Current Psychology
https://doi.org/10.1007/s12144-024-06481-9

Testing effect in L2 discourse comprehension: importance of retrieval-based learning

Dandan Liu[1] · **Tong Zheng**[2] · **Yu Song**[3]

Accepted: 26 July 2024

Abstract
In recent years, a large body of research has found that testing proves to be a more efficacious method for improving learning when juxtaposed with simply restudying material. Nevertheless, there remains a lack of understanding regarding whether the testing effect extends to the comprehension of L2 discourse, as well as the factors that mediate this effect. This study sought to examine the impact of different relearning strategies and text complexity on varying test formats in L2 discourse comprehension. A total of 150 EFL university students were divided into four groups: rereading group, self-explanation group, free recall group and free recall combined with self-explanation group. All participants studied two texts with differing levels of complexity. After one week, all of them underwent a final test. In both verbatim and inference questions, the free recall group and the free recall combined with self-explanation group proved to be more effective in enhancing learning performance than the other two restudying groups, which highlights the significance of retrieval-based learning in the context of L2 discourse comprehension. Furthermore, the superior performance of the free recall combined with self-explanation group in answering inference questions of the complex text validated the advantages of combining elaborative strategy with testing. This finding is consistent with the complementary view, indicating that retrieval-based elaboration prompts both constructive and consolidative processing of complex information, thereby fostering a deeper comprehension of the text compared to retrieval alone. These results suggest that relearning strategies, text complexity and test formats are all important boundary conditions for the testing effect in L2 discourse comprehension.

Keywords Relearning strategies · Text complexity · Test formats · L2 discourse comprehension · Testing effect

Introduction

For an extended period of time, the prevailing belief has been that learning constitutes the process of encoding novel information, whereby the freshly encoded information becomes stored within memory and might be subsequently retrieved as needed. Thus, encoding is considered to be a crucial factor affecting the effectiveness of memory, while retrieval is simply the process of accessing the stored information and a neutral method for assessing the quality of encoding. However, a large body of recent research (see Adesope et al., 2017; Agarwal et al., 2021; Rowland, 2014, for review and meta-analyses) has shown that, compared to restudying material, taking a test after the initial learning of knowledge can lead to better long-term memory retention. This phenomenon is termed as testing effect (Roediger & Karpicke, 2006). This "testing effect" is driven by the retrieval processes that learners engage in when they take tests, and thus the key phenomenon is referred to as retrieval-based learning (Karpicke, 2017).

The conventional research paradigm for investigating the testing effect encompasses three sequential phases: an initial learning phase, an intervening phase, and a final testing phase (Rowland, 2014). In the initial learning phase, participants across all groups assimilate the study material for the first time. Subsequently, in the intervening phase, participants employ diverse learning strategies to reinforce their understanding of the material. While individuals in the restudying group repetitively study or elaborate on the

✉ Dandan Liu
443614849@qq.com

[1] School of Foreign Studies, Jiangnan University, Wuxi, China

[2] College of Pharmacy and Health Sciences, Yancheng Polytechnic College, Yancheng, China

[3] Business English Department, Wuxi Vocational College of Science and Technology, Wuxi, China

Published online: 29 August 2024

learning material, those in the testing group engage in an initial test to retrieve information from memory. In the final testing phase, all participants undergo an evaluative test to assess their learning outcomes. The testing effect manifests when the testing group outperforms the restudying group on the final test. This established paradigm has served as a framework for researchers to extensively investigate the testing effect. Indeed, research has consistently shown this effect across a wide range of age groups, diverse learning materials, varying retention periods, and different test formats (e.g., Blunt & Karpicke, 2014; Carpenter, 2012; Greving & Richter, 2018; Karpicke et al., 2016; Rawson & Zamary, 2019; Toppino & Cohen, 2009).

Despite this tremendous effort and progress, there is considerable room for additional progress in addressing several puzzles, inconsistencies, and gaps in the research on the testing effect (Karpicke, 2017). Current investigations into the testing effect are centered on exploring various relearning strategies during the intervening phase. Support for the testing effect primarily stems from comparisons between testing and repetitive learning paradigms (e.g., Karpicke et al., 2016; Shobe, 2021). Although some studies suggest that testing enhances learning outcomes more effectively than traditional elaborative learning techniques (e.g., Blunt & Karpicke, 2014; Lechuga et al., 2015), further empirical investigation is warranted to substantiate these findings. Additionally, research on the combined impacts of testing with other learning strategies, such as elaborative encoding activities, remains relatively scarce and the results are subject to debate (e.g., Endres et al., 2017; Roelle & Nückles, 2019). Another crucial area of inquiry involves whether the benefits of retrieval practice vary depending on the types of learning materials. Numerous studies have demonstrated the testing effect with simple materials, such as word pairs (e.g., Barcroft, 2007; Yang et al., 2017). However, there is less understanding of the power of retrieval to enhance the arguably more important outcomes of understanding and comprehension of complex, educationally relevant learning content (Endres et al., 2017), including texts. Some studies suggest that the effect persists with more complex materials (e.g., Karpicke & Aue, 2015), while others argue that it diminishes or even disappears (e.g., Gog & Sweller, 2015). To address these gaps, there are several potential avenues for further research on retrieval-based learning. Firstly, instead of aiming for a broad generalization of the testing effect across learners, materials, and assessments, it is crucial to make more efforts to identify whether testing effects are invariant or interact with various factors. These factors may include types of learning strategy, features of materials or aspects of final assessments. Additionally, a deeper understanding of the mechanisms underlying retrieval-based learning is crucial (Karpicke, 2017).

Discourse comprehension is an ability to understand ideas and the relationships between ideas conveyed via written information or oral texts (McNamara, 2007). It involves the concurrent extraction and construction of meaning as individuals engage with textual material (Watter et al., 2016). During text comprehension, individuals are capable of forming three distinct levels of representation: surface representation, text-based representation, and situational model representation (Kintsch, 1994, 1998). The surface representation consists of the wording and syntax of the clauses extracted from the text. This level of representation usually lasts for a few seconds. At the text-based level, the meaning of words and sentences is explicitly processed and stored in the reader's working memory. Situational model representations refer to a coherent non-linguistic mental representation of the state of affairs described in the text, rather than a mental representation of the text itself (Zwaan & Radvansky, 1998). Constructing a situation model relies not only on the textual content but also on the integration of distant segments of the text and the inferences drawn from prior knowledge. These representations, therefore, reflect readers' comprehensive and in-depth understanding of the text. Essentially, discourse comprehension is a complex activity that relies upon a range of perceptual, cognitive, and linguistic skills and processes (Watter et al., 2016). In the field of second language (L2) acquisition, discourse serves as a vital resource, providing authentic linguistic content and facilitating the effective acquisition and practical application of the language. Compared to the first language (L1) discourse comprehension, L2 discourse comprehension is more complex due to the dynamic interplay of a dual language system, which requires the simultaneous activation of both L1 and L2 comprehension mechanisms. Specifically, challenges in L2 discourse comprehension include navigating vocabulary and grammar nuances, grasping context, understanding cultural differences, managing cognitive load, and addressing inadequate learning strategies. Consequently, achieving a profound understanding of discourse, especially at higher-level representation, presents a significant obstacle for L2 learners.

Despite numerous studies examining the testing effect in foreign language vocabulary acquisition or native language texts, its application and characteristics in L2 discourse comprehension remain relatively unexplored. This study aims to explore the presence of a testing effect in L2 discourse comprehension and identify factors that may either diminish or enhance this effect. By doing so, it seeks to provide valuable insights with practical implications for L2 learning and education, as well as to achieve a more in-depth examination of the mechanisms underlying the testing effect. Specifically, this study investigates whether and

how testing can effectively facilitate and optimize various dimensions of L2 discourse comprehension across different relearning strategies and levels of text complexity. Drawing from prior research, the hypotheses of this study are formulated as follows: (1) Testing will yield superior L2 discourse comprehension compared to traditional methods such as repeated reading and elaborative learning. (2) Integrating testing with an elaborative learning strategy will enhance L2 discourse comprehension more effectively than using a single learning approach. (3) The effectiveness of the testing effect in L2 discourse comprehension varies depending on the interaction with the complexity of the text.

Relevant theories and studies

Effects of relearning strategies on testing effect

The critical factor determining the testing effect lies in the relearning strategies employed during the second phase, i.e., the intervening phase. The testing effect is evident if performance on tests exceeds that of other learning methods during this phase. A significant body of research has focused on the comparison of testing and repetitive learning and found a memory advantage of testing (e.g., Agarwal et al., 2008; Karpicke et al., 2016; Li et al., 2022; Roediger & Karpicke., 2006; Shobe, 2021). According to the Desirable Difficulties Theory (Bjork, 1994, 1999; Bjork & Bjork, 1992), deliberately introducing challenges and decelerating the learning process can optimize long-term memory and facilitate knowledge transfer. These demanding conditions are referred to as desirable difficulties. The deliberate effort devoted to actively retrieving information is regarded as a valuable task since it boosts the strength and durability of knowledge, thereby raising the likelihood of future accessibility and application.

In contrast to repetitive learning, elaborative learning is widely recognized as an effective and efficient encoding strategy and frequently applied in real-world educational settings. This approach goes beyond mere surface-level memorization and places emphasis on the formation of higher-level mental representations by connecting new information with prior knowledge. Through the creation of meaningful connections and associations, elaborative learning is thought to enhance understanding, boost memory retention, and enable the transfer of knowledge in diverse contexts (Fiorella & Mayer, 2016). Considering that both retrieval practice and elaborative learning require substantial engagement and effort from the learner and are recognized as effective learning strategies, an important question emerges: Does the testing effect persist when contrasting retrieval practice with elaborative learning?

Presently, there are divergent theoretical perspectives and empirical findings on this issue. Carpenter's Elaborative Retrieval Hypothesis (2009) proposes that the process of retrieval not only activates the target information but also prompts the activation of related semantic knowledge. This integrated activation helps construct a more extensive and elaborative semantic network that serves as a potent cue for subsequent retrieval. This hypothesis receives support from several empirical studies. For instance, Carpenter's (2011) study confirmed the elaborative function of retrieval by demonstrating that cued-recall practice tests not only improved retention of the target information (e.g., 'child' in the pair 'mother: child'), but also bolstered retention for words strongly associated semantically with the pair (e.g., 'father'). Additionally, Carpenter's (2009) study showed that initial recall with weak associates produced a more pronounced testing effect than with strong associates, suggesting that scenarios characterized by restricted retrieval support and heightened effort encourage greater elaboration, which in turn promotes long-term retention. Thus, in line with the Elaborative Retrieval Hypothesis, it can be postulated that there is no discernible distinction between the testing and elaborative learning strategies, as the mechanism facilitating learning through retrieval is essentially characterized by elaborate processing.

In contrast, according to the Episodic Context Account (Karpicke et al., 2014), testing and elaborative learning contribute to memory enhancement through distinct mechanisms. The successful retrieval of an item during the initial testing leads to the updating of its contextual representation, encompassing features from both the original study context and the current test context. Subsequently, the refined context can narrow the search scope during subsequent retrievals, empowering learners to eliminate competing candidates and effectively retrieve target items. Karpicke and Blunt (2011) demonstrated that retrieval practice (recalling material) outperformed elaborative studying (concept mapping) in a final short-answer test conducted a week later. This unexpected finding challenged the traditional emphasis on elaboration in education and sparked increased research into testing strategies. Further studies highlighted the potential superiority of retrieval practice over elaborative methods, such as concept mapping (Blunt & Karpicke, 2014; Lechuga et al., 2015; Zhou et al., 2013), the image-based keyword approach (Karpicke & Smith, 2012), and verbal elaboration (Goossens et al., 2014).

Given that both testing and elaboration are effective learning techniques, another question emerges: Can the combination of retrieval practice and elaborative learning optimize the testing effect? The integration of retrieval practice and elaborative learning can occur simultaneously or sequentially, with retrieval practice preceding elaborative

processing or vice versa. This article specifically focuses on instances where retrieval practice and elaborative learning strategies are combined simultaneously.

In theoretical terms, the effectiveness of combining retrieval practice and elaborative learning depends largely on whether their underlying cognitive processes complement each other or redundantly overlap. As posited by Carpenter's Elaborative Retrieval Hypothesis (2009), the act of retrieval engages cognitive processing akin to elaboration in the learning context, leading to a richer and more varied encoding of the studied information and thereby establishing multiple pathways for subsequent retrieval. However, the redundancy view posits that if the cognitive processes involved in elaborative encoding overlap with or duplicate those activated during retrieval practice, combining these two approaches may not necessarily yield additional or cumulative benefits (Cummings et al., 2023; McDaniel, 2023; Miyatsu & McDaniel, 2019).

On the contrary, some scholars argue that elaborative learning functions to construct coherent and robust mental representations by activating prior knowledge and establishing connections with new material (McDaniel, 2023; Roelle et al., 2022). Retrieval practice, on the other hand, primarily functions to strengthen later retrievals (McDaniel, 2023) or fosters the consolidation of learners' mental representations and hence long-term retention (Roelle et al., 2022). Thus, integrating both strategies can improve the quality and accessibility of mental representations. This synergy is expected to bring more benefits than retrieval practice alone (Cummings et al., 2023; Fritz et al., 2007; McDaniel, 2023; Roelle et al., 2022), especially promoting meaningful learning with deep comprehension and knowledge application in various contexts (Roelle et al., 2022). Hinze et al. (2013) showed this complementary view for complex learning situations with text and introduced the Constructive Retrieval Hypothesis. This hypothesis suggests that retrieval's primary function is to focus rather than broaden relevant information. In contrast, elaboration is seen as a generative learning activity involving the assimilation of information in the text and its adaptation to existing schemata. Therefore, while retrieval may not inherently promote the construction of solid and coherent situational model representations, elaboration can enhance the integration of mental representations with learners' prior knowledge (Fiorella & Mayer, 2016) by facilitating rich inferences and connections.

Although the integration of elaborative processing and retrieval practice may offer complementary effects due to their distinct underlying mechanisms, this does not necessarily guarantee consistent improvements in learning outcomes. According to the Desirable Difficulties Theory (Bjork, 1994, 1999; Bjork & Bjork, 1992), when learners engage in both retrieval practice and elaborative activities simultaneously during the initial testing stage, the retrieval task becomes more challenging. This potentially amplifies the testing effect through increased cognitive effort and desirable difficulties. Nevertheless, it is crucial to note that retrieval effort and retrieval success often oppose each other. Emphasizing an increase in retrieval effort in learning activities is likely to lead to lower retrieval success (O'Day & Karpicke, 2021). In fact, not all difficulties in learning are desirable. Some learning conditions that make the process more challenging only impose an extraneous cognitive burden on learners (Richter et al., 2022). When the cognitive demands of the learning task exceed an individual's total cognitive resources, there is an issue of improper cognitive resource allocation, resulting in cognitive overload that adversely affects learning (Kahneman, 1973). Therefore, when considering the integration of retrieval practice and elaborative activities, it is essential to explore ways to enhance retrieval success without compromising retrieval effort and to establish conditions within the cognitive load that learners can handle (O'Day & Karpicke, 2021).

Empirical discoveries about the impact of the combined approach are mixed. Echoing the complementary view outlined earlier, a limited body of research (e.g., Endres et al., 2017; Hinze et al., Exp. 3, 2013; Lachner et al., Exp. 2, 2021) suggests that when handling textual information, the combination of tests with elaborative learning strategies, like free recall plus self-explanation with examples, may lead to superior learning performance on comprehension-based questions that demand participants to make inferences, compared to testing alone. Contrarily, an emerging body of research appears to align with the redundancy view, asserting that the effective processes elicited by the elaborative technique likely overlap with those associated with retrieval practice. For instance, studies by Fritz et al. (2007, Exp. 3) and Karpicke and Smith (2012, Exp. 1 and 2) found no discernible advantages in combining the keyword method with retrieval practice compared to practicing retrieval alone in the context of vocabulary learning. Similarly, in the domain of text learning, Blunt and Karpicke (2014, Exp. 2) and O'Day and Karpicke (2021, Exp. 1) demonstrated that two different retrieval formats, specifically free recall and creating a concept map without reference to texts, proved to be equally effective as learning tools. Furthermore, Rummer et al. (2017) failed to discover benefits arising from the combination of note-taking with free recall. Supporting this perspective, Roelle and Nückles (2019) similarly found no apparent advantages linked to the incorporation of highlighting connections of main content and providing examples during the retrieval process. Additionally, Larsen et al. (2013) and Lachner et al. (2021, Exp. 1) suggested that involving students in additional explaining activities is not more effective than engaging in retrieval practice.

Effects of material complexity on testing effect

The quality of learners' mental representations at the end of the initial study phase is expected to significantly moderate the benefits of the learning activities in the second study phase (Roelle & Nückles, 2019). Specifically, the degree of difficulty or complexity of the learning materials directly influences how learners mentally construct and represent the information during the first study phase. This mental construction and representation, in turn, then determines the degree of success they achieve in the testing during the second study phase and affects the extent of the testing effect.

A recent debate has arisen regarding the occurrence of testing effect with complex materials. While existing studies have predominantly utilized simple learning materials, such as word pairs (e.g., Barcroft, 2007; Cho et al., 2017; Karpicke & Roediger, 2007; Yang et al., 2017), to elucidate the testing effect, some investigations propose a potential decrease or even disappearance of the testing effect as the complexity of learning materials escalates (e.g., De Jonge et al., 2015; Hanham et al., 2017; Leachy et al., 2015; Van Gog & Sweller, 2015). Element interactivity, a metric quantifying the number of elements that a learner must simultaneously process in their working memory (Sweller, 1988; Sweller & Chandler, 1994), serves as the criterion employed by Van Gog and Sweller (2015) to define the complexity of learning materials. According to certain scholars (e.g., Hanham et al., 2017; Van Gog & Sweller, 2015), elevated levels of element interactivity within learning materials may impede the efficacy of the testing effect. To optimize learning, it's crucial to reduce element interactivity, which allows each element to be learned independently, eliminating the need for reference to others and thus leading to a deeper understanding of the material and a solid foundation for successful information retrieval. Nonetheless, Karpicke and Aue (2015) argued that the study by Van Gog and Sweller (2015) failed to provide a quantitative and measurable definition of element interactivity, and the complexity of the materials was not effectively distinguished from that of the task. Additionally, several empirical studies that opposed the testing effect in complex materials employed problem-solving as a retrieval task, such as how to interpret and use a bus table (Leachy et al., 2015). However, it proved difficult to ensure that participants truly retrieved information from worked examples before solving the problem. In contrast, a considerable amount of literature using textual materials suggests that the testing effect persists in complex contexts (e.g., Eglington & Kang, 2018; Hinze & Wiley, 2011; Karpicke & Aue, 2015; Karpicke & Blunt, 2011; Rawson & Katherine, 2015; Rowland, 2014).

Direct comparisons of retrieval practice effects with different materials are infrequent but could provide theoretical insights into mechanisms of retrieval-based learning (Karpicke, 2017). Roelle and Nückles (2019) employed a research design featuring two expository texts. One text exhibited high cohesion and elaboration, while the other displayed low cohesion and elaboration. Regarding the highly cohesive and elaborated text, only the group engaged in retrieval practice (free recall) exclusively outperformed the rereading group in comprehension questions. Presumably, this is because learners created a coherent mental representation of the text in the first study stage, and retrieval was able to consolidate this mental representation. For the expository text marked by low cohesion and elaboration, distinct results emerged. Elaboration alone (highlighting the connections of the main content and providing examples) had a significant effect on comprehension and transfer, likely because the elaborative learning activity encouraged learners to utilize their prior knowledge to enhance cohesion and elaborate on the learning material. However, involving learners in elaborative activities proved ineffective when they were simultaneously engaged in retrieval practice.

Research methods

Objectives of the research

The primary aim of this research was to explore the phenomenon and underlying mechanisms of the testing effect in L2 discourse comprehension. Specifically, three crucial objectives were established. Firstly, through comparing testing with traditional learning approaches, namely repeated reading and elaborative learning strategy during the intervening phase, the aim was to determine the presence of a testing effect in L2 discourse comprehension. Secondly, it was explored whether the combination of testing and elaborative learning strategy could facilitate L2 discourse comprehension more effectively than using a single learning strategy. Finally, it was examined whether the potential findings were dependent on text complexity.

Research design

To address these objectives, the experimental design in the second learning phase involved the use of diverse relearning strategies. These included two traditional restudying strategies: a repetitive learning condition (rereading, denoted as RR), an elaboration-only condition (self-explanation, denoted as SE), and two retrieval-based strategies: a retrieval practice-only condition (free recall, denoted as FR), as well as a combined condition integrating both elaborative encoding and retrieval practice (the combination of free recall with self-explanation, denoted as FR + SE).

Additionally, the study utilized two texts with varying levels of complexity.

Participants

The study included first-year students majoring in English from six parallel classes at Jiangnan University in Jiangsu Province, China. The study received approval from the School of Foreign Studies at Jiangnan University, China. Participants volunteered for the experiment and were rewarded with six credits. Informed consent was obtained from all participants, and for those aged 18 and below, consent was also acquired from their parent(s) or guardian(s).

Due to various reasons such as mid-term withdrawals from the experiment and sub-optimal audio recording quality, the final number of participants included in the analysis was 150. The participants, aged between 17 and 20, shared Chinese as their mother tongue and had no prior overseas experience. All had a minimum of six years of English language study and had successfully passed the Test for English Majors-Band 4 (Mean = 73.61, SD = 6.31). As per China's Standards of English Language Ability (2018), they were categorized as intermediate English learners, possessing certain language comprehension and self-assessment abilities. To ensure homogeneity, participants were randomly assigned to four groups: 37 in the rereading group, 38 in the self-explanation group, 38 in the free recall group, and 37 in the free recall combined with self-explanation group. According to Endres et al. (2017), an interaction effect exists between relearning strategies and final test type. A priori power analysis was performed using G*Power version 3.1 for sample size estimation. With $a = 0.05, 1\text{-}\beta = 0.95, f = 0.25$ (medium effect size), a minimum of 76/4 = 19 individuals are required for each group to reach the interaction effect. Thus, the sample size was adequate for the study. Before the main experiment, all participants underwent a language proficiency test. A one-way ANOVA test indicated no significant difference in L2 proficiency among the four groups, $F_{(3, 146)} = 0.038$, $p = .990$.

Materials

The learning materials were two texts of different levels of complexity. In line with the work of Karpicke and Aue (2015), referential cohesion, which measures the degree of overlap and connectivity of ideas across sentences within a text, was employed to quantify text complexity. Coh-Metrix, an automatic text analysis tool developed by Graesser et al. (2004), was used to measure various text characteristics, including referential cohesion. Karpicke and Aue (2015) suggested that if referential cohesion scored above the 70th percentile, the material could be considered complex. Analysis using Coh-Metrix revealed a referential cohesion percentile of 6.43 for the simple text and 82.89 for the complex text, indicating a significant difference in complexity between the two texts.

The simple text, comprising 492 words, explored the "Learning paradox," emphasizing that the more one struggles or even fails when attempting to grasp new information, the more likely they are to remember and apply it later. The complex text, with 475 words, delved into "Decision fatigue," illustrating that our brain has a limited capacity for decision-making, leading to a decline in our ability to make sound choices. To mitigate the influence of prior knowledge on the experimental results, participants were asked to rate their familiarity with the texts on a scale from 0 to 10 after the main study. Those participants who demonstrated a high level of familiarity with the learning materials were excluded from the data analysis. Results indicated that both topics were perceived as fresh by participants (Simple text: Mean = 1.73; Complex text: Mean = 1.77). No significant differences were observed among the four groups regarding their prior knowledge of the simple text, $F_{(3, 146)} = 0.048$, $p = .986$ and the complex text, $F_{(3, 146)}$, $p = .841$.

Instruments

Initial test

The initial test for the FR group and the FR + SE group in the intervening phase was a free recall test. Two experienced college English teachers divided the simple text into 45 idea units and the complex text into 43 idea units. Each idea unit was worth 1 point. Two experimental assistants counted the number of idea units recalled, and the scores of two assistants were averaged. An inter-rater reliability test ($r = .903$, $p > .001$) demonstrated that the two assistants' scores were highly correlated to each other.

Final test

The study employed two distinct test formats—verbatim questions and inference questions—to assess text-based retention and situational-level comprehension in the final evaluation. Verbatim questions were intentionally designed to gauge lower-level cognitive processing, requiring a straightforward recall of information from the text. These questions could be answered by executing a basic search of memory. In contrast, inference questions were formulated to necessitate the synthesis of information from both inside and outside the provided text, as the answer was not explicitly stated in the passage. These questions might prompt learners to predict outcomes in a novel situation, choose an appropriate explanation for a phenomenon, or

select the correct sequence of events in a scientific process. Therefore, these inference items were crafted to demand the types of inferences crucial for achieving situation-level comprehension of texts (Graesser et al., 2002; Hinze et al., 2013; Kintsch, 1994; Wiley et al., 2005). Both types of tests have been employed in prior studies to gain insights into the impact of tests on different levels of text comprehension (e.g., Endres et al., 2017; Hinze et al., 2013).

窗体底端

The final assessment comprised 16 short answer questions, with 8 questions corresponding to each text and two distinct test types. Among these, six questions were verbatim in nature, specifically designed to evaluate participants' memory retention. An illustration of a verbatim question from the simple text was, "What's the function of struggle in learning?" The answer, directly retrievable from the text, was, "It leads people to understand the deep structure of problems, not simply their correct solutions. When these students encounter a new problem of the same type on a test, they're able to transfer the knowledge they've gathered more effectively than those who receive someone else's expertise passively." The remaining two questions in each set were inference questions crafted to assess participants' transfer skills. For instance, an inference question from the complex text was, "If two cake shops are promoting their new product, one shop provides 6 types of cake while another shop provides 24 types of cake samples to shoppers, which shop do you think will have better selling? Why?" Since the answer was not explicitly provided in the text, participants were tasked with making inferences based on their profound understanding of the topic, providing well-founded interpretations to support their answers.

The aggregate score for the 8 short answer questions corresponding to each text was set at 16 points, with each individual question carrying a value of 2 points. To accommodate variations in participants' comprehension levels and to better analyze differences in learning conditions, we adopted a partially correct scoring approach. Specifically, we divided the response to each verbatim question into four main parts, assigning 0.5 points to each part. Participants earned a score based on the number of accurately mentioned parts, with potential scores ranging from 0.5 to 2 points. For each inference question, the score was determined by assessing the rationality of the participant's choice and explanation. Two experimental assistants independently evaluated the scores of the final short-answer questions, and the scores assigned by the two assistants were averaged. A test of inter-rater reliability ($r = .917$, $p > .001$) indicated a high correlation between the scores provided by the two assistants.

Post-experimental interview

A post-experimental interview was conducted to primarily assess the cognitive state of learners during both the initial learning phase and the intervening phase under different learning conditions. Key areas of focus included text comprehensibility, the adequacy of reading time, levels of mental effort, difficulties encountered during reading, and the solutions applied.

Procedures

The two-week main experiment took place in a computer laboratory. All participants were provided with a computer and headphones to accomplish all tasks solely through the use of the computer. The entire experimental process was orchestrated through software developed in the Python language. Instructions for the experiment were displayed on the computer screen, and participants were instructed to click "Next" to advance to the subsequent step after comprehending the instructions thoroughly. Before initiating the primary experimental procedures, participants in both the SE group and the FR + SE group underwent training on how to generate self-explanations through practical exercises. All participants were tasked with learning two distinct texts featuring varying degrees of complexity, and the sequence of presenting the simple and complex texts was counterbalanced.

The experiment adhered to the conventional research paradigm of the testing effect, encompassing three distinct phases: the initial learning phase, the intervening phase, and the final testing phase. In the initial learning phase, participants were presented with a text for 8 min and informed about a final test scheduled for one week later. During the intervening phase, participants underwent a 4-minute relearning session utilizing different learning strategies. Specifically, participants in the RR group reread the text, while those in the SE group explained the content of the text with examples from their own lives, studies, or work during the rereading process. On the other hand, the two testing groups, the FR group and the FR + SE group, did not have access to the text. Participants in the FR group were tasked with recalling as much of the previously learned text as possible, while the FR + SE group was instructed to not only recall the text but also to simultaneously explain the recalled information with examples from their own lives, studies, or work. The free recall and/or explanation for the SE group, FR group, and FR + SE group were conducted individually and orally, with the recorded oral protocols later transcribed for analysis.

After learning the first text, participants engaged in a 15-minute video distraction task before being introduced to

the second text. The experimental procedures for the second text mirrored those for the first text. Upon completing the initial learning phase and intervening phase, participants received the post-experimental interview. The final testing phase occurred one week later, during which all participants assessed their memory retention and inference for the two texts they had previously learned.

Results

A mixed-design analysis of variance (ANOVA) was performed using SPSS 18.0, incorporating a 4 (relearning strategy: RR, SE, FR, or FR + SE) x 2 (text complexity: simple or complex) x 2 (test format: verbatim question or inference question) factorial structure. Relearning strategy was considered a between-subjects factor, while text complexity served as a within-subjects factor. The statistical significance threshold was set at an alpha level of 0.05. Subsequent to identifying significant interaction effects, post hoc pairwise comparisons were conducted for further decomposition. Effect sizes were quantified using Cohen's d, where 0.2 denoted a small effect, 0.5 signified a moderate effect, and 0.8 represented a large effect.

To directly compare the two types of testing, we calculated the percentage of correct answers based on the test scores, taking into account the varying numbers of verbatim and inference questions. The percentage of free recalled idea units in the initial test is presented in Fig. 1. Given that the initial test scores demonstrated homogeneity of variance across comparison groups, as indicated by a Levene's test, and the obtained initial scores on the continuous variables were normally distributed based on measures of skewness, an independent samples t-test was utilized to compare the disparities among the two retrieval groups on the initial test. The result showed that there were no significant differences between the FR group and the FR + SE group with respect to the percentage of free-recalled idea units of the simple text ($t = 6.97$, $p = .996$), the complex text ($t = 5.67$, $p = .98$) and total texts ($t = 5.23$, $p = .773$).

Figure 2 displays the means for the percentage of correctly answered verbatim questions within each group. In Fig. 3, the data pertaining to inference questions is reported. Since the final test scores for both verbatim questions and inference questions indicated homogeneity of variance across comparison groups via a Levene's test, and the scores obtained on the continuous variables were normally distributed as evaluated by skewness, a $4 \times 2 \times 2$ ANOVA was employed to explore the boundary conditions of testing effect in L2 discourse comprehension. Table 1 illustrates the Bonferroni post hoc pairwise comparisons of results for verbatim questions, and Table 2 outlines the Bonferroni post hoc pairwise comparisons of results for inference questions.

A 4 (relearning strategy) x 2 (text complexity) x 2 (test format) mixed ANOVA analysis revealed that the main effect

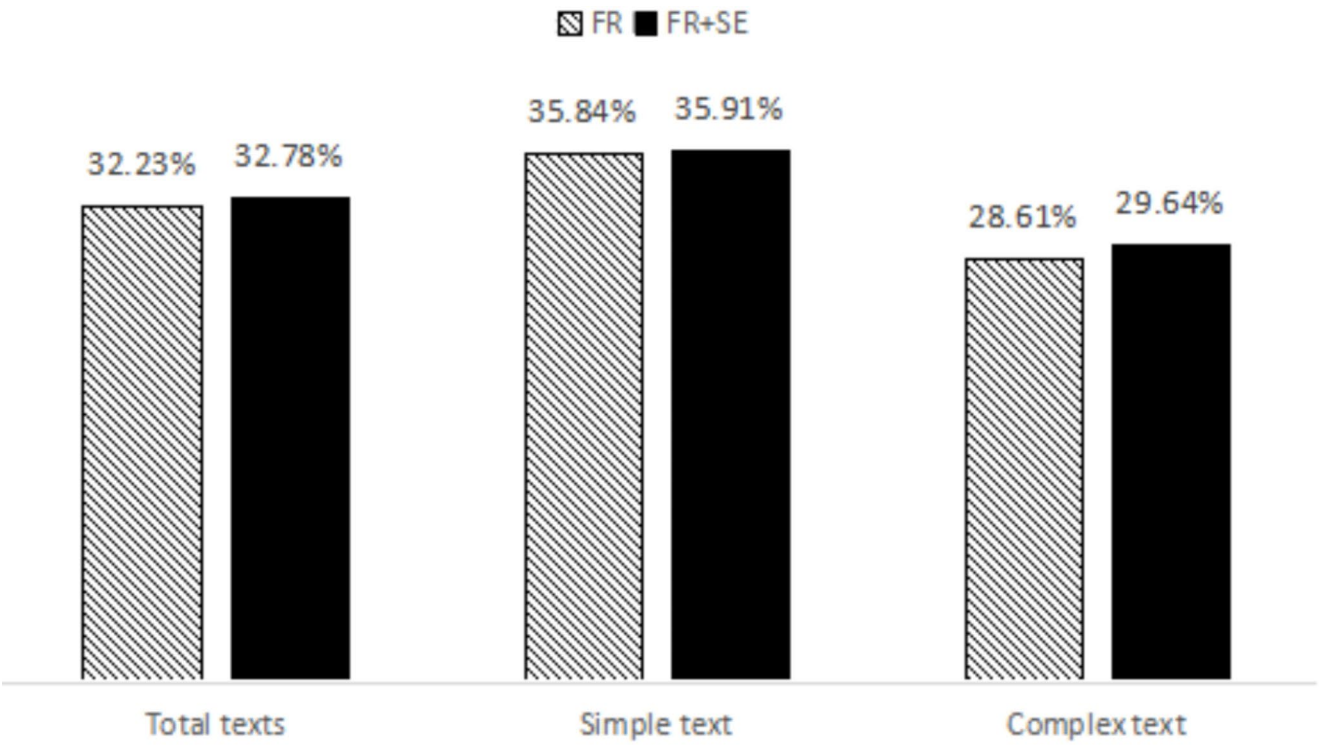

Fig. 1 Percentage of free-recalled idea units in the initial test

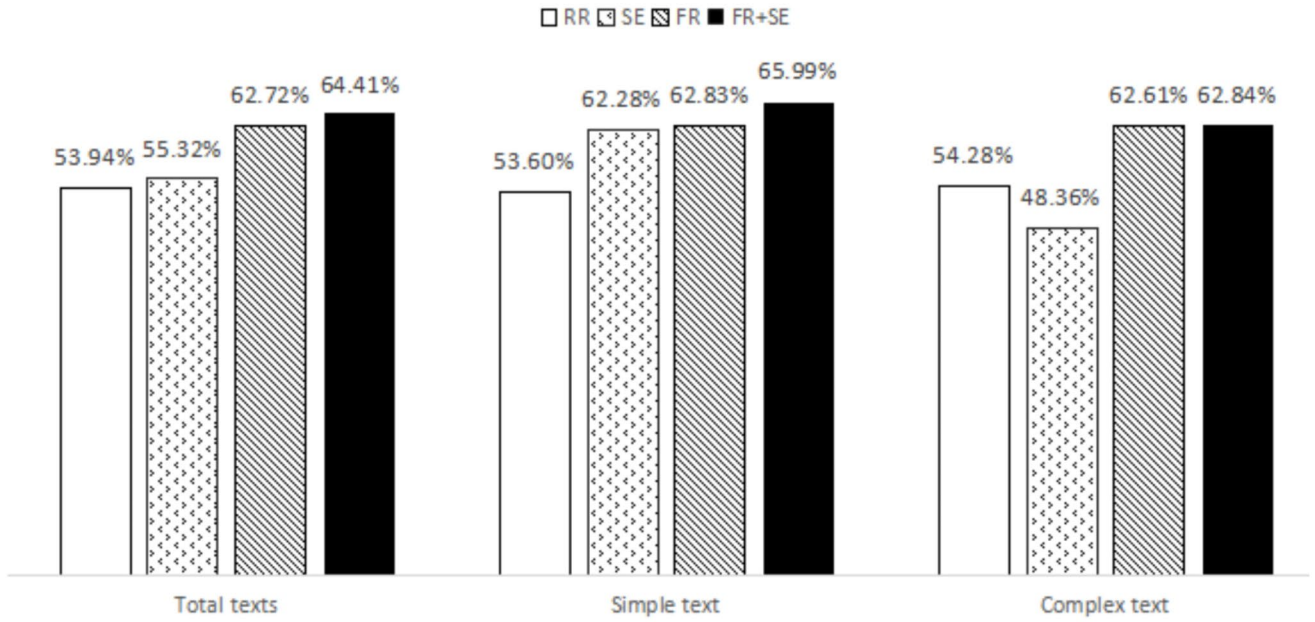

Fig. 2 Percentage of correctly answered verbatim questions in the final test

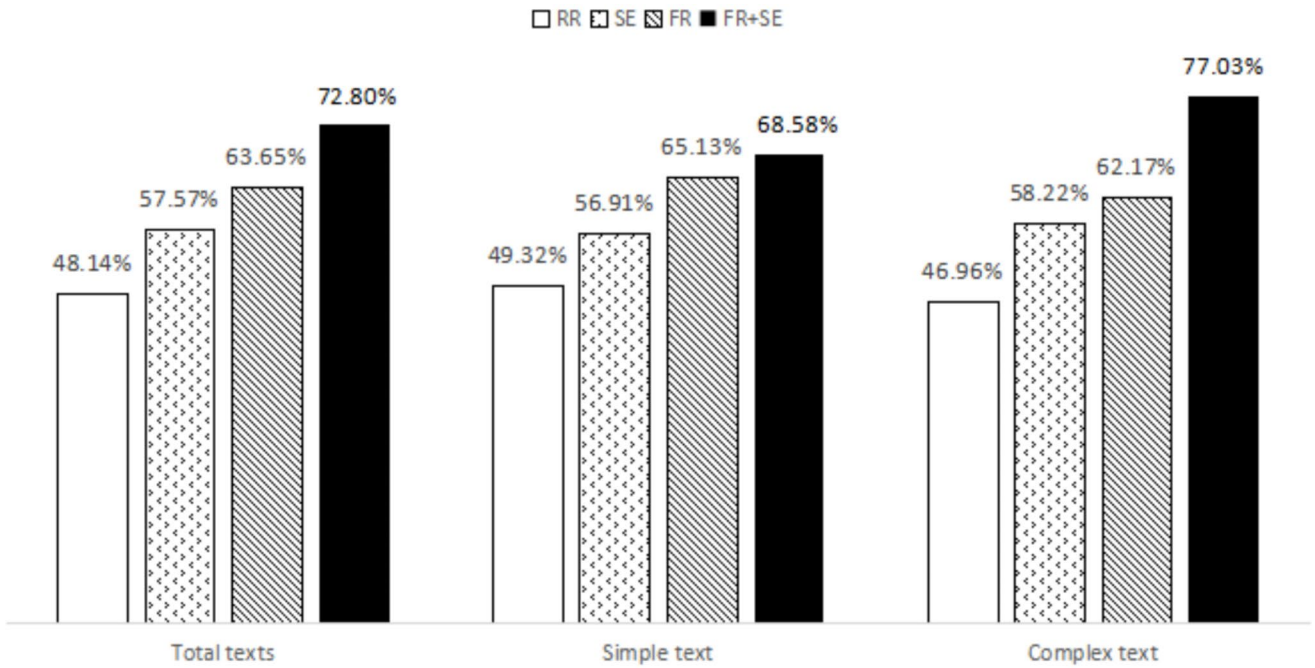

Fig. 3 Percentage of correctly answered inference questions in the final test

of relearning strategy was significant, $F_{(3,\ 146)} = 20.165$, $p > .001$. There was a significant two-way interaction between relearning strategy and test format, $F_{(3,\ 146)} = 6.559$, $p > .001$, and between text complexity and test format, $F_{(1,\ 149)} = 5.496$, $p = .02$.

In terms of the performance of verbatim questions, Bonferroni post hoc pairwise comparisons revealed that the FR group performed significantly better than the RR group. The FR + SE group was significantly better than the RR group and the SE group. Nevertheless, no statistically significant differences were identified between the RR group and the SE group, the SE group and the FR group, as well as the FR group and the FR + SE group.

In the inference questions, the score of the RR group was significantly lower than that of the SE group, the FR group and the FR + SE group. The FR + SE group performed

Table 1 Bonferroni post hoc pairwise comparisons of results of verbatim questions

		Total texts		Simple text		Complex text	
Group(I)	Group(J)	Sig.	Cohen's d	Sig.	Cohen's d	Sig.	Cohen's d
RR	SE	1.000	0.07	1.000	-0.57	1.000	-0.39
	FR	0.033*	0.42	0.790	0.60	0.920	0.56
	FR + SE	0.006**	0.50	0.102	0.81	0.826	0.56
SE	FR	0.108	0.36	1.000	0.04	0.008**	0.93
	FR + SE	0.024*	0.44	1.000	0.24	0.007**	0.94
FR	FR + SE	1.000	0.08	1.000	0.21	1.000	0.01

** $p < .01$, * $p < .05$

Table 2 Bonferroni post hoc pairwise comparisons of results of inference questions

		Total texts		Simple text		Complex text	
Group(I)	Group(J)	Sig.	Cohen's d	Sig.	Cohen's d	Sig.	Cohen's d
RR	SE	0.003**	0.45	0.071	0.49	0.230	0.73
	FR	< 0.001**	0.74	< 0.001	1.03	0.011*	0.99
	FR + SE	< 0.001**	1.18	< 0.001	1.25	< 0.001**	1.96
SE	FR	0.132	0.29	0.029	0.54	1.000	0.26
	FR + SE	< 0.001**	0.73	< 0.001	0.76	< 0.001**	1.22
FR	FR + SE	0.004**	0.44	1.000	0.22	0.015*	0.97

** $p < .01$, * $p < .05$

significantly better than the SE group and the FR group. But there was no significant difference between the SE group and the FR group.

Additionally, a significant three-way interaction was found between relearning strategy, text complexity and test format, $F_{(3, 146)} = 4.519$, $p = .005$. In the simple text, Bonferroni post hoc pairwise comparisons revealed that for the performance of verbatim questions, there were no significant differences among all groups. In the complex text, the score of the FR group was significantly higher than that of the SE group and the score of the FR + SE group was significantly higher than that of the SE group. But no statistically significant difference was observed between the RR group and the remaining three groups, nor between the FR group and the FR + SE group.

With regard to the inference questions, Bonferroni post hoc pairwise comparisons of the scores on the simple text showed that there was no significant difference between the RR group and the SE group. However, the RR group was significantly lower than the FR group and the FR + SE group. Similarly, the SE group was significantly lower than the FR group and the FR + SE group, but no significant difference was found between the FR group and the FR + SE group. For the complex text, there was no significant difference between the RR group and the SE group, the SE group and the FR group. However, the RR group exhibited significantly lower performance compared to both the FR group and the FR + SE group. Additionally, the FR + SE group demonstrated superior performance, significantly outperforming both the SE group and the FR group.

Discussion

Effects of relearning strategy on L2 discourse comprehension

Comparisons between rereading and retrieval-based learning

Consistent with some previous studies that have revealed the testing effect in discourse comprehension (e.g., Endres & Eitel, 2024; Karpicke & Roediger, 2011; Roediger & Butler, 2011), the results indicated that both retrieval-based learning strategies, namely free recall and free recall plus self-explanation, outperformed repetitive learning when addressing verbatim and inference questions. This discovery provides further empirical support to the notion that testing not only enhances memory retention but also facilitates the transfer of information (e.g., Barenberg et al., 2021; Butler, 2010; Eglington & Kang, 2018; Karpicke & Blunt, 2011).

Rereading primarily consists of a direct re-encoding of the surface-level information within the text. Sweller (1999) suggested that low task requirements could give rise to a "low cognitive load effect," potentially lessening learners' motivation to participate in the learning process. Only when learning activities have an appropriate level of difficulty and learners are willing to make an effort can the germane cognitive load be generated. Thus, even if the extraneous cognitive load in rereading scenarios was low, participants showed an unwillingness to allocate surplus cognitive resources for detailed processing of previously obtained knowledge. A post-experimental

interview reflection by a participant in the RR group revealed a decreased investment of time and effort during the second reading stage compared to the initial learning stage. The participant concentrated solely on understanding the information that was not fully comprehended before. As a consequence, the participants' episodic memory displayed minimal change, due to the similarity in cognitive processing between the two learning stages.

In contrast, according to the Episodic Context Account (Karpicke et al., 2014), the updated episodic context, formed through the integration of representations from encoding and retrieval contexts, plays a crucial role in effectively narrowing down the search set and distinguishing the target item from interfering information. These processes of discrimination and strengthening elevate the accessibility of individual memory traces by highlighting their distinctiveness, likely contributing to improvements in retention and inference compared to mere re-exposure (Hinze et al., 2013). This effect becomes particularly pronounced when there is a significant disparity between the encoding and retrieval contexts, making the updated context more distinct and accessible for subsequent retrieval. In both the free recall and the free recall plus self-explanation conditions, participants were instructed to generate as many cues as possible without external assistance. This deliberate process facilitated the acquisition of the maximum amount of updated context and thereby enhance the diagnostic value of their retrieval cues.

Comparisons between self-explanation and free recall

The absence of a significant differentiation between self-explanation group and free recall group in responses to both verbatim and inference questions contradicts established research advocating the benefits of testing over elaboration (e.g., Goossens et al., 2014; Lechuga et al., 2015; Karpicke & Blunt, 2011; Karpicke & Smith, 2012; Zhou et al., 2013).

On one hand, different elaborative learning strategies have distinctive features (Nückles et al., 2009) and correspondingly yield varying outcomes. For instance, the concept mapping method, employed in studies demonstrating testing advantages over elaboration (e.g., Karpicke & Blunt, 2011; Zhou et al., 2013), is an organizational strategy focusing on the main ideas and structure of the text. This approach does not necessitate participants to deduce additional information during organizational tasks. In contrast, it is posited that self-explanation coupled with the generation of appropriate examples serves to enhance the integration of new content with learners' extant knowledge structures and actively construct understanding through inferences and knowledge revision. Such an approach has purportedly led to a more exhaustive comprehension and update of the mental models of the text (Fiorella & Mayer, 2016; McNamara, 2007). Additionally, self-explanation enables learners to discern gaps or misconceptions in their understanding, fostering awareness of areas requiring further study and remediation in the intervening phase.

On the other hand, the absence of a substantial retrieval success during the initial free recall test might have contributed to a diminished or inconsistent testing effect within the free recall group. The proportion of freely retrieved idea units in the initial test was only 32.23% due to troubles in encoding L2 discourse. According to Rowland (2014), if the success rate of the initial retrieval is not sufficiently high, it raises the possibility that a dependable testing effect may not be produced.

The findings seem to align with the Elaborative Retrieval Hypothesis by Carpenter (2009), which claims that retrieving information activates elaborately encoded semantic knowledge that serves as effective cues in the final test. But caution is needed in interpreting the testing mechanism, as similar results don't necessarily signify identical cognitive processing. It's possible that self-explanation and free recall had comparable outcomes via different mechanisms. This is further supported by the performance difference between the SE and the FR group when dealing with texts of different complexities, as detailed in the section of "Effects of text complexity on L2 discourse comprehension".

Superiority of the combination of self-explanation and free recall

Unlike the FR group, the FR + SE group demonstrated superior performance than the SE group on both verbatim and inference questions. Moreover, in line with earlier research (e.g., Endres et al., 2017; Hinze et al., Exp. 3, 2013; Lachner et al., Exp. 2, 2021) and the complementary view, especially the Constructive Retrieval Hypothesis proposed by Hinze et al. (2013), the current study found that using a combination of free recall and self-explanation tactics led to superior performance on inference questions compared to only employing retrieval practice.

Consistent with the findings of studies by Endres et al. (2017) and Lachner et al. (2021), no statistically significant differences were detected in the quantity of recalled idea units during the initial test between the FR group and the FR + SE group. Nevertheless, it is probable that the qualitative aspects of the two tasks varied significantly. The initial free recall task primarily entailed retrieving memorized content from the original text, potentially falling short in promoting a profound understanding of its underlying meaning. In contrast, an elaborative

learning strategy can enhance the quality of retrieval. Notably, self-explanation, by activating related schemas stored in long-term memory, enables individuals to attain a deeper comprehension of concepts and integrates conceptual units into coherent mental representations. For example, while both the FR group and the FR + SE group might recall a sentence such as "People can only make some limited good choices during a day, so they should focus their limited energy on more important things," participants in the FR + SE group can integrate new information with existing knowledge and strengthen the memory trace by offering illustrative examples. A participant articulated, "Take Einstein for example, he did not deliberate on what to wear daily because he directed all his energy toward more significant decisions, such as choosing his research topics." The enhanced generative processing and inference-making contribute to the formation of a coherent and lasting representation, specifically situational model representations. Furthermore, practicing free recall allows learners to establish their own organizational structure, which they can employ during retrieval practice, resulting in a better understanding of the text (Smith et al., 2013). Free recall also reinforces subsequent retrievals and aids in consolidating learners' mental representations. Thus, the interplay between active free recall and the creation of meaningful explanations highlights their combined impact on fostering a more robust and durable comprehension of the material.

However, Larsen et al. (2013) did not identify a clear advantage in combining both retrieval practice and self-explanation. Several reasons might explain this difference. Firstly, in their study, both retrieval practice and elaborative processing are fully utilized, along with provided external aids. In the retrieval condition, students engaged in repeated tests(four times)with feedback, while in the elaboration condition, they were furnished with review sheets. Consequently, these conditions facilitated the efficacy of each strategy in isolation. Secondly, the successful completion of self-explanations differed between conditions. In the self-explanation alone condition, students successfully explained 96% of the key information, as the review sheet provided all necessary details. However, in the retrieval plus self-explanation condition, self-explanations were completed for only 71% of the information recalled, and overall recall was about 75%. Given that the study material consisted of intricately detailed lecture, the decreased quantity of recall and explanation generation had a crucial negative impact on the effectiveness of these integrated strategies.

Effects of text complexity on L2 discourse comprehension

Effects of text complexity on verbatim questions

The consistency in responses to verbatim questions was evident across all four experimental groups when engaging with the simple text. This uniformity resulted from the text's inherent simplicity, which made it easily understandable due to its low complexity. According to the survey conducted among participants regarding the perceived difficulty of the text, it was found that simple texts had a mean score of 3.21 on a scale ranging from 0 (very easy) to 10 (very difficult). The simplicity of the questions, which focused on factual information directly extractable from the text, also contributed to this uniform performance. Therefore, each strategy independently proved effective in accomplishing the task. And the combination of free recall and self-explanation showed redundancy, highlighting the overlap when employing two strategies simultaneously. Similarly, in Roelle & Nückles' (2019, Exp. 1) investigation, where learners interacted with expository text distinguished by high cohesion and elaboration, within the uncomplicated final recognition test, the outcomes likewise indicated no discernible effects of rereading, retrieval practice, elaboration, or combination of elaboration and retrieval practice.

When confronted with complex text, testing did not exhibit superiority over rereading in the assessment of fact-based comprehension. Many previous studies confirming the testing effect have employed foreign language vocabulary pairs or texts in the native language as materials. In contrast, this study employed more complex L2 texts. Notably, rereading has been proven advantageous for enhancing the comprehension of the text, especially when it comes to the detailed understanding of facts and information at the textual level. Thus, this research finding seems to suggest that the complexity of learning materials may adversely affect the efficacy of the testing effect.

However, both the FR group and the FR + SE group manifested a remarkable superiority over the SE group. Complex materials frequently entail numerous details and interrelated concepts. The SE group, which was tasked with re-encoding highly interactive texts and concurrently generating relevant explanations within a restricted time frame, experienced a cognitive strain. Such competition led to a diminished capacity for retaining intricate details and facts. Moreover, as intermediate L2 learners, the SE group lacked proficiency in discriminating between major and minor ideas and details. Many participants in the SE group endeavored to comprehend all information and offered abundant examples and explanations. Nevertheless, dealing with an overwhelming number of interactive elements that do not directly contribute

to learning triggered a redundancy effect, as suggested by Sweller and Chandler (1994). Consequently, the elaboration process might impair memory because less time is allocated to the main ideas and facts, despite the considerable overall time expenditure involved (Daley & Rawson, 2018, 2020). On the contrary, the FR + SE group, despite coming across the dual-task challenge of free recall and self-explanation, merely needed to direct their efforts towards explaining the recalled information. Although the quantity of information processed is relatively limited, it is the information that the learner can remember and has received further refinement, so that the memory can be effectively retained. The different performance of self-explanation and testing in simple and complex texts clarifies the distinct processing mechanisms of elaboration and retrieval practice.

Additionally, this research revealed that there was no discernible discrepancy in the retention of facts between the two retrieval groups. This lack of distinction can be attributed to the efficacy of free recall as a retrieval mechanism for factual information. Besides, the inherent meaningfulness derived from self-explanation does not substantially augment factual comprehension. Therefore, incorporating self-explanation appears to be superfluous and redundant.

Effects of text complexity on inference questions

The performance of both the FR group and the FR + SE group in addressing inference questions exceeded that of the RR group, a distinction noticed in both simple and complex texts. Although rereading can increase familiarity with the text and improve retention, it might not necessarily foster the ability to transfer that knowledge to different situations, which often demands a deeper understanding and more active engagement with the text. This result contradicts earlier studies, such as Van Gog and Sweller (2015), which asserted the absence of a testing effect in more complex learning materials. The disparity can be ascribed to the utilization of different learning materials. Although the research materials in the present study comprise challenging L2 discourse, the perception of difficulty among participants indicates that even though the complex text, with an average rating of 5.27 (0 = very easy, 10 = very difficult), might be more demanding than the simpler one (Mean = 3.21), it does not present a significant challenge for learners. When contrasted with the task of grasping problem-solving principles presented in the worked example in Van Gog and Sweller's study (2015), the level of complexity is relatively lower. Instead, it offers an ideal degree of difficulty, enhancing the reinforcing and consolidating effects of retrieval practice in knowledge transfer. The design of the retrieval task could also account for the disparities witnessed in the studies cited by Van Gog and Sweller (2015) that failed to detect a testing effect. For example, when a fill-in-the-blank test is utilized as a retrieval task, it demands less retrieval and organizational processing compared to free recall and short-answer questions. As a result, these experiments might have been unable to observe retrieval practice effects not because of the complexity of the materials, but rather due to the nature of the retrieval tasks involved (Karpicke & Aue, 2015).

The FR group exhibited superior performance compared to the SE group on inference questions of the simple text. The finding of this study aligns with those of Roelle and Nückles (2019). They argue that the benefits of elaborative activities and retrieval practice should substantially depend on the quality of learners' mental representations before they engage in the follow-up retrieval practice. When the text demonstrates high cohesion and elaboration, learners' mental representations are already relatively coherent and integrated with their prior knowledge by the end of the initial study phase. Consequently, the engagement in elaborative learning is deemed redundant. Conversely, given that the benefits of retrieval practice are due to memory consolidation, it is reasonable to find that the benefits increase with an enhanced quality of learners' mental representations (Roelle & Nückles, 2019).

As the complexity of the text increased, however, the distinction between the SE group and the FR group in answering inference questions diminished. The strength of the testing effect is primarily determined by retrievability, which measures how successfully information can be retrieved at the intervening stage (Rowland, 2014). Retrievability is primarily influenced by the quality of initial encoding. When the complexity of the text intensifies, a superficial initial reading may prove inadequate for achieving a profound understanding of the text. A participant from the FR group noted during the post-experimental interview, "I could only recall limited information because I didn't fully understand the text." This implies that retrievability may have been compromised. Consequently, this deficiency in the initial test performance might have a successive impact, ultimately reducing the final understanding of the discourse. This discovery provides partial support for the notion that the impact of testing might be negated by material complexity.

The inference questions within the simple text revealed no substantial disparity between the two testing groups, indicating that incorporating an additional elaborative strategy may not be necessary and could be redundant in uncomplicated contexts. Participants in the FR group demonstrated a good understanding of the text and thereby possessed sufficient cognitive resources. Many of them not only recalled information from the text but also provided their own interpretations, similar to the FR + SE group. Some participants even went a step further by reorganizing and summarizing the information, stating, for example, "This

article emphasizes considering the deep structure of a question rather than focusing solely on the surface information," which likely contributed to their deeper comprehension of the simple text. This finding aligns with the study conducted by Blunt and Karpicke (2014), where no significant difference was observed between the free recall and retrieval-based concept mapping groups.

Notably, the FR + SE group demonstrated outstanding inference abilities when dealing with complex texts. The efficacy and efficiency of elaboration strategies are closely related to prior knowledge (Zhou et al., 2013). As a result of the participants' limited topic knowledge, the SE group struggled to access relevant information directly from long-term memory. Moreover, the limited time for detailed processing of complex texts strained cognitive resources and led to relatively limited effectiveness. Due to the decreased retrievability caused by the text complexity, the FR group also performed poorly in demanding inferences. In contrast, the combined FR + SE group, though restricted in the amount of information they could recall, employed a method that filtered out minor and redundant details from the text. Therefore, this strategy not only addressed the challenges associated with achieving a desirable level of difficulty but also avoided surpassing cognitive load limits. By effectively using both retrieval and elaborate processing strategies, the approach played a complementary role in achieving a nuanced balance between retrieval effort and retrieval success. More specifically, participants actively constructed schemas and episodic contexts by reorganizing and integrating various elements from the retrieved information with their existing cognitive framework. The elaboration during the intervening phase compensates for the inadequacies of comprehension in the initial encoding stage, thereby enhancing the quality of mental representation and facilitating optimal information retrieval. Through actively retrieving the updated information from memory, participants strengthened the connections between different pieces of information, thus increasing the durability of memory traces. This reinforcement and consolidation boosts the likelihood that they can effectively apply what they have learned in new and varied contexts.

Our research findings contrast with those of Roelle and Nückles (2019). They discovered that when the learning material lacked cohesion and elaboration, only involving learners in generative learning activities was helpful, while engaging them in retrieval practice or in both generative activities and retrieval practice wasn't. The main cause of this disparity lies in the varying levels of matches between the conditions of material complexity and the methods of elaborative processing in the two studies. Roelle and Nückles (2019) deliberately decreased the degree of cohesion and elaboration of the target text. For instance, they removed all illustrative examples and figures to lower the level of elaboration. Their generative activities entailed highlighting the main content items of the text and illustrating these main points with examples. Consequently, engaging in such generative processing effectively dealt with the lack of coherence and elaboration in the text. In our present study, the complexity was measured through referential cohesion, and it evaluates the degree to which ideas within a text are interlinked and integrated across sentences. However, the text still contains quite a few examples. Hence, interpreting and exploring the text based on personal examples and experiences might not have the same effect as demonstrated in Roelle and Nückles (2019).

Another notable difference relates to the text characteristics. Although the texts in both studies aim to introduce a concept, theirs is more complex, featuring more words, subtopics, figures, as well as a comprehensive and structured analysis. This difference led to dissimilar degrees of understanding when extracting limited information via the combined approach. In Roelle and Nückles (2019), the generative group outperformed the generative-and-retrieval group in the degree of organization, the number of elaborations, and the number of covered idea units, the crucial mediators of transfer performance in their study. In contrast, in the present study, even if the combined group only explained and exemplified the limited extracted information, it was sufficient for deepening the understanding of the concept and could be effectively applied to transfer questions.

Overall, this study reveals that testing can be effective in complex educational settings. This investigation further validates that testing is not just an assessment means but also an efficient learning method. Theoretically, the exploration of the direct learning function of tests enriches the research perspectives in L2 acquisition. Meanwhile, the research emphasizes the cognitive disparities between elaborate processing and retrieval practice and shows their synergy in enhancing the testing effect, effectively uncovering the mechanism of the testing effect and facilitating the advancement of related research. Pedagogically, the results of this study offer insights into how to effectively utilize the testing effect to enhance L2 discourse comprehension. More importantly, it provides genuine and objective data along with persuasive reasons for educational policymakers, teachers, learners, schools, publishers, and software developers to alter the conventional notion of testing and highlight the learning function of tests. Formative tests should be rationally organized within the curriculum syllabus, teaching schedules, self-directed learning, and textbook layout to enhance the quality of teaching and learning. However, this study is limited to the comprehension of L2 texts in universities. To guarantee the generalizability of the research

findings and further enhance its educational significance, researchers can undertake additional studies in various educational fields and with diverse groups of people to validate and broaden the comprehension of the testing effect.

Limitations and future research

During the current study, several limitations were identified, from which suggestions for future research were subsequently provided.

Firstly, the study observed that testing could enhance learning more effectively than elaborative self-explanation. However, to ensure that each experimental group has the same amount of learning time, the elaborative processing time may not be adequate. Future research could explore whether the duration of time imposes limitations on the effectiveness of the self-explanation strategy. Moreover, future studies can investigate the influence of more tests and traditional learning methods as well as their characteristics on the testing effect.

Secondly, the distinction in text complexity within this study was insufficiently marked. Even when learners dealt with highly complex texts, they did not consider them overly challenging. As a consequence, the examination regarding the impact of complexity was undermined. Future research should aim at widening the gap in the complexity of learning materials. Additionally, complexity is likely a multi-faceted concept that is influenced not only by the materials themselves but also by the learners' proficiency and prior knowledge. However, the exploration of learners' characteristics in the testing effect was seriously inadequate. Furthermore, it is necessary to conduct additional research on the influence of the nature of the material, such as its genre, length, and topic, on the testing effect.

Finally, in our study, learners engaged in generative activities simultaneously with retrieval practice. However, recent research by McDaniel (2023) and Roelle et al. (2023) suggests that the efficacy of incorporating elaborative encoding techniques alongside retrieval practice might be heightened when applied in a sequential manner rather than concurrently. Therefore, future investigations into the combined effects of elaborative activities and retrieval practice could benefit from further exploration and validation of sequence effects.

Conclusion

This study aimed to explore how different relearning strategies during the intervening phase and the complexity of texts influence L2 discourse comprehension. It is crucial to note that the testing effect is highly dependent on the interaction among relearning strategies, the complexity of the studied material, and the eventual testing formats. For lower-level verbatim questions and simpler material, all techniques yield comparable results. However, when it comes to high-level inference questions, retrieval outperforms rereading. In complex language learning circumstances (complex text or/and inference questions), retrieval can have effects that are comparable to or even superior to self-explanation. These findings indicate that testing and elaboration are distinct mechanisms in information processing. When the text is straightforward or the final retrieval task involves factual questions, the combination of free recall and self-explanation might be redundant. But in more complex scenarios, a synergistic approach that combines elaborate processing and testing leads to better outcomes. It appears that this combined approach not only plays a complementary role but also establishes a delicate balance between retrieval effort and retrieval success.

In conclusion, a comprehensive exploration of various influencing factors, such as instructional methods, learning materials, learner characteristics, and learning environments, is indispensable for a better understanding of the mechanisms underlying the testing effect, particularly in complex educational domains. By considering these elements, researchers can unravel the complexities surrounding how and why testing enhances learning, thus advancing our knowledge in the field of education and cognitive psychology.

Author contributions The first author designed the experiment, analyzed the data and wrote this report. The second author collected and analyzed the data and wrote this report. The third author analyzed the data and wrote this report. All authors have read and agreed to the published version of the manuscript.

Data availability The raw data supporting the conclusions of this article will be made available by the authors without undue reservation.

Declarations

Ethics statement This study was approved by School of Foreign Studies at Jiangnan University, China. Participants volunteered to participate in the experiment and all of them were given six credits as rewards. Informed consent was obtained from participants before the experiment.

Conflict of interest The authors declare that there are no competing interests relevant to this journal submission.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing.

Review of Educational Research, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open and closed book tests. *Applied Cognitive Psychology, 22*(7), 861–876. https://doi.org/10.1002/ACP.1391

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*, *33*, 1409–1453. https://doi.org/10.1007/s10648-021-09595-9

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*(1), 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

Barenberg, J., Berse, T., Reimann, L., & Dutke, S. (2021). Testing and transfer: Retrieval practice effects across test formats in English vocabulary learning in school. *Applied Cognitive Psychology*, *35*(3), 700–710. https://doi.org/10.1002/acp.3796

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

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher, & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (Vol. 459, p. 435). MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Erlbaum.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology*, *106*(3), 849–858. https://doi.org/10.1037/a0035934

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(5), 1118–1133. https://doi.org/10.1037/a0019902

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning Memory and Cognition*, *37*(6), 1547–1552. https://doi.org/10.1037/a0024140

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*(5), 279–283. https://doi.org/10.1177/0963721412452728

Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (2017). Testing enhances both encoding and retrieval for both tested and untested items. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *70*, 1–60. https://doi.org/10.1080/17470218.2016.1175485

Cummings, E. L., Reeb, A., & McDaniel, M. A. (2023). Do not forget the keyword method: Learning educational content with arbitrary associations. *Journal of Applied Research in Memory and Cognition*, *12*(1), 70–81. https://doi.org/10.1037/mac0000031

Daley, N., & Rawson, K. A. (2018). Elaborations in expository text impose a substantial time cost but do not enhance learning. *Educational Psychology Review*, *31*, 197–222. https://doi.org/10.1007/s10648-018-9451-9

Daley, N., & Rawson, K. A. (2020). Effects of elaborations included in textbooks: Large time cost, reduced attention, and lower memory for main ideas. *Educational Psychology Review*, *33*, 1–25. https://doi.org/10.1007/s10648-020-09553-x

De Jonge, M. O., Tabbers, H. K., & Rikers, R. M. J. P. (2015). The effect of testing on the retention of coherent and incoherent text material. *Education Al Psychology Review, 27*(2), 305–315. https://doi.org/10.1007/s10648-015-9300-z

Eglington, L. G., & Kang, S. H. K. (2018). Retrieval practice benefits deductive inference. *Educational Psychology Review*, *30*, 215–228. https://doi.org/10.1007/s10648-016-9386-y

Endres, T., Carpenter, S., Martin, A., & Renkl, A. (2017). Enhancing learning by retrieval: Enriching free recall with elaborative prompting. *Learning and Instruction*, *49*, 13–20. https://doi.org/10.1016/j.learninstruc.2016.11.010

Endres, T., & Eitel, A. (2024). Motivation brought to the test: Successful retrieval practice is modulated by mastery goal orientation and external rewards. *Applied Cognitive Psychology*, *38*(1), 1–16. https://doi.org/10.1002/acp.4160

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Fritz, C. O., Morris, P. E., Acton, M., Etkind, R., & Voelkel, A. R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology*, *21*, 499–526. https://doi.org/10.1002/acp.1287

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative rest udy in primary school vocabulary learning. *Journal of Applied Research in Memory and Cognition*, *3*, 177–182. https://doi.org/10.1016/j.jarmac.2014.05.003

Graesser, A. C., et al. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods Instruments and Computers*, *36*(2), 193–202.

Graesser, A. C., Leon, J. A., & Otero, J. (2002). Introduction to the psychology of science text comprehension. In J. Otero, J. A. León, & A. C. Graesser (Eds.), *The psychology of Science text comprehension* (pp. 1–15). Erlbaum.

Greving, S., & Ritchter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology*, *9*, 1–10. https://doi.org/10.3389/fpsyg.2018.02412

Hanham, J., Leahy, W., & Sweller, J. (2017). Cognitive load theory, element interactivity, and the testing and reverse testing effects. *Applied Cognitive Psychology, 31*(3), 265–280. https://doi.org/10.1002/acp.3324

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using short answer tests. *Memory (Hove, England)*, *19*(3), 290–304. https://doi.org/10.1080/09658211.2011.560121

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language*, *69*(2), 151–164. https://doi.org/10.1016/j.jml.2013.03.002

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Karpicke, J. D. (2017). Retrieval-based learning: a decade of progress. In *Learning and Memory: A Comprehensive Reference* (Vol. 2, pp. 487–514). https://doi.org/10.1016/b978-0-12-809324-5.21055-9

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3

Karpicke, J. D., & Blunt, J. B. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., Blunt, J. R., & Smith, M. A. (2016). Retrieval-based learning: Positive effects of retrieval practice in elementary school children. *Frontiers in Psychology*, *7*(350), 1–9. https://doi.org/10.3389/fpsyg.2016.00350

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: an episodic context account. In B. H. Ross (Ed.),

Psychology of Learning and Motivation (Vol. 61, pp. 237–284). Elsevier Academic Press.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*(2), 151–162. https://doi.org/10.1016/j.jml.2006.09.004

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language*, *67*(1), 17–29. https://doi.org/10.1016/j.jml.2012.02.004

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*(4), 294–303. https://doi.org/10.1037/0003-066x.49.4.294

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Lachner, A., Jacob, L., & Hoogerheide, V. (2021). Learning by writing explanations: is explaining to a fictitious student more effective than self-explaining? *Learning and Instruction, 74*, 101438. https://doi.org/10.1016/j.learninstruc.2020.101438

Larsen, D. P., Butler, A. C., & Roediger, H. L. III. (2013). Comparative effects of test-enhanced learning and self - explanation on long-term retention. *Medical Education, 47*(7), 674–682. https://doi.org/10.1111/medu.12141

Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review*, *27*, 291–301. https://doi.org/10.1007/s10648-015-9296-4

Lechuga, M. T., Ortega-Tudela, J. M., & Gomez-Ariza, C. J. (2015). Further evidence that concept mapping is not better than repeated retrieval as a tool for learning from texts. *Learning and Instruction*, *40*, 61–68. https://doi.org/10.1016/j.learninstruc.2015.08.002

Li, J. X., Zhang, E. H., He, X. Y., Zhang, H. H., Gou, H. C., Wang, X. Y., Wang, S. R., & Cao, H. W. (2022). Retrieval practice enhances learning and memory retention of French words in chinese-english bilinguals. *Lingua*, 272. https://doi.org/10.1016/j.lingua.2022.103294

McDaniel, M. A. (2023). Combining Retrieval Practice with Elaborative Encoding: Complementary or redundant? *Educational Psychology Review*, *35*(3), 75. https://doi.org/10.1007/s10648-023-09784-8

McNamara, D. S. (2007). *Reading comprehension strategies: Theories, interventions and technologies*. Lawrence Erlbaum Associates.

Miyatsu, T., & McDaniel, M. A. (2019). Adding the keyword mnemonic to retrieval practice: A potent combination for foreign language vocabulary learning? *Memory & Cognition, 47*, 1328–1343. https://doi.org/10.3758/s13421-019-00936-2

Nückles, M., Hübner, S., & Renkl, A. (2009). Enhancing self-regulated learningby writing learning protocols. *Learning and Instruction*, *19*(3), 259–271. https://doi.org/10.1016/j.learninstruc.2008.05.002

O'Day, G. M., & Karpicke, J. D. (2021). Comparing and combining retrieval practice and concept mapping. *Journal of Educational Psychology*, *113*(5), 986–997. https://doi.org/10.1037/edu0000486

Rawson, K. A., & Katherine, A. (2015). The status of the testing effect for complex materials: Still a winner. *Educational Psychology Review*, *27*(2), 327–331. https://doi.org/10.1007/s10648-015-9308-4

Rawson, K. A., & Zamary, A. (2019). Why is free recall practice more effective than recognition practice for enhancing memory? Evaluating the relational processing hypothesis. *Journal of Memory and Language*, *105*, 141–152.

Richter, T., et al. (2022). Using interleaving to promote inductive learning in educational contexts: Promise and challenges. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 54*(4), 164–175. https://doi.org/10.1026/0049-8637/a000260

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Science, 15*(1), 20–27.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roelle, J., Endres, T., Abel, R., Obergassel, N., Nückles, M., & Renkl, A. (2023). Happy together? On the relationship between Research on RetrievalPractice and Generative Learning using the case of Follow-Up Learning tasks. *Educational Psychology Review*, *35*, 102–129. https://doi.org/10.1007/s10648-023-09810-9

Roelle, J., Froese, L., Krebs, R., Obergassel, N., & Waldeyer, J. (2022). Sequence matters! Retrieval practice before generative learning is more effective than the reverse order. *Learning and Instruction*, *80*(3), 1–12. https://doi.org/10.1016/j.learninstruc.2022.101634

Roelle, J., & Nückles, M. (2019). Generative learning vs. retrieval practice in learning from text: The cohesion and elaboration of the text matters. *Journal of Educational Psychology*, *111*(8), 1341–1361. https://doi.org/10.1037/edu0000345

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Rummer, R., Schweppe, J., Gerst, K., & Wagner, S. (2017). Is testing a more effective learning strategy than note-taking. *Journal of Experimental Psychology: Applied*, *23*(3), 293–300. https://doi.org/10.1037/xap0000134

Shobe, E. (2021). Achieving testing effects in an authentic college classroom. *Teaching of Psychology*, *49*, 164–175. https://doi.org/10.1177/00986283211015669

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1712–1725. https://doi.org/10.1037/a0033569

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J. (1999). *Instructional design in technical areas*. ACER Press.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction*, *12*(3), 185–233. https://doi.org/10.1207/s1532690xci1203_1

The Ministry of Education and the State Language Commission of the People's Republic of China. (2018). *China's Standards of English Language Ability.*

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*(4), 252–257.

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review, 27*(2), 247–264. https://doi.org/10.1007/s10648-015-9310-x

Watter, K., Copley, A., & Finch, E. (2016). Discourse level reading comprehension interventions following acquired brain injury: A systematic review. *Disability and Rehabilitation*, *39*, 315–337. https://doi.org/10.3109/09638288.2016.1141241

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *Journal of General Psychology, 132*(4), 408–428. https://doi.org/10.3200/GENP.132.4.408-428

Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied*, *23*(3), 263–277. https://doi.org/10.1037/xap0000122

Zhou, A. B., Ma, X. F., Jing, M., Li, J., & Cui, D. (2013). The advantage effect of retrieval practice on memory retention and

transfer: Based on explanation of cognitive load theory. *Acta Psychologica Sinica*, *45*, 849–859. https://doi.org/10.3724/SP.J.1041.2013.00849

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. https://doi.org/10.1037/0033-2909.123.2.162