐ) means "water". However, the word "水果" (shuǐ guǒ), which consists of two characters, means "fruit". This highlights how the Chinese reading system operates on two levels - character

level and word level. Two-character Chinese words undergo a constituent character assembly process that increases processing complexity and is unnecessary for single characters (Li & McBride-Chang, 2014; Tan & Perfetti, 1999). Therefore, findings based on character recognition may not explain word recognition in Chinese.

Despite the importance of word recognition for fluent reading, research has been mostly performed on Chinese character recognition (e.g. Ke, 1998; Tan et al., 2005; Xu et al., 2013; Hsiung et al., 2017). Few studies have focused on Chinese word recognition. Guan et al. (2011) compared the effects of hand-writing, reading, and typing with Pinyin on CFL learning and determined that the hand-writing condition outperformed the reading condition in word recognition. However, they ignored the difference between Chinese characters and words, and they used characters as testing material to draw conclusions on word recognition. Lyu et al. (2021) reviewed 43 comparison studies of e-writing and hand-writing in Chinese language learning and found only one study (Lu et al., 2019) comparing the learning results at word level. Lu et al. (2019) reported that reading and typing practice was a more efficient way of learning word recognition than hand-writing.

Reading Does not Depend on Hand-Writing

Language reading and hand-writing are two distinct functional systems in the brain that interact predictably (Berninger et al., 2002). Although Chinese word recognition and hand-writing are correlated, they are separable. Bi et al. (2009) noted a Chinese individual with brain damage whose capacities challenged that connection, concluding that "reading does not depend on writing, even in Chinese" (p.1198). From a practical perspective, these findings indicate that learners can learn how to read without knowing how to handwrite.

The Pinyin input method employs pronunciation information of words to write digitally (i.e., typing via a keyboard with Roman letters), allowing L1 and L2 users alike to retrieve word information through sound when composing text via technology. Many studies have shown that sound plays a critical role in word recognition and it is automatically activated (Kilpatrick, 2020). Zhu et al. (2016) recommended that CFL beginners should rely on the Pinyin input method instead of hand-writing practice for more efficiency in essay writing tasks. Zhang (2021, also this volume) described the Pinyin input method in detail, summarizing the theories related to Pinyin-based e-writing methods from a cognitive psychology perspective and noting that Pinyin-based e-writing employs the holistic processing of phonological–visual chunks, while hand-writing is a sub-lexical process that does not necessitate activation, encoding, or retrieval of sound information. She performed a longitudinal study on the learning results of e-writing versus hand-writing, which provided additional evidence that e-writing improves word recognition at the sentence level.

Research Gaps and the Present Chapter

When trying to improve CFL word recognition, one must consider efficiency. As Yue (2017) demonstrated how most Chinese teachers face the challenge of raising language ability to sustain students' motivation while avoiding intimidating them with vocabulary learning and boring them with memorization. To address this issue, it is crucial to consider efficiency when measuring learning results. Given that students' learning time is limited, the efficiency of CFL word recognition practice has become increasingly important. Would e-writing without hand-writing be a more efficient way of learning Chinese for lower-level learners? Although studies have compared CFL online and hand-writing learning results (Lyu et al., 2021), many of these failed to consider efficiency by not measuring the learning results with a time dimension.

Another issue to consider is whether hand-writing practice is the best way to improve CFL word recognition when e-writing is widely used by proficient L2 users and all L1 users. Substantial research has found some evidence of a relationship between hand-writing practice and character recognition in the contexts of native language learning (such as Tan et al., 2005; Tan et al. 2013; Packard et al. 2006) and foreign language learning (such as Osborne et al., 2020; Xu et al., 2013). However, few studies have focused on the effects of hand-writing on word recognition, especially in the CFL field. The present study considered ways to improve student word recognition and the role (or possible lack thereof) of hand-writing for this purpose.

Finally, it is worth noting that learning strategies for CFL lower-level learners may differ from those of higher-level learners (Shen, 2005). Most research related to CFL character or word recognition has been conducted at the advanced level or in a study-abroad program in Chinese-speaking regions, and most of those participants have already mastered hand-writing skills, and also have substantial oral language skills and mental lexicons. Little word recognition research has been performed in lower-level CFL classrooms with students unfamiliar with the Chinese writing system.

The present series of studies aimed to measure the efficiency of CFL word recognition by comparing a WH condition to a NH condition in an online learning environment, in addition to considering the effects of supplemental hand-writing practice. Collectively, these studies helped identify the factors that could contribute to developing more efficient methods of learning CFL. The current study attempted to answer the following research questions:

1. Which condition is more efficient for lower-level learners' Chinese word recognition, WH or NH?

2. Our previous study indicated that students in the WH condition spent at least 30% of their learning time on hand-writing, but the 30% of hand-writing time did not result in learning gains to match or outperform the non-hand-writing group. Would a difference exist between including or not including the 30% hand-writing time? In other words, could students avoid the 30% hand-writing time and still achieve the same learning goal with only 70% NH practice?

Experiment 1: Intermediate Learners

Method

Participants

Participants were 48 undergraduate students enrolled in an Intermediate Level Chinese class from a private university in the U. S. They were non-heritage learners on their third semester of CFL learning, had prior knowledge of Pinyin and were knowledgeable about typing Chinese with a computer.

Setting

This experiment as well as the following three were conducted in class using ASSISTments, an online learning platform that supports student learning with hints and immediate feedback (Heffernan & Heffernan, 2014). Two in-class quizzes on ASSISTments were assigned before the experiment began to help participants become familiarized with the system. The pretest, word practice assignments, and post-tests were delivered using ASSISTments during class time. The hand-writing pretest and post-tests, as well as the supplemental hand-writing practice during the experiment, were delivered on paper during class time.

Materials

Two sets of five novel Chinese words (Appendix A) were selected for target words: “Set X” (with words 1-5) and “Set Y” (with words 6-10). It was verified that the words could not be found in preliminary course textbooks. All ten target words contained two characters and were selected from the Second-Class Vocabulary listed in the Outline of the Graded Vocabulary for the HSK [汉语水平考试] (HSK Department, 1992).

For the practice content, the following materials were developed: (a) Two 85-second introductory videos for Set X and Set Y, which introduced each new word by reading it twice in Chinese, reading its English meaning twice, and then reading it twice again in Chinese, with the characters, Pinyin, and English meanings displayed on the screen. (b) Seven online question types were developed for word practice content as shown in Table 1. Question types varied by providing one of three elements (orthographic, phonological, or semantic) in the word acquisition task and required one additional element. While practicing within ASSISTments, participants were able to access hints and could ultimately obtain the answer to move on. To help participants become familiar with the new words at the start, practice in each set were designed using two parts: the first contained question types one to four, while the second contained question types five to seven. At the first part in the first round, each word for question types one and two appeared in a linear order. The second part question types and orders were randomized. If participants were able to finish the first round with both parts within 13 minutes, they would be offered a second round of practice with all the question types in a randomized order. If they could finish the second round within the given time limit, they would then be offered additional rounds until the time is up. (c) A hand-writing practice sheet (see Appendix B) was developed for participants to practice hand-writing in the WH condition. In the WH condition, the hand-writing practice sheet was provided at the beginning of each round of practice. No matter

receiving hand-writing practice or not, all participants spent an equal amount of time on the practice content across conditions.

Table 1

Question types, answers, hint types, and correct responses for a sample word 稳定 in all the four experiments

	Question Type	Answer Type	Hint Type	Correct Response
1	Choose the correct meaning of the word.	Multiple Choice	wěndìng	稳定 → stable
2	Choose the correct Pinyin of the word.	Multiple Choice	stable	稳定 → wěndìng
3	Listen and choose the word that matches the sound.*	Multiple Choice	stable	"wěndìng" → 稳定
4	Type out the English meaning of the word.	Entry	wěndìng	稳定 → stable
5	Choose the word that matches the meaning.	Multiple Choice	wěndìng	stable → 稳定
6	Type out Pinyin of the Chinese word. Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out "ni3hao3".	Entry	stable	稳定 → wen3ding4
7	Type out words based on meaning.	Entry	wěndìng	stable → 稳定

Note. Audio was provided through a YouTube video showing the question, with the word read aloud twice. participants could replay the video if needed.

The pretest and the online part of all the post-tests contained the same ten problems which were presented in random order to each participant in each test. Each problem contained two sub-questions which focused on word recognition by showing participants a Chinese word and requiring them to enter the meaning and pronunciation using Pinyin. An example of these problems was shown in Table 2.

All post-tests included a hand-writing task right after the online recognition task. Participants were expected to produce characters from memory on a worksheet, as prompted by the word's Pinyin and English meaning. The presenting order of the ten target words on worksheets was randomized for each post-test. An example of post-test worksheet questions was shown in Figure 1:

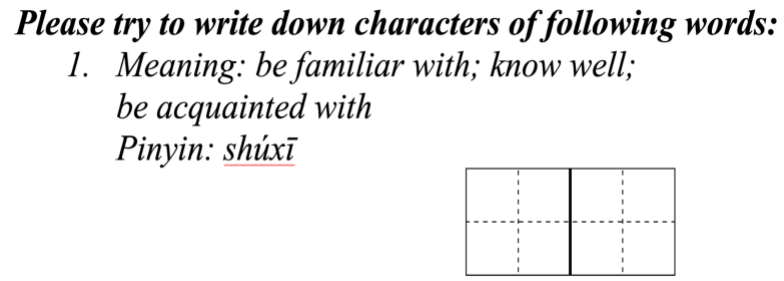
Table 2

An example of problems with two sub-questions in the online pretest and post-tests in Experiment 1

Question 1a	Write down the English meaning of Chinese word "稳定". If you do not remember, please feel free to answer "dnk".
Question 1b	Type out Pinyin of the Chinese word: 稳定. Please type out tones as well, and use 1 for first tone, 2, for second tone, 3 for third tone, 4 for fourth tone and 0 for neutral tone. For example: if you see 你好, you should type out "ni3hao3". If you do not remember, please feel free to answer "dnk".

Figure 1

Example of hand-writing post-test questions



Design

This experiment was a 2*3 within-subject design with two conditions and three test points. An online pretest was conducted to test participants' prior knowledge of the target words. The two conditions that all participants experienced in a crossover fashion ere: no-hand-writing (NH) and with-hand-writing (WH). The crossover design means that all participants were exposed to both conditions but in a different order. In the current study, some participants started with the NH condition and then switched to the WH condition, while others started with the WH condition and then switched to the NH condition. This was done to control for individual differences in the participants' responses and to increase the statistical power of the study. Appendix C figure a illustrated the crossover design used in the experiment. Both conditions had a time limit of 13 minutes. The only difference between the two conditions was the inclusion of hand-writing practice. Participants in the NH condition spent all of their practice time in the online environment answering the seven question types (Table 1) round after round while participants in the WH condition started each practice round by completing a hand-writing practice sheet (see Appendix B), before proceeding to the same online practice within the given

time limit. Participants were randomly assigned to four subgroups for the purpose of the crossover design, and these four subgroups were counterbalanced in a way that each subgroup experienced the two conditions in a different order (Appendix C figure a). Participants received Set X or Set Y in one condition, followed by a post-test, before alternating to the other set and the remaining condition. Three post-tests were assigned on the first and second day, in addition to one week later.

Procedure

Piloting. Before the experiment, a small-scale pilot experiment was conducted to determine the ideal number of target words and the amount of time necessary to complete at least one round of practice. A student from the same university who was not enrolled in the course, but had a similar Chinese language learning background to those enrolled, was recruited to test the online and hand-writing practice. The instructor recorded the amount of time she spent in each type of practice and decided to give 13 minutes for each set of question practice (question set X or Y).

Block participants and randomization. Participants were anonymized and rank-ordered based on daily quiz grades. Then, the top two performing participants were paired, and each paired member was randomly assigned to a condition progression, thereby increasing the chance that participants in each subgroup had equal variance. The same method was repeated to divide each condition progression into two subgroups.

Experimental process. On the first day of the experiment, the instructor explained the entire procedure to participants prior to the experiment, reminding them to take advantage of different types of practices by continuing to progress. Then, participants began with an online pretest to assess knowledge of the ten target words. We verified that the words could not be

found in preliminary course textbooks. However, since students may learn from different sources on their own, we expected that some participants might know certain aspects of some of the target words. Upon completion of the pretest, participants began practicing content by watching the introductory video that provided them with a specific word set. After the video, participants in the NH condition worked through the practice questions online at their own pace round after round until the time limit was reached, while participants in the WH condition were provided hand-writing practice worksheets prior to moving on to their first round of online content. If they were able to finish the first round before the time was up, they would start a second round again with another hand-writing practice worksheet prior to the online content. A post-test (post-test 1-a) on the assigned word set was provided immediately following the 13-minute first word set practice session. This process was then followed by a second practice session in which participants were given the alternate word set and condition. Another post-test (post-test1-b) on the alternative word set was provided immediately following the second practice session. A second post-test which included all the ten target words was given on Day Two of the experiment during the next course meeting. A third post-test which included all the ten target words was given during class one week later. The post-tests were self-paced, and all students were able to complete in ten minutes. The procedure of the experiment can be found in Appendix C figure a.

Training and delivery. The course instructor oversaw the procedure and monitored the whole experiment together with two teaching assistants. The two teaching assistants were trained before the experiment to familiarize themselves with the procedure and to distribute and collect worksheets during practices. A stopwatch was used to ensure that all participants started and ended concurrently.

Scoring. All online pretest and post-test scores were collected by the ASSISTments system. Those who used the correct Pinyin with wrong tones received partial credit scores (.5 out of 1) for the questions asking for the correct Pinyin with tones. All the hand-writing post-tests were anonymized and graded by one Chinese instructor, and scores were calculated as the rates of the correctly written characters.

Results

44 students completed the experiment. Most participants received 0 on the pretest of 20 questions and participants' average grades were 4%. Paired *t*-test did not find differences within participants on the pretest divided across conditions. No pretest was offered for hand-writing because it was expected that participants' scores on hand-writing would be very low. Table 3 presented the mean scores and standard deviations for all tests in both the NH condition and WH condition.

Table 3

Mean scores and standard deviations for all tests in NH and WH conditions in Experiment 1

	Online word recognition results			Hand-writing results		
Test time	NH condition mean (*SD*)	WH condition mean (*SD*)	Difference by Condition *p* value (*SE*)	NH condition mean (*SD*)	WH condition mean (*SD*)	Difference by Condition *p* value (*SE*)
Pretest	.04 (.07)	.05 (.08)	.36 (.01)			
Post-test 1	.77 (.22)	.67 (.27)	.003** (.03)	.11 (.11)	.14 (.14)	.09 (.02)
Post-test 2	.50 (.28)	.40 (.29)	.02* (.04)	.10 (.10)	.11 (.13)	.56 (.02)
Post-test 3	.39 (.27)	.32 (.24)	.04* (.04)	.12 (.10)	.12 (.12)	.94 (.02)

Notes. *. The mean difference is significant at the .05 level.

**. The mean difference is significant at the .01 level.

A two-way repeated measure ANOVA was conducted on the basis of two independent variables on the online word recognition post-test scores. Both main effects were statistically significant at the .005 significance level. The main effect for "condition" yielded an *F* ratio of *F* (1, 39) = 10.60, *p* = .002, partial eta squared = .21, observed power = .89, the effect size was large, indicating a significant difference between online-only practice (*M* = .55, *SD* = .03) and WH practice (*M* = .46, *SD* = .04). This main effect was further explored with post hoc paired *t* tests, and Figure 2a shows the online word recognition scores of each condition broken out by test intervals. It showed a clear downward trend that represented forgetting, and at the same time, showed reliable differences split out by conditions. There was a significant difference in the scores for post-test 1 NH condition (*M* = .77, *SD* = .22) and post-test 1 WH condition (*M* = .67, *SD* = .27) , *t* (43) = 3.1, *p* = .003, 95% CI [.04, .18], Cohen's *d* = .47; a significant difference in the scores for post-test 2 NH condition (*M* = .50, *SD* = .28) and post-test 2 WH condition (*M* = .40, *SD* = .29) , *t* (41) = 2.46, *p* = .02, 95% CI [.02, .16], Cohen's *d* = .38; and a significant difference in the scores for post-test 3 NH condition(*M* = .39, *SD* = .27) and post-test 3 WH condition (*M* = .32, *SD* = .24) , *t* (41) = 2.11, *p* = .04, 95% CI [.003, .15], Cohen's *d* = .33. These results suggested that the NH condition led to better post-test results in all online post-tests compared to the WH condition.

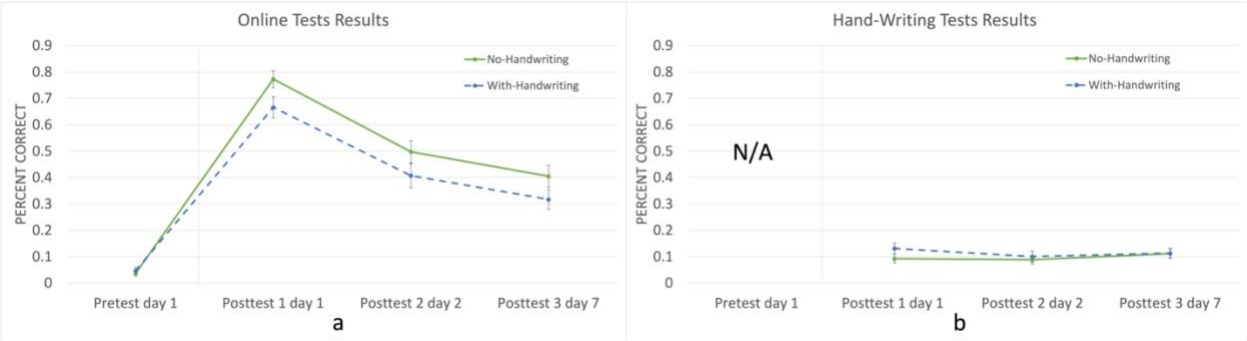
Another two-way repeated measure ANOVA was conducted to compare the main effects of conditions and test time on the *hand-writing* post-test scores in (Figure 2b). Both main effects were not statistically different, *p* >.05. The post hoc paired *t* tests results did not find difference between conditions (Table 3). These results suggested that the hand-writing post-test results did

not reveal significant hand-writing improvement, the hand-writing practice time is inadequate for hand-writing output.

All 45 participants who attended class on the first day engaged in both online and hand-writing practices. The word numbers seen, attempts tried, hint count, time spent per word, as well as time spent on hand-writing practice can be found in Table 5. All participants completed the first round of hand-writing practice in the WH condition except for one. Twenty participants started the second-round hand-writing practice and 11 of them completed it as well by the end of the practice session. Three participants started the third round and two of them completed it, and one student also completed the fourth and fifth rounds. When students practiced hand-writing in the WH condition, the time spent on hand-writing was recorded by the platform. To account for personal error, actions recorded less than 20 seconds for the hand-writing practice were removed. Each student spent an average of 3.95 minutes (236.98 seconds, *SD* = 95.82) on hand-writing practice in WH condition, which took 30.38% of their overall practice time during the learning phase of the experiment.

Figure 2

Scores for each condition in each post-test in Experiment 1



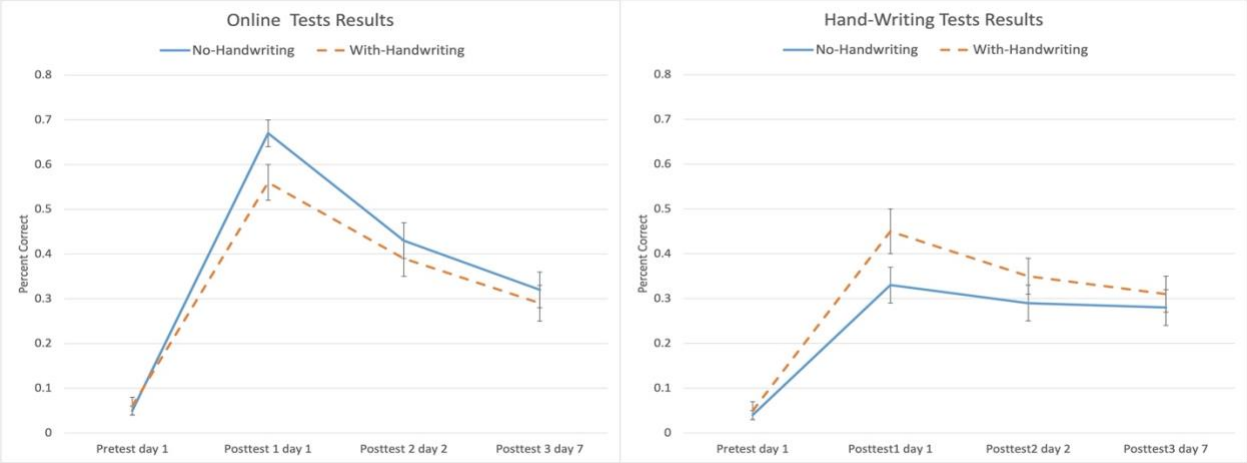
Experiment 2: L Replication of Experiment 1

Experiment 2 (Lu et al., 2019) served as a replication of Experiment 1, and the results and data have been published. Experiment 2 was run on the same campus with students at the same language level using the same material. The same experimental procedure was repeated. The only change was including a hand-writing pretest to enhance the validity of the experiment. The purpose of this replication was to add information about the reliability of the conclusions, and we hypothesized that, as observed in Experiment 1, students would score significantly higher on word recognition post-tests when they were not subjected to hand-writing exercises.

This experiment observed a significant difference between The NH condition and the WH condition on the immediate post-test (post-test 1) on online word recognition, with students in the NH condition (M = .67, SD = .25) outperforming those in the WH condition (M = .56, SD = 028), with a t-value of 3.05, p = .004, 95% CI [.04, .18], and Cohen's d of .41. On Post-test 2, there was a marginally significant difference between the NH condition (M = .43, SD = .29) and the WH condition (M = .39, SD = .29) with t-value of 1.95, p = .058, 95% CI [−.001, .09], and Cohen's d of .14. However, no significant differences were found between the conditions on Post-test 3, with similar scores for students in both the NH condition (M = .32, SD = .27) and the WH condition (M = .29, SD = .26), t (48) = .87, p = .391. Although the results were not as significant as in Experiment 1, Experiment 2 trended in the same direction as Experiment 1 and reaffirmed that practice without hand-writing in the same amount of time led to better word recognition in the short term.

Figure 3

Scores for each condition in each post-test in Experiment 2



Note. This graph demonstrated the online word recognition and hand-writing test results of Experiment 2. Adapted from "Save Your Strokes: Chinese Handwriting Practice Makes for Ineffective Use of Instructional Time in Second Language Classrooms," by X., Lu, et al., 2019, *AERA Open*, *5*(4). https://doi.org/10.1177/2332858419890326. CC BY-NC 4.0.

Experiment 3: Novice Learners

Experiment 3 assessed the external validity of the study by replicating Experiment 1 at a different university, at a different language level, and with similar material. It aimed to confirm the results of Experiments 1 and 2.

Method

Participants

Participants were 34 undergraduate students enrolled in a Novice-high Level Chinese class from a public university in the U.S. They were non-heritage learners on their second semester of CFL learning, had prior knowledge of Pinyin and were knowledgeable about typing Chinese on the computer.

Materials

Selection standards of the ten target words were identical to Experiment 1, however, there was word variation to ensure that the target words were not listed in their preliminary course textbooks (Appendix A). The introductory videos, seven question types of practice, hand-writing practice sheets, and assessments were identical in design to Experiment 1.

Design & Procedure

Experiment 3 shared the same design and procedure with Experiment 1 and 2. Due to various class schedules of the two Chinese programs, the second post-test was given on Day Three of the experiment during course meetings, rather than on the second day. The procedure chart was shown in Appendix C figure b.

Results

The experiment included 28 participants. The online pretest average grade was .02 (*SE* = .01). The average grades of each post-test were .39 (*SE* = .05), .10 (*SE* = .03), and .09 (*SE* = .03), respectively. A two-way repeated measure ANOVA was conducted based upon two independent variables on the online word recognition scores. Both main effects (online test time and conditions) were statistically significant, and there was a substantial difference in the interaction. The main effect for "conditions" yielded an *F* ratio of $F(1, 22) = 7.04$, $p = .01$, partial eta squared = .24, observed power = .72, the effect size was large, indicating a significant variance between online-only practice ($M = .19$, $SE = .03$) and WH practice ($M = .11$, $SE = .02$). This main effect was further explored with post hoc paired *t* tests, and Figure 4a shows the online word recognition scores of each condition broken down by test intervals. It showed a clear downward trend that represented forgetting, and at the same time showing reliable differences split out by conditions. No difference was found in the pretest for NH condition ($M = .04$, *SD*

= .06) and WH condition (*M* = .02, *SD* = .06), *t* (26) = .85, *p* = .40, 95% CI [-.02, .04], Cohen's *d* = .16. There was a significant difference in the scores for post-test 1 NH condition (*M* = .44, *SD* = .32) and WH condition (*M* = .30, *SD* = .27) , *t* (28) = 2.46, *p* = .02, 95% CI [.02, .25], Cohen's *d* = .46; a significant difference in the scores for post-test 2 NH condition (*M* = .14, *SD* = .16) and WH condition (*M* = .07, *SD* = .12) , *t* (27) = 2.64, *p* = .01, 95% CI [.02, .13], Cohen's *d* = .50. No significant difference was found in the scores for post-test 3 NH condition (*M* = .11, *SD* = .17) and WH condition (*M* = .07, *SD* = .10), *t* (25) = 1.82, *p* = .08, 95% CI [-.006, .10], Cohen's *d* = .36. These results were consistent with Experiment 1 and 2, and suggested that NH condition led to better post-test results in short-term online word recognition post-tests compared to WH condition.

The average grade of the hand-writing pretest was .02 (*SE* = .01). The average grades of the three hand-writing post-tests were .06 (*SE* = .02), .02 (*SE* = .01), and .02 (*SE* = .01), respectively. A two-way repeated measure ANOVA was conducted to compare the main effects of conditions and test time on the hand-writing scores. The main effect "conditions" between NH practice (*M* = .02, *SE* = .01) and WH practice (*M* = .04, *SE* = .02) was not significantly different, *F* (1, 18) = .63, *p* = .44, partial eta squared = .03, observed power = .12. These results suggested that WH condition did not lead to better post-test results compared to the NH condition (Figure 4b). Table 4 presented the mean scores and standard deviations for all tests in both the NH and WH conditions.

Table 4

Mean scores and standard deviations for all tests in NH and WH conditions in Experiment 3

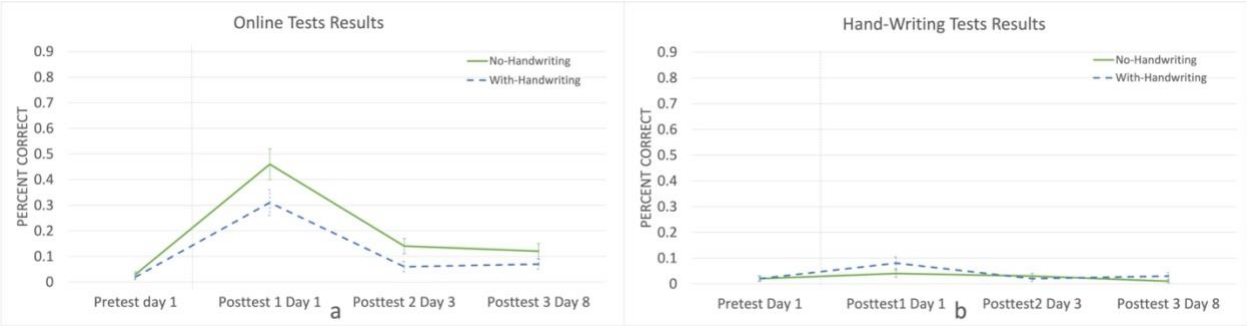
	Online word recognition results			Hand-writing results		
Test time	NH condition mean (*SD*)	WH condition mean (*SD*)	Difference by Condition *p* value (*SE*)	NH condition mean (*SD*)	WH condition mean (*SD*)	Difference by Condition *p* value (*SE*)
Pretest	.04 (.06)	.02 (.06)	.40 (.02)	.02 (.05)	.02 (.05)	.60 (.01)
Post-test 1	.44 (.32)	.30 (.27)	.02* (.06)	.04 (.06)	.08 (.14)	.57 (.03)
Post-test 2	.14 (.16)	.07 (.12)	.01** (.03)	.03 (.07)	.02 (.05)	.77 (.01)
Post-test 3	.11 (.17)	.07 (.10)	.08 (.03)	.01 (.03)	.03 (.07)	.38 (.02)

Notes. *. The mean difference is significant at the .05 level.

**. The mean difference is significant at the .01 level.

Figure 4

Scores for each condition in each post-test in Experiment 3



The word (item) numbers seen, attempts tried, hint count, time spent per word, and time spent on hand-writing practice can be found in Table 5. All participants completed the first round of hand-writing practice during the WH condition. Eight participants were able to start the second round of hand-writing practice, however, only two of them could complete the second round by the end of the practice session. When considering all the completed hand-writing problems, each student spent an average of 6.87 minutes (412.19 seconds, *SD* = 140.01) on hand-writing practice, which took 52.84% of total practice time. The results indicated that these participants spent more time on hand-writing compared to the first two experiments, but scored lower in the hand-writing post-tests. Since the Experiment 3 class were in the second semester while the other two experiment classes were in the third semester, the difference might account for the difference in learning experiences and prior knowledge. Another explanation could be that spending more time on hand-writing resulted in seeing fewer problems in online practices, and the lower repetition frequency hindered learning. More discussion can be found in the discussion section.

Table 5

In-Experiment Practice Phase Data Comparison of Experiments 1 and 3

	Experiment 1 (Intermediate Low-Mid learners) mean (*SD*)	Experiment 3 (Novice High learners) mean (*SD*)
Items (words) seen per student	37.46 (12.12)	33.76 (11.94)
Attempts made per student	1.27 (.22)	1.37 (.25)
Hint count per student	.15 (.11)	.19 (.06)
Time spent per item (seconds)	16.92 (62.21)	17.07 (52.25)
	Median = 7.11	*Median* = 6.71
Time on hand-writing practice (seconds)	236.98 (95.82)	412.19 (140.01)
	30.38% of overall practice time	52.84% of overall practice time

Experiment 4: Accounting for Time Spent Hand-Writing

The first three experiments revealed a significant difference between the NH and WH conditions in short-term results. These results strongly indicated that lower-level learners using online-only questions to practice new words recognized words more effectively than students who practiced hand-writing for the same time. In the previous experiments, participants in the WH condition spent at least 30% of their time on hand-writing practice, and this amount of time yielded unsuccessful results. In the current experiment, the results were further explored by adding one more condition to the comparison: 70% NH practice without the remaining 30% WH.

The goal of this experiment was to investigate whether exempting students from the 30% of hand-writing practice time during word learning would result in similar learning outcomes on the post-tests. Specifically, the experiment aimed to determine whether the 30% amount of hand-writing practice has an effect on the efficiency of word recognition among lower-level learners. Would a difference exist with or without the 30% hand-writing time? If the 30% hand-writing time does not have an effect on the learning outcomes, the experiment would further confirm that students can save time and effort by focusing solely on online practice, rather than dividing their time between online practice and hand-writing practice. The experimental procedure is shown in Appendix C figure c.

Method

Participants

Participants included 60 students enrolled in three sections of an Intermediate-level Chinese class from the same university as Experiment 1. They were non-heritage learners on their third semester of CFL learning.

Materials

Eight, two-character target words were selected following the same standards as Experiment 1, 2, and 3, which can be seen in Appendix A. The introductory videos, seven question practice types while learning the words, hand-writing practice sheets, and the pretest and post-tests were identical in design to Experiment 1, 2, and 3.

Different to the first three experiments, one more type of practice during the practice section was added besides the online receptive practice and hand-writing practice: non-related online practice. Non-related online practice was designed similarly to the online practice. The online practice, however, focused on the target words while the non-related online practice

emphasized non-target words. Including the non-related online practice helped to ensure that the post-test timing was the same across all conditions.

Design

This experiment was a between-subject design and had two independent variables: the test point and condition. The test point referred to a nominal variable indicating post-test 1, 2, and 3, and the between-subject variable condition included three levels: A: No-Hand-Writing with 70% of practicing time condition (NH70%), B: No-Hand-Writing condition (NH), and C: With-hand-writing condition (WH) (Appendix C figure c). All the three conditions started with 16-minute NH online practice focused on eight target words. After the 16-minute online practice, participants in condition A spent another 8 minutes taking non-related online practice, participants in condition B took more online practice focused on the target words, while participants in condition C spent time completing hand-writing worksheets to practice.

Procedures

This experiment was conducted over the course of seven days. Participants randomization, scoring protocol, training provided and delivery of treatment were identical to Experiment 1. The procedure of the experiment was also similar to Experiment 1 except for the practice session on the first day (Appendix C figure c). Right after watching the video on the first day, all participants had 16 minutes to do an online NH practice at their own pace. When 16 minutes approached, all participants were instructed to immediately quit the first practice session and move on to their second practice session with a duration of 8 minutes. There were three conditions and each received different content in the second practice session.

Results

The experiment included 41 participants. There were four outliers whose online pretest accuracy rates were higher than 20%. These were also excluded. The mean scores and standard deviations for the three conditions in different test points in the various test mediums are shown in Table 6. The three conditions were essentially the same in the pretest.

The online word recognition and hand-writing results were processed separately. Figure 5 shows both the online word recognition and the hand-writing test results. Factorial mixed ANCOVAs were conducted for both online and hand-writing data to examine the main effects of the three levels of each condition (NH70%, NH, WH), three levels of post-test administration (1, 2, 3), and the interaction effects of the independent variables, using pretest results as covariates. ANCOVAs were administered for each post-test to ensure accuracy, again using pretest accuracy as covariates (Dimitrov & Rumrill, 2003) to obtain a more profound understanding of the variations between conditions in each post-test admission for each portion.

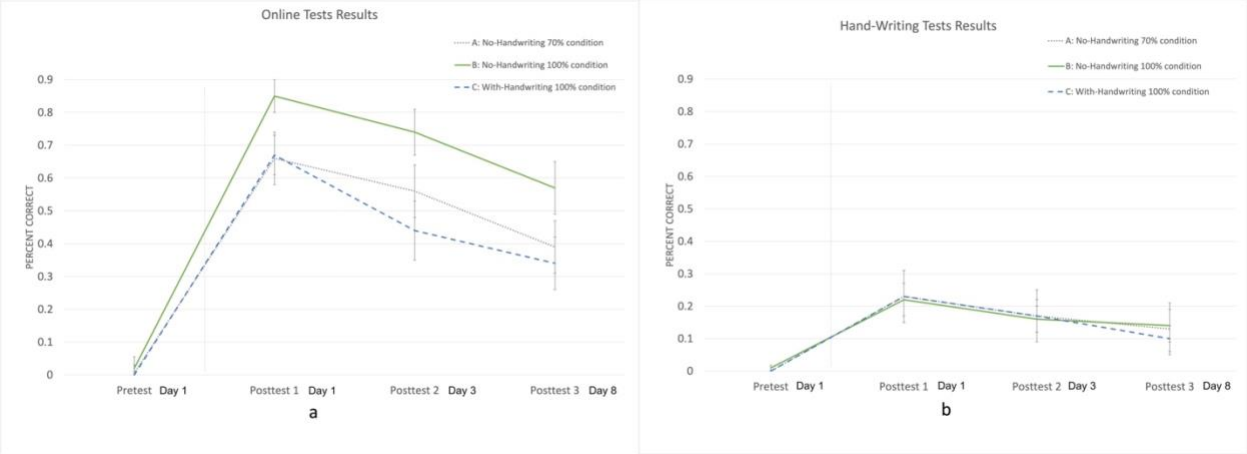
Table 6

Mean scores and standard deviations for all tests in the three conditions in Experiment 4

	Online word recognition results			Hand-writing results		
Test time	A: No-Hand-Writing 70% condition mean (*SD*)	B: No-Hand-Writing 100% condition mean (*SD*)	C: With-hand-Writing 100% condition mean (*SD*)	A: No-Hand-Writing 70% condition mean (*SD*)	B: No-Hand-Writing 100% condition mean (*SD*)	C: With-hand-Writing 100% condition mean (*SD*)
Pretest	.005 (.02)	.02 (.03)	.00 (.00)	.00 (.00)	.01 (.03)	.00 (.00)
Post-test 1	.64 (.26)	.85 (.15)	.65 (.24)	.23 (.28)	.22 (.16)	.23 (.25)
Post-test 2	.53 (.28)	.74 (.20)	.47 (.25)	.17 (.29)	.16 (.14)	.17 (.16)
Post-test 3	.39 (.30)	.57 (.26)	.34 (.25)	.13 (.27)	.14 (.16)	.10 (.12)

Figure 5

Scores for each condition in each post-test in Experiment 4



Word Recognition Results

Factorial mixed ANCOVA did not find differences between conditions in the *online word recognition*, F (2, 31) = 2.10, p = .14, partial η2 = .12, observed power = .40. To observe the differences between conditions in each post-test, three ANCOVAs were implemented for the post-test 1, 2, and 3 results, each using pretest accuracy as a covariate. The covariate was independent in the three post-tests. A significant difference between conditions was found in the post-test 1, F (2, 44) = 4.04, p = .03, partial η2 = .16, ω^2 = -.11. The contrast results (*K* Matrix) indicated that after controlling for the pretest accuracy, NH condition (M = .85, SD = .15) received significant higher scores than WH condition (M = .65, SD = .24), p = .02. No difference was found between the WH condition and the NH70% condition (M = .64, SD = .26), p > .05. A significant difference between conditions was found in the post-test 2, F (2, 42) = 4.64, p = .02, partial η2 = .18, ω^2 = -.14. The contrast results (*K* Matrix) indicated that after controlling for the pretest accuracy, NH condition (M = .74, SD = .20) reached a significantly higher score than WH condition (M = .47, SD = .25), p = .006. No difference was found between the WH condition and

the NH70% condition (M = .53, SD = .28), p > .05. No significant difference between the three conditions was found in the post-test 3, F (2, 33) = 1.80, p = .18, partial $\eta 2$ = .10, ω^2 = -.04. Covariance of online post-tests accuracy with online pretest accuracy as covariates can be found in Table 7.

Table 7

Covariance of online post-test accuracy as a function of conditions with online pretest accuracy as covariates in Experiment 4

Source	*df*	*SS*	*MS*	*F*	*p*	η 2
Post-test 1						
Pretest (Covariate)	1	.15	.15	3.09	.09	.07
Condition	2	.39	.20	4.04	.03*	.16
Error	44	2.13	.05			
Corrected Total	47	2.71				
Post-test 2						
Pretest (Covariate)	1	.15	.15	2.52	.12	.06
Condition	2	.55	.28	4.64	.02*	.18
Error	42	2.51	.06			
Corrected Total	45	3.24				
Post-test 3						
Pretest (Covariate)	1	.06	.06	.92	.34	.03
Condition	2	.25	.12	1.80	.18	.10
Error	33	2.27	.07			
Corrected Total	36	2.72				

Notes. *. The mean difference is significant at the .05 level.

**. The mean difference is significant at the .01 level.

Hand-Writing Results

Factorial mixed ANCOVA did not find differences between conditions in the hand-writing. An ANOVA with the pretest accuracy as the dependent variable, and with condition as the independent variable, was run to check the independence of the pretest accuracy. The main effect was significant, F (2, 44) = .6.03, p = .005 and demonstrated that pretest accuracy was not independent of the conditions, and that it could not be used as a covariate when comparing the results of the three conditions in each post-test. When analyzing the data, the NH condition's pretest accuracy was higher than the other two conditions because participants in that condition already knew several of the words. Learning gains were used for analysis that further explored the effects of different conditions, defined by post-test accuracy and reduced by pretest accuracy. ANOVAs were conducted for the learning gains of the three post-tests. The main condition effect (p > .05) was insignificant in terms of any post-test learning gain. Pairwise comparisons showed no variation between the three conditions in the post-tests.

General Discussion

Summary of Main Findings

The studies reported in this chapter aimed to test the learning efficiency of CFL word recognition among lower-level learners with or without hand-writing, replicating and extending previous research to further our understanding of the relative roles of receptive learning and hand-writing practice in CFL word learning at multiple proficiency levels.

The online word recognition results of all four experiments strongly indicated that the NH condition led to better results in the online word recognition post-tests than the WH condition. Experiments 1, 2, and 3 all showed significant differences between the two practice

conditions on the online word recognition post-test results, supporting that hand-writing practice is an ineffective use of practice time. Participants scored considerably lower on word recognition post-tests after spending more than 30% of their practice time on hand-writing exercises. These significant differences appeared immediately following practice and persisted with repeated testing for at least one day after the test. Additionally, experiments 1, 3, and 4 found significant differences in their second posttests, and in experiment 1, the significant difference lasted up to the seventh day. One additional condition was considered in Experiment 4 based on the prior experiments' results. Experiment 4 also revealed that the NH condition received significantly higher scores than the WH condition on the online word recognition post-tests 1 and 2. No difference in any online word recognition post-test level was found between the WH and 70% NH conditions. Experiment 4 reaffirmed that participants who spent their entire time practicing online performed better on word recognition tasks than those spending partial time practicing hand-writing on paper. Furthermore, 30% of hand-writing time resulted in inefficient word recognition. The student performance of those who spend 70% of their time practicing online and 30% of their time practicing hand-writing did not show improvement over that of students who spent only 70% of their time practicing online. These findings reaffirmed Lu et al.'s results (2019). They also aligned with previous classroom-based studies that have found that e-writing was more efficient for word recognition and literacy development among CFL beginners (Jiang 2007; Zhang, 2021). In conclusion, online-only practice is a more efficient way for lower-level CFL learners to recognize new words than hand-writing practice. Teachers should consider encouraging students to focus their practice time on online activities to improve word recognition efficiency.

Hand-Writing Practice Is Inefficient for Word Recognition

The online word recognition results demonstrated that NH practice without hand-writing is more efficient than WH, while the hand-writing post-test results specified that 30–50% of hand-writing practice time is insufficient for word recognition and inadequate for hand-writing output. The hand-writing post-test results in each experiment did not reveal significant hand-writing improvement, even for those who spent time hand-writing words. This finding aligns with Jiang's (2007, 2017) study, which showed that reduced hand-writing practice led to better performance on character hand-writing tests. It also further supported Zhang's (2021) finding that practice with NH produced the same level of character accuracy as practice WH.

Although correlated, word recognition and handwritten word production are two distinct psychological processes. Word recognition involves searching for and matching pronunciation and meaning to a target written form. In contrast, word production by hand involves recalling and retrieving the written form of a target word and then writing it down by hand, without necessarily activating the same (amount of) meaning and/or sound information. Interestingly, both historically and contemporarily, many CFL teachers insist that students learn to handwrite while learning reading, listening, and speaking, purportedly because this aids in literacy development and word retention. This practice can be attributed to the mistaken belief that hand-writing is a good use of time for increasing lower-level learners' Chinese language knowledge. This study demonstrated that hand-writing exercises negatively impact word recognition when they are part of the limited amount of learning time allocated to language practice, adding empirical evidence to Allen's claim that 'learning to write Chinese is a waste of time' (2008).

Efficiency Leads to Success

Pedagogical efficiency, namely, allocation of limited resources (e.g., time) to obtain maximal results (i.e., learning; see Chu, this volume) was the main goal of the studies reported in

this chapter. When considering efficiency, our results have shown that WH is not as efficient as NH when preparing for word recognition tasks.

Among the few studies that have controlled learning time, Guan et al. (2011) reported two studies comparing the effects of hand-writing, reading, and typing with Pinyin on CFL learning and demonstrating that hand-writing conditions outperformed reading conditions on word recognition. However, the learning time for each condition in their study was determined by the time students required during hand-writing conditions. In the NH condition, students could not do anything except passively look at words and characters. Other studies comparing the effects of WH and NH have adopted a similar methodology (e.g., Xu et al., 2013). Staring at a word without practice is an inefficient learning method. Meaningful practices related to the word would impact efficiency and result in a significant learning improvement.

The current study's objective is not to prove that hand-writing is completely useless or that it should be completely abandoned. When considering efficiency, though, current experiments have demonstrated that hand-writing practice is not as efficient as engaged reading and e-writing practice when preparing for word recognition tasks. Therefore, the role of hand-writing (in terms of amount, type, and timing—what characters are hand-written, when, for how long, and under what conditions, e.g., recalled from memory or traced or copied) is a topic that merits further exploration and scholarly conversation (see also, Coss, this volume).

High Repetition Frequency Improves Lower-Level Learners' Word Recognition

Many studies have shown that word repetition and the frequency of its occurrence play an essential role in second language word learning (Saragi et al., 1978; Webb, 2007; Uchihara et al., 2019). Lu et al. (2019) observed that in the Experiment 2 condition, the difference between NH and WH conditions did not last as long as in Experiment 1, possibly because students in

Experiment 2 experienced fewer repetitions of target words during practice time. Experiment 3 further supported this conjecture. When comparing Experiments 1 and 3, it was found that experimental data from the first-day practice exercises demonstrated that participants' online use of the interventions in the two experiments was balanced. This data included metrics documenting the average attempts tried, the hint count, and the time spent on each problem, as shown in Table 5. The two participant groups exhibited different learning habits, as evidenced by the participants in Experiment 3 spending more time on hand-writing than those in Experiment 1. Participants in Experiment 1 spent less time hand-writing, allowing them to view more problems than those in Experiment 3. A higher repetition frequency of the target words could have been responsible for higher scores.

Conclusions and Pedagogical Implications

This study tested the impact of WH and NH on word recognition for novice and intermediate CFL learners. Experiment 1 was replicated twice in Experiments 2 and 3, confirming that NH practice was a more efficient way for lower-level learners to learn Chinese word recognition. Second, Experiment 4 results indicated that 30% of hand-writing practice time was inefficient for word recognition. Students could save 30% of that time by not practicing and achieving the same results in word recognition as those who practiced WH. It is worth mentioning that 30% is only the minimum time a learner would spend on hand-writing without strict hand-writing requirements. Lu et al. (2019) reported that instructors at the college level with strict hand-writing requirements expected students to devote an average of 44% of their practice time on hand-writing, and this number can reach as high as 70%. Thus, our study lends further credibility to the proposition that word recognition with NH is a more efficient learning method than a learning routine which includes both receptive learning and hand-writing.

There has been considerable debate in the literature about the best teaching practices for CFL with regard to Chinese characters. It is commonly accepted that Chinese takes considerably longer to learn than other languages; consequently, the reading and hand-writing of Chinese words should be taught starting on the first day of class. However, with the development of a deeper understanding of CFL learning and the increasing amount of evidence supporting e-writing (not to mention the ecological validity of e-writing as a communication method), more CFL teachers have considered decreasing hand-writing requirements in favor of word recognition practice when attempting to increase learning efficiency. Cui (1999) advocated that students should first learn listening and speaking. Zhang (2021) found that compared with hand-writing-primary method, the typing-based approach is a more efficient way for lower-level learners to learn and maintain Hanzi, thus facilitating the development of their reading and e-writing skills. Our study fills the gap in the literature by considering efficiency and further supporting that lower-level CFL learners can learn word recognition more efficiently with NH.

The findings of this study should assist CFL teachers in making informed decisions about best practices for teaching lower-level CFL learners to recognize Hanzi. Suggested pedagogical implications include that, rather than spending time on hand-writing exercises, CFL teachers should be spending valuable time focusing on listening, speaking, reading and e-writing skills if the teaching objective is to make Chinese learning a more productive experience for early-stage learners. Hand-writing may be inefficient and have limited value in terms of character and word learning. Within the same amount of learning time, lower-level (novice and intermediate) CFL teachers are encouraged to offer different kinds of reading and e-writing word recognition tasks and demonstrate words in various contexts to enhance word recognition without requiring learners to write or memorize characters or words via hand-writing.

Limitations and Future Work

Despite the success of this study, there were certain limitations. Firstly, Experiment 1 did not assess participants' handwriting ability prior to the study, as it was deemed illogical to require participants to write down characters of unknown words. However, the lack of testing in this aspect might impact the validity of the experiment. Experiment 2 and Experiment 3 made improvements upon Experiment 1 by addressing this limitation. Secondly, Participants in the NH conditions of the experiments received handwriting practice prior to online practice. However, all post-tests included a handwriting task after the online recognition task. Although the online tests and the handwriting tests assessed different skills and their results were analyzed separately, aligning the test order to the practice order would have improved the study's design. In addition, the drop-out rate of participants in Experiment 3 affected the results due to relatively small sample sizes.

Future research could examine the long-term effects of online word recognition learning, investigate the most effective question types for word recognition, and explore the interaction between online learning effects and learning platform functions such as retry and immediate feedback. Additionally, more studies comparing the differences between character recognition and word recognition and how these two learning processes interact with each other would be of interest.

Acknowledgements

We acknowledge funding from NSF (2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, R305A120125 & R305R220012), GAANN (P200A180088 & P200A150306), EIR (U411B190024

S411B210024, & S411B220024), ONR (N00014-18-1-2768), NHI (via a SBIR R44GM146483), Schmidt Futures, BMGF, CZI, Arnold, Hewlett and a $180,000 anonymous donation. None of the opinions expressed here are those of the funders.

References

Allen, J. R. (2008). Why learning to write Chinese is a waste of time: a modest proposal. *Foreign Language Annals*, *41*(2), 237–251. https://doi.org/10.1111/j.1944-9720.2008.tb03291.x

Berninger, V. W., Abbott, R. D., Abbott, S. P., Graham, S., & Richards, T. (2002). Writing and reading: connections between language by hand and language by eye. *Journal of Learning Disabilities*, *35*(1), 39–56. https://doi.org/10.1177/002221940203500104

Bi, Y., Han, Z., & Zhang, Y. (2009). Reading does not depend on writing, even in Chinese. *Neuropsychologia*, *47*(4), 1193–1199. https://doi.org/10.1016/j.neuropsychologia.2008.11.006

Brown, J. (2020). Perseverance. In E. M. Sanchez (Ed.), *Merriam-Webster*. Merriam-Webster. https://www.merriam-webster.com/dictionary/word

Cui, Y. (1999). About the reforms in patterns of basic Chinese teaching. *Chinese Teaching in the World. 1999*(1), 3–8.

Dimitrov D. M., & Rumrill P. D. (2003). Pretest-posttest designs and measurement of change. *Work*, *20*, 159–65. Retrieved January 15, 2020, from https://content.iospress.com/articles/work/wor00285

Everson, M. E. (2009). Literacy development in Chinese as a foreign language. In M. E. Everson & Y. Xiao (Eds.), *Teaching Chinese as foreign language: Theories and applications* (pp. 97–113). Boston: Cheng & Tsui.

Garnett, K. (2011). Fluency in learning to read: conceptions, misconceptions, learning disabilities, and instructional moves. In J. R. Birsh (Ed.), *Multisensory teaching of basic language skills* (p. 293-320). Baltimore, MD: Brookes Publishing.

Grabe, W. (1991). Current developments in second language reading research. *TESOL quarterly*, *25*(3), 375-406. https://doi.org/10.2307/3586977

Guan, C. Q., Liu, Y., Chan, D. H. L., Ye, F., & Perfetti, C. A. (2011). Writing strengthens orthography and alphabetic-coding strengthens phonology in learning to read Chinese. *Journal of Educational Psychology,103*, 509–52. https://doi.org/10.1037/a0023730

Hanban. (2010). 汉语国际教育用音节汉字词汇等级划分 [The graded Chinese syllables, characters and words for the application of teaching Chinese to the speakers of other languages]. Beijing: Beijing Language and Culture University Press.

Hayes, L., & Flanigan, K. (2014). *Developing word recognition.* The Guilford Press.

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497. https://doi.org/10.1007/s40593-014-0024-x

Hsiung, H., Chang, Y., Chen, H., & Sung, Y. (2017). Effect of stroke-order learning and handwriting exercises on recognizing and writing Chinese characters by Chinese as a foreign language learners. *Computers in Human Behavior*, *74*, 303–310. https://doi.org/10.1016/j.chb.2017.04.022

HSK Department. (1992). 汉语水平（词汇）等级大纲 [The handbook of outline of the graded vocabulary for HSK]. Beijing, China: Beijing Language Institute Press.

Jiang, X. (2007). “认写分流、多认少写”汉字教学方法的实验研究 [An experimental study on the effect of the method of 'Teaching the learner to recognize characters more than writing]. 世界汉语教学 [Chinese Teaching in the World], 2, 91–97.

Jiang, X. (2017). 键盘时代的汉字教学 [Teaching Chinese in the keyboard age]. 国际中文教学期刊 [Journal of International Chinese Teaching], 2, 4–10.

Ke, C. (1998). Effects of strategies on the learning of Chinese characters among foreign language students. *Journal of the Chinese Language Teachers Association*, *33*(2), 93–112. https://doi.org/10.1111/j.1944-9720.1998.tb01335.x

Kilpatrick, D. A. (2020). How the phonology of speech is foundational for instant word recognition. *Perspectives on Language and Literacy*, *46*(3), 11-15. Retrieved from: https://www.literacyhow.org/wp-content/uploads/2020/09/The-Phonology-of-Speech-in-WR-Kilpatrick.pdf

Koda, K. (1996). L2 Word recognition research: a critical review. *The Modern Language Journal*, *80*(4), 450-460. https://www.jstor.org/stable/329725

Li, X., Bicknell, K., Liu, P., Wei, W., & Rayner, K. (2014). Reading is fundamentally similar across disparate writing systems: a systematic characterization of how words and characters influence eye movements in Chinese reading. *Journal of Experimental Psychology: General*, *143*(2), 895–913. https://doi.org/10.1037/a0033580

Li, T., & McBride-Chang, C. (2014). How character reading can be different from word reading in Chinese and why it matters for Chinese reading development. In: Chen, X., Wang, Q., Luo, Y. (eds) *Reading development and difficulties in monolingual and bilingual Chinese Children. Literacy Studies*, 8. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-7380-6_3

Lu, X., Ostrow, K. S., & Heffernan, N. T. (2019). Save your strokes: Chinese handwriting practice makes for ineffective use of instructional time in second language classrooms. *AERA Open*. https://doi.org/10.1177/2332858419890326

Lyu, B., Lai, C., Lin. C-H, & Gong, Y. (2021). Comparison studies of typing and handwriting in Chinese language learning: a synthetic review. *International Journal of Educational Research, 106*. https://www.sciencedirect.com/science/article/pii/S0883035521000100

Osborne, C., Zhang, Q., & Zhang, G. X. (2020). Which is more effective in introducing Chinese characters? An investigative study of four methods used to teach CFL beginners. *The Language Learning Journal, 48*(4), 385-401. https://doi.org/10.1080/09571736.2017.1393838

Packard, J. L., Chen, X., Li, W., Wu, X., Gaffney, J. S., Li, H., & Anderson, R. C. (2006). Explicit instruction in orthographic structure and word morphology helps Chinese children learn to write characters. *Reading and Writing, 19*(5), 457–487. https://doi.org/10.1007/s11145-006-9003-4

Pan, D. J., Yang, X., Lui, K. F. H., Lo, J. C. M., McBride, C., & Ho, C. S. H. (2021). Character and word reading in Chinese: why and how they should be considered uniquely vis-a-vis literacy development. *Contemporary Educational Psychology, 65*, 101961. https://doi.org/10.1016/j.cedpsych.2021.101961

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System,6*, 72–78. https://doi.org/10.1016/0346-251X(78)90027-1

Shen, H. H. (2005). An investigation of Chinese-character learning strategies among non-native speakers of Chinese. *System, 33*(1): 49–68. https://doi.org/10.1016/j.system.2004.11.001

Su, X. (2001). 关于《现代汉语词典》 词汇计量研究的思考 [Quantitative studies of the dictionary of modern Chinese]. 世界汉语教学 [Chinese Teaching in the World]. 58(4): 39–47.

Tan L. H., & Perfetti, C. A. (1999). Phonological activation in visual identification of Chinese two–character words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 382–393. https://doi.org/10.1037/0278-7393.25.2.382

Tan, L. H., Spinks, J. A., Eden, G. F., Perfetti, C. A., & Siok, W. T. (2005). Reading depends on writing, in Chinese. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(24), 8781–8785. http://doi.org/10.1073/pnas.0503523102

Tan L. H., Xu, M., Chang, C. Q., & Siok, W. T. (2013). China's language input system in the digital age affects children's reading development. *Proceedings of the National Academy of Sciences of the United States of America*, *110* (3), 1119–1123. https://doi.org/10.1073/pnas.1213586110

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: a meta-analysis of correlational studies. *Language Learning*, *69*, 559–599. https://doi.org/10.1111/lang.12343

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*, 46–65. https://doi.org/10.1093/applin/aml048

Xu, Y., Chang, L., Zhang, J., & Perfetti, C. (2013). Reading, writing, and animation in character learning in Chinese as a foreign language. *Foreign Language Annals*, *46*(3), 423–444. https://doi.org/10.1111/flan.12040

Yang, X., Pan, D.J., Lo, C.M. et al. (2022). Same or different: Chinese character reading and word reading of young readers with development. *Reading and Writing*. https://doi.org/10.1007/s11145-022-10255-9

Yue, Y. (2017). Teaching Chinese in K–12 schools in the United States: what are the challenges? *Foreign Language Annals*, *50*(3), 601–620. https://doi.org/10.1111/flan.12277

Zhang, P. N. (2021). Typing to replace handwriting: effectiveness of the typing-primary approach for L2 Chinese beginners. *Journal of Technology and Chinese Language Teaching, 12*(2), 1-28. Retrieved from: http://www.tclt.us/journal/2021v12n2/zhangn.pdf

Zhao, R., Fan, R., Liu, M., Wang, X., & Yang, J. (2017). Rethinking the function of brain regions for reading Chinese characters in a meta-analysis of fMRI studies. *Journal of Neurolinguistics, 44*. 120–133. https://doi.org/10.1016/j.jneuroling.2017.04.001

Zhu, Y., Shum, S. M., Tse, S. B., & Liu, J. J. (2016). Word-processor or pencil-and-paper? A comparison of students' writing in Chinese as a foreign language. *Computer Assisted Language Learning. 29*(3), 596–617. https://doi.org/10.1080/09588221.2014.1000932

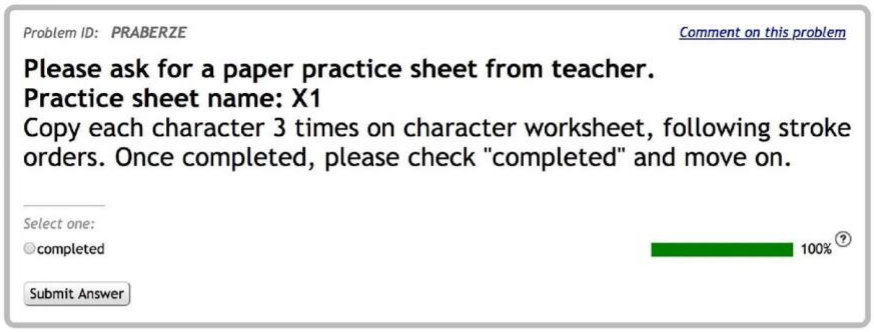
Appendix A

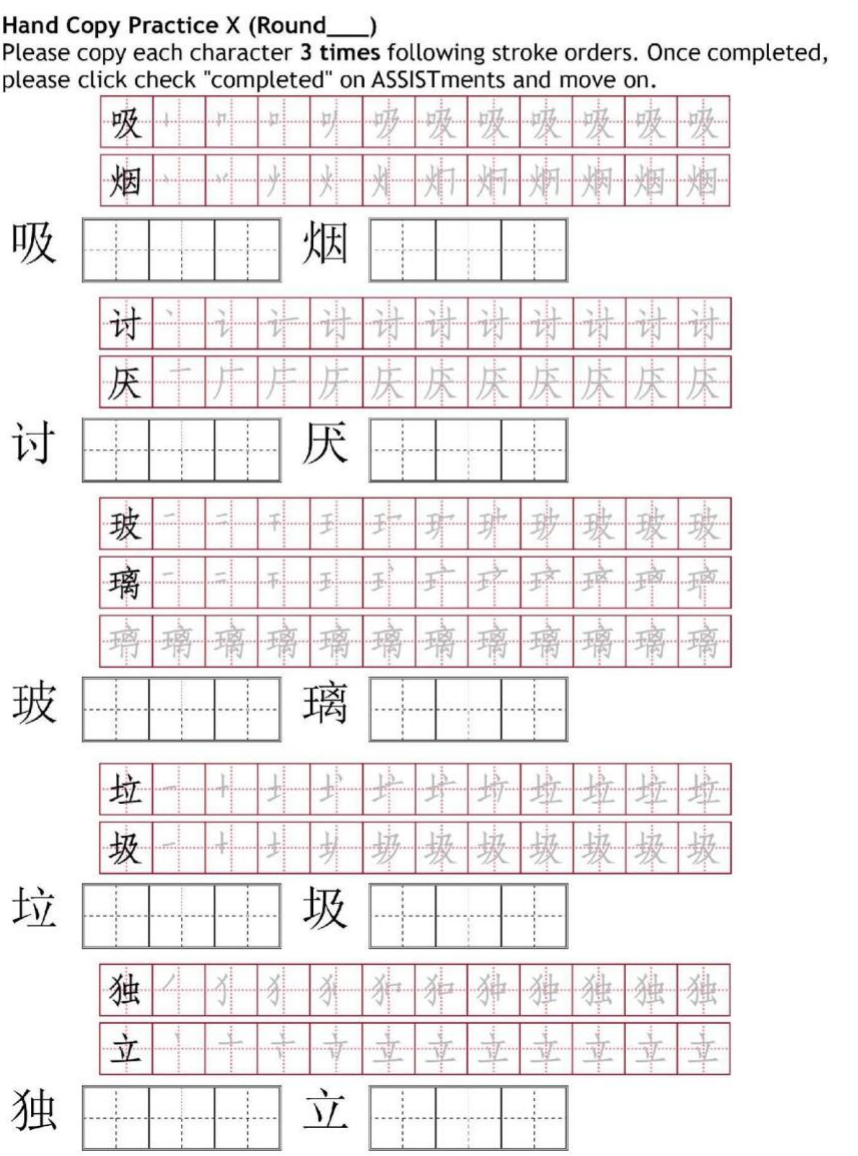
Target words, Pinyin pronunciations, and English meanings for the four experiments

Experiment 1 & 2			Experiment 3			Experiment 4		
Target Word	**Pinyin**	**English Meaning**	**Target Word**	**Pinyin**	**English Meaning**	**Target Word**	**Pinyin**	**English Meaning**
Set X			Set X			建议	jiànyì	suggest; advise
稳定	wěndìng	stable	吸烟	xīyān	to smoke	玻璃	bōlí	glass
仔细	zǐxì	tentative	讨厌	tǎoyàn	sick of; disgusting	独立	dúlì	independent
标准	biāozhǔn	standard	玻璃	bōlí	glass	改革	gǎigé	reform
原因	yuányīn	reason	垃圾	lājī	trash; waste	拒绝	jùjué	turn down; refuse; reject
其实	qíshí	as a matter of fact	独立	dúlì	independent	婚姻	hūnyīn	marriage
Set Y			Set Y			目标	mùbiāo	target; goal
距离	jùlí	distance	污染	wūrǎn	pollute; contaminate	轻松	qīngsōng	relaxed; take it easy
后悔	hòuhuǐ	regret	拒绝	jùjué	turn down; refuse; reject			
熟悉	shúxī	be familiar with	环境	huánjìng	environment; surroundings			
主动	zhǔdòng	initiative	性格	xìnggé	personality			
于是	yúshì	hence	轻松	qīngsōng	relaxed; take it easy			

Appendix B

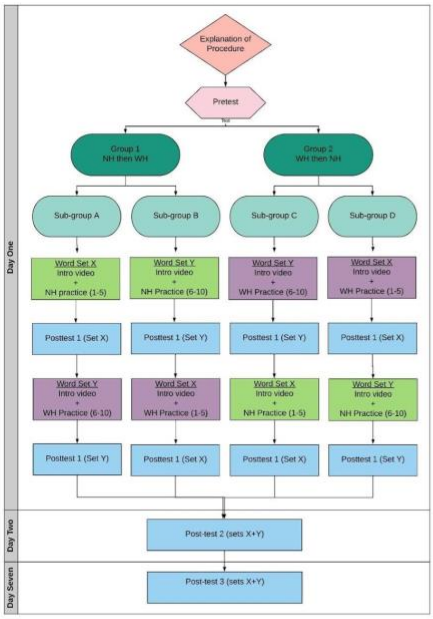
Hand-writing practice prompt and worksheet example for the four experiments



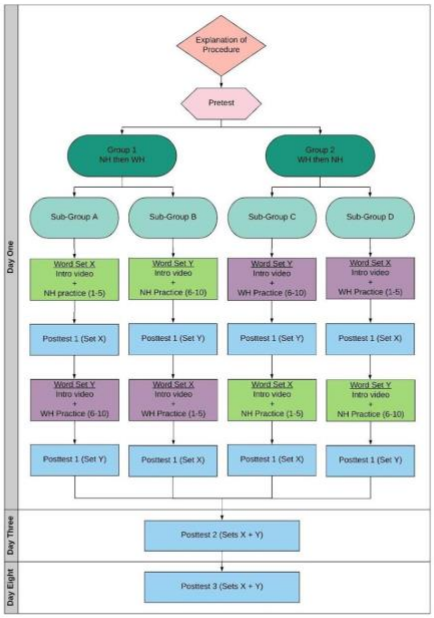


Appendix C

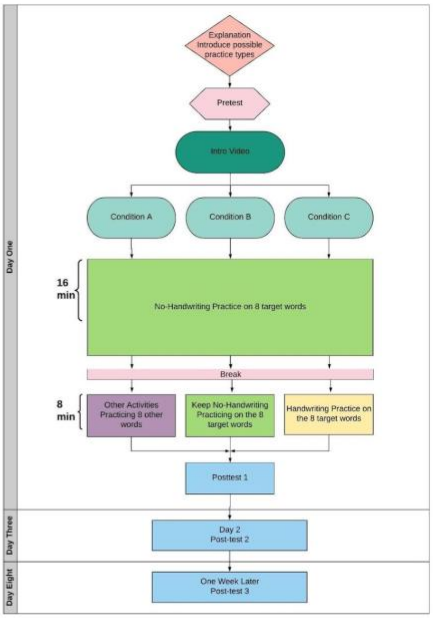
Experimental design for the four experiments



a. Experiment 1 & 2



b. Experiment 3



c. Experiment 4

Chapter 5: Immediate Versus Delayed Feedback on Learning: Do People's Instincts Really Conflict with Reality?

This chapter presents the following manuscript:

Lu, X., Sales, A., & Heffernan, N. T. (2021). Immediate Versus Delayed Feedback on Online Learning: Do people's Instincts Really Conflict with Reality? *Journal of Higher Education Theory and Practice, 21*(16). https://doi.org/10.33423/jhetp.v21i16.4925

Abstract

Researchers have held differing views on the effects of feedback timing for decades. A closer reading of the timing of feedback literature that favored delayed feedback revealed that this conclusion may have been reached prematurely, because the results might have been confounded by the time interval between feedback and a posttest. This study differs from previous feedback timing studies in three distinct ways: First, this study addressed the limitations of previous studies by holding time interval between the feedback (either immediate or delayed) and the posttest constant. Second, this study included various types of knowledge and investigated the interaction between feedback timing and different knowledge types. Third, most studies that investigate the comparative effectiveness of immediate and delayed feedback on written assignments were conducted in the STEM fields, whereas few studies can be found in the second language learning field. Results revealed that the immediate feedback condition significantly outperformed the delayed feedback condition on conceptual knowledge learning, however, no difference between the two conditions was found on situational knowledge learning.

Key words: feedback timing, immediate feedback, delayed feedback, knowledge types, second language learning

Introduction

Feedback in educational contexts has been considered crucial to learning (Narciss & Huth, 2004). Feedback effectiveness is affected by a range of factors such as the purpose, specificity, information provided, forms, as well as the timing (Shute, 2008). Among all of the above-mentioned features, the preference for immediate or delayed feedback has long been a debate. The cognitive psychology literature suggests that the timing of feedback may affect learning (Butler et al., 2007; Kulik & Kulik, 1988), and each preference has their own theories to support it (Fu & Li, 2020). Researchers have been holding opposing conclusions on the effects of feedback timing for decades (Butler et al. 2007; Clariana, 1999; Kulik & Kulik, 1988; Pound & Bailey, 1975). Most educators and students assume that immediate feedback improves learning. There has been a significant shift toward the use of immediate feedback in online courses, and learning platforms are designed to improve learning through its use. Not everyone, however, understands the effectiveness of immediate feedback the same way. Some researchers have underscored the assumption that immediate feedback is intrinsically better as derived from the behaviorist approach to learning (Holland & Skinner, 1961). However, behaviorism is far from explaining human learning behavior based on current cognitive research results. Butler & Woodward (2018) reviewed many studies and concluded that task-level delayed feedback outperformed immediate feedback. Corral et al. (2021) conducted three experiments and argued that longer intervals in delayed feedback might enhance the processing of feedback and enhance conceptual knowledge learning. Furthermore, with the development of the cognitive revolution, many laboratory studies reported that delayed feedback can be more effective in promoting long-term retention than immediate feedback (e.g. Smith & Kimball, 2010, Mullet et al. 2014). The

ManyClasses experiment (Fyfe et al., 2019) was a vigorous attempt in this regard and asked the question about when delayed feedback is better than immediate feedback across a variety of authentic educational contexts. They pre-registered a complete experiential analysis plan in which several instructors tested the two conditions of immediate and delayed feedback. They did not, however, find any significant difference between the two feedback conditions.

Does the desire to provide immediate feedback conflict with reality - that delayed feedback improves learning? A closer reading of the timing of feedback literature which favored delayed feedback reveals that past studies related to delayed feedback have reached premature conclusions. Many of these studies contain methodological flaws that undermine their results. For instance, the uncontrolled time interval between different kinds of feedback and posttests might have misled the results (Metcalfe et al., 2009; Nakata, 2015). For example, Fyfe et al. (2019) only required a set time period between the assignment and the posttest, and designated delayed feedback provided a confounder between the assignment and the posttest, as students in the delayed feedback condition received a shorter time interval between the exposure of the question and answer to the posttest. However, the immediate feedback condition did not. Other similar experimental designs include Mullet et al. (2014). In other words, previous studies were designed inherently to support the opinion that delayed feedback is preferable for a "test with a set date". This contrast learning in which immediate feedback is given, because the delayed feedback appears closer to the posttest, hence improving students' test grades. The time interval confounds the results, potentially caused by the different time interval rather than feedback timing *per se*.

The types of information that students learn might also affect the feedback results, as different types of knowledge indicate various informational processing methods in the brain.

Educational psychologists De Jong, T., & Ferguson-Hessler, M. (1996) summarized previous research on different types and qualities of knowledge, providing four major categories of knowledge from the practical perspective: 1. situational, 2. conceptual, 3. procedural, and 4. strategic. The first two types of knowledge, situational knowledge (such as a cultural fact) and conceptual knowledge (such as the usage of a grammar pattern), are the most commonly acknowledged in all learning fields, relatively easy to quantize, and each possess unique features. Hence, we chose the first two types of knowledge as the research subjects of the current experiment.

There have been studies that focus on the feedback timing of situational knowledge such as vocabulary learning, especially in the first language learning field. In the second language learning field, Nakata (2015) investigated the effectiveness of feedback timing on Japanese college students learning English vocabulary. They were able to hold the time interval between feedback and posttests constant to avoid confounding. However, they discovered little variance between immediate and delayed feedback on vocabulary learning. As for conceptual knowledge such as grammar patterns, it is not yet clear whether the learning of this kind of knowledge may be affected by feedback timing in the field of second language learning.

In the second language learning field, feedback for oral communication and written assignments are explored separately due to their distinct features. Most studies that focus on the corrective feedback for oral communication favored immediate feedback (Fu & Li, 2020, Li, 2020). Although there have been many studies in the STEM field on the effects of feedback timing on written assignments, few studies have been done in the second language (L2) learning field. The “feedback” mentioned in the current study referred to written feedback, not oral corrective feedback. Nakata (2015) compared the learning effects of immediate feedback

(immediately after each response) and delayed feedback (withheld until the end of the entire assignment) and found no difference between the immediate and delayed feedback. However, the delayed feedback they studied was not typical delayed feedback in a real classroom: in a traditional classroom with paper worksheets, teachers would typically not provide feedback until a few days later.

In the current study, immediate feedback refers to feedback provided immediately following each question and prior to the next one, while delayed feedback refers to feedback provided a few days after students have completed the entire assignment. This study focused on applied research in the classroom with computers rather than in laboratory research. As a result, the definition of immediate and delayed feedback underscored the difference between online assignment feedback, which can be provided right after each question, rather than the traditional method of providing assignment feedback which usually takes a fews days. Feedback delay by a few seconds is considered immediate feedback, rather than delayed feedback in the laboratory (Carpenter & Vul, 2011).

This study differs from previous feedback timing studies in three keyways: First, this study addressed the limitations of previous studies by holding the time interval between the feedback (either immediate or delayed) and the posttest constant. By doing this, students in different conditions had an equal length of time to forget information, and the posttest results were not confused with a lag to posttest. Second, this study included different types of knowledge and investigated the interaction between feedback timing and different types of knowledge. Third, most studies on the comparative effectiveness of immediate and delayed feedback on written assignments were conducted in the STEM fields, and not much support can

be found in the second language learning field. This study sought to fill this gap by conducting studies in a second language classroom.

This study hypothesized that immediate feedback is at least as effective as delayed feedback when the same time interval between feedback and posttest are the same. The research question is: how does feedback timing affect students' learning of situational and conceptual knowledge in a second language classroom?

Methods

Participants

Participants were 41 undergraduate students enrolled in an Intermediate Level Chinese class in the Fall 2019 semester at a private university. The participants included 24 males and 17 females, with 3 freshmen, 20 sophomores, 13 juniors, and 5 seniors. Students participated in the course for credit.

Experimental Design

This study was a pre-post randomized experiment with a within-subject design: feedback timing (Immediate versus Delayed) was the independent variable and students' posttest results were the dependent variable. Controlled variables included test contents, pretest score, and test time.

Participants were randomly divided into two groups. To increase the chance that participants in each group had equal variance, participants were anonymized and rank-ordered based on the pretest scores. Then, the top two performing participants were paired, and each pair member was randomly assigned to a group, as shown in Figure 1.

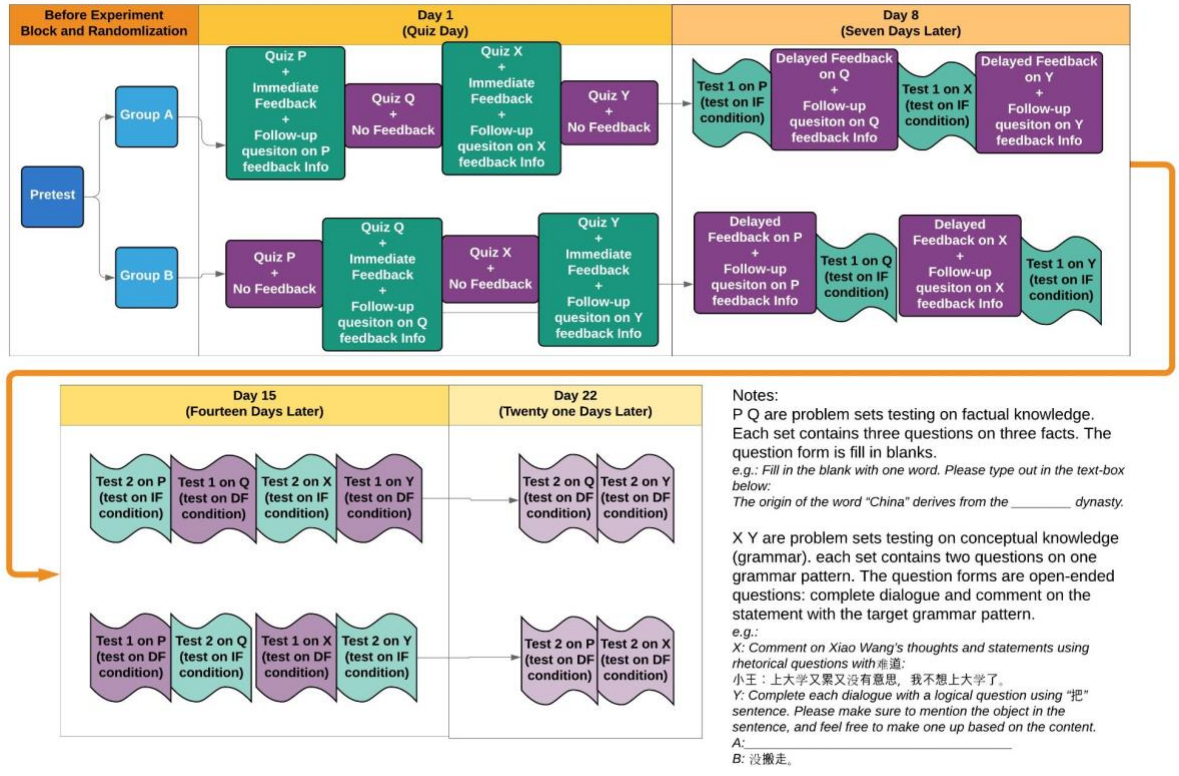
Both groups received all the situational and conceptual knowledge questions in the same order of P, Q, X, Y, yet that contained different feedback conditions. Group A received question

sets P and X with immediate feedback, Q and Y with delayed feedback, while Group B received question sets P and X with delayed feedback and Q and Y with immediate feedback.

All the assignments and tests took place during class time. The time interval between feedback and posttests was a period of seven days and that was determined by the course schedule.

Figure 1

Experimental design and procedure



Materials

Assignments

Four sets of assignments were designed: P, Q, X, and Y. Quiz P and Q each contained three situational questions. Situational knowledge included a series of fun facts related to China and Chinese culture rarely known by students and challenging to recall after a single reading. Quiz X and Y each contained three conceptual questions for one grammar pattern. Two grammar patterns were implemented that students learned previously, yet had a difficult time mastering: making rhetorical questions for set X and *ba* (把) pattern for set Y. The questions for each assignment were listed in Appendix A. P and X were grouped together, Q and Y were grouped together when assigning quizzes to groups.

The Pretest and Posttests

Each test contained all the situational questions in P and Q, as well as two questions on conceptual knowledge 1 and two questions on conceptual knowledge 2. P and Q were represented situational knowledge, so the pretest and the two posttests contained the same questions on the assignments. X and Y represented conceptual knowledge so the pretest and the two posttests contained new questions about the same concepts in the assignments. The questions for all the tests can be found in Appendix B. Identical questions were used in both the pretest and posttests to better compare the learning results. Since no feedback was provided after the pretest, and the time lag between the pretest and posttest was at least 7 days, it is assumed that the pretest would not cause a significant difference in posttest learning. Target grammar patterns and cultural facts were not mentioned in the class nor in any of the assignments during the entire experimental process.

Feedback

All the situational questions could be graded automatically by the system and both types of feedback were provided: The feedback included confirmation of accuracy as well as the correct answers. For the feedback of conceptual questions, a sample answer, as well as a few sentences explaining the usage of the grammar pattern, were provided as feedback. This kind of feedback was provided because there were numerous variations for grammar pattern usage, and students' answers could not be graded automatically. All students enrolled in the course accessed and completed the assignments and received the same feedback. However, the timing of feedback varied depending on the conditions. To ensure that the feedback was read, students were required to view the feedback for each question in order to receive credit for completing the assignment. A follow-up question appeared immediately following students reading of the feedback, whether immediate or delayed, and the instructions prompted the following response: "The correct answer is … Did you get it right?"

Procedure

This experiment was conducted during regular class hours with the use of ASSISTments embedded in the Moodle platform. ASSISTments is an online learning platform that supports student learning with hints and immediate feedback (Heffernan, N.T. & Heffernan, C.L., 2014). The assignments, immediate or delayed feedback, follow-up questions on feedback information, and the posttest were all delivered using ASSISTments. Since the course gives all participants daily in-class quizzes using student's individual laptops, all participants were comfortable taking quizzes online. Two in-class quizzes on ASSISTments were assigned prior to the experiment to help students familiarize themselves with the system.

ASSISTments anonymized the research data before providing it to researchers. All experiments were overseen by WPI's IRB team.

At the beginning of the semester, instructors announced the opportunity to participate in the study, clarifying to students that participation in the research study would not change their course experiences, but rather, stated that their course data would be provided to the researchers. Before the experiment started, students had been assigned three homework assignments on ASSISTments to familiarize themselves with the platform. A pretest was assigned during class time with no feedback given. Then, students were assigned randomly into two groups based on their pretest scores. On the first day of the experiment, the course instructor asked all students to login to the Moodle platform. Instructors provided ample class time for all students to complete their assignments. Depending on the groups, students received immediate feedback for two sets of questions and no immediate feedback for the remaining two. Immediate feedback was provided immediately following each question. If they received immediate feedback, students would be asked to answer a follow-up question to confirm that it was read.

On the eighth day, the teacher again asked all students to login to the Moodle platform and complete all the questions. Students were asked to take post-test 1 for the question sets on which they had received immediate feedback seven days ago, as well as read delayed feedback for the two sets of questions they hadn't received feedback on from seven days ago. The delayed feedback is also followed by questions to confirm the reading of the feedback. On the sixteenth day, students were required to take test 2 for the assignments with immediate feedback, and test 1 for those with delayed feedback. On the twenty-second day, students were asked to take test 2 for the assignments with delayed feedback. A procedure flow chart can be found in Figure 1.

Preregistration

The experimental design was pre-registered on OSF and the fully anonymized data can be found in Lu & Heffernan (2020). The sole analysis change from the plan was using R instead

of SPSS. During pre-registration t-tests and ANCOVAs in SPSS were suggested, however, given the fact that we paired our students up during randomization, the stronger and more appropriate analysis accounts for that pairing. So, we began implementing R and ran the linear mixed effect model below with the pairing variable (called Pair Number) as a fixed effect.

Results

One participant did not attend on the posttest days and was excluded from the study. Conceptual knowledge data and situational knowledge data were both analyzed separately.

Conceptual knowledge X, Y

A linear mixed model fit by REML in R (RStudio version 1.3.959) is used to perform linear mixed-effects analysis of the relationship between posttest scores and conditions. T-tests used Satterthwaite's method [lmerModLmerTest]. As fixed effects, we entered the condition, test content, pretest score, and test time. As random effects, we had intercepts for subjects, and converted the variable "pair number" into a factor but preserved the variable and value label attributes. The experimental model was:

$$Score_{ij} = \beta_{0pair[i]} + \beta_1 Condition_{ij} + \beta_2 TestContent_{ij} + \beta_3 PreAveXorY_{ij} + \beta_4 TestTime_{ij} + \eta_i + \varepsilon_{ij} \quad (1)$$

Where $Score_{ij}$ is the score for student *i* at measurement time *j=1,2,3* (?), $\beta_{0pair[i]}$is a fixed intercept for *i*'s randomization pair, and η_i is a random intercept for student *i*. The REML criterion at convergence was 100.1. The model's intercept remained at 0.45 (SE = 0.21, 95% CI [0.05, 0.70]). Within this model: The effect of condition was significant, β_1 = 0.16, SE = 0.04, 95% CI [0.07, 0.24], *t* (110.65) = 3.71, $p < .001$***, which indicated the immediate condition performed significantly better than the delayed condition. The effect of test content was significant (beta = -0.23, SE = 0.05, 95% CI [-0.32, -0.12], *t* (117.72) = -4.70, $p < .001$***).

When looking into the test content, participants performed significantly better in X than in Y, which might be caused by differing question types or the varying difficulty levels of the two grammar patterns. The effect of the pretest was significant, (beta = 0.39, SE = 0.11, 95% CI [0.25, 0.68], t (128.96) = 3.44, p < .001***), which indicated that higher pretest scores led to higher posttest scores. The effect of test time was insignificant, (beta = -0.001, SE = 0.04, 95% CI [-0.08, 0.09], t (110.31) = -0.03, p = 0.97. None of the effect of pair numbers was significant. Variance explained by the random effect subject was 0.03, SD = 0.19. See Table 1.

Table 1

Linear mixed-effects analysis for conceptual knowledge using posttest scores as the criterion

Predictor	***β***	**SE**	***df***	***t***	***p***	**95%CI**
(Intercept)	0.448	0.207	39.262	2.160	0.037*	[0.053, 0.704]
Condition	0.163	0.044	110.652	3.714	0.0003 ***	[0.069, 0.241]
TestContent	-0.233	0.050	117.721	-4.704	7.01e-06 ***	[-0.315, -0.124]
Pretest Average	0.388	0.113	128.956	3.443	7.75e-04***	[0.252, 0.677]
Test Time	-0.001	0.043	110.313	-0.032	0.974	[-0.083, 0.086]

Notes: N=154; * p<.05; ** p<.001; *** p<.0001.

Situational knowledge P, Q

The same model as the conceptual knowledge dataset was used for situational knowledge data. The REML criterion at convergence was -1.5. The model's intercept was at 0.28 (SE = 0.12, 95% CI [0.04, 0.41]). Within this model: The effect of condition was insignificant, beta = -0.05, SE = 0.09, 95% CI [-0.02, 0.10], *t* (110.9) = -0.55, which indicated that the condition did not explain much of the variance, and no difference was found between immediate and delayed conditions. The effect of test content was insignificant (beta = -0.04, SE = 0.03, 95% CI [-0.03, -0.09], *t* (113.4) = 1.23, *p* =0.22). Alternatively, the pretest effect was significant, (beta = 0.54, SE = 0.10, 95% CI [0.38, 0.74], *t* (127.7) = 5.51, *p* < .001***), which indicated that higher pretest scores led to higher posttest scores. The effect of test time was insignificant, (beta = 0.003, SE = 0.04, 95% CI [-0.02, 0.09], *t* (110.5) = 0.007, *p* = 0.99. None of the effect of pair numbers was significant. Variance explained by the random effect "subject" was 0.02, SD = 0.13. See Table 2.

Table 2

Linear mixed-effects analysis for situational knowledge using posttest scores as the criterion

Predictor	***β***	**SE**	***df***	***t***	***p***	**95%CI**
(Intercept)	0.232	0.123	27.629	1.888	0.070	[0.045, 0.407]
Condition	0.042	0.029	111.776	1.452	0.149	[-0.016, 0.097]
Test Content	0.038	0.031	114.430	1.233	0.220	[-0.027, 0.093]
Pretest Average	0.541	0.099	128.831	5.488	2.08e-07 ***	[0.381, 0.744]
Test Time	0.032	0.029	111.741	1.101	0.273	[-0.023, 0.090]

Note: N=155; SE = standard error of B; * *p*<.05; ** *p*<.001; *** *p*<.0001.

Discussion

The purpose of this study is to identify the optimal feedback timing for student learning in the second language field. This study included two informational types to study the impact of feedback timing with different knowledge types. This study also kept the time interval constant, between feedback and posttests in both immediate and delayed conditions. Results of this study suggested that the immediate condition performed significantly better than the delayed condition when learning conceptual knowledge (grammar), no difference was found between the two conditions when learning situational knowledge (cultural facts).

The results of this study contradicted the findings that claim a significant delay-retention effect, and supported the effectiveness of immediate feedback when learning conceptual knowledge such as grammar. This study partially supported that immediate feedback outperformed delayed feedback by clearly defining immediate and delayed feedback in a classroom setting while holding the time interval between the feedback and posttests as constant in both conditions.

Conceptual knowledge and situational knowledge reacted differently to the feedback timing. No variation was found between the immediate and delayed feedback conditions while students were learning situational knowledge, and these results further supported Nakata (2015)'s findings. However, a significant difference was found in the two conditions in the learning of conceptual knowledge. According to Anderson (1983), conceptual knowledge is encoded declaratively first, which is identical to situational knowledge, then translated into procedures. If this is the case, then the different results on the two types of knowledge in our study can be indirect evidence that immediate feedback has a positive impact on the procedures of translating declarative information into conceptual knowledge.

In the second language learning field, most research related to immediate feedback is focused on vocabulary learning. Effects of feedback timing on the learning of grammar is not explored thoroughly. One reason might be caused by the difficulties of providing immediate feedback to open-ended questions, such as for grammar usage. For example, there are usually several correct ways to say one sentence with the same grammar pattern due to language variations. This study tries to fill this gap in the research by providing suggested correct answers. Results indicated that the provision of correct answers improves students' online learning for open-ended questions.

It should be noted that the two different conceptual knowledge test content of X and Y made a difference in the results, and R analysis indicated that students performed significantly better in the posttest of X than in Y. In our data analysis, we discovered that X also received significantly higher pretest results compared to Y. This might be caused by varying levels of difficulty of the two grammar patterns or different question types. Test contents were selected based on the course schedule to best serve students' learning purposes. To prevent students from learning the content of the second condition while completing the first, different contents were required.

Pedagogically, the results imply that immediate feedback may be preferred when learning conceptual knowledge such as grammar in the second language field, and both delayed and immediate feedback may be used when learning situational knowledge such as cultural facts. Since most educators and students have the assumption that immediate feedback improves learning, providing immediate feedback for situational and conceptual knowledge might be more desirable. The advantage of immediate feedback also indicated that online homework assignments with feedback provided immediately after each question may be a more effective

way of learning, as compared to traditional handwritten homework assignments which typically take a few days for students to receive any feedback. Although it is challenging to immediately correct open-ended questions such as forming sentences with a grammar pattern, it is nevertheless beneficial for instructors to provide immediate feedback with suggested answers.

Acknowledgement

This work was supported in part by NSF (e.g., 2118725, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (e.g., R305D210031, R305A170137, R305A170243, R305A180401, & R305A120125), GAANN (e.g., P200A180088 & P200A150306), EIR (U411B190024), ONR (N00014-18-1-2768) and Schmidt Futures.

References

Anderson, J. R., (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22* (3), 261-295. https://doi.org/10.1016/S0022-5371(83)90201-3.

Butler, A., Karpicke, J., & Roediger, H. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal ofExperimental Psychology: Applied, 13*(4), 273–281. https://doi.org/10.1037/1076-898X.13.4.273

Butler, A. C., & Woodward, N. R. (2018). Toward consilience in the use of task-level feedback to promote learning. *Psychology of Learning and Motivation, 69*. https://doi.org/10.1016/bs.plm.2018.09.001

Carpenter, S. K., & Vul, E. (2011). Delaying feedback by three seconds benefits retention of face–name pairs: the role of active anticipatory processing. *Memory & Cognition 2011 39*(7), 1211–1221. https://doi.org/10.3758/S13421-011-0092-1

Clariana, R.B. (1999). *Differential Memory Effects for Immediate and Delayed Feedback: A Delta Rule Explanation of Feedback Timing Effects*. Paper presented at the Association of Educational Communications and Technology annual convention, Houston, TX.

Corral, D., Carpenter, S. K., & Clingan-Siverly, S. (2021). The effects of immediate versus delayed feedback on complex concept learning. *Quarterly Journal of Experimental Psychology, 74*(4), 786–799. https://doi.org/10.1177/1747021820977739

De Jong, T., & Ferguson-Hessler, M. (1996). Types and qualities of knowledge. *Educational Psychologist, 31*(2), 105-113.

Fu, M., & Li, S. (2020). The effects of immediate and delayed corrective feedback on L2 development. *Studies in Second Language Acquisition*, 1-33. https://doi.org/10.1017/S0272263120000388

Fyfe, E., de Leeuw, J. R., Carvalho, P. F., Goldstone, R., & Motz, B. (2019, May 28). ManyClasses 1: Assessing the generalizable effect of immediate versus delayed feedback across many college classes. https://doi.org/10.31234/osf.io/4mvyh

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497. https://doi.org/10.1007/s40593-014-0024-x

Holland, J. G., & Skinner, B. F. (1961). *The analysis of behavior: A program for self-instruction.* New York, NY, US: McGraw-Hill.

Kulik, J. A., & Kulik, C-L. C. (1988). Timing of Feedback and Verbal Learning. *Review of Educational Research, 58*(1), 79–97. https://doi.org/10.3102Timing of Feedback and Verbal Learning - James A. Kulik, Chen-Lin C. Kulik, 1988/00346543058001079

Li, S. (2020). What is the ideal time to provide corrective feedback? Replication of Li, Zhu & Ellis (2016) and Arroyo & Yilmaz (2018). *Language Teaching, 53*(1), 96-108. https://doi.org/10.1017/S026144481800040X

Lu, X., & Heffernan, N. T. (2020, July 14). Data. Retrieved from osf.io/wvfxk

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition, 37*, 1077–1087. https://doi.org/10.3758/MC.37.8.1077

Mullet, H. G., Butler, A. C., Verdin, B., von Borries, R., & Marsh, E. J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition,* 3, 222-229. https://doi:10.1016/j.jarmac.2014.05.001

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multi-design for multimedia learning. In H. M. Niegemann, D. Leutner, & R. Brunken (Ed.), *Instructional design for multimedia learning* (pp. 181-195). Munster, NY: Waxmann.

Nakata, T. (2015). Effects of feedback timing on second language vocabulary learning: Does delaying feedback increase learning? *Language Teaching Research, 19*(4), 416–434. https://doi.org/10.1177/1362168814541721

Pound, L.D., & Bailey, G.D. (1975). Immediate Feedback Less Effective than Delayed Feedback for Contextual Learning. *Reading Improvement, 12*(4), 222-224.

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay–retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 80-95. https://doi.org/10.1037/a0017407

Appendix A

Assignment Questions for P, Q, X, Y

Assignment P:

The origin of the word “China” derives from the Qin dynasty.

China is considered to be the oldest civilization with some historians marking 6000 BC as the beginning of the Chinese civilization. Also, it has the world’s longest-used written language.

The full name of the current leader/Chairman of China is Xi Jinping.

Assignment Q:

The population of China now is around 1.4 billion.

China has been the source of many innovations, scientific discoveries and inventions. This includes the Four Great Inventions: papermaking, the compass, gunpowder, and printing.

The people's republic of china was founded in 1949.

Assignment X:

Comment on Xiao Wang’s thoughts and statements using rhetorical questions with 难道:

小王：我知道明天有考试，可是今晚我还是要去看电影。

Your comment:

A sample answer: 难道你复习好了吗？

The subject “你” can be put in front of “难道” or right after it.

Did you get it right?

小王：我三天没有睡觉了。

Your comment:

A sample answer: 难道你不累吗？

Generally speaking, rhetorical questions that are affirmative in form carry an emphatically negative meaning; rhetorical questions that are negative in form carry an emphatically positive meaning. The meaning of the sample answer is: “Don’t you feel tired?”

Did you get it right?

You: 你的钱包丢了，这很糟糕。但你的钱包不一定是小张偷的。

小王：肯定是小张偷走的！

You:

A sample answer: 你怎么知道？难道你看见了吗？

The underline meaning of the sample answer is “Since you didn’t see it by yourself, you cannot say that Xiao Zhang stole it.”

Did you get it right?

Assignment Y:

Complete each dialogue with a logical question using “把” sentence. Please make sure to mention the object in the sentence, and feel free to make one up based on the content.

A: ____

B: 写完了。

Sample answer: 你把作业写完了吗？

The object should be known. You have to make one up which goes with the verb.

把字句 are most often used to describe what happened to the object in some detail.

Remember to use “了” as in this dialogue the action has taken place.

The verb is not just "bare"; there's "more stuff" after it. In this case, a resultative complement and “了” is needed.

Did you get it right?

A: ____

B: 冰箱里。

Sample answer: 你把蛋糕放在哪儿了？

把 sentences are most often used to describe location change of the object. When talking about location, you should think about 在 in front of the question word. And if it has happened, “了” is also needed.

Did you get it right?

A: ____

B: 送给妈妈。

Sample answer: 你想把这个礼物送给谁？

When there are two objects in one sentence, 把 sentence is often used.

把字句 are not tied to any particular time. When talking about plans, you do not need “了” in the question.

Did you get it right?

Appendix B

Pretest and Posttest Questions for P, Q, X, Y

Tests for P (same to the assignment questions)

1. The origin of the word “China” derives from the Qin dynasty.
2. China is considered to be the oldest civilization with some historians marking 6000 BC as the beginning of the Chinese civilization. Also, it has the world’s longest-used written language.
3. The full name of the current leader/Chairman of China is Xi Jinping.

Tests for Q (same to the assignment questions)

1. The population of China now is around 1.4 billion.
2. China has been the source of many innovations, scientific discoveries and inventions. This includes the Four Great Inventions: papermaking, the compass, gunpowder, and printing.
3. The people's republic of china was founded in 1949.

Tests for X (different from the assignment questions)

Comment on Xiao Wang’s thoughts and statements using rhetorical questions with 难道:

1. 小王：上大学又累又没有意思，我不想上大学了。

Your comment:

2. 小王：我的老板很糟糕，我每天工作时心情都不好。

Your comment:

Tests for Y (different from the assignment questions)

Complete each dialogue with a logical question using “把” sentence. Please make sure to mention the object in the sentence, and feel free to make one up based on the content.

1.

A: ____

B: 没搬走。

2.

A: ____

B: 卖给了一个学生。

Chapter 6: Immediate Text-Based Feedback Timing on Foreign Language Online Assignments: How Immediate Should Immediate Feedback Be?

This chapter presents the following manuscript:

Lu, X., Wang, W., Motz, B., Ye, W., Heffernan, N. T. (2023). Immediate Text-Based Feedback Timing on Foreign Language Online Assignments: How Immediate Should Immediate Feedback Be? *Computers and Education Open, 100148*.

https://doi.org/10.1016/j.caeo.2023.100148.

Abstract

Immediate feedback has been considered a cornerstone of online language learning platforms. A closer reading of the research related to immediate feedback, however, reveals that the definition of "immediate feedback" is inconsistent. Findings from the STEM literature were not well supported from other fields. As a result, clarification is needed in order to assess which type of immediate feedback leads to improved performance in a computer-assisted learning environment. Research related to the effects of immediate feedback outside of STEM classes is meaningful in order to better understand whether the findings can be generalized. This study investigated the effects of immediate feedback timing in online language learning exercises. Three conditions were examined: no feedback, end-of-question feedback, and end-of-assignment feedback. A Planned Contrasts test revealed that with a pretest as the covariate, the end-of-question feedback condition received significantly higher grades in the posttest compared to the end-of-assignment feedback condition, and students' learning improved significantly while taking assignments in the end-of-question feedback condition. Students with lower pretest scores used more attempts, although their learning progress was not significantly better as compared to students with higher prior knowledge. The findings of this study provide insights into the use of immediate feedback to improve learning as part of foreign language classroom instruction.

Keywords: Online learning; Immediate feedback; End-of-question feedback; End-of-assignment feedback; Experiments and methodologies

Introduction

With the rapid growth of learning technologies, educational systems have swiftly adopted online learning technologies over the past decade (Liu et al., 2020). Similar to other disciplines, language learning has embraced the potential of online tools for improving efficiency and flexibility (Kannan & Munday, 2018). Increasing evidence suggests that, when properly integrated into the curriculum, online learning tools (e.g., online tutoring systems and computer-mediated communication) can facilitate improvements in foreign language (FL) development and aptitude (Blake, 2011; Escueta et al., 2017). Among all of the language learning features that online technology may provide, offering feedback on assignments is one of the most promising. Therefore, the current study focused on the provision of text-based feedback on online textual assignments, as it is not only ubiquitous in language learning compared with oral assignments but also affordable (Bahari, 2021).

Shute (2008) used the term "formative feedback" to define the feedback provided to learners during the learning process. In the learning sciences, Butler and Woodward (2018) defined feedback as "information about the gap between actual and desired performance" (p. 2). Furthermore, feedback's positive effects on student performance have been considered the cornerstone of learning (Lipnevich & Panadero, 2021; Wisniewski et al., 2020), which is a commonality across all educational systems. Many studies have supported this contention, although only from multidimensional perspectives (Butler & Woodward, 2018). Hattie and Timperley (2007) summarized the common features of definitions of feedback, underscoring that the type of feedback and how it is provided can be differentially effective. Moreover, Shute (2008) argued that the effectiveness of feedback depends on multiple features, such as its specificity, complexity, length, and timing. Among these features, a prominent one is the timing

of feedback, which is defined as the length of time following a learning action before feedback is presented.

Determining whether differences exist in the effect of various timings of feedback on learning could help researchers to improve the consistency of definitions, reduce the confusion of concepts, tailor feedback to more effectively meet the needs of learners, and maximize the impact of feedback on learning. Therefore, this study aimed to investigate the potential impact on student FL learning by providing immediate feedback at different timings, and also to provide empirical evidence as a point of reference for learners, educators, and platform designers to be able to distinguish the two types of immediate feedback.

The feedback in this study emphasized the provision of textual explanations and demonstrations of language and grammar knowledge for exercises, instead of individualized feedback to student responses. An online textual assignment refers to an assignment offered on an online platform in the form of text, requiring students to type answers or choose them from a list. The assignment in this study focused on improving language skills rather than on being a comprehensive project requiring in-depth research and analysis.

Literature Review

Theoretical Perspectives on Immediate Feedback

Behaviorism

Numerous studies have explored the effects of feedback timing, each backed up by different theories. The systematic study of using immediate feedback in FL classrooms can be traced back to the mid-20th century, when behaviorist theories of learning were prevalent in the field of education (Budiman, 2017). According to behaviorists, the acquisition of language occurs through the formation of habits, and the crucial factors in developing and reinforcing

these habits are immediate positive feedback and correction. This led to an emphasis on immediate feedback as a crucial part of language instruction.

Cognitive Load Theory

In addition to behaviorist theories, cognitive load theory (Sweller, 2011) supports the notion that immediate feedback can enhance learning. This theory suggests that working memory has limited capacity, and providing learners with detailed feedback in manageable amounts prevents information overload. By offering immediate feedback, learners can better process and integrate the information, leading to enhanced learning outcomes. Cognitive load theory supports the notion that immediate feedback can be a powerful tool for facilitating language learning.

Lag Effects

Conversely, the leg effects (Bjork, 2018) posits that increasing the time interval between practice and feedback can lead to improved retention in subsequent tests. The lag effects imply that incorporating a time delay before providing feedback introduces increased learning complexity, which has the potential to enhance retention and learning. Serfaty and Serrano (2022) investigated the impact of lag effects on second language grammar learning and found that longer lags between learning sessions were associated with significantly higher scores among faster learners and learners with higher proficiency levels. On the other hand, shorter lags were found to promote significantly higher scores among slower learners and learners with lower proficiency levels.

Immediate Feedback Timing on Online Assignments

Regarding the timing of feedback, the effectiveness of immediate and delayed feedback has been largely discussed, but conflicting research findings have been reported. Many educators and students, especially students, favor immediate feedback over delayed feedback (see, e.g., van

der Kleij et al., 2012; Lefevre & Cox, 2017), assuming that immediate feedback improves learning. As a result, many online courses, learning platforms, and applications have been designed to improve learning through the use of immediate feedback. Deeva et al. (2021) investigated the use of feedback timing in selected online learning systems. They discovered that over 74% of systems provided immediate feedback, either based on student actions or after the task was completed. Providing immediate feedback immediately after students' actions or errors was found to be optimal.

However, a closer reading of the research related to immediate feedback revealed that the definitions of "immediate feedback" have been inconsistent. To avoid confusion, instead of immediate feedback, we defined the types as end-of-question feedback (EQF) and end-of-assignment feedback (EAF). Text-based feedback refers to feedback provided in the form of text, as opposed to the form of audio and video. Shute (2008) defined immediate feedback as feedback provided "right after a student has responded to an item or problem or, in the case of summative feedback, right after a quiz or test has been completed" (p. 163), which included both EQF and EAF. Butler and Woodward (2018) reviewed many studies related to feedback and highlighted the inconsistency of the definition of immediate feedback. Various authors have used the term "immediate feedback" to refer to the following: (a) EQF, which indicates feedback provided immediately after each question (see, e.g., Dihoff et al., 2004; Metcalfe et al., 2009; van der Kleij et al., 2012; van der Kleij et al., 2015); (b) EAF, which indicates feedback provided immediately after the entire assignment (see, e.g., Fyfe et al., 2021; Qi et al., 2020); and even (c) feedback provided after a short delay, such as immediately after the assignment deadline (Mullet et al., 2014). On the other hand, some studies have classified EAF as delayed feedback (see, e.g., Attali & van der Kleij, 2017; van der Kleij et al., 2015). In a meta-analysis of feedback timing, van der Kleij et

al. (2015) defined immediate feedback as feedback provided immediately after each question, while all other feedback was called delayed feedback, including EAF. Crucially, without a clear definition, one cannot draw any conclusions regarding the effectiveness of immediate feedback timing.

Furthermore, few studies have addressed the difference between EQF and EAF. Dihoff et al. (2004) compared various types of feedback, including EQF, EAF, 24-hour delay, and a control group. They demonstrated that EQF promoted the most retention and the most accurate identification of initial responses. Attali and van der Kleij (2017) investigated the effects and interactions of feedback types; their unique timings, such as immediate or delayed; and item format on learning. They found that EAF outperformed EQF when the option for a later review was provided. However, which type of immediate feedback leads to improved learning retention remains ambiguous.

A variety of factors must be considered when attempting to measure the effect of feedback on student learning and retention, including incentivization and retry. For example, researchers must validate that students have indeed read the feedback in the first place. Timmers and Veldkamp (2011) found that half of their participants only paid attention to feedback on incorrect answers, while a quarter did not pay any attention to feedback. Mullet et al. (2014) also found that a mere half of the feedback provided on all homework assignments was actually viewed by students. Additionally, assignments did not include incentives for students to view feedback. This suggests that incentives to view feedback should be employed if researchers wish to examine differences in feedback manipulation.

Another decision to make is whether to allow the "retry" function in a classroom experiment design. This is widely provided by learning platforms immediately after students

answer incorrectly and before feedback is provided. Some researchers considered the retry function on assignments to design a confounder in a classroom experiment (Zhang et al., 2021). They argued that different processing times and the engagement of different strategies would confound the results or induce anxiety and frustration. Other studies have indicated that providing choices and retry opportunities to students could enhance their learning (Culpepper, 2014; Ostrow & Heffernan, 2014; Attali & Powers, 2010). After reviewing automated feedback technologies, Deeva et al. (2021) emphasized the potential positive influences of a higher degree of control over feedback on learners' performance. It is reasonable to acknowledge each method of feedback as having its own unique features for a classroom experiment aimed at measuring learning and guiding best teaching practices. When considered holistically, all features may impact the learning and retention of material.

Immediate Feedback on Chinese language Learning

Panadero and Lipnevich (2022) underscored the importance of feedback research within different learning contexts. They noticed that different learning contexts impact the effects of feedback timing. Feedback research in the FL learning field might provide a different perspective (Lahcen & Mahmoud, 2022). Feedback in FL acquisition gained more research attention after Long proposed the interaction hypothesis in 1983 (Gass & Mackey, 2014), which indicates that corrective feedback might be a crucial method for enhancing language acquisition. Regardless of the popularity of textual assignments and text-based feedback, most studies related to feedback timing in the FL education field have focused on oral feedback in classroom communication (Henderson, 2021; Li, 2022) due to the nature of the field. Rassaei (2019) highlighted that both oral and text-based feedback are beneficial to FL learners, whereas Kang and Han (2015) found that text-based feedback led to greater grammatical accuracy in FL writing. Therefore, the current

study aimed to fill this gap.

Computer-mediated text-based feedback has been widely studied in the field of English as an FL (Barrot, 2021). However, most studies have investigated the efficacy and accuracy of providing text-based feedback (also called written corrective feedback), not the timing of feedback itself. Among the researchers who have studied feedback timing, most have compared immediate feedback (either EQF or EAF) with delayed feedback (either EAF or with a delay). Their findings have been controversial, with some studies favoring immediate feedback (Shintani & Aubrey 2016) and others favoring delayed feedback (Salajegheh et al., 2022).

Notably, only a few studies in the FL education field have offered empirical results on the effects of immediate feedback timing on online text homework assignments; however, their conclusions have still been immature. Nakata (2015) compared the effects of EQF and EAF on FL vocabulary learning, finding nonsignificant differences between these feedback types. Henshaw (2011) have found similar results. However, Kılıçkaya (2022) compared EQF (defined as immediate feedback) and EAF (defined as delayed feedback) in computer-supported FL grammar instruction and discovered EQF to be more efficient. With limited experimental studies comparing the differences in immediate feedback timing in the FL learning field, it remains unclear which type of immediate feedback is most effective when learning an FL.

Research Question and Hypothesis

Overall, providing immediate feedback is widely accepted to enhance learning, and therefore, an increasing number of online tools that provide immediate feedback to language learners have been offered. However, a more focused inquiry should be evaluated, as exemplified in the following question: When should immediate feedback be provided to benefit learners the most – after each question is answered or after an assignment is completed? Therefore, the present

study built on currently available evidence regarding the effectiveness of immediate feedback in online learning environments, aiming to clarify the inconsistent definition of immediate feedback and to examine the possible impact on student learning by providing EQF and EAF. Specifically, it investigated the differences between the two types of immediate feedback using an experimental design in the teaching of Chinese as a FL. The following research question was addressed:

> If immediate feedback improves FL learning, which type of immediate feedback is preferable – EQF with retries (immediately after each question) or EAF (at the end of the entire quiz assignment)?

Due to the limited research findings in this area, we hypothesized that no differences exist between the EQF and EAF conditions, and that both feedback conditions would receive higher scores in the posttest compared with the control condition.

Methods[1]

Participants

This study employed a convenience sampling method to recruit participants. The participants were 81 undergraduate students enrolled in intermediate-level Chinese classes in the fall 2020 semester at two private universities in the United States. Chinese was not their native language, and students were placed into the classes based on placement tests at the beginning of the semester for credits. There were 40 male and 41 female participants with an age range of 18–25 years. Four students did not complete the experiment as a result of their absence.

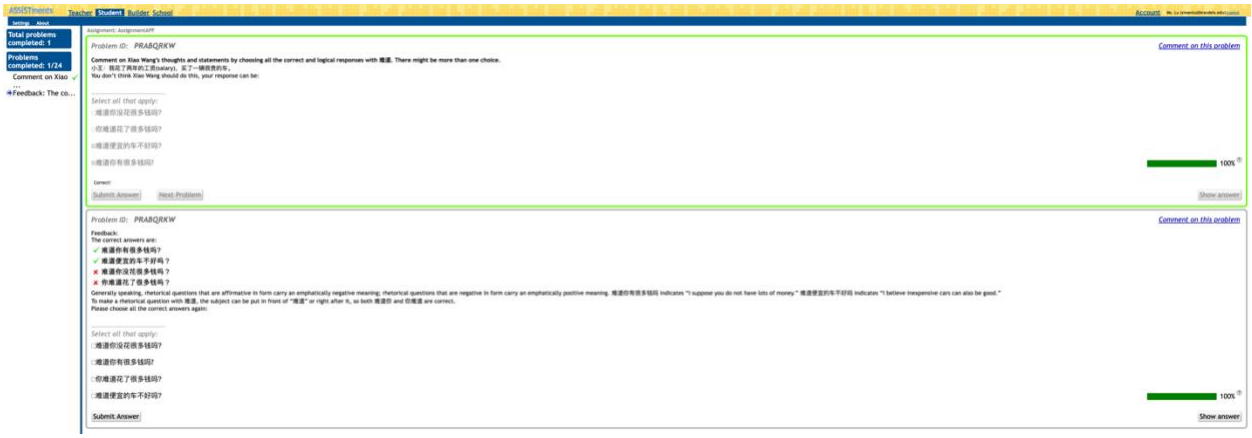
The Online Learning Environment

[1] A time-stamped, independent, read-only registration of this manuscript and data is available at https://osf.io/h95pv/?view_only=0b8495850fdd43d6929f24ca775ca1be.

This study's experiment was conducted during in-person classes with the use of the ASSISTments platform. Similar to many other learning platforms, ASSISTments is an online learning platform that supports student learning with hints, feedback, and scaffolding. The system supports research, such as the current study, by sharing learner data with fellow data scientists (Heffernan & Heffernan, 2014). Figure 1 presents the ASSISTments interfaces for students, teachers, and researchers. At least one in-class quiz was assigned on ASSISTments before the experiment began to familiarize students with the system. The quiz assignments, immediate feedback, follow-up questions on feedback information, and posttest were all delivered using ASSISTments, and all assignments, pretests, and posttests in the current experiment were conducted during class time. The research data were anonymized by the platform before being provided to the researchers. Each course provided all students with daily in-class quizzes determined by the nature of the courses using the students' personal laptop computers; as a result, the participants were comfortable taking online quizzes. Both courses provided access to laptops at the beginning of the semester to support inclusivity. Students could also use their smartphone to take the quizzes.

Figure 1

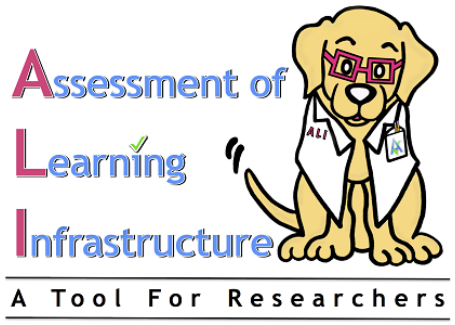
ASSISTments Interfaces for Students (a), Teachers (b), and Researchers (c)



a. Student Interface

Student/Problem [Unanonymize]	Average Data driven	PRABQRSE Data driven	PRABQRSF Data driven	PRABQRSG Data driven	PRABQRSH Data driven	PRABQRSJ Data driven	PRABQRSK Data driven	PRABQX7F Data driven	Total Hints	Time Spent
Problem Average Graph	43%	43%	19%	24%	44%	44%	27%	100%		
Common Wrong Answers		我把鞋放到了床的旁边。,16% 我把鞋放到床的旁边了。,16%	我把书放进书包里了。,我把书放了在水杯的旁边。,32% +feedback 我把书放进书包里了。,我把书放了进书包里。,17% +feedback 我把书放进书包里了。,我把书放了进书包里。,我把书放了在水杯的旁边。,11% +feedback	你把钱找错了。,你把钱找得错。,37% +feedback 你把钱找错了。,28%	他今天得把中文作业做完。,他今天得把中文作业做。,30% +feedback 他今天得把中文作业做完。,21% 今天他得做完中文作业。, 他今天得把中文作业做。,17% +feedback	把你的功课都做完。,把你的功课都做。,34% +feedback 把你的功课都做完。,把一些功课都做完。,26% +feedback	先把书都看完，再出去玩。,53% 先把书都看了，再出去玩。,先看了都的书，再出去玩。,16% +feedback			
Correct Answer(s)		*我把鞋放到床的旁边了。，我把鞋放到了床的旁边。*	*我把书放进书包里了。*	*你把钱找错了。，你找错了钱。*	*他今天得把中文作业做完。，今天他得做完中文作业。*	*把你的功课都做完。*	*先把书都看完，再出去玩。，先把书都看了，再出去玩。*	*16，17，18，19，20，21，22，23，24，25，26，27，28，29，30*		
XXXXXXXX	29%	✓ 我把鞋放到床的旁边了。，我把鞋放到了床的旁边。 100%	× 我把书放进书包里了。,我把书放了在水杯的旁边。 0%	× 你把钱找错了。,你把钱找得错。 0%	× 他今天得把中文作业做完。 0%	× 把你的功课都做完。,把一些功课都做完。,你做完都功课。 0%	× 先把书都看了，再出去玩。,先看了都的书，再出去玩。 0%	✓ 20 100%	0	00:04:03
XXXXXXXX	57%	× 我把鞋放到床的旁边了。 0%	✓ 我把书放进书包里了。 100%	× 你把钱找错了。 0%	✓ 他今天得把中文作业做完。,今天他得做完中文作业。	✓ 把你的功课都做完。 100%	× 先把书都看完，再出去玩。 0%	✓ 20 100%	0	00:04:07

b. Teacher Interface



Data Record for PSABM23U - Logs prior to December 10, 2020

Dear Researcher,
Welcome to the data record for problem set PSABM23U. You have received this record based on your recent data request. Automated data analysis is featured below, offering a preliminary overview of your sample and a selection of analyses for your consideration. The latter portion of this report contains the raw data files from which you can conduct your own thorough analyses. When publishing your work, please reference this report as a stable location for readers to access your data for review and replication.

By clicking any link to download content from this page, you are agreeing to our Terms of Use.

Automated Data Analysis

Completion Rates
Students that have started PSABM23U : 42
Students that have completed PSABM23U : 41

Bias Assessment
Before analyzing learning outcomes, we suggest first assessing potential bias introduced by your experimental conditions (i.e., examine differential dropout). The table below reports the number of students that have completed PSABM23U, split out by experimental condition.

There is no data on completion rates for this study

c. Researcher Interface

Materials

Three sets of questions were created for a particular grammar topic – the first set for a pretest, the second set for a quiz assignment, and the third set for a posttest. Each set contained six target questions. A Chinese grammar topic, namely *ba* sentences, was chosen as the learning object. *ba* sentence was selected as the target topic because it is considered one of the most challenging grammar topics to learn in Chinese as a FL. The difficulty of the *ba* sentences reduced the possibility of hitting the ceiling effect, which refers to a large percentage of participants receiving

the full score or coming near this upper limit. Furthermore, it enabled more expansive possibilities for material development. Three *ba* sentence grammar rules were chosen, and six questions around each rule were developed for a total of 18 questions. Every six questions relating to one rule were assigned randomly to one of the three question sets. As a result, each question set contained six questions, with every two questions being related to one grammar rule. Then, the three question sets were assigned randomly to the three conditions. The order of questions was randomized by the platform in both the pre- and posttests. All of the questions used in the experiment can be found in the Appendix.

A 4-minute teaching video was created to introduce the three grammar rules of a *ba* sentence. The video introduced the grammar topic and listed all of the grammar rules that the students needed to learn with examples provided. A third instructor, unaffiliated with the two universities, recorded the video to standardize the classroom instruction.

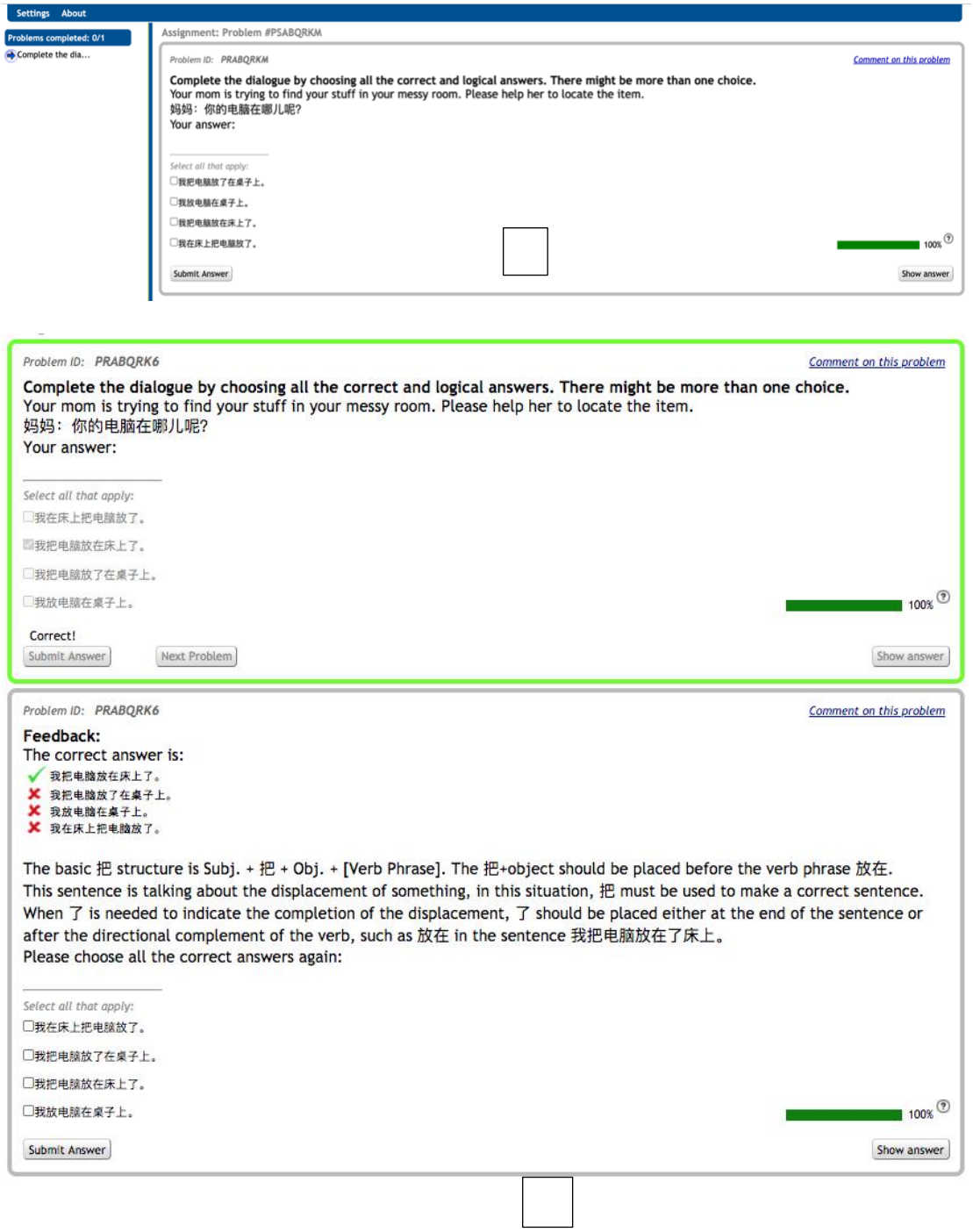
Figure 2 demonstrates the questions and different types of feedback that the students saw. The assignment questions were identical in all three conditions (Figure 2a), and all questions included a "choose all that apply" option for students to select potentially more than one correct answer. The students were required to select all correct answers for a given question for their response to be considered correct. The question answer choices were randomized by the platform. During the experiment, three questions of an unrelated grammar pattern preceded the quiz assignment, followed by three questions of another unrelated grammar pattern.

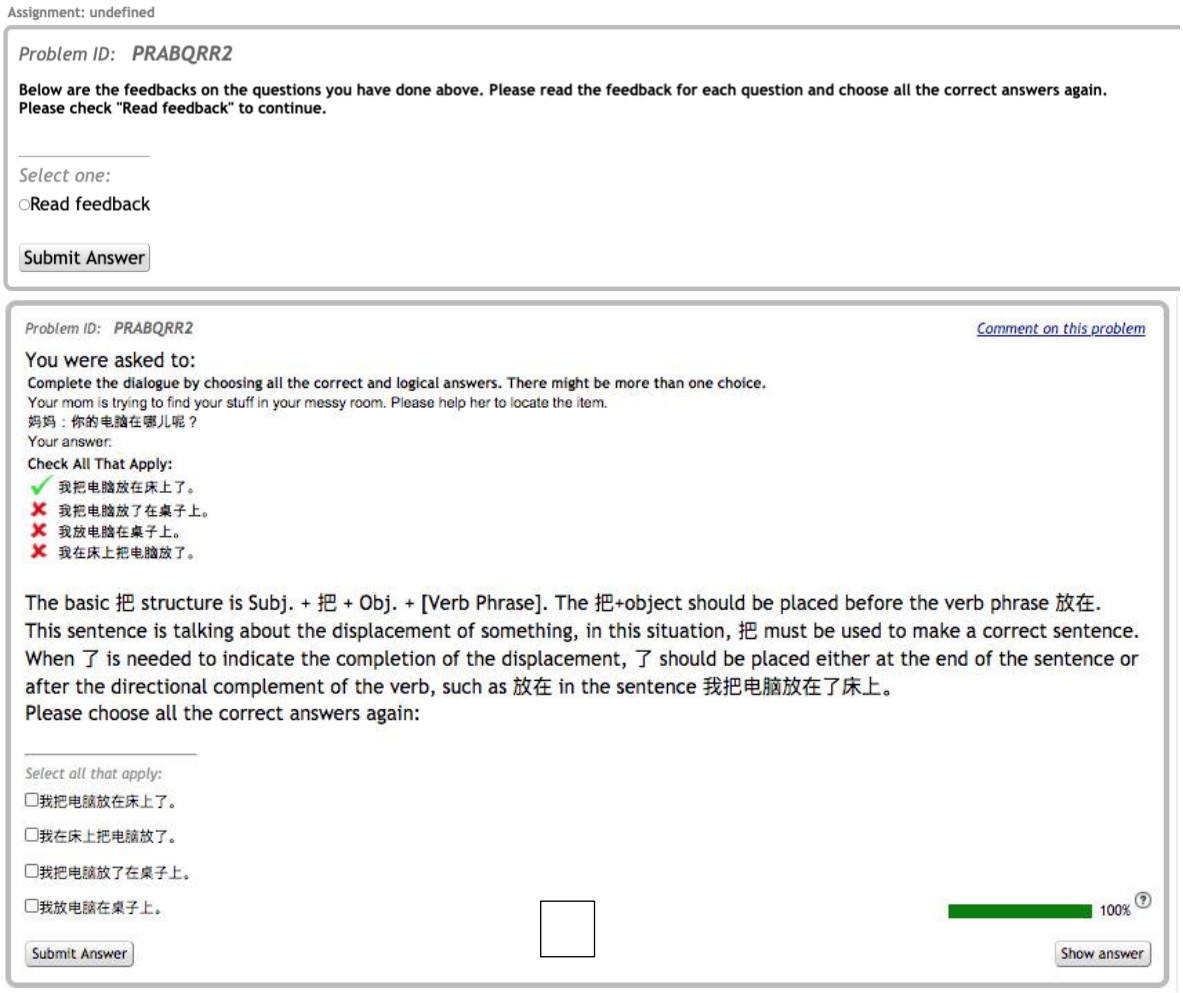
During the experimental process, the control group did not receive any feedback. The feedback in both the EQF and EAF conditions was identical in content and comprised the following two parts: (1) an indication of whether the students' answers were correct or incorrect

along with the correct answer; and (2) a recap of the grammar rules and an explanation of the reason for an answer being marked correct or incorrect (Figures 2b & 3c).

Figure 2

Feedback Examples





Note. a. Question with no feedback in Condition 0; b. question and EQF in Condition 1; and c. EAF in Condition 2.

Experimental Design

With respect to students' prior knowledge and learning, the present study employed a between-subject design with the following three conditions in the context of Chinese language learning:

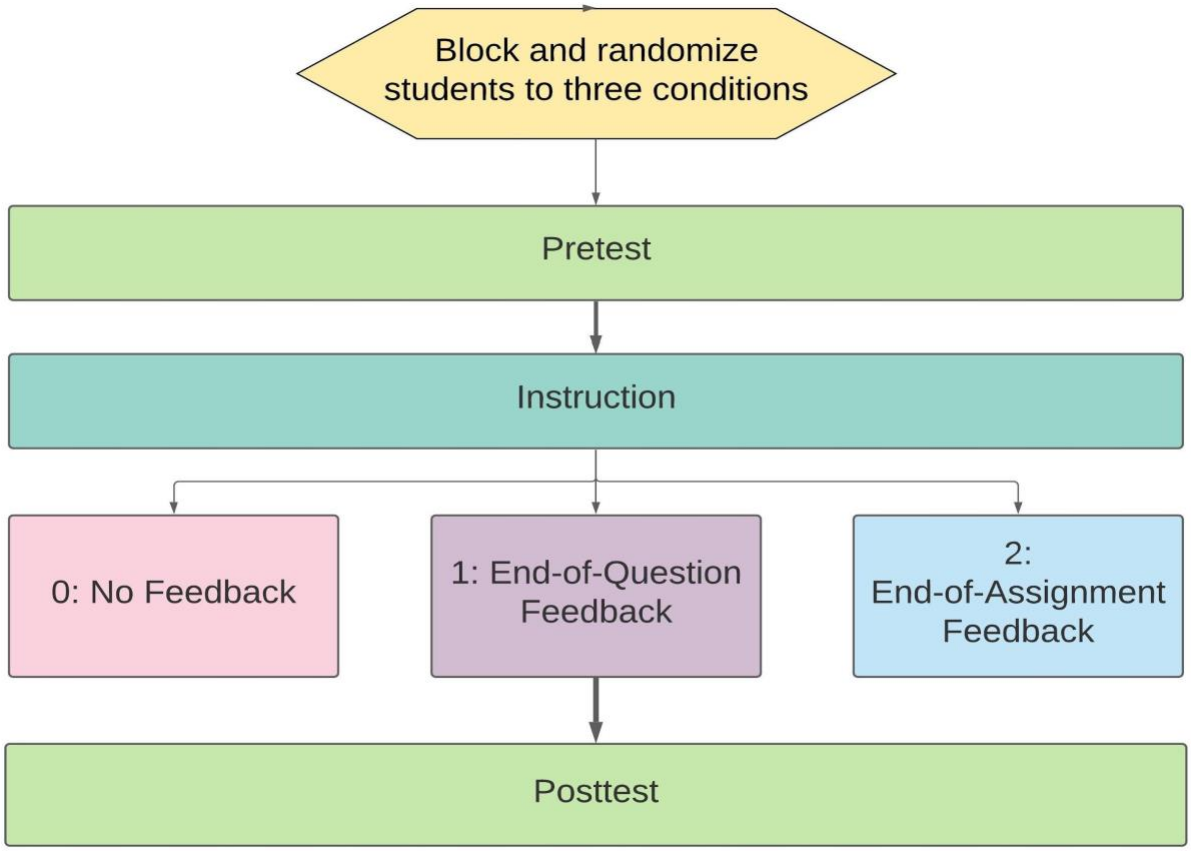
Condition 0: No feedback (NF) condition;

Condition 1: EQF condition with retries;

Condition 2: EAF at the end of the quiz assignment.

Figure 3

Flowchart of the Experimental Design



The independent variable was the condition, and the dependent variable was the posttest results. The pretest results were used as a covariate.

Procedure

The first author was responsible for implementing the procedure and monitoring the entire experiment in one university. The first author then trained an instructor at the other university using the flowchart in Figure 3. During the training, this additional instructor also performed exercises on ASSISTments to ensure that he could seamlessly gain access to the assignments and assign questions to all students accordingly.

This study followed a classic pretest–posttest design. Prior to the experiment, students in each university were ranked based on their temporal course grades and then grouped by trios. There were 27 trios in total. The purpose of blocking students was to distribute them evenly between conditions to minimize the baseline difference. Each trio was then assigned randomly to one of the three conditions, as indicated in Figure 3. Each condition had 27 students. Four trios that contained absent students were dropped.

Prior to the experiment, students took the pretest with NF and then watched an instructional video. The video was played only once and was controlled by the course instructor. On the day of the experiment, the students were asked to use a learning platform to complete all assigned questions, which differed according to their condition. They were told that they may or may not receive feedback. Students in Condition 0 did not receive feedback (until after the posttest and the experiment was complete). Students in Condition 1 received feedback immediately after answering a question. If they submitted incorrect answers, they were allowed to retry up to three times or to directly check the correct answer; furthermore, they could decline to retry the question and check the feedback. Lastly, students in Condition 2 did not receive any feedback until they had completed the entire assignment. After completing a question in Condition 2, students were automatically directed to the next question. Once all of their answers had been submitted, a report page was generated that presented all of the questions that they had completed, along with the same feedback for each question in Condition 1. Feedback was incentivized by asking students to select all of the correct answers again immediately after they received each piece of feedback to ensure that the students had reviewed the feedback (Figures 2b & 2c). Since the order of choices within each question was randomized in every feedback incentivization, it was necessary for the students to read each of the choices again to select all correct answers. Three days after the

assignment, the posttest was conducted. The pretest, video viewing, assignment, and posttest were all conducted in the classroom during class time. Students from both universities who participated in this study used the same materials and followed an identical experimental procedure.

Furthermore, students' interactions with all of the experimental materials were logged by ASSISTments, including their scores, start and end time, duration, attempts, and hint usage, for each question. Students' average scores on their in-class quizzes were also collected to represent their prior knowledge in addition to the data logged on ASSISTments.

Moreover, each participant was informed in detail about the aim and purpose of the study and was required to complete an informed consent form. The ASSISTments platform anonymized the research data prior to providing it to the researchers. To ensure that the study met ethical requirements, students in the NF condition (Condition 0) received feedback after the posttest to guarantee equal learning opportunities. Course instructors also went over the use of the target grammar in the classroom after the experiment to eliminate any remaining confusion related to the course material. The experiment was overseen by an Institutional Review Board team.

Data Analysis

First, descriptive statistics were computed related to participants' performance on the pre- and posttests. An ANOVA was then performed on the pretest scores of the three conditions to determine their homogeneity. To answer the research question, a planned contrast test that included a pretest as a covariate was conducted to investigate the impact of whether immediate feedback was provided as well as its timing.

Descriptive statistics were computed during the experiment, and the in-experiment data were analyzed to further explore how each type of immediate feedback impacted student learning. We first calculated the gain score (i.e., the difference between the posttest and pretest scores) for

each student before fitting a multilinear model to predict them based on prior knowledge and conditions. Then, a two-way ANOVA was run to determine whether conditions, prior knowledge, and the interaction of the main effects impacted students' learning gains. Lastly, a regression was employed to check the relationship between attempts used in the EQF condition and students' prior knowledge.

Preregistration

A time-stamped, independent, read-only registration of this manuscript and data is available at https://osf.io/h95pv/?view_only=0b8495850fdd43d6929f24ca775ca1be.

Results

Immediate Feedback Timing

Table 1 lists the means and standard deviations of the pre- and posttest scores of the three conditions.

Table 1

Means and Standard Deviations of Pretest and Posttest Scores of the Three Conditions (N = 69)

	Pretest			Posttest			
	n	*M*	*SD*	*n*	*M*	*M* adjusted	*SD*
Condition 0 (NF)	23	0.229	0.235	23	0.347	0.376	0.269
Condition 1 (EQF)	23	0.319	0.194	23	0.486	0.448	0.219
Condition 2 (EAF)	23	0.258	0.183	23	0.318	0.326	0.233

Before a planned test was run to determine the differences in the posttests between conditions, all assumptions were checked. (1) First, we checked that the linearity assumption was met and (2) that there was no significant interaction between the covariate and the grouping variable (homogeneity of regression slopes). (3) Then, since we had a small sample size, determining the distribution of the residuals was critical for choosing an appropriate statistical method. A Shapiro–Wilk test was performed, which did not reveal evidence of nonnormality ($W = 0.976$, $p = 0.214$). Based on this outcome, we were able to assume the normality of residuals and used a parametric test. (4) Next, Levene's test was conducted and the result was deemed nonsignificant ($F = 1.540$, $p = 0.222$); therefore, we were able to assume homogeneity of variance. (5) Subsequently, outliers were identified by examining standardized residuals and delineated as the number of standard errors away from the regression line. Observations greater than 3 with standardized residuals in absolute value were possible outliers. There were no outliers in the data, as determined by no cases having standardized residuals greater than 3 in absolute value. (6) Then, an ANOVA was performed on the pretest scores of the three conditions to check the conditions' homogeneity. The interaction between pretest scores and conditions was nonsignificant ($p = 0.649$), and homogeneity of the regression slopes existed. No difference was found in pretest scores between the three conditions (F [2, 63] = 2.214, $p = 0.118$, generalized eta squared = 0.066).

To answer the research question, a planned contrast test that included a pretest as a covariate was conducted to investigate the impact of whether immediate feedback was provided as well as the impact of feedback timing, (see Table 2). Condition 2's contrast results revealed that students who received EQF earned significantly higher grades in the posttest than those who received EAF (t [65] = −2.111, $p = 0.039$ [two-tailed], $r = 0.004$, 95% CI [−0.664, 0.508]), and the effect size was small. Condition 1's contrast results revealed that whether feedback was provided

did not affect the posttest results (t [65] = 0.228, $p > .05$, r = 0.002, 95% CI [−0.502, 0.516]), and the effect size was small. This effect was likely a result of the score difference between the two feedback conditions, since the adjusted mean of the EAF condition (adjusted M = 0.326) was lower than that of the NF condition (adjusted M = 0.376), while the adjusted mean of the EQF condition (adjusted M = 0.448) was higher than that of the NF condition. The adjusted posttest means of each condition are illustrated in Figure 4. The covariate pretest scores were significantly related to the posttest scores (F [1, 65] = 40.746, $p < .001$, r = 0.014), and the effect size was small. This result is reasonable as the students' pre- and posttest scores were naturally related.

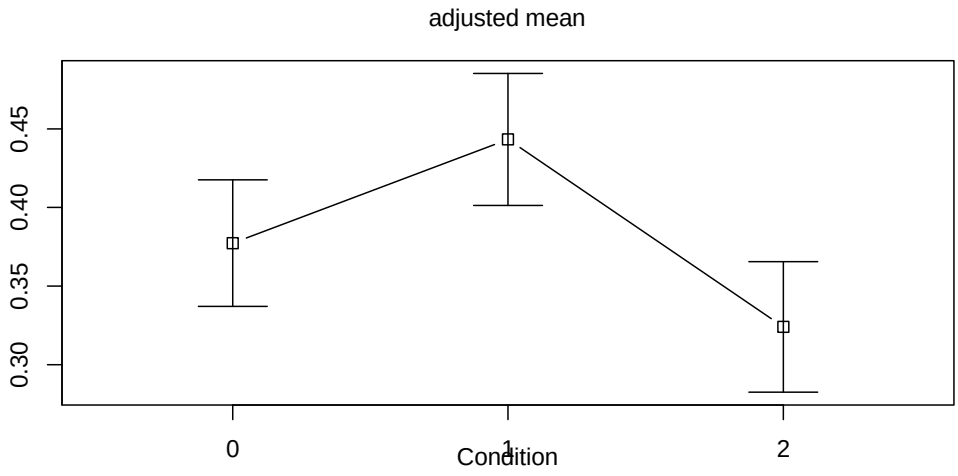
Table 2

Planned Contrast Results Using the Pretest as a Covariate (N = 65)

	Estimate	Std. Error	*t* value	Pr (>|*t*|)	*r*	95% CI
(Intercept)	0.186	0.039	4.817	9.1e-06 ***		
Pretest Average	0.735	0.115	6.383	2.1e-08 ***	0.014	
Condition 1 Contrast (NF = −2, EQF = −1, EAF =1)	0.004	0.016	0.228	0.8202	0.002	[−0.502, 0.516]
Condition 2 Contrast (NF = 0, EQF = −1, EAF =1)	−0.061	0.029	−2.111	0.0386 *	0.004	[−0.664, 0.508].

Figure 4

Adjusted Posttest Means of Each Condition (N = 69)



Note. 0 = NF condition; 1 = EQF condition; 2 = EAF condition.

The students answered six questions during the experiment, and their average time spent on each question was 53.5 seconds (*SD* = 37.9), 59.1 seconds (*SD* = 22.5), and 69.4 seconds (*SD* = 72.2) in the NF, EQF, and EAF conditions, respectively. Time spent on feedback was not considered. ANOVA and Tukey multiple comparisons revealed no differences between the conditions in the average time spent on problems.

Immediate Feedback Timing and Students' Learning Process

The in-experiment data were analyzed to further explore the interrelationship between each type of immediate feedback and its subsequent impact on students' learning. We first calculated the gain scores for each student before fitting a multilinear model to predict them based on prior knowledge and conditions. Students' prior knowledge was categorized into a binary variable with two levels (high and low) based on the average score of their prior knowledge. Each university

was categorized separately. The multilinear model indicated that students' prior knowledge did not accurately predict learning gains across conditions (adjusted $R^2 = -0.006$, F [1, 67] = 0.617, p = 0.435).

Then, a two-way ANOVA was run to determine whether conditions, prior knowledge, and the interaction of the main effects impacted students' learning gains. The results, presented in Table 3, indicated that students' prior knowledge did not have a significant effect on their learning gains (F [63, 1] = 1.396, p = 0.242, η_p^2 = 0.018), and the effect size was small. The interaction of students' prior knowledge and condition also did not have a significant effect on their learning gains (F [63, 2] = 1.109, p = 0.336, η_p^2 = 0.034), and the effect size was small. The results indicated that feedback was equally beneficial for lower and higher prior knowledge students.

Table 3

Two-Way ANOVA with Conditions and Prior Knowledge as Independent Variables and Learning Gains as the Dependent Variable

	df	Sum Sq	Mean Sq	*F*	*p*	η_p^2
Prior Knowledge	1	0.055	0.055	1.396	0.242	0.018
Condition	2	0.118	0.059	1.504	0.230	0.230
Prior Knowledge *Condition	2	0.087	0.044	1.109	0.336	0.034
Residuals	63	2.479	0.039			

Nevertheless, when analyzing the learning process and verifying the use of attempts in the EQF condition through a regression, we found that instead of checking the correct answer without further retries, students tended to use the retry function to learn when their first answer was incorrect; the average number of attempts used in the EQF condition was 2.469. Furthermore, the regression revealed that students' prior knowledge strongly predicted their attempts ($\beta = -1.947$, $p = 0.003$, adjusted $R^2 = 0.288$), and the effect size was small to medium. Notably, students with lower prior knowledge used more attempts.

Discussion

End-of-Question Feedback Is More Efficient

This study investigates the effects of two different time points for providing immediate feedback to FL learners while they complete online language assignments. With respect to students' prior knowledge and learning, this study aims to examine the possible impact on students' learning by providing EQF and EAF. The results of planned contrast tests indicate that EQF is more efficient at improving learning outcomes than EAF; however, the EAF condition does not outperform the NF condition.

A plausible explanation for EQF having a more significant impact on students' learning and retention could be cognitive load theory (Sweller, 2011), which posits that the capacity of working memory is limited. This theory suggests that providing detailed feedback in manageable amounts can prevent the information provided through feedback being disregarded due to information overload. This finding is also consistent with the results of Opitz et al. (2011), who investigated the role of visual immediate and delayed feedback in an artificial grammar learning task. They suggested that EQF is more effective as it makes the task requirements less demanding and hence leads to higher perceptual gain.

Overall, the present study suggests that learning can be enhanced through the provision of EQF. When students receive feedback immediately following the completion of a question, they can assess the accuracy of their grammar understanding. This practice proves effective in helping students consolidate knowledge and identify any misunderstandings. EQF has also been discussed as beneficial for students' ability to apply feedback to future practice (Hatziapostolou & Paraskakis, 2010). Moreover, Lee et al. (2021) believed that EQF also increases students' confidence and motivation to learn.

In addition, the present study reveals that the adjusted mean of the EAF condition is lower than that of the NF condition. Since providing feedback is generally considered a positive teaching method compared with NF, this finding is beyond our expectation. One possible explanation can be found in the theory of errorless learning (e.g., Clare & Jones, 2008), which suggests that by providing learners with immediate feedback and minimizing errors, errorless learning can reduce the negative emotional experiences that may be associated with learning. If feedback is not provided immediately after a response, the students' errors might be consolidated, making them more difficult to correct in the future. Another potential explanation could be attributed to the one-time test design, which may have influenced the lower adjusted mean of the EAF condition. Mullet et al. (2014) proposed that providing feedback with a delay promotes long-term retention. However, the sole posttest in our study was administered three days following the assignment, limiting our ability to assess longer-term learning outcomes.

More Attempts improve Learning of Students with Lower Prior Knowledge

According to Fyfe et al. (2012), prior knowledge is often considered a predictor of learning from feedback. The current study does not support the two immediate feedback conditions being more beneficial for students with lower prior knowledge. Nevertheless, upon further analysis of

the data in the EQF condition, a strong correlation is found between students' prior knowledge and retry attempts. Moreover, students with lower prior knowledge tend to use more attempts, indicating that the retry function may be a key resource for the success of students with limited prior knowledge. With more retries, students with lower prior knowledge can learn as much as students with higher prior knowledge.

Although the present study supports the notion that allowing students to retry may be beneficial, especially for those with lower prior knowledge, some researchers (Zhang et al., 2021; Mullet et al., 2014) have expressed concerns that within an experimental design, permitting students varying degrees of freedom for analyzing feedback, such as retries, may confound the results. However, we argue that retrying should be considered an integral part of the immediate feedback process. Students should still have a certain degree of freedom in terms of processing their feedback; for instance, if they receive EQF immediately following each question, then they should be allowed to retry before the answers are disclosed. This rationale is supported if one considers a retry function to be part of EQF, which is similar to representing problems prior to EAF as an essential part of the experimental process. When presenting feedback at the end of an assignment, it is necessary to re-illustrate the questions, as depicted in Figure 2c. Each feedback method possesses its own unique features, and we believe that all of these characteristics take effect when considered holistically. We do not propose eliminating certain features simply because they do not appear to be on par with the other condition. Since the intent of feedback is to encourage students to correct their mistakes and to motivate them to explore and learn more, it seems unreasonable to remove the retry function from the EQF condition merely because it encourages more curiosity and potentially more time spent on feedback assessment.

Limitations and Recommendations

A significant limitation of this study is that it did not investigate the long-term retention of materials. A possible explanation for why some studies do not find conclusive variations between EQF and EAF is that they do not provide enough time for variations between the two conditions to appear. Therefore, future studies should consider including a long-term condition or a retest after a considerable amount of time has passed. A larger sample size from more schools could also help to increase the validity of the results.

The low posttest scores observed in the EAF condition could potentially be attributed to the limitations of the one-time test design. To enhance the validity and accuracy of the results, more tests should be implemented. It is also recommended to incorporate additional posttests, particularly long-term posttests. This methodological adjustment would provide a more comprehensive assessment and allow for a more robust evaluation of the effects of the EAF condition on learning outcomes over an extended period.

In addition, future work should expand this experiment to other learning domains to further test its validity. Since individuals may debate whether to include a retry in the EQF condition, it would be appropriate to perform an investigation that focuses on this variable. Further exploration is required regarding the psychological reasons behind the results.

As for the generalizability of the results, the following points should be noted: First, as the study was conducted in two private universities, generalizing its findings to young learners may not be appropriate. Further investigations are required to examine how the two types of immediate feedback take effect in younger age groups while using educational computer systems. Second, although our results are supported by part of the literature from other learning fields (Attali & van der Kleij, 2017; Dihoff et al., 2004), this study was conducted in FL classrooms. Therefore, one should be cautious when generalizing its results to other fields. Finally, since the current study

focused on a classroom study, whether the effects of the two immediate types of feedback would be the same in laboratories should be further examined. Chen et al. (2018) and Zhu et al. (2020) have highlighted that delayed feedback appears to be more effective in laboratory studies, while applied studies in the classroom usually favor immediate feedback. This is potentially because immediate feedback reduces the preservation of initial misconceptions. Furthermore, as the current study picked only one language as the research object, we hope to expand the research to other language classrooms in the near future. Various experimental results in the field of language learning could benefit from the study of immediate feedback timing to further inform FL learning.

Conclusions and Implications for Practice

Feedback has been considered fundamental to students' learning as well as an essential learning tool (Jensen et al., 2021). This study underscored the importance of differentiating two types of immediate feedback – EQF and EAF – and explored the impact of various types of immediate feedback. The results support that, in the context of designing online grammar learning assignments, EQF outperforms EAF, and also that educators should provide students with choices to resubmit answers. Rather than focusing on the development of novel educational technologies, this study aimed to serve as a verification of learning outcomes, which will guide learners, educators, and platform designers in exploring the best options for enhancing learning.

This study encourages a discussion related to a more clearly defined classification of feedback timing while using educational computer systems. A univocal definition of immediate feedback in computer-based assessments should be proposed to facilitate future studies. Not only is this critical for ensuring the validity and reliability of future research but it can also serve as a point of departure for educators to improve their online teaching implementation and design. Lastly, this study underscores the interconnections between immediate feedback and other related

factors, such as retries. Additional considerations include the impact of retries on student learning and the measurement of immediate feedback's impact on retention.

Acknowledgements

We acknowledge funding from NSF (2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, R305A120125 & R305R220012), GAANN (P200A180088 & P200A150306), EIR (U411B190024 S411B210024, & S411B220024), ONR (N00014-18-1-2768) and NHI (via a SBIR R44GM146483).

References

Attali, Y., & Powers, D. (2010). Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement, 70* (1), 22-35. https://doi.org/10.1177/0013164409332231

Attali, Y., & van der Kleij, F. (2017). Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers & Education, 110,* 154-169. https://doi.org/10.1016/j.compedu.2017.03.012

Bahari, A. (2021). Computer-mediated feedback for L2 learners: Challenges versus affordances. *Journal of Computer Assisted Learning, 37*(1), 24-38. https://doi.org/10.1111/jcal.12481

Barrot, J. S. (2021). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*, 1-24. https://doi.org/10.1080/09588221.2021.1936071

Blake, R. J. (2011). Current trends in online language learning. *Annual Review of Applied Linguistics*, *31*, 1–17. https://doi.org/10.1017/S026719051100002X

Bjork, R. A. (2018). Being suspicious of the sense of ease and undeterred by the sense of difficulty: Looking back at Schmidt and Bjork (1992). *Perspectives on Psychological Science, 13*(2), 146–148. https://doi.org/10.1177/1745691617690642

Budiman, A. (2017). Behaviorism and foreign language teaching methodology. *English Franca: Academic Journal of English Language and Education*, *1*(2), 101-114. http://dx.doi.org/10.29240/ef.v1i2.171

Butler, A. C., & Woodward, N. R. (2018). Toward consilience in the use of task-level feedback to promote learning. *Psychology of Learning and Motivation - Advances in Research and Theory*, *69*, 1–38. https://doi.org/10.1016/bs.plm.2018.09.001

Chen, X., Breslow, L., & DeBoer, J. (2018). Analyzing productive learning behaviors for students using immediate corrective feedback in a blended learning environment. *Computers & Education, 117,* 59–74. https://doi.org/10.1016/j.compedu.2017.09.013

Clare, L., & Jones, R. S. (2008). Errorless learning in the rehabilitation of memory impairment: A critical review. *Neuropsychology review, 18,* 1-23. https://doi.org/10.1007/s11065-008-9051-4

Culpepper, S. (2014). If at first you don't succeed, try, try again: Applications of sequential IRT models to cognitive assessments. *Applied Psychological Measurement, 38* (8), 632-644. https://doi.org/10.1177/0146621614536464

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education, 162.* https://doi.org/10.1016/j.compedu.2020.104094

Dihoff, R. E., Brosvic, G. M., & Epstein, M. L. (2004). Provision of feedback during preparation for academic testing: Learning is enhanced by immediate but not delayed feedback. *Psychological Record, 54*(2), 207–231. http://dx.doi.org/10.1007/BF03395471

Escueta, M., Quan, V., Nickow, A., & Oreopoulos, P. (2017, August). *Education Technology: An Evidence-Based Review* (NBER Working Paper No. 23744). http://dx.doi.org/10.3386/w23744

Fyfe, E. R., de Leeuw, J. R., Carvalho, P. F., Goldstone, R. L., Sherman, J., Admiraal, D., Alford, L. K., Bonner, A., Brassil, C. E., Brooks, C. A., Carbonetto, T., Chang, S. H., Cruz, L., Czymoniewicz-Klippel, M., Daniel, F., Driessen, M., Habashy, N., Hanson-Bradley, C. L., Hirt, E. R., … Motz, B. A. (2021). ManyClasses 1: Assessing the

generalizable effect of immediate feedback versus delayed feedback across many college classes. *Advances in Methods and Practices in Psychological Science, 4*(3). https://doi.org/10.1177/25152459211027575

Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology*, *104*(4), 1094–1108. https://doi.org/10.1037/a0028389

Gass, S. M., & Mackey, A. (2014). Input, interaction, and output in second language acquisition. In B. VanPatten, & J. Williams (Eds.), *Theories in Second Language Acquisition* (2nd Ed., pp. 194-220). Routledge. https://doi.org/10.4324/9780203628942

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hatziapostolou, T., & Paraskakis, I. (2010). Enhancing the impact of formative feedback on student learning through an online feedback system. *Electronic Journal of E-Learning, 8*(2), 111. https://eric.ed.gov/?id=EJ895699

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497. https://doi.org/10.1007/s40593-014-0024-x

Henderson, C. (2021). The effect of feedback timing on L2 Spanish vocabulary acquisition in synchronous computer-mediated communication. *Language Teaching Research, 25*(2), 185–208. https://doi.org/10.1177/1362168819832907

Henshaw, F. G. (2011). Effects of feedback timing in SLA: A computer assisted study on the Spanish subjunctive. In C. Sanz, & R. P. Leow (Eds.), *Implicit and Explicit Language*

Learning: Conditions, Processes, and Knowledge in SLA and Bilingualism (pp. 85-100). Georgetown University Press. https://www.jstor.org/stable/j.ctt2tt7k0.12

Jensen, L. X., Bearman, M., & Boud, D. (2021). Understanding feedback in online learning – A critical review and metaphor analysis. *Computers & Education, 173*. https://doi.org/10.1016/j.compedu.2021.104271

Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The Modern Language Journal (Boulder, Colo.), 99*(1), 1–18. https://doi.org/10.1111/modl.12189

Kannan, J. & Munday, P. (2018). New trends in second language learning and teaching through the lens of ICT, networked learning, and artificial intelligence. *Círculo de Lingüística Aplicada a la Comunicación, 76,* 13-30. http://dx.doi.org/10.5209/CLAC.62495

Kılıçkaya, F. (2022). Pre-service language teachers' online written corrective feedback preferences and timing of feedback in computer-supported L2 grammar instruction. *Computer Assisted Language Learning, 35*(1-2), 62-87. https://doi.org/10.1080/09588221.2019.1668811

Lahcen, B., & Mahmoud, B. (2022). Perspectives on error and written corrective feedback in second language acquisition and composition studies. *International Journal of English Literature and Social Sciences, 7* (2). https://dx.doi.org/10.22161/ijels

Lee, S., Choi, Y. I., & Kim, S. W. (2021). Roles of emotions induced by immediate feedback in a physics problem-solving activity. *International Journal of Science Education, 43*(10), 1525-1553. https://doi.org/10.1080/09500693.2021.1922778

Lefevre, D., & Cox, B. (2017). Delayed instructional feedback may be more effective, but is this contrary to learners' preferences? *British Journal of Educational Technology, 48*(6),

1357–1367. https://doi.org/10.1111/bjet.12495

Li, S. (2022). Oral corrective feedback. In H. Mohebbi & C. Coombe (Eds.), *Research Questions in Language Education and Applied Linguistics: A Reference Guide* (pp. 353-358). Springer Texts in Education.

Lipnevich, A. A., & Panadero, E. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education, 6*, Article 720195, https://www.frontiersin.org/articles/10.3389/feduc.2021.720195/full

Liu, Q., Geertshuis, S., & Grainger, R. (2020). Understanding academics' adoption of learning technologies: A systematic review. *Computers & Education, 151*. https://doi.org/10.1016/j.compedu.2020.103857

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition 37*, 1077–1087. https://doi.org/10.3758/MC.37.8.1077

Mullet, H. G., Butler, A. C., Verdin, B., von Borries, R., & Marsh, E. J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition, 3*(3), 222–229. https://doi.org/10.1016/j.jarmac.2014.05.001

Nakata, T. (2015). Effects of feedback timing on second language vocabulary learning: Does delaying feedback increase learning? *Language Teaching Research, 19*(4), 416-434. https://doi.org/10.1177/1362168814541721

Opitz, B., Ferdinand, N. K., & Mecklinger, A. (2011). Timing matters: The impact of immediate and delayed feedback on artificial language learning. *Frontiers in Human Neuroscience, 5*, Article 8. https://doi.org/10.3389/fnhum.2011.00008

Ostrow, K. S., & Heffernan, N. T. (2014). Testing the multimedia principle in the real world: A comparison of video vs. text feedback in authentic middle School math assignments. In Stamper, J., Pardos, Z., Mavrikis, M., McLaren, B.M. (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 296-299). London, United Kingdom.

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, *35*, 100416. https://doi.org/10.1016/j.edurev.2021.100416

Qi, D., Rajab, A., Haladin, N. B., Wang, W., & Fu, X. (2020). The effect of immediate and delayed feedback on the achievement of Chinese EFL learners on reading comprehension. *European Journal of Molecular & Clinical Medicine, 7*(6), 476-491.

Rassaei, E. (2019). Computer-mediated text-based and audio-based corrective feedback, perceptual style and L2 development. *System*, *82*, 97–110. https://doi.org/10.1016/j.system.2019.03.004

Salajegheh, S., Khomeijani Farahani, A. A., & Shahabi, H. (2022). The role of explicit corrective feedback timing in second language structure accuracy. *Journal of Language, Culture, and Translation, 4*(2), 1-21. https://doi.org/10.30495/LCT.2022.1947452.1050

Serfaty, J., & Serrano, R. (2022). Lag effects in grammar learning: A desirable difficulties perspective. *Applied Psycholinguistics*, *43*(3), 513-550. https://doi.org/10.1017/S0142716421000631

Shintani, N., & Aubrey, S. (2016). The effectiveness of synchronous and asynchronous written corrective feedback on grammatical accuracy in a computer-mediated environment. *The Modern Language Journal, 100*(1), 296-319. https://doi.org/10.1111/modl.12317

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Sweller, J. (2011). Cognitive load theory. In J. P. Mestre, & B. H. Ross (Eds.), *Psychology of learning and motivation* (55, pp. 37-76). Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Timmers, C., & Veldkamp, B. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education, 56*(3), 923-930. https://doi.org/10.1016/j.compedu.2010.11.007

van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education, 58* (1), 263-272. https://doi.org/10.1016/j.compedu.2011.07.020

van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: a meta-analysis. *Review of Educational Research, 85*, 475-511. https://doi.org/10.3102/0034654314564881

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.03087

Zhang, S., Bergner, Y., DiTrapani, J., & Jeon, M. (2021). Modeling the interaction between resilience and ability in assessments with allowances for multiple attempts. *Computers in Human Behavior, 122*. https://doi.org/10.1016/j.chb.2021.106847

Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers &*

Education, 143. https://doi.org/10.1016/j.compedu.2019.103668.

Appendix

Experiment Materials: pretest, assignment, and posttest

Pretest

Complete each dialogue by choosing ALL correct and logical answers. There might be more than one choice.
(P)G1-3. Your mom is trying to find your stuff in your messy room. Please help her to locate the item.
妈妈：你的水杯在哪儿呢？
Your answer:
@@ 我把水杯放在了书的旁边。
我把水杯放了书的旁边。
我把水杯放了在书的旁边。
我在书的旁边把水杯放了。
(P)G1-5. Your mom is trying to find your stuff in your messy room. Please help her to locate the item.
妈妈：你的篮球在哪儿呢？
Your answer:
@@ 我把篮球放到桌子下了。
我把篮球放了在桌子下。
我放篮球在桌子下了。
我在桌子下把篮球放了。
(P)G2-3. You have a stomachache and go to see a doctor.
医生：你怎么了？
Your answer:
@@ 我吃西瓜把肚子吃坏了。
@@ 我吃西瓜吃坏了肚子。
我把西瓜吃坏了肚子。
我吃西瓜把肚子坏了。
(P)G2-4. You are talking about your cat with a friend.
朋友：你的猫为什么这么胖?
Your answer:
@@ 它常常把我家狗的饭也吃了。
@@ 它常常吃我家狗的饭。
它常常把我家狗也吃饭了。
它常常吃把我家狗的饭。
(P)G3-1. Your mom is very strict with you. Please select what your mom will say in the following situation.
You: 我吃饱了，你帮我吃吧。
Your mom:
@@: 把你的饭都吃完。
把你的饭都吃。
吃完都你的饭。
都把你的饭吃完。
(P)G3-4. Your mom is very strict with you. Please select what your mom will say in the following situation.
You: 我怎么收拾(clear up)房间？
Your mom:
@@: 先把房间里的东西都放回去。
先都把房间里的东西放回去。
都先把房间里的东西放回去。
先都房间里的东西放回去。

Assignment

Complete the dialogue by choosing all the correct and logical answers. There might be more than one choice.
Your mom is trying to find your stuff in your messy room. Please help her to locate the item.
(E)G1-1. 妈妈：你的电脑在哪儿呢？
Your answer:
@@ 我把电脑放在床上了。
我把电脑放了在桌子上。
我放电脑在桌子上。
我在床上把电脑放了。
Feedback:
The correct answer is:
我把电脑放在床上了。
The basic 把 structure is Subj. + 把 + Obj. + [Verb Phrase]. The 把+object should be placed before the verb phrase 放在.
This sentence is talking about the displacement of something, in this situation, 把 must be used to make a correct sentence.

When 了 is needed to indicate the completion of the displacement, 了 should be placed either at the end of the sentence or after the directional complement of the verb, such as 放在 in the sentence 我把电脑放在了床上。

Please choose all the correct answers again:

Your mom is trying to find your stuff in your messy room. Please help her to locate the item.
(E)G1-6. 妈妈：你的手机在哪儿呢？
Your answer:
@@ 我把手机放到电视前面了。
@@ 我把手机放到了电视前面。
我把手机放了在电视前面。
我放手机在电视前面了。
Feedback:
The correct answers are:
我把手机放到电视前面了。
我把手机放到了电视前面

This sentence is talking about the displacement of something, in this situation, 把 must be used to make a correct sentence. When 了 is needed to indicate the completion of the displacement, 了 should be placed either at the end of the sentence or after the directional complement of the verb, such as 放到 in this sentence. So both 放到电视前面了 and 放到了电视前面 are correct.

Please choose all the correct answers again:

(E)G2-1. It's not your day. Everything goes wrong.
朋友：你的女朋友为什么不高兴？
Your answer:
@@ 我看电影的时候把她的电脑弄坏了。
@@ 看电影的时候我弄坏了她的电脑。
我把她的电脑看电影看坏了。
我看电影坏了她的电脑。
Feedback:
The correct answers are:
我看电影的时候把她的电脑弄坏了。
看电影的时候我弄坏了她的电脑。

Resultative complements work well with 把 structure, but 把 structure is not required in this situation. With 把 structure, the sentence focuses on the result or influence of action on the object.

The verb in 把 sentence is not just "bare"; there's "more stuff" after it. Often the "stuff" is related to some kind of manipulation of the object, such as a resultative complement 坏 in this sentence, a phrase to describe the degree of the action, or 了.

Please choose all the correct answers again:

(E)G2-5. Your friend is waiting for Wang Peng.
朋友：王朋为什么不打球？
Your answer:
@@ 他昨天打球把球拍(racket)打坏了。
@@ 他昨天打球打坏了球拍(racket)。
他昨天打球把球拍(racket)坏了。
他把打球打坏了球拍(racket)。
Feedback:
The correct answers are:
他昨天打球把球拍(racket)打坏了。
他昨天打球打坏了球拍(racket)。
昨天打球 is a topic, this sentence still needs a main verb which happens to be 打. So 打 cannot be dropped.

Resultative complements work well with 把 structure, but 把 structure is not required in this situation. With 把 structure, the sentence focuses on the result or influence of action on the object.

Please choose all the correct answers again:

(E)G3-3. Your mom is very strict with you. Please select what your mom will say in the following situation.
You: 老师要我们看的书太多了，我能不能只看一些？
Your mom:
@@: 把老师要你们看的书都看完。
@@把老师要你们看的书都看了。
都把老师要你们看的书看完。
都老师要你们看的书看了。
Feedback:
The correct answers are:
把老师要你们看的书都看完。
把老师要你们看的书都看了。
When 都 is used in front of the verb to indicate that "all" the object should follow the action, 把 structure must be used.

The verb in 把 sentence is not just "bare"; there's "more stuff" after it. Often the "stuff" is related to some kind of manipulation of the object, such as a resultative complement 完 in this sentence, a phrase to describe the degree of the action, or 了.

Please choose all the correct answers again:

(E)G3-6. Your mom is very strict with you. Please select what your mom will say in the following situation.
You: 我不喜欢洗衣服。
Your mom:
@@: 把你的衣服都洗完。
洗完都你的衣服。
都洗完你的衣服。
把都你的衣服洗完。
Feedback:
The correct answer is:
把你的衣服都洗完。

When 都 is used in front of the verb to indicate that “all” the object should follow the action, 把 structure must be used. 都 should be placed right before the verb.

Please choose all the correct answers again:

Posttest

(Post)G1-2. 妈妈：你的鞋在哪儿呢？
Your answer:
@@ 我把鞋放到床的旁边了。
@@ 我把鞋放到了床的旁边。
我把鞋放了到桌子上。
我在桌子的旁边把鞋放了。

(Post)G1-4. 妈妈：你的书在哪儿呢？
Your answer:
@@ 我把书放进书包里了。
我把书放了进书包里。
我把书放了在水杯的旁边。
我在电脑的旁边把书放了。
(Post)G2-2. 服务员：你买完东西怎么又回来了？
Your answer:
@@ 你把钱找错了。
@@ 你找错了钱。
你把钱找了。
你把钱找得错。
(Post)G2-6. 朋友：高晓明为什么没有来？
Your answer:
@@ 他今天得把中文作业做完。
@@ 今天他得做完中文作业。
他今天得把中文作业做。
今天他得完中文作业。
(Post)G3-2. You: 功课太多了，我想先去睡觉。
Your mom:
@@: 把你的功课都做完。
把你的功课都做。
把一些功课都做完。
你做完都功课。
(Post)G3-5. You: 我能不能先出去玩，再看书？
Your mom:
@@: 先把书都看完，再出去玩。
@@先把书都看了，再出去玩。
先把都的书看了，再出去玩。
先看了都的书，再出去玩。

Chapter 7: The Effect of Immediate Feedback on Semi-Open-Ended Questions in Online Foreign Language Learning

Lu, X., Heffernan, N. T.

Abstract

Semi-open-ended (SOE) questions are commonly utilized in online foreign language assignments; however, the effects of online feedback on SOE questions have received limited attention in research. To address this gap, the current study employed an instructor survey and conducted an in-class experiment to investigate the effects of online feedback on SOE questions. The survey findings revealed that although most instructors had access to online platforms enabling immediate feedback, only a small proportion of instructors utilized this function to provide immediate feedback to SOE questions. Furthermore, the majority of instructors expressed the belief that providing both correct responses and elaborated feedback would be the most effective approach towards learning. Contrary to the findings of the instructor survey, our experiment indicated that the impact of feedback on learning outcomes for SOE questions cannot be solely attributed to feedback complexity. An in-class experiment was conducted using the ASSISTments platform to examine the influence of various online feedback on students' performance in responding to SOE questions, as well as to investigate the impact of feedback on students' judgments of learning. It was demonstrated that providing only correct responses for SOE questions resulted in the highest learning gains during the second-day posttest, while providing elaborated feedback without correct responses yielded the lowest learning gain. Furthermore, the results supported the notion that providing correct responses necessitated the least amount of time for feedback and enhanced students' judgment of learning.

Keywords: semi-open-ended questions, online foreign language learning, immediate feedback, judgment of learning

1. Introduction

Given the ongoing digital transformation, computer-based automated feedback has become readily available across diverse educational contexts, and its positive effects on learning outcomes have been widely recognized (Swart et al., 2019). A substantial body of research (e.g., (Ma et al., 2014) has demonstrated the effectiveness of immediate computer-based feedback in promoting learning. Among the current available research on immediate computer-based feedback, most focuses on close-ended questions in the STEM field, there is a paucity of studies on immediate feedback for semi-open-ended (SOE) questions in the realm of online learning as well as in the field of foreign language (FL) learning.

The scarcity of research in this area can be attributed to the variability of correct answers to SOE questions in terms of sentence form and word usage. Instructors are faced with the challenge: either acquiring expertise in natural language processing and machine learning techniques to design assignments capable of autonomously providing immediate feedback to SOE questions (Zhang et al., 2022), or resorting to pre-designed assignment questions within platforms tailored to specific textbooks.

At the same time, SOE questions are commonly employed in FL assignments and are used more frequently than multiple-choice questions or questions with concise answers. However, grading these assignments subjectively is a time-consuming task. As a result of the grading challenges associated with SOE questions, instructors may be hesitant to utilize them. Nevertheless, the provision of immediate feedback during online learning should not be restricted solely to question types with single correct answers. The present project aims to explore the effects of immediate feedback beyond close-ended questions, with a specific focus

on SOE questions in the context of FL learning. This research sheds light on the need to alleviate the workload of instructors while ensuring timely and consistent feedback for students.

2. Literature Review

A comprehensive literature review on the complexity of feedback in student learning and the effects of feedback on the judgment of learning is presented, along with the selection of four feedback types for the present study.

2.1 Online Immediate Feedback Types and Complexity

The provision of immediate feedback via online platforms for student learning holds immense importance in today's digital era. With the increasing prevalence of online education and remote learning, the availability of immediate feedback has become even more crucial. Immediate feedback online allows students to receive timely guidance about their performance, comprehension, and progress, enabling them to correct errors, reinforce their understanding, and adjust their learning strategies accordingly (Epstein et al., 2002).

Shute (2008) summarized six types of feedback: no feedback (NF), verification (also called knowledge of results), correct response (CR), try again, error flagging, and elaborated feedback (EF). Swart et al. (2019) provided a summary of previous studies on feedback types and emphasized the categorization of feedback into two sub-categories, namely verification and elaboration, based on an information processing perspective. In our study, verification is utilized to refer to feedback that entails confirming the accuracy or correctness of a task. Feedback types that do not necessitate verification of students' responses have been employed to provide immediate feedback to SOE questions, assuming that instructors do not have access to online auto-grading engines. When considering the absence of verification, the feedback types commonly utilized for SOE questions, as outlined by Shute (2008), included NF, CR, and EF

addressing the target concept. These feedback types were arranged in ascending order of information complexity, ranging from the simplest to the most complex.

Different types of feedback have varying complexities and effects on learning, as supported by previous research (Enders et al., 2021). Existing studies have generally supported the notion that immediate feedback of different complexities enhances error correction compared to the absence of feedback (Kuklick et al., 2023). A large number of studies have supported the idea that the effectiveness of feedback increases with its complexity (Van der Kleij et al., 2015). Wisniewski et al. (2020) ran a meta-analysis on educational feedback research and found that the greater the amount of information contained in feedback, the more effective it tended to be. Swart et al. (2019) found that elaboration feedback outperformed verification feedback in learning from text. Enders et al. (2021) conducted a study and found that elaborated feedback yielded superior results compared to verification plus correct response in an assessment involving close-ended questions, with Kuklick et al. (2023) further supporting this finding. However, a body of research has yielded findings contrary to the commonly held belief that more elaborate feedback is advantageous for learning (Golke et al., 2015). Kulhavy et al. (1985) demonstrated that feedback versions with higher complexity have a minor impact on students' self-correction of errors. Conversely, feedback with lower complexity offers greater advantages to learners in terms of efficiency and outcomes compared to complex feedback. Ruth et al. (2021) compared the effects of verification plus correct response feedback and elaborated feedback on learning scores with mobile quiz apps but did not find a difference. Kuklick et al. (2023) found that although all feedback enhanced learning compared to no feedback, EF did not outperform CR. Jaehnig and Miller (2007) observed that students might necessitate additional time to process more complex messages, potentially diminishing the efficiency of elaborated

feedback. While a longer duration of time spent can serve as a positive indicator of motivational attentiveness, it does not necessarily increase learning.

It is important to note that in the majority of the reviewed studies, both elaborated feedback and correct response feedback were accompanied by verification (Van der Kleij et al., 2015). One major reason is that most studies have focused on close-ended questions. Providing verification before presenting more complex information for close-ended questions is a common practice to establish a solid foundation for learning. Feedback that includes verifications has been observed to outperform feedback without verifications (Kuklick & Lindner, 2021). Unfortunately, feedback to open-ended questions typically lacks verification.

The complexity of feedback in relation to open-ended questions has received relatively little research attention. The majority of research on feedback for open-ended questions has primarily centered on natural language processing methods, which facilitate the automatic evaluation of students' answers, and we did not come across any studies specifically focusing on the provision of feedback for open-ended questions in the absence of auto-grading mechanisms. To obtain a more comprehensive understanding of the influence of feedback types provided following SOE questions, the current study employed four distinct feedback types with varying levels of complexity. These types included (1) No Feedback (NF), which did not provided feedback except for telling students “answer recorded”; (2) Correct Response (CR), which illustrated one correct response example to the question; (3) Elaborated Feedback (EF), which provided a detailed explanation of the required learning target by listing all the attributes; and (4) All Feedback (AF), which encompassed both CR and EF.

2.2 Feedback and Judgment of Learning

Beyond its effects on learning outcomes, feedback has also been suggested as a valuable instrument for enhancing students' metacognitive skills related to performance (Butler et al., 2008; Labuhn et al., 2010; Stone, 2000; Urban & Urban, 2021). Judgment of learning (JoL) is a metacognitive process through which individuals predict or assess their own learning and memory performance on a particular task or material. It involves estimating the likelihood or certainty of being able to recall or recognize information in the future (Rhodes, 2016). By examining JoL, we can gain a better understanding of how students assess their own learning from different types of feedback and how different types of feedback affect students' JoL.

There are several learning theories that help explain the concept of JoL and its relationship to feedback. Among these is cognitive load theory, which suggests that learners have limited cognitive resources and that the allocation of these resources can impact their ability to make accurate judgments about their learning (Sweller, 1994). Based on this theory, feedback that is concise, clear, and well-structured can reduce cognitive load and facilitate more accurate JoL.

Another relevant theory that contributes to the understanding of JoL and feedback is metacognitive theory. According to this theory, individuals possess metacognitive awareness and monitoring abilities that allow them to assess their own learning and make judgments about their future performance. Within this framework, feedback plays an important role by providing learners with valuable information about their current level of understanding or performance. Armed with this feedback, learners can then evaluate their progress and make informed decisions about how to adjust their learning strategies accordingly (Flavell, 1979).

Furthermore, self-regulated learning theory emphasizes the role of feedback in self-regulation processes such as goal setting, monitoring, and adjusting strategies (Zimmerman,

2000). Feedback provides learners with information about the effectiveness of their strategies and helps them make informed decisions about how to improve their learning.

Given the existing body of research indicating a positive correlation between higher confidence levels and enhanced information retrieval from memory, we incorporated students' JoL into our investigation and examined whether different types of feedback had an impact on students' JoL. In the context of this study, JoL refers to students' self-evaluated scores indicating their predictions of their own learning outcomes following the receipt of feedback. To minimize the potential impact of JoL on learning outcomes, the cues of our JoL questions were intentionally designed differently from the test questions (Myers, et al., 2020).

2.3 Semi-Open-Ended Questions in Foreign Language Learning

Online teaching and learning in the FL field have been extensively examined from various perspectives, including design and teaching guidelines, particularly in response to the COVID-19 pandemic. Moreover, an increasing application of technology has been witnessed in the context of Chinese FL learning (Liu et al., 2022). Research has shown that open-ended questions develop students' problem-solving skills (Bonotto, 2013). The utilization of open-ended items allows FL students to engage in real-life problem-solving scenarios where certain information may be missing and where there is not a single predetermined solution. This approach encourages FL students to employ reasoning skills and actively contribute to the discussion by making assumptions and providing comments about the missing information.

Compared to multiple-choice questions that provide an answer list with options for students to select from, short-answer questions require respondents to construct a response. Responses to short-answer questions are usually between one and a few sentences in length. Ye and Manoharan (2018) classified the answers to short answer questions with two categories,

concise and descriptive, which are also often referred to as close-ended questions and open-ended questions. Concise answers are fixed, and students are required to provide a number, a word, or a phrase that is unique. For example, for the question “When was the Qin dynasty founded?” the sole answer should be “221 BC.” A descriptive answer consists of one or a few sentences. “What majors are you interested in? Please explain your reasons” is one example in which the answers are open-ended and students can answer based on their own experiences. Zhang et al. (2022) pointed out that the third category of short answer questions, SOE short answer questions, falls between the two categories described thus far. SOE short-answer questions prompt students to express their subjective opinions within a given context. For example, for the question “Give a command to a robot with 把 (bǎ) structure and ask it to do a chore for you,” a sample answer could be a sentence using the grammar structure “把桌子擦干净” (Bǎ zhuōzi cā gānjìng), which translates to "Clean the table."

In the present study, our focus was specifically on SOE questions that elicit simple answers with a restricted range of acceptable responses. We did not include open-ended questions that require more complex answers involving critical thinking, information analysis, and detailed explanations or arguments. For the sake of brevity and clarity, we refer to our research targets as SOE questions that require the learner to provide a short, concise answer to a question while still allowing for some variability in the response. SOE questions often begin with an open-ended prompt but then provide specific instructions or parameters for the response. These questions are often used in FL education settings to assess a student's understanding of a specific topic or usage of language. Commonly used SOE question types in FL learning include

translating, making sentences with a provided structure, completing dialogues, and answering questions based on the text.

Some studies have used constructed-response questions to indicate open-ended questions. We do not use constructed-response questions because this is a broader concept that ranges from fill-in-the-blank questions to essay writing questions (Kuechler & Simkin, 2010), and it is more often used for essay writing assessment (Livingston, 2009). Due to the fact that certain scoring engines have the technical capability to grade SOE questions (Livingston, 2009), we do not use non-computer-gradable questions to define the question type we are investigating either, although these technologies are not yet widely accessible to frontline instructors.

3. Study Aims

3.1 Research Questions

Based on previous literature, the primary objective of the current study is to examine the influence of various online feedback on students' performance in responding to SOE questions, as well as to investigate the impact of feedback on students' JoL. An instructor survey and an experiment were conducted to investigate the following research questions:

1. Do instructors believe that feedback with high complexity provided for SOE questions would lead to greater learning gains?
2. Do different types of feedback with varying complexity have significantly different effects on students' performance when answering SOE questions?
3. Are there significant differences in students' JoL based on different types of feedback provided for SOE questions?
4. Do students' performances in response to different types of feedback align with instructors' perspectives on the effectiveness of feedback types?

3.2 Research Hypotheses

Given the previous literature results, we hypothesized that instructors believe that high complexity feedback (AF) to SOE questions would lead to greater learning gains (H1), and students' performances in response to different types of feedback align with instructors' perspectives on the effectiveness of feedback types (H2). We also assumed that CR, EF, and AF would outperform NF (H3), and AF would potentially be the most effective in enhancing learning (H4). For the relationship between time spent on feedback and learning, we anticipated that the duration of time spent on feedback messages would positively predict learning gains (H5). We also hypothesized that feedback with high complexity (AF) would yield greater JoL scores compared to CR and EF (H6).

4. Instructor Survey

4.1 Survey Objectives

Despite the popularity of SOE questions in the field of FL learning and the recognized role of immediate feedback in online learning, the provision of immediate feedback on SOE questions during online learning is limited, probably due to the challenges associated with subjective grading. It is important to note that although many instructors do not have access to ready-made question banks that offer auto-graded immediate feedback for SOE questions, this does not imply that feedback for SOE questions cannot be provided or that instructors are unable to offer timely responses. Two common approaches for providing immediate feedback to SOE questions without the support of auto-grading systems are CE, which offers sample answers, and EF, which provides detailed explanations such as grammar pattern rules.

Prior to delving deeper into the experiment, we first gathered the perspectives of instructors on this topic. Instructors are the ones directly responsible for delivering the assignments and feedback. Their understanding of the students, teaching styles, and learning

needs is invaluable. By conducting an instructor survey, we can gain insights into the challenges they face, their preferences, and their expectations for the experiment. An instructor survey (see appendix A) was designed to explore the attitudes and opinions of K–16 Chinese FL instructors regarding feedback on online assignments and to investigate their views on providing immediate feedback, especially for SOE questions.

4.2 Instrument

The survey used for data collection was developed by the author and reviewed by two Chinese language instructors. The survey was written in English and included four sections with 15 questions.

Section 1 included two general questions asking about the type(s) of institutions where the instructors taught and at what language level they taught.

Section 2 included six questions related to homework assignments and perspectives on providing immediate feedback to close-ended questions, open-ended questions, and SOE questions. To ensure that instructors understood the definition of SOE questions, an example was provided. The ninth question asked instructors which method they used more often when assigning homework: an online platform with the option to provide immediate feedback or another format that could not provide immediate feedback (e.g., written homework or email/upload answer sheet). Based on the feedback, instructors were asked to answer questions either in Section 3 or Section 4.

Section 3 was designed for instructors who used an online platform with the option to provide immediate feedback for homework assignments. This section consisted of three questions: How often do you use online tools to provide immediate auto-feedback to close-ended questions, open-ended questions, and SOE questions when assigning homework through an

online platform? How do you provide immediate feedback to questions that are not auto-gradable? When assigning homework without immediate auto-feedback, how long does it usually take for students to receive feedback?

Section 4 was designed for instructors who didn't use an online platform with the option to provide immediate feedback, and three questions were asked: How would you rate the helpfulness of providing immediate feedback to questions that are not auto-gradable for students? Which type of immediate feedback for non-auto-gradable questions do you believe would be most effective? When assigning homework, how long does it usually take for students to receive feedback?

4.3 Participants

Participants were Chinese language instructors in the United States who teach K-16.

4.4 Data Collection and Analysis

The survey was conducted using a web-based platform called Qualtrics that allowed participants to respond to a series of questions via their personal electronic devices. After receiving official IRB approval in late March 2023, the recruitment of participants for the survey was carried out using a combination of email, social network platform WeChat, and personal or professional contacts. The survey was designed to take approximately 10 minutes to complete and was available to interested Chinese instructors for a period of two months. The data were analyzed with Qualtrics built-in tools and Excel.

4.5 Results

A total of 60 Chinese language instructors took the survey, of which 17 were K–12 instructors and 43 were college instructors. A total of 35 (58.33%) of the participants taught

beginning-level language courses, 20 (33.33%) taught intermediate courses, and 5 (8.33%) taught advanced courses.

In total, an overwhelming majority of the respondents (83.94%) indicated that they assign homework to their students at least once per week. 30.36% of the participants reported assigning homework every day, 26.79% a few times a week, and 26.79% every week. When asked about the components of the assignments given to students, on average, 38.11% of assignments were close-ended questions, 37.96% open-ended questions, and 33.53% SOE questions. The top three question types used most often in assignments were essay writing, multiple choice, and translation; two out of three were open-ended questions.

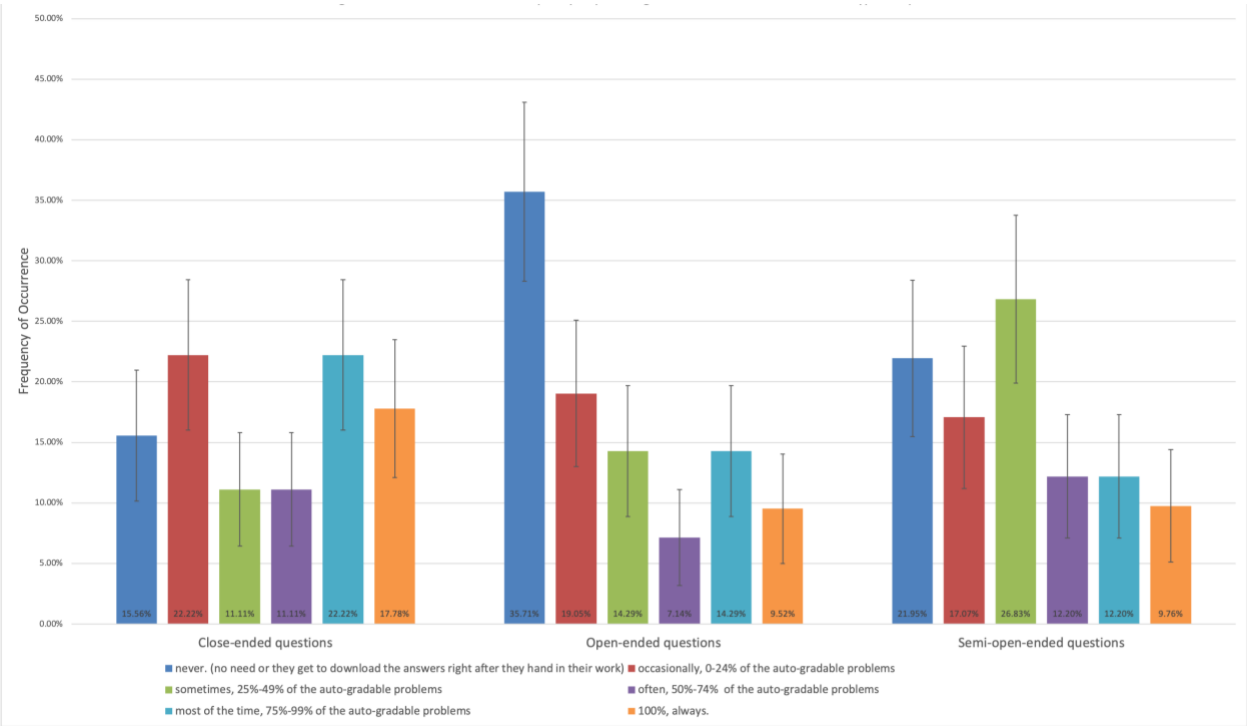
The survey suggested that most instructors have access to online platforms with the option to provide immediate feedback. Of the 56 instructors who answered this survey question, 26 (46.43%) used an online platform with the option to provide immediate feedback; 10 (17.86%) used other formats that cannot provide immediate feedback, such as on-paper practice; and 20 (35.71%) reported using both.

When asked to rate on a 0-100% scale how helpful it would be for students to receive immediate feedback on questions that are not auto-gradable, such as SOE questions, 55 out of 60 instructors answered the question. On average, they expressed 74.89% confidence that it would be helpful. However, even with the immediate feedback option available and instructors' belief that immediate feedback for open-ended and SOE questions is helpful, only a low percentage of instructors utilized the immediate feedback function, even for close-ended questions. Among the instructors who could use an online platform with the option to provide immediate feedback, only 17.78% always provided immediate feedback to close-ended questions all of the time, and 15.56% never provided immediate feedback to close-ended questions or other question types.

The percentage of respondents providing immediate feedback was even lower for open-ended questions and SOE questions. Only 9.76% of instructors always provided immediate feedback to SOE questions. Figure 1 presents the percentage of instructors who provided immediate feedback to different types of questions with varying frequencies.

Figure 1

Frequency of instructor answer occurrence for frequency of providing immediate feedback across three types of questions



Note. The total answer occurrences for the close-ended questions, open-ended questions, and semi-open-ended questions were 45, 42, and 41, respectively. The errors bars stand for standard error.

In terms of providing immediate feedback when assigning homework online, the respondents were mostly positive (76.61%) about its helpfulness/necessity. Fifty-five instructors who answered the question expressed 74.89% confidence that immediate feedback is helpful for questions that are not auto-gradable, such as SOE questions. Among the 56 instructors who answered the question, the majority of them (85.71%) believed that feedback containing both CR and EF is the most effective for SOE questions. Only two (3.57%) of them believed that CR is the most effective, while four (7.14%) believed that EF is the most effective.

In summary, our survey found that most instructors had access to an online platform that could provide immediate feedback, and most of them held positive attitudes towards providing immediate feedback to all types of questions. However, only a low percentage of them actually provided immediate feedback, even with the option available, and even fewer instructors provided feedback for SOE questions. When considering different types of immediate feedback for SOE questions, a majority (85.71%) believed that AF is the most effective for SOE questions. The survey results offered backing to Hypothesis 1, suggesting that instructors generally hold the belief that providing feedback with high complexity (AF) for SOE questions would result in higher learning gains.

5. Experiment

5.1 Methods

5.1.1 Design

The experiment incorporated two factors and four conditions as well as four test points. The two factors examined were the provision of CR and the provision of EF. These factors were manipulated between participants, resulting in four distinct between-subject conditions: the NF condition (no EF, no CR), the CR condition (with CR, no EF), the EF condition (with EF, no

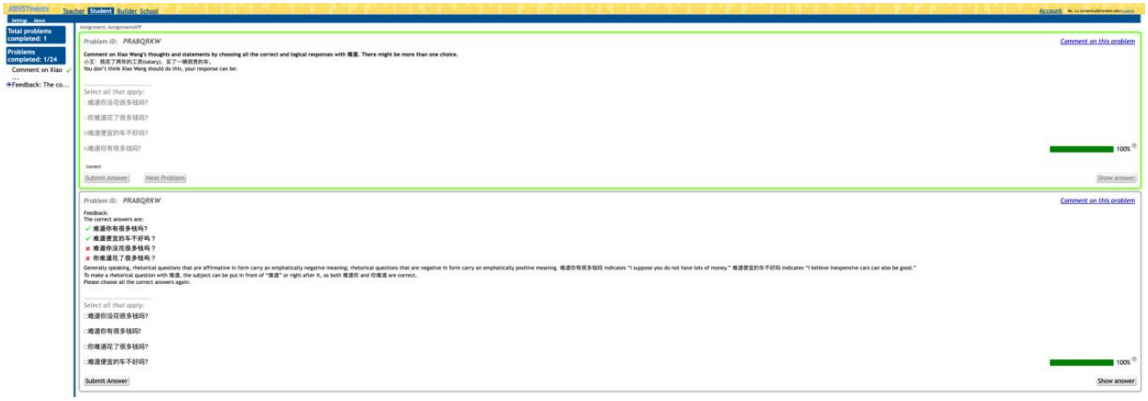
CR), and the AF condition (with CR, with EF). All participants underwent the pretest as well as the three subsequent posttests.

5.1.2 The Online Learning Platform

The experiment was conducted in-person using the ASSISTments platform, which is an online learning platform designed to provide support to students through hints, feedback, and scaffolding (Heffernan & Heffernan, 2014). All assignments, pretests, and posttests were administered during class time. Figure 2 displays the interfaces of ASSISTments as seen by students, instructors, and researchers. The pretest, quiz assignments, immediate feedback, and posttest were all delivered through the ASSISTments platform. Prior to analysis, the research data was anonymized by the platform to ensure confidentiality. For each course, daily in-class quizzes were administered to all students using their personal laptop computers. Students were already accustomed to taking online quizzes and were provided with laptops at the start of the semester to ensure inclusivity. Additionally, students had the option to use their smartphones to complete the quizzes.

Figure 2

ASSISTments Interfaces for Students (a) and Instructors (b)



a. Student Interface

Student/Problem [Unanonymize]	Average Data driven	PRABQRSE Data driven	PRABQRSF Data driven	PRABQRSG Data driven	PRABQRSH Data driven	PRABQRSJ Data driven	PRABQRSK Data driven	PRABQX7F Data driven	Total Hints	Time Spent
Problem Average Graph	43%	43%	19%	24%	44%	44%	27%	100%		
Common Wrong Answers		我把鞋放到了床的旁边。,16% 我把鞋放到床的旁边了。,16%	我把书放进书包里了。,我把书放了在水杯的旁边。,32% +feedback 我把书放进书包里了。,我把书放了进书包里。,17% +feedback 我把书放进书包里了。,我把书放了进书包里。,我把书放了在水杯的旁边。,11% +feedback	你把钱找错了。,你把钱找得错。,37% +feedback 你把钱找错了。,28%	他今天得把中文作业做完。,他今天得把中文作业做。,30% +feedback 他今天得把中文作业做完。,21% 今天他得做完中文作业。,他今天得把中文作业做。,17% +feedback	把你的功课都做完。,把你的功课都做。,34% +feedback 把你的功课都做完。,把一些功课都做完。,26% +feedback	先把书都看完，再出去玩。,53% 先把书都看了，再出去玩。,先看了都的书，再出去玩。,16% +feedback			
Correct Answer(s)		*我把鞋放到床的旁边了。，我把鞋放到了床的旁边。*	*我把书放进书包里了。*	*你把钱找错了。，你找错了钱。*	*他今天得把中文作业做完。，今天他得做完中文作业。*	*把你的功课都做完。*	*先把书都看完，再出去玩。，先把书都看了，再出去玩。*	*16，17，18，19，20，21，22，23，24，25，26，27，28，29，30*		
XXXXXXXX	29%	✓ 我把鞋放到床的旁边了。，我把鞋放到了床的旁边。 100%	× 我把书放进书包里了。,我把书放了在水杯的旁边。 0%	× 你把钱找错了。,你把钱找得错。 0%	× 他今天得把中文作业做完。 0%	× 把你的功课都做完。,把一些功课都做完。,你做完都功课。 0%	× 先把书都看了，再出去玩。,先看了都的书，再出去玩。 0%	✓ 20 100%	0	00:04:03
XXXXXXXX	57%	× 我把鞋放到床的旁边了。 0%	✓ 我把书放进书包里了。 100%	× 你把钱找错了。 0%	✓ 他今天得把中文作业做完。,今天他得做完中文作业。	✓ 把你的功课都做完。 100%	× 先把书都看完，再出去玩。 0%	✓ 20 100%	0	00:04:07

b. Instructor Interface

5.1.3 Participants

A total of 91 undergraduate students participated in the experiment as part of their in-class learning. Participants were recruited from four intermediate-level Chinese courses in the spring 2023 semester at two universities. Chinese was not their native language, and students were placed into the for-credit classes based on placement tests at the beginning of the semester. Seven students who achieved an average score of 100 on the pretest were excluded from the analysis as this indicated that they had already mastered the material.

5.1.4 Materials

The Chinese grammatical structure *ba* (把) within the scope of intermediate-level Chinese FL courses was selected to provide the content of questions, and the three primary usages of the *ba* pattern were chosen as the specific learning targets for the study. Seven SOE questions of the same difficulty but in different communicative situations were created for each usage of the *ba* pattern. Out of these seven questions, four were intentionally designed to be highly similar, with only variations in the usage of situations. These four questions were randomly selected to be pretest and posttests prior to the commencement of the experiment, in order to effectively minimize any testing differences. The rest three questions were used for learning practice. An example of the four test questions, the three practice questions can be found in Appendix B.

Three types of feedback were developed for each practice question. The first type was CR, which provides an example of the correct answer for the question. The second type was EF, which presents the necessary grammar instruction required to answer the question. The third type is AF, which combines both CR and EF. An example question with its CR, EF, and AF feedback can be found in Appendix B.

In calculating the content validity of the grammar tests and feedback, we collected expert opinions. Three CFL instructors at the college level reviewed the content to ensure the questions, situations provided, examples, and grammar explanations were appropriate and at the same difficulty level.

After receiving a feedback (except for no feedback) in the learning practice question, students were asked a JoL question to rate how likely they were to be able to use the grammar pattern correctly on a later test based on the feedback they received within a range from 0% to 100%. See the example below:

With the feedback given above, how likely will you be able to use this structure correctly on a later test? Indicate from 0% to 100%.

0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%

5.2 Procedure

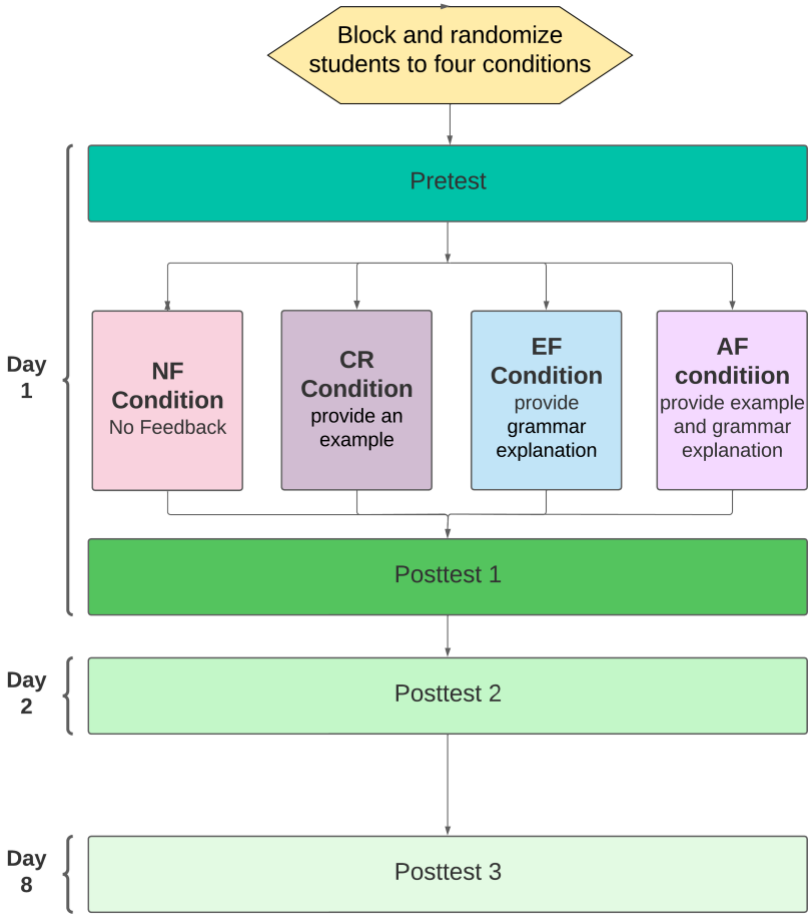
We oversaw the procedure and trained instructors before the experiment to help them familiarize themselves with the procedure of delivering the experiment. Participants were randomly assigned to one of the four conditions before the experiment started. On experiment day, before providing consent, participants were informed that they would study grammar patten *ba* and would be asked to take tests later on the same grammar patterns. Participants were not told what type of feedback they would receive prior to studying the grammar pattern, although they were told they might not be given feedback. They were told that the testing results were not connected to their course grades. All students logged into ASSISTments with laptops/computers/iPads to complete the tasks.

On the first day of the experiment, after signing the consent form, each participant responded to three pretest questions in a randomized order; completed nine practice questions covering the three usages in a randomized presentation order; and answered three questions for Posttest 1, also in a randomized order. The order of the three questions in each usage was also randomized. If assigned to a CR, EF, or AF condition, each feedback was immediately followed by a JoL question in practice. On the second day, participants completed three questions for Posttest 2 in a randomized order on ASSISTments. One week later (eighth day), participants took three questions for Posttest 3, again in a randomized order on ASSISTments. No feedback was provided for either test.

To ensure that the study met ethical requirements, all students in the four conditions received all types of feedback by the end of the experiment to guarantee equal learning opportunities. Instructors also went over the target grammar usage in the classroom after the experiment to eliminate any remaining confusion. All the experiments took place in the classrooms to guarantee that students did not receive extra support beyond the feedback. The experimental procedure is displayed in Figure 3.

Figure 3

Experimental procedure



Throughout the experimental process, all participants' attempts on the learning platform were logged automatically, which served as a statistical analysis of the data. Participants from all universities that participated in this study used the same materials and followed an identical experimental procedure.

5.3 Scoring and Analysis

All questions, including the pretest, practice questions, and the three posttests, underwent anonymization and were manually assessed by three raters using a grading rubric. The grading rubric utilized in the study is provided in Table 1. Each question was assigned a maximum of 100 points and evaluated based on three criteria: structure (50 points), word usage (20 points), and communication (30 points). In cases where typing errors occurred (e.g., "我闷" instead of "我们"), these errors were considered as incorrect word usage. If an answer did not employ the designated structure at all, it received a score of 0. Likewise, if the answer utilized the structure correctly but lacked logical coherence with the question, indicating a lack of understanding of the question or grammar usage and a failure in communication, the answer also received a score of 0.

Table 1

Grading rubric

Criteria	Unsatisfactory	Needs Improvement	Good
Structure	**0** Did not use the structure or used with wrong order (e.g., 把吃那碗汤) or major mistakes.	**25** Used the structure correctly with a minor mistake, such as the wrong usage of or missing 了 in 把 structure, missing 要, or missing resultative complement.	**50** Used the structure correctly.
Word usage	**0** Three or more words not properly used in the sentence, such as using 桌 instead of 桌子.	**10** One or two words in the sentence not properly used, such as using 桌 instead of 桌子, or with one or two typing errors, e.g., 把-吧.	**20** All words properly used.
Communication/ meaning delivery	**0** Answer not logically connected to the question and cannot be used to answer the question.	**15** Student showed that they understood the question, but the answer does not directly answer the question and leads to confusion.	**30** Student successfully answered the question and reached the communication goal.

5.4 Results

5.4.1 Effects of Feedback Types on Learning Between Conditions

First, to investigate the effects of the feedback types on learning, we compared the learning gains of students in each posttest between conditions. Robust ANOVA tests were conducted for each posttest learning gain. By employing robust ANOVA tests, specifically using

the WRS2 package in R, we were able to overcome the limitations associated with violations of normality and equal variance assumptions. Robust ANOVA tests are designed to handle non-normal and heteroscedastic data, making them suitable for analyzing our dataset. The analysis included the learning gain results from each of the three posttests, with an effective number of 8,000 bootstrap samples. The summary statistics of the three posttest learning gains in different conditions are included in Table 2.

Table 2

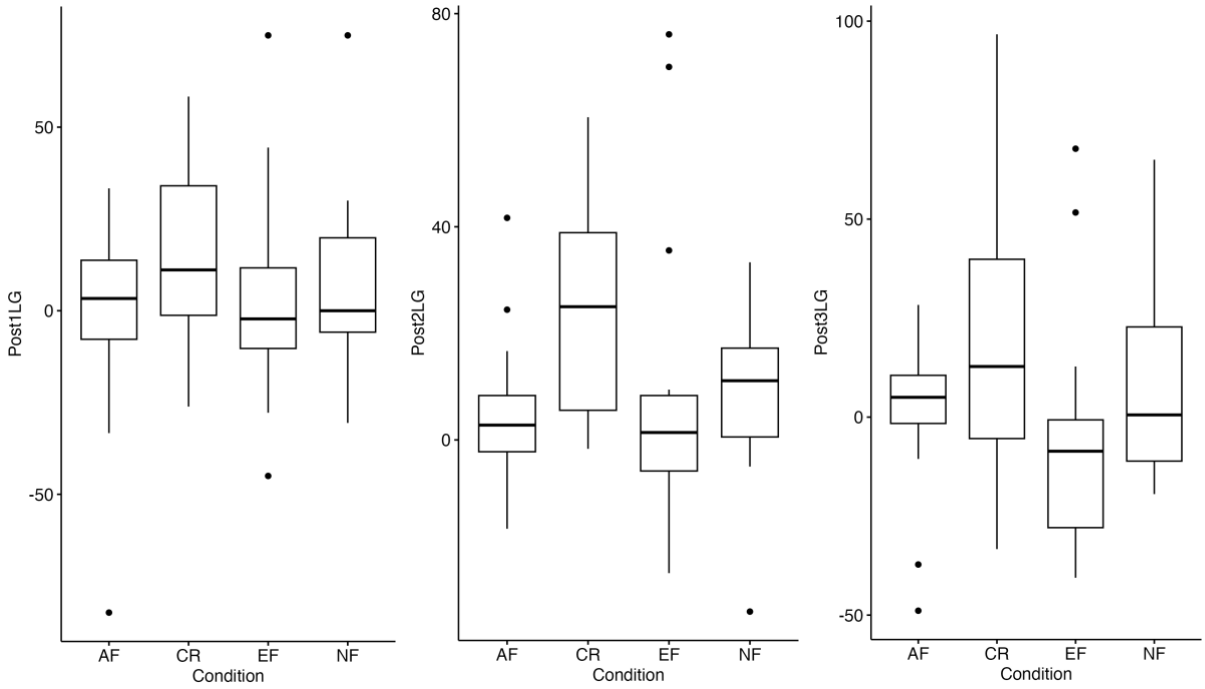
Means and standard deviations of the three posttests' learning gain scores of the four conditions as well as Robust ANOVA results for each posttest

	Condition	Learning Gain			Robust ANOVA		
		n	*Mean*	*SD*	effective size	statistic	p-value
Posttest 1	No Feedback (NF)	16	7.64	23.90	0.31	1.31	0.29
	Correct Response (CR)	20	16.90	25.00			
	Elaborated Feedback (EF)	19	1.810	25.90			
	All Feedback (AF)	22	1.67	25.60			
Posttest 2	No Feedback (NF)	15	8.93	16.10	0.56	4.03	0.03
	Correct Response (CR)	17	25.00	20.80			
	Elaborated Feedback (EF)	18	8.27	26.80			
	All Feedback (AF)	17	5.16	13.40			
Posttest 3	No Feedback (NF)	15	8.15	24.10	0.50	3.76	0.05
	Correct Response (CR)	16	18.20	33.00			
	Elaborated Feedback (EF)	18	-7.19	28.70			
	All Feedback (AF)	16	1.13	20.00			

In Posttest 1, the test statistic was 1.31, resulting in a p-value of 0.29. The effect size was 0.31. No significant difference was found between the conditions. For Posttest 2, the test statistic was 4.03, yielding a *p*-value of 0.03, and the effect size was 0.56, indicating a significant difference between the conditions. A robust post hoc test using trimmed means revealed a significant difference between the CR and EF conditions, 95% CI of [-37.78, -2.68], test statistics value of -19.95. Providing CR was more effective than providing EF. No significant difference was found between the AF and CR conditions (test statistics value of 13.14, CI [-4.77, 33.46]), suggesting that after CR is provided, providing EF or not does not make a difference. In Posttest 3, the test statistic was 3.76, resulting in a p-value of 0.05. The effect size was 0.50, indicating a marginal significant difference between the conditions. However, a robust post hoc test using trimmed means did not find significant difference between conditions. Figure 4 displays the boxplot of the three posttest learning gains for different conditions, which supported the observed differences between the conditions. In summary, robust ANOVA findings did not indicate that providing feedback to SOE questions resulted in better performance compared to the absence of feedback, but providing CR was more effective than providing EF in the Posttest 2 results.

Figure 4

Box plots for the three posttest learning gains with four conditions



Posttest 1 Posttest 2 Posttest 3

5.4.2 Effects of CR and EF on learning

To further investigate how different factors affect learning, we conducted an analysis to examine the impact of two factors, namely CR (with or without) and EF (with or without), on the learning outcome. We aimed to determine the extent to which these factors could predict students' learning gains. To achieve this, we employed robust linear regression models for each posttest result. Effect coding was employed in our analysis to enable a meaningful assessment of the effects of the predictor variables on the outcome variable. The model used to analyze the learning gain for Posttest 1 did not account for a significant amount of the variance: $F(3, 73) = 1.63$, $p = 0.19$, $R^2 = 0.06$, adjusted $R^2 = 0.02$. The analysis further revealed that neither the two factors nor their interaction significantly predicted the Posttest 1 learning gain ($p > 0.05$ for both

cases). In the case of Posttest 2, the model yielded a significant explanation for a portion of the variance in the learning gain: $F(3, 63) = 3.73$, $p = 0.02$, $R^2 = 0.14$, adjusted $R^2 = 0.10$. These findings indicated that the model demonstrated a meaningful relationship between the predictor variables and the value of the Posttest 2 learning gain. Specifically, the analysis revealed that the EF factor significantly predicted the value of the Posttest 2 learning gain ($\beta = -10.25$, t [63] = -2.12, $p = 0.04$, 95% CI [-19.89, -0.60]). This suggests that the presence or absence of EF had a significant impact on the learning gain for Posttest 2, with negative values indicating a decrease in learning gain. CR, on the other hand, did not significantly predict the value of the Posttest 2 learning gain ($p > 0.05$). The interaction between the two factors of CR and EF did not significantly predict the value of the Posttest 2 learning gain ($p > 0.05$), indicating that the combined effect of the two factors did not have a statistically significant influence on the learning gain for Posttest 2. The model for Posttest 3 did not find a significant variance either: $F(3, 61) = 2.15$, $p = 0.10$, $R^2 = 0.12$, adjusted $R^2 = 0.07$. However, the factor of EF significantly predicted the value of Posttest 3 learning gain: $\beta = -16.20$, t (61) = -2.43, $p = 0.02$, 95% CI [-29.53, -2.88]. The factor CR and the interaction did not significantly predict the Posttest 3 learning gain ($p > 0.05$ for both cases). Table 3 summarizes the statistics of the three models.

Table 3

Robust linear regression models for the three posttests' learning gains regressed on two factors: CR and EF

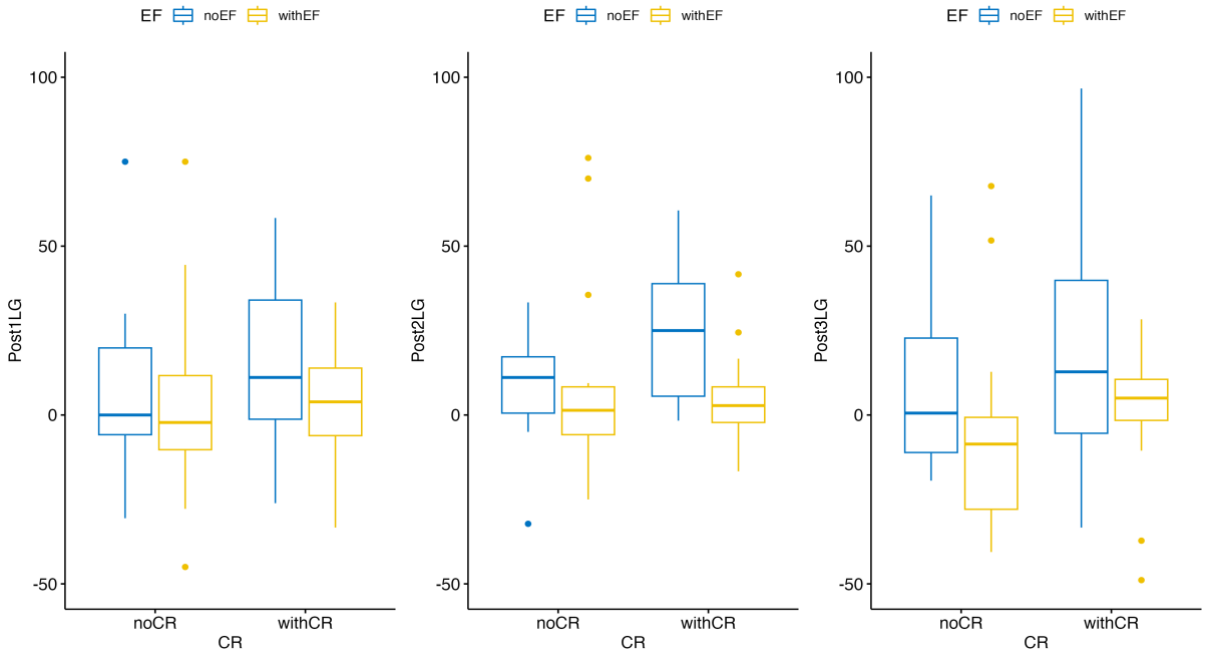
	Factors	Coefficient (β)	Robust standard error	*t*-statistic	*p* - value	95% Confidence Interval for β		*df*
						Lower bound	Upper bound	
Posttest 1	(Intercept)	7.01	2.87	2.44	0.02	1.29	12.73	73
	CR	4.57	5.74	0.80	0.43	-6.88	16.01	73
	EF	-10.54	5.74	-1.83	0.07	-21.98	0.91	73
	CR * EF	-9.43	11.49	-0.82	0.41	-32.32	13.47	73
Posttest 2	(Intercept)	11.84	2.41	4.91	6.893 e-06	7.02	16.66	63
	CR	6.48	4.83	1.343	0.18	-3.16	16.13	63
	EF	-10.25	4.83	-2.12	0.04	-19.89	-0.60	63
	CR * EF	-19.18	9.66	-1.99	0.051	-38.48	0.11	63
Posttest 3	(Intercept)	5.07	3.33	1.52	0.13	-1.59	11.73	61
	CR	9.18	6.66	1.38	0.17	-4.14	22.51	61
	EF	-16.20	6.66	-2.43	0.02	-29.53	-2.88	61
	CR * EF	-1.73	13.33	-0.13	0.90	-28.37	24.92	61

The results revealed that providing EF or not played a significant role in the posttest learning gains. Regardless of whether CR was provided or not, the provision of EF negatively affected learning. The difference was not significant in Posttest 1, but it emerged in both Posttest

2 and Posttest 3. A boxplot for the three posttest learning gains with the two factors is presented in Figure 5.

Figure 5

Box plots for the three posttest learning gains with the two factors



5.4.3 Time Spent on Learning

To further explore how students interacted with feedback in different conditions, the in-experiment data was analyzed. We first examined the time spent on problems as a potential factor. The time spent on each problem was determined by subtracting the problem start time from the problem end time. We constructed a robust linear model, clustering the data by student ID and employing the CR^2 standard error type, to assess the relationship between the time spent on each problem and the condition. Twenty-three problems that took longer than five minutes were deemed errors and subsequently excluded from the data. The model did not yield

significant results (F[3, 95] = 1.32, p = 0.27), indicating no significant differences in time spent on problems were observed between the conditions.

We then examined the time spent on feedback as a potential factor. After each problem, students spent some time reading the provided feedback (if any). The time spent reading feedback was calculated as the feedback end time minus the feedback start time. Five attempts with feedback durations longer than four minutes were taken as outliners and removed from the data. The summary statistics can be found in Table 4.

Table 4

Summary statistics for in-experiment time spent on feedback by condition

Condition	*n*	*Mean (in seconds)*	*SD*
AF	228	18.80	26.30
CR	191	11.40	10.70
EF	210	15.60	19.50

Note: n indicates the number of feedback attempts collected for analysis in each condition.

We constructed a robust linear model of time spent reading the content each feedback as a function of condition, clustered by student ID, and employed the CR^2 standard error type. The model has a significant effect: $F(2, 76) = 4.11$, $p = 0.02$, Multiple $R^2 = 0.02$, Adjusted $R^2 = 0.02$. Compared to providing all feedback (condition AF), providing only CR took significantly less time, $\beta = -7.45$, $t = -2.59$, $p = 0.01279$. The regression results can be seen in Table 5. The result obtained appeared to be reasonable given that the AF condition contained a higher amount

of information. However, it was worth noting that despite the greater amount of information provided in the AF condition, it did not outperform the CR condition in the subsequent tests. Therefore, based on the findings mentioned earlier, it can be concluded that providing CR may be the most efficient approach among the four conditions.

Table 5

Robust linear model of predictor condition on feedback time with 95% confidence intervals

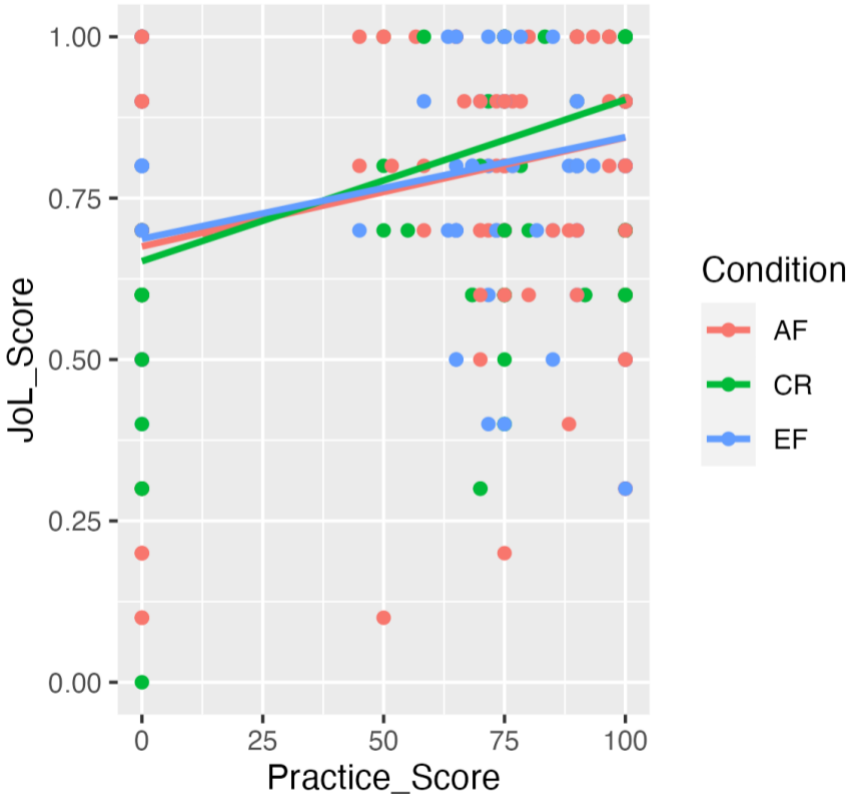
	Coefficient	Std. Error	*t*-value	Pr (>|t|)	CI Lower	CI Upper	*df*
(Intercept)	18.83	2.62	7.20	1.47e-07	13.44	24.21	25.18
Condition CR	-7.45	2.87	-2.59	0.01	-13.24	-1.67	44.44
Condition EF	-3.24	3.39	-0.96	0.34	-10.06	3.58	48.27

5.4.4 Effects of Feedback types on JoL

After receiving each feedback text (if applicable) in the learning practice questions, students were asked to provide a JoL score. To investigate the impact of different types of feedback on students' JoL scores, we analyzed the JoL scores across the three conditions where feedback was provided, taking students' learning gains into consideration. Figure 6 displays the plot depicting the relationship between students' JoL scores and learning gain scores across different feedback types.

Figure 6

Students' JoL confidence scores and learning gain scores across different feedback types



A robust linear model was employed with the CR2 standard error type, and data was clustered by student ID. The model yielded a significant effect, indicating that the type of feedback had a significant influence on students' JoL scores ($F[3, 621] = 35.34$, $p < 0.001$). The multiple R^2 value was 0.15. When holding students' problem scores constant, a significant difference in confidence levels was observed between the CR condition and the AF condition. Specifically, the CR condition received significantly higher confidence ratings compared to the AF condition ($\beta = 0.04$, $t = 2.15$, standard error = 0.02, $p = 0.03$). The regression results can be seen in Table 6. These findings suggest that, when controlling for students' problem scores, the

provision of CR led to significantly higher levels of JoL scores compared to the provision of both CR and EF. Given that CR feedback had the lowest complexity in the experiment and AF feedback had the highest, the results indicate that simpler feedback was more effective than more complex feedback for SOE questions.

Table 6

Robust linear model of predictor score and condition on JoL scores with 95% confidence intervals

	Coefficient	Std. Error	*t*-value	Pr(>|t|)
(Intercept)	0.66	0.018	37.17	<2e-16
Score	0.002	0.0002	9.93	<2e-16
Condition CR	0.04	0.02	2.15	0.03
Condition EF	0.003	0.02	0.16	0.87

6. General Discussion

The purpose of this study is to gain a better understanding of the effectiveness of different types of immediate feedback on students' learning of SOE questions in FL classrooms. Our survey results reveal that most instructors hold positive attitudes towards providing immediate feedback to SOE questions; however, very few of them put it into practice. Most instructors believe that providing immediate feedback with both a correct response and elaboration would be the most effective response to SOE questions; however, this result is not supported by our experiment on student learning gains. Our experiment results indicate that CR and EF act differently on students' learning, with CR more effective than EF. Surprisingly, providing EF negatively affected SOE learning regardless of whether CR is provided or not. The difference was not significant in Posttest 1 but emerged in Posttest 2 and Posttest 3. What's

more, students spent significantly longer time in the AF condition of feedback compared to the CR condition. No difference was found between the time spent on feedback in the CR versus EF conditions. Lastly, when controlling for students' performance on practice problems, the provision of CR resulted in significantly higher levels of JoL scores compared to the condition where both CR and EF were provided. However, providing only EF did not significantly raise students' JoL scores. Further elaboration will be provided in the subsequent sections.

6.1 More Complexity Does Not Mean Greater Effectiveness in SOE questions

In contrast to the results of the instructors' survey, the provision of both CR and EF for SOE questions did not yield the best learning outcomes. Instead, CR proved to be a superior choice compared to EF feedback; providing EF feedback actually had a negative impact on learning with SOE questions. These findings contrast with previous research that generally supported the idea that feedback providing more information is more effective.

It is important to note that most of the research ((Van der Kleij et al., 2015) supporting the increase in learning performance with feedback complexity is based on feedback for close-ended questions, in which all complex feedback (CR, EF, AF) includes verifications. One key difference between feedback for SOE questions and feedback for close-ended questions is the lack of verification. Our study strongly indicates that without verification, learning outcomes may not improve with an increase in feedback complexity for SOE questions. Upon examining the CR and EF conditions in our experiment, we observe that while CR feedback does not directly provide verification of responses, it still allows students to compare their answers to a correct response, implicitly aiding them in verifying their answers. In contrast, the EF condition only provides bullet points of grammar explanations, which may make it more difficult for students to verify their answers. Feedback that includes verifications has been observed to

outperform feedback without verifications (Kuklick & Lindner, 2021). Longer explanations without the provision of verification may lead to students' uncertainty regarding their performance on the task as well as uncertainty about how to respond to the feedback, potentially confusing students and creating learning difficulties (Weaver, 2006). Hattie and Timperley (2007) also assume that forms of feedback are "most useful when they assist students in rejecting erroneous hypotheses and provide direction for searching and strategizing" (pp. 91–92). When providing feedback for SOE questions, the answers provided by students are not verified, making it challenging to reject erroneous hypotheses through feedback. In such cases, longer explanations may lead to confusion rather than provide clear guidance.

Support for this viewpoint can be substantiated by theories from the field of psychology. Cognitive load theory suggests that when learning materials cause confusion, cognitive resources are diverted towards tasks that are not directly relevant to the learning process (Sweller, 2010). EF might have provided overwhelming feedback to students compared to CR, especially when an example which can be used as verification is not provided.

One remaining question to address is why the AF condition did not outperform the CR condition in our experiment. The AF condition encompassed both CR and EF information, yet the addition of EF to CR did not yield improved learning outcomes. One possible explanation is that our study had a relatively small sample size, which may have hindered our ability to detect an effect. The results could be attributed to individual differences. Additionally, since we did not observe a difference between the AF and CR conditions, another reason could be that the provided CR ensured students' learning outcomes in the AF condition, while EF did not have an impact. This discrepancy might be due to the length of the EF or other factors that necessitate further investigation.

6.2 Correct Response is More Efficient for SOE Questions

One notable finding of our study is that the provision of AF resulted in significantly more time spent on feedback compared to providing CR; the provision of EF did not. This finding is in line with expectations as the AF condition contained the most information. However, interestingly, we did not observe a significant increase in the time spent on EF feedback despite it also containing significantly more information compared to CR feedback; this implies that simply including more information in the feedback does not necessarily lead to a longer learning time. On the other hand, the combination of CR and EF feedback may significantly increase the time students spend on feedback as CR helps students verify their answers in a certain way. Without verification, students may feel confused and may not invest additional time in examining detailed feedback. They may be unsure of what to focus on in the feedback or may become overconfident in their answers, leading them to spend less time on the feedback details. Hattie and Timperley (2007) emphasized that when feedback fails to establish a clear goal for students, they may not perceive the need to bridge the gap between the goal and their current status through feedback, thereby limiting their engagement with detailed feedback information. A significant body of research has consistently demonstrated that learning outcomes are enhanced with an increased investment of time in receiving feedback (Kuklick et al., 2023). However, our study on SOE questions did not yield results indicating that the AF condition outperformed the other conditions in terms of posttest learning outcomes. Consequently, our findings did not align with the proposition that extended time spent on feedback contributes to enhanced learning in the context of SOE questions.

When considering learning gains in relation to the amount of time spent reading feedback, the provision of only CR feedback was more efficient than the provision of EF or AF

because providing CR led to the highest scores with the shortest time spent on feedback. These findings suggest that for SOE questions, CR feedback is more efficient in facilitating learning.

6.3 Correct Response Improves Judgment of Learning Outcomes

Besides learning outcomes, we also discovered that the provision of CR led to higher JoL outcomes compared to the condition where both CR and EF were provided. This finding is consistent with prior research demonstrating that varying feedback content had a significant impact on cognitive load (Taxipulati & Lu, 2021). The finding could be explained by a substantial growth in cognitive load within both the EF and AF conditions, which likely played a mediating role in lowering the JoL level. An alternative explanation for the observed results is rooted in the concept that participants tend to base their JoLs on the fluency of information or ease of processing (Castel et al., 2007; Yue et al., 2013). In this context, it is plausible that participants perceived one-sentence examples as more fluent compared to detailed grammar points, despite the latter containing a greater volume of information.

6.4 Implications for Practice

The idea for the present study originated from frontline teaching, and it is our intention that the findings be applied to teaching practices. The results of this study offer encouragement for instructors to incorporate more immediate feedback to SOE questions. Specifically, it is recommended to provide immediate feedback in the form of a correct response, such as a working example, rather than more detailed explanations. If providing EF is desired, ensure that the correct response is provided alongside the EF. It is believed that students who receive correct responses for SOE questions not only experience enhanced learning outcomes but also exhibit improved judgment of their own learning, thereby positively influencing future learning endeavors.

6.5 Limitations and Future Research

First, the participant number is relatively small, which hindered the ability to detect a difference between conditions in Posttest 1. Future research should increase the number of participants to obtain more robust results.

Secondly, since our research primarily focused on feedback, JoL questions were administered solely after the feedback conditions and not in the NF condition. This design choice precluded us from comparing JoL between the feedback and NF conditions. Future studies may consider incorporating JoL measures within the NF condition without unduly interfering with the structure of assessment.

Extending the current findings through the investigation of additional types of feedback with varying complexities would be beneficial. This could involve exploring EF with differing levels of information, or considering EF both with and without verification components. Although providing verification feedback to SOE questions poses challenges, it would be intriguing to investigate whether the inclusion of verification would alter students' learning behaviors and outcomes.

It might be useful to survey these students about their perceptions of the feedback. Moreover, considering the utilization of focus groups and individual interviews could offer a more comprehensive understanding of their perspectives.

Additionally, it is noteworthy that various types of SOE questions exist beyond the scope of the questions examined in this study. Subsequent research endeavors should consider including different question types to investigate potential variations in findings. Furthermore, future studies should expand to encompass a wider range of open-ended questions as they may also demonstrate the potential benefits of immediate online feedback. Considering the limited

number of studies investigating feedback provided for open-ended questions, it is our aspiration that our research will offer valuable insights into this matter.

Surveying these students about their perceptions of the feedback could provide valuable insights. Additionally, conducting focus groups and individual interviews might offer even deeper understanding.

7. Conclusion

This research serves as an initial endeavor in examining immediate online feedback for SOE questions in FL learning. The current study contributes to our comprehension of the impacts of feedback on SOE questions and offers preliminary recommendations on delivering immediate feedback while administering SOE questions online. Although the generalizability of the present findings should be established through future research, this study nonetheless provides robust evidence supporting the provision of immediate correct responses to SOE questions. It is our hope that this study will inspire further exploration in this crucial domain.

Acknowledgements

We acknowledge funding from NSF (2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, R305A120125 & R305R220012), GAANN (P200A180088 & P200A150306), EIR (U411B190024 S411B210024, & S411B220024), ONR (N00014-18-1-2768) and NHI (via a SBIR R44GM146483).

References

Bonotto, C. (2013). Artifacts as sources for problem-posing activities. *Educational studies in Mathematics, 83*, 37-55. https://doi.org/10.1007/s10649-012-9441-7

Butler, A. C., Karpicke, J. D., & Roediger, H. L. III. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 918–928. https://doi.org/10.1037/0278-7393.34.4.918

Castel, A. D., McCabe, D. P., & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review 14*, 107–111. https://doi.org/10.3758/BF03194036

Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., & Brosvic, G. M. (2002). Immediate Feedback Assessment Technique Promotes Learning and Corrects Inaccurate first Responses. *Psychol Rec 52*, 187-201 https://doi.org/10.1007/BF03395423

Enders, N., Gaschler, R., & Kubik, V. (2021). Online quizzes with closed questions in formal assessment: How elaborate feedback can promote learning. *Psychology Learning & Teaching, 20*(1), 91-106. https://doi.org/10.1177/1475725720971205

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906-911. https://doi.org/10.1037/0003-066X.34.10.906

Golke, S., Dörfler, T., & Artelt, C. (2015). The impact of elaborated feedback on text comprehension within a computer-based assessment. *Learning and instruction, 39*, 123-136. https://doi.org/10.1016/j.learninstruc.2015.05.009

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497. https://doi.org/10.1007/s40593-014-0024-x

Jaehnig, W., & Miller, M. L. (2007). Feedback Types in Programmed Instruction: A Systematic Review. *Psychol Rec 57*, 219–232. https://doi.org/10.1007/BF03395573

Kuechler, W. L., & Simkin, M. G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education*, *8*(1), 55-73. https://doi.org/10.1111/j.1540-4609.2009.00243.x

Kuklick, L., Greiff, S., & Lindner, M. A. (2023). Computer-based performance feedback: Effects of error message complexity on cognitive, metacognitive, and motivational outcomes. *Computers & Education*, *200*, 104785. https://doi.org/10.1016/j.compedu.2023.104785

Kuklick, L., & Lindner, M. A. (2021). Computer-based knowledge of results feedback in different delivery modes: Effects on performance, motivation, and achievement emotions. *Contemporary Educational Psychology*, *67*, 102001. https://doi.org/10.1016/j.cedpsych.2021.102001

Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedback complexity and corrective efficiency. *Contemporary educational psychology*, *10*(3), 285-291. https://doi.org/10.1016/0361-476X(85)90025-6

Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010) Enhancing students' self-regulation and mathematics performance: the influence of feedback and self-evaluative standards. *Metacognition Learning 5*, 173–194. https://doi.org/10.1007/s11409-010-9056-2

Liu, S., Wang, Y., Zhan, H. (2022). Perspectives of Instructors and Students on Online Chinese Teaching and Learning in 2020: Preliminary Findings. In: Liu, S. (Ed.), *Teaching the Chinese Language Remotely* (pp. 349–372). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-87055-3_15

Livingston, S. A. (2009). Constructed-Response Test Questions: Why We Use Them; How We Score Them. R&D Connections. *Educational Testing Service*, 11. https://eric.ed.gov/?id=ED507802

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901–918. https://doi.org/10.1037/a0037123

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory & Cognition, 48,* 745–758. https://doi.org/10.3758/s13421-020-01025-5

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky, & S. K. L. Tauber (Eds.), *Oxford handbook of metacognition* (pp. 65-80). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199336746.013.4

Rüth, M., Breuer, J., Zimmermann, D., & Kaspar, K. (2021). The effects of different feedback types on learning with mobile quiz apps. *Frontiers in Psychology, 12*, 665144. https://doi.org/10.3389/fpsyg.2021.665144

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189. https://doi.org/10.3102/0034654307313795

Stone, N. J. (2000). Exploring the Relationship between Calibration and Self-Regulated Learning. *Educational Psychology Review 12*, 437–475. https://doi.org/10.1023/A:1009084430926

Swart, E. K., Nielen, T. M., & Sikkema-de Jong, M. T. (2019). Supporting learning from text: A meta-analysis on the timing and content of effective feedback. *Educational Research Review*, *28*, 100296. https://doi.org/10.1016/j.edurev.2019.100296

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, *4*(4), 295-312. https://doi.org/10.1016/0959-4752(94)90003-5

Sweller, J. (2010). Cognitive load theory: Recent theoretical advances. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 29–47). Cambridge University Press. https://doi.org/10.1017/CBO9780511844744.004

Taxipulati, S., & Lu, H. D. (2021). The influence of feedback content and feedback time on multimedia learning achievement of college students and its mechanism. *Frontiers in Psychology*, *12*, 706821. https://doi.org/10.3389/fpsyg.2021.706821

Urban, K., & Urban, M. (2021). Anchoring Effect of Performance Feedback on Accuracy of Metacognitive Monitoring in Preschool Children. *Europe's Journal of Psychology, 17*(1), 104-118. https://doi.org/10.5964/ejop.2397

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. *Review of Educational Research, 85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment and Evaluation in Higher Education, 31*(3), 379–394. https://doi.org/10.1080/02602930500353061

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in psychology, 10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Ye, X., & Manoharan, S. (September 2018) Machine learning techniques to automate scoring of constructed-response type assessments, *proceedings of 28th EAEEIE Annual Conference, Hafnarfjordur, Iceland.* https://doi.org/10.1109/EAEEIE.2018.8534209

Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is--and is not--a desirable difficulty: the influence of typeface clarity on metacognitive judgments and memory. *Memory & cognition, 41*(2), 229–241. https://doi.org/10.3758/s13421-012-0255-8

Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2022). An automatic short-answer grading model for semi-open-ended questions. *Interactive learning environments, 30*(1), 177-190. https://doi.org/10.1080/10494820.2019.1648300

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

Appendix A
Survey questions

Section 1: Background

1. Are you a K-12 teacher or a college instructor?
   - K-12
   - College
   - both
2. At what language level do you teach? If you teach more than one level, please select the lowest level and answer all the following questions based on this level.
   - Beginning
   - intermediate;
   - advanced
   - high advanced

Section 2: homework assignments and perspectives on providing immediate feedback to close-ended questions, open-ended questions, and semi-open-ended questions

3. How often do you assign homework to students?
   - every day
   - a few times a week
   - about every week
   - about every other week
   - about once a month
   - never (we do all the work in class)
4. What question types do you use most often in your assignments? Please check the top 3:
   - multiple choice
   - true or false
   - matching
   - fill in the blanks
   - translate
   - make sentences to complete dialogues
   - essay writing
   - listen and type
   - listen and record audio/video
   - collaborative/group project
   - forum discussion
   - other: _____
   - other: _____
   - other: _____
5. (1) In general, what percentage of the assignments you give to students are close-ended questions, such as multiple choice and fill in the blanks?
   - choose from the scale: 0%-100%

   (2) What percentage of the assignments you give to students are semi-open-ended short-answer questions? Semi-open-ended short-answer questions require students to express their descriptive answers based on a specified context. Examples in CFL learning are: using the 把 pattern to complete the dialogue, and summarizing the main idea of a text.
   - choose from the scale: 0%-100%

   (3) What percentage of the assignments you give to students are open-ended questions, such as essay writing?

- choose from the scale: 0%-100%

6. When assigning practice online, immediate feedback is an assessment technique that gives students immediate feedback on their answers to a question. To what extent do you think this technique is helpful/necessary?
   - choose from the scale: 0%-100%
7. Which type of immediate feedback to semi-open-ended questions do you believe would be most effective?
   - a working example
   - some explanation
   - both an example and explanation
   - no feedback
   - other, please explain_____
8. How helpful do you think it would be for students to provide immediate feedback to questions that are not auto-gradable, such as semi-open-ended questions?
   - choose from the scale: 0%-100%
   - Optional: please explain_____

9. When assigning homework, which method do you use more often: an online platform with the option to provide immediate feedback, or another format that cannot provide immediate feedback (ex. written homework or email/upload answer sheet)?
   - online platform with the option to provide immediate feedback
   - other formats that cannot provide immediate feedback
   - both
   - neither (we do all the work in class)

Section 3: If assigning homework online or both, they will answer the following questions

10. When assigning homework through an online platform:

(1) Do you use online tools to provide immediate auto feedback to close-ended answer questions that are auto-gradable, such as showing the correct answers to a multiple-choice question immediately after students submit their answers?
   - never, (no need or they get to download the answers right after they hand in their work)
   - occasionally, 0-24% of the auto-gradable problems
   - sometimes, 25%-49% of the auto-gradable problems
   - often, 50%-74% of the auto-gradable problems
   - most of the time, 75%-99% of the auto-gradable problems
   - 100%, always

(2) Do you use online tools to provide immediate suggestions to questions that are not auto gradable, such as completing the dialogue with a required structure? The suggestions can be in the form of a suggested usage of the grammar structures, or a working example answer.
   - never, (no need, or students get to download the answers right after they hand in their work)
   - occasionally, 0-24% of the auto-gradable problems
   - sometimes, 25%-49% of the auto-gradable problems
   - often, 50%-74% of the auto-gradable problems
   - most of the time, 75%-99% of the auto-gradable problems
   - 100%, always

11. How do you provide immediate feedback to questions that are not auto-gradable?
    - a working example
    - some explanation
    - both example and explanation
    - no feedback
    - other, please explain_____

12. When assigning homework without immediate auto feedback, how long does it usually take for students to receive feedback?
    - choose from the scale: never -4 weeks
    - unsure- they get access to the answers after handing in their work

Section 4: If assigning homework without an immediate feedback option, they will answer the following questions

13. How would you rate the helpfulness of providing immediate feedback to questions that are not auto-gradable for students?
    - not at all helpful
    - somewhat unhelpful
    - neutral
    - somewhat helpful
    - extremely helpful
    - Optional: please explain_____

14. Which type of immediate feedback to non-auto-gradable questions do you believe would be the most effective?
    - a working example
    - some explanation
    - both example and explanation
    - no feedback
    - other, please explain

15. When assigning homework, how long does it usually take for students to receive feedback?
    - choose from the scale: 1 day – 2 weeks, Never-they are supposed to correct by themselves

Appendix B

Experimental Materials

Grammar pattern usage 1: Talk about result on an object with 把…… complement/了/一下 /verb reduplication.

Pretest and three posttests:

1. Your parent has just returned home and is checking with you on the progress of your laundry. Tell your parent what you have done or what happened with a 把 sentence.
2. Your parent has just returned home and is checking with you on your birthday party planning progress. Tell your parent what you have done or what happened with a 把 sentence.
3. Your parent has just returned home and is checking with you on the progress of your homework. Tell your parent what you have done or what happened with a 把 sentence.
4. Your parent has just returned home and found that the fridge is all empty, so your parent is checking with you what happened to the food in the fridge. Tell your parent what you have done or what happened with a 把 sentence.

Practices:

1. Give a command to a robot with 把 and ask it to do a chore for you.

    An example of correct response: 把这件衣服洗了。

2. Your friend asks you if you have found the ball you lost yesterday. Use a 把 sentence to answer it.

    An example of correct response: 我把球找到了。

3. Give a command to your little sister with 把 and ask her to finish eating something.

    An example of correct response: 把面条吃完。

Elaborated feedback with grammar explanation for the three practice questions:

a. The basic construction of 把 is as follows: Subject + 把 + Object + Verb + Other Element (Complement/了, etc.) In a 把 sentence, the verb cannot stand alone, and must be followed by another element that describes the detail or the result of the action.

b.把 sentences are most often used to describe what happened to the object in some detail, with the result of the action indicated by the “other element”. Resultative complements, directional complements, 一下, verb reduplication (e.g., 洗洗), and 了 are often used following the verb.

d. When giving commands with 把, the subject is often omitted.

e. The object of a 把 sentence must be something specific and definite.

Grammar pattern usage 2: Talk about placement of an object with 把…… 放在/到

Pretest and three posttests:

1. Your dorm room is a mess. Tell your roommate what to do to help clean it up. A few items for your reference: books, laptop, cell phone, …
2. Your dorm room is a mess. Tell your roommate to help clean it up by putting one of his clothes away. A few items for your reference: pants, shoes, shirts, …
3. Your dorm room is a mess. Tell your roommate to help clean it up by putting one of his food away. A few items for your reference: beverage, cake, apple…
4. Your dorm room is a mess. Tell your roommate what to do to help clean it up. A few items for your reference: chairs, photos, pillows, water bottle, …

Practices:

1. Your friend went grocery shopping with you and helped you to bring all the things back home. Tell your friend where to put one of your grocery items with a 把 sentence.

An example of correct response: 把牛奶放到冰箱里。

2. Your friend went to a shopping mall with you and helped you to bring all the things back home. Tell your friend where to put one of your shopping items with a 把 sentence.

An example of correct response: 把衬衫放在衣柜里。

3. Your friend went to an Apple store with you and helped you to bring all the things back to your home. Tell your friend where to put one of your electronic products with a 把 sentence.

An example of correct response:把电脑放在桌子上。

Elaborated feedback with grammar explanation for the three practice questions:

a. To talk about placement of an object (i.e. to put something in/at a place), 把 structure must be used instead of the Subject+Verb+Object sentence.

b. The construction of 把 sentence that indicating placement is as follows: Subject + 把 + Object + 放在/放到+location.

c. The object of a 把 sentence must be something specific and definite.

d. When 了 is needed to indicate the completion of the displacement, 了 should be placed either at the end of the sentence or after the directional complement of the verb.

Grammar pattern usage 3: Use 把 with two objects

Pretest and three posttests:

1. You are moving to another city and would like to give away your tableware. Tell your friend to whom you are planning to give one of your dorm supplies using a 把 sentence.
2. You are moving to another city and selling furniture. Tell your friend to whom you sold one of your pieces of furniture using a 把 sentence.
3. You are moving to another city and gave your furniture away. Tell your friend to whom you gave one of your pieces of furniture using a 把 sentence.
4. You are moving to another city and selling your dorm supplies online. Tell your friend to whom you sold one of your dorm supplies using a 把 sentence.

Practices:

1. You are eating dinner with your friend. Ask your friend to pass something to you with a 把 sentence.

An example of correct response: 把盘子拿给我。

2. Your friend saw you holding a gift, he asks whom you are giving the gift to. Answer the question with a 把 sentence.

An example of correct response: 我要把礼物送给我的妈妈。

3. You sold some of your used furniture, and your friends asks whom you sold them to. Please answer the question with a 把 sentence.

An example of correct response: 我把桌子卖给一位大学生了。

Elaborated feedback with grammar explanation for the three practice questions:

a. For certain verbs, you can have two objects in a 把 sentence. Their use in a 把 sentence will also involve prepositions, and 给 is the most often used. They use the following structure: Subj. + 把 + Obj. 1 + Verb + 给 + Obj. 2

b. Common verbs that take two objects include: 送 (sòng, send), 拿 (ná, take), 递 (dì, pass), 卖 (mài, sell), 借 (jiè, borrow/lend), 还 (huán, return), 介绍 (jièshào, introduce).

c. When 了 is needed to indicate the completion of the action, 了 can be placed either at the end of the sentence or right after 给. When indicating that an action will happen, 要/想/会 should be used before the verb. When used as a command, no 了 or 要/想/会 is needed.