2022

WORD GAMES FOR EDUCATION: INVESTIGATING THE EFFECTIVENESS OF ADDING ELABORATION TASKS TO CROSSWORDS FOR LEARNING TECHNICAL VOCABULARY

Warat Khaewratana
wkhaewra@mtu.edu

WORD GAMES FOR EDUCATION: INVESTIGATING THE EFFECTIVENESS OF ADDING ELABORATION TASKS TO CROSSWORDS FOR LEARNING TECHNICAL VOCABULARY

By

Warat Khaewratana

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Applied Cognitive Science and Human Factors

MICHIGAN TECHNOLOGICAL UNIVERSITY

2022

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Applied Cognitive Science and Human Factors.

Department of Cognitive and Learning Sciences

Dissertation Co-Advisor:	*Elizabeth Veinott*
Dissertation Co-Advisor:	*Shane Mueller*
Committee Member:	*Shari Stockero*
Committee Member:	*Amanda Gonczi*
Committee Member:	*Qian Zhang*
Department Chair:	*Kelly Steelman*

Table of Contents

Abstract

One challenge in STEM education is the learning of technical terms. In order to reason about higher-order scientific concepts, knowledge of technical vocabulary is often a prerequisite. Improving this knowledge may enhance the learning of higher-order concepts because it reduces cognitive load students experience while learning. To that end, we need innovative learning-aid tools that help students not only in learning and remembering technical terms, but also in applying the learned knowledge to broader concepts. This dissertation investigates the hypothesis that learning gained from crosswords can be used to teach technical terms. Furthermore, I am also examining the hypothesis that additional elaboration techniques will enhance the effect of crosswords. In a series of seven experiments, I investigated the effect of crossword puzzles with an add-on elaboration on students' ability to remember learned technical terms and to provide more in-depth explanations of those terms. Across the experiments, I investigated (a) three different types of elaboration techniques, (b) collaboration vs. individual participation, (c) in-person vs. online training, and (d) short vs. long delay. Across experimental variations, results indicated that using a crossword puzzle alone produced a statistically significant learning effect relative to a control condition. Although adding structured elaboration did provide benefits when added to a crossword puzzle, it did not consistently improve retention compared to the crossword puzzle alone. Also, different elaboration techniques did not provide specific enhancement on memory retention. Implications for theoretical perspectives on learning technical vocabulary and best practices to implement crossword in educational settings are discussed.

1 Introduction to the Use of Crosswords in Education

1.1 From leisure activity to learning-aid tool

One of the critical challenges in the education of science, technology, engineering, and mathematics (STEM) is the learning of field-specific technical terms (Wong, Chu, & Yap, 2014; Yager, 1983; Yuriev, Capuano, & Short, 2016). Addressing such challenges may reduce barriers to entry of potential students into technical fields, because students need to be able to speak and understand the basic language of a discipline before they can engage in higher-order scientific reasoning, and understanding technical terms will reduce cognitive load and enable them to engage in more complex domain-specific problem solving (Sweller, 1994). To support the teaching of technical vocabulary, instructors need to be creative in their delivery method to ensure that their students can absorb such complex materials and develop an initial conceptual understanding of technical terms, as well as maintain interest in doing so (Rosenthal, 1995). Among many innovative delivery methods that are suitable for such tasks, crosswords receive strong support from many educational researchers (e.g., Gaikwad & Tankhiwale, 2012).

Since the invention of the first crossword puzzle by Arthur Wynne in 1913, people have played crosswords while holding a belief that doing so increases their intelligence even if it was not scientifically proven (Saxena, Nesbitt, Pahwa, & Mills, 2009). Therefore, crosswords might be perceived as a mere word game that exercises a player’s brain and not as an effective learning support tool. Regardless, crosswords were popular among educators and schoolteachers, and many began implementing the fad in the classroom shortly after the game was introduced, hoping to benefit their students in gaining and retaining knowledge provided in those courses (Van Vleet, 1925). The widespread usage of crosswords in school curriculums led educators to investigate (a) whether or not working on crossword puzzles could enhance learning or, more specifically, retention of knowledge gained, especially in science; and (b) whether or not teachers should consider implementing crossword puzzles in their curriculum (Pruitt, 1927). Consequently, various studies to validate the effectiveness of puzzles for learning have been conducted. For example, Crossman and Crossman (1983) studied the effect of crosswords on learning scientific vocabulary and associated facts in psychology, and they reported statistically significant improvement in students' test scores. In other studies, Whisenand and Dunphy (2010) investigated a crossword's effect on an information system; Gaikwad and Tankhiwale (2012) in pharmacology; Coticone (2013) in biochemistry; Yuriev, Capuano, and Short (2016) in chemistry; Mueller and Veinott (2018) in psychology; and Gilani et al. (2020) in dental education. These studies are only a few examples of a much larger number of cases that show the implementation of crosswords in practical educational settings, which was reviewed by Mueller and Veinott (2018). This suggests the potential benefits of crossword learning-aid tools that can help students learn the scientific vocabulary necessary for them to advance their discussions and understanding in various STEM fields.

Aside from the delivery method of technical terms, there is also the issue of how technology might be integrated into word-game based learning. As we move into the 21st century, access to online resources and remote learning are becoming more feasible options. Yet, some schools still face a challenge in accessing good Wi-Fi (e.g., rural), while other schools lack guidelines for integrating technology into the school curriculum (Morgan, 2020). Concern over the accessibility issue grows even stronger, especially after the strike of the COVID-19 pandemic, during which many schools must leverage online or remote learning and adopt a variety of pedagogical strategies for the new normal (Baber, 2020). All of these challenges highlight the need to develop methods and tools that not only work across different educational settings and platforms but also can achieve the goal of providing students with active and constructive learning opportunities (Chi & Wylie, 2014). Crossword learning seems to fit many classrooms as they can be easily generated in both printed and computerized formats.

Consequently, this dissertation examines the efficacy of crosswords for improving memory retention of learned technical terms. Across seven experiments, I examine the effectiveness of and practice of using crossword learning to teach scientific vocabulary. In addition, in an effort to enhance the learning benefits of crosswords for learning, I also investigate the effect of adding different elaboration strategies to crossword puzzles and the resulting learning outcomes.

1.2 Potential benefits of crosswords in education

To solve a crossword puzzle, a person first reads a given hint (which can be the word's definition or its associated factual information), then counts up the number of blank boxes, and then finds a word whose meaning and number of letters match. This simple set of procedures requires several skills including spelling, reasoning, making inferences, evaluating choices, and drawing conclusions (Whisenand & Dunphy, 2010). When it comes to crosswords, oftentimes the person has already learned the word and its meaning from somewhere else beforehand and thus uses crossword hints to search his or her memory for the correct word. Also, the person knows not just the target word but a bank of words with similar meanings for one particular puzzle, so that he or she must select the most correct word based on the letters and number of blank boxes.

One might argue that a person can also learn vocabulary and utilize the associated learning skills through other means besides crossword puzzles, such as textbooks and lists of words. However, crossword puzzles have characteristics that provide greater utility over other options. First, solving a crossword puzzle is regarded as an active learning strategy that has been widely accepted to produce better knowledge representations and retention than passive learning strategies in education (Bonwell & Sutherland, 1996; Chi & Wylie, 2014; Rosenthal, 1995; Rowles, 2013; Rubin & Herbert, 1998). Active learning, as operationalized by Chi and Wylie (2014), refers to a mode of engagement in a learning task in which “some form of overt motoric action or physical manipulation is undertaken” (p. 221). In the case of crossword puzzles, the active learning process involves the action of writing or typing correct words in the provided spaces (Saran & Kumar, 2015; Weisskirch, 2006). Specifically, a person must initiate an overt finger movement to write or type the missing letters in blank boxes. This differentiates the learning task from one that involves simply reading and memorizing each word’s spelling, which would be considered passive learning that has been shown to be less productive (Chi & Wylie, 2014).

Second, the need to find words to fill in puzzle boxes adds some degree of difficulty to the learning task as opposed to reading a book or a list. However, the added challenge (at the appropriate level) is considered *desirable difficulty*, which has been shown to result in more benefits than comparatively easier learning tasks (Bjork & Bjork, 2011). Note that, in the case of crossword puzzles, the notion of desirable difficulty only applies if a person either has previously learned the missing words from somewhere else or has access to outside resources (e.g., textbook and dictionary) to help learn new words. Otherwise, the difficulty in trying to search for unfamiliar missing words might exceed the desirable level, and amount to irrelevant cognitive load (Sweller, 1994). Third, people often perceive crossword puzzles as a game, which is a type of activity for leisure rather than an act of learning. Solving crossword puzzles is playful, yet challenging and fun learning (Rambli, Matcha, & Sulaiman, 2013). Given its potential benefits for learning, people may choose a crossword puzzle alternative as it is more fun and less intimidating than studying from textbooks or lectures, so even if it is no more effective than studying term-definition lists, it may have a greater benefit because students are willing to do it.

1.3 Lack of quantitative evidence in crossword studies

As already mentioned, there have been a several number of published studies that report learning improvement as a result of solving crossword puzzles containing vocabularies from STEM disciplines (Crossman & Crossman, 1983; Gaikwad & Tankhiwale, 2012; Gilani et al., 2020; Mueller & Veinott, 2018; Whisenand & Dunphy, 2010; Yuriev et al., 2016). However, some studies validated the crossword effect but found no learning improvement. For example, Coticone (2013) studied the effect of crossword puzzles which were generated by students themselves on learning scientific vocabularies in biochemistry. The study found no statistically significant difference in the final test scores between the intervention and control groups. There have also been cases where crossword improves learning inconsistently. In their study, Davis, Shepherd, and Zwiefelhofer (2009) found that the student group that did a crossword for the exam review performed better on the exam (compared to a group that used a word-list alternative). The student group that simply reviewed the crossword did not improve performance. These mixed results suggest that crossword puzzles do not guarantee vocabulary acquisition. Perhaps there are multiple factors involved, and it might also depend on how exactly the crossword is implemented.

To validate the credibility of those findings, it is important to look closely at the design of experiments, as well as a means of evaluation. According to the review by Mueller and Veinott (2018) covering more than 200 crossword studies, the majority of those studies evaluated crossword effectiveness in facilitating learning only by assessing participants' perceived usefulness. The method of assessment involved using a questionnaire to ask questions such as "How much do you think you learned?", "How do you think you did on the test?", and "How fun was it?" Such assessment provides a finding that portrays a person's opinion toward the use of crosswords in aiding learning, which does not necessarily reveal his or her actual learning gain. To measure the true benefit of learning, the main focus should be on the collection and analysis of quantitative data, with perceived usefulness as supplementary data (Mueller & Veinott, 2018).

Table 1 shows the portion of crossword studies that evaluated crossword effectiveness by measuring quantitative learning gains, such as a direct comparison between pretest and posttest scores. Among those studies, some of them did not conduct controlled experiments (Crossman & Crossman, 1983; Yuriev et al., 2016). Failure to include proper control conditions in an experiment makes it difficult to conclude that crosswords add value beyond what students do normally. In fact, only a handful of studies that involved randomized, controlled experiments, used statistical approaches to assess the effectiveness of crosswords or found statistically significant results showing a benefit (Gaikwad & Tankhiwale, 2012; Mueller & Veinott, 2018). This small number of studies raises the question of the validity of crossword puzzles as a learning tool. Thus, this dissertation will provide a series of randomized, controlled experiments to compare what students do naturally for learning (baseline control) and whether adding a crossword and elaboration improves learning. In addition, we use quantitative measures of learning to examine the benefits of implementing them in classroom settings. While quantitative evidence may be lacking, the majority of crossword studies have shown that crosswords provide both pragmatic and educational value. However, it remains unclear what the best methods for implementing crosswords in the classroom. Consequently, one of the goals of this dissertation is to contribute to principles of best practices based on the results from the experiments.

Table 1. Review of experimental research validating crossword's effectiveness (chronologically ordered)

Field of Study	Control Condition	Delay Length	Validation	Result
Psychology (Crossman & Crossman, 1983)	No	Weeks	Pre- vs Post- Tests	Improved test score
Athletics (Berry & Miller, 2008b)	Yes*	Within a week	Pre- vs Post- Tests	Improved test score Not better than control
Sociology (Davis et al., 2009)	Yes*	Within a week	Course Exam	Not better than control
Information System (Whisenand & Dunphy, 2010)	Yes*	1 Week	Periodic Tests	Better than control
Pharmacology (Gaikwad & Tankhiwale, 2012)	Yes*	Immediate	Pre- vs Post- Tests	Improved test score Better than control
Veterinary (Abuelo et al., 2016)	Yes*	Immediate	Pre- vs Post- Tests	Improved test score Better than control
Chemistry (Yuriev et al., 2016)	No	Weeks	Pre- vs Post- Tests	Improved test score
Social Science (Ismayenti et al., 2018)	No	Weeks	Periodic Tests	Improved test score
Psychology (Mueller & Veinott, 2018)	Yes*	Immediate Days	Post-Tests	Better than control (only without delay)
Electronics Engineering (Shetgar & Thalange, 2018)	No	Immediate	Pre- vs Post- Tests	Improved test score
Dental Education (Gilani et al., 2020)	Yes*	Immediate	Pre- vs Post- Tests	Improved test score
Biology (Ruron & Hastuti, 2020)	No	Weeks	Post-Tests	Improved test score
Pharmacology of Epilepsy (Shawahna & Jaber, 2020)	Yes*	Immediate 4 Weeks	Pre- vs Post- Tests Delayed vs Immediate Post Tests	Improved test score No improvement
Astronomy (Kruawong & Soontornwipast, 2021)	No	1 Week	Pre- vs Post- Tests	Improved test score
Fourth-Grade Science (Pratiwi, 2021)	No	1 Week	Pre- vs Post- Tests	Improved test score

Note: * indicates studies that compare treatment condition against control condition

1.4 Control conditions for assessing crossword effectiveness

Crossword studies rarely use control conditions as mentioned above, and when they do they tend to use one of two forms. One set of control conditions compares learning using a crossword to learning the vocabulary using another strategy, such as a key term list (Davis, Shepherd, & Zwiefelhofer, 2009) or self-learning module (Gaikwad & Tankhiwale, 2012). This comparison assesses usefulness of crossword puzzles against other alternatives. It helps answer the question of whether students will benefit from allocating their study time for crosswords, given they can also learn the same materials using other strategies. If the goal is to optimize study time, then this control strategy will be important for assessment.

In this dissertation, however, our goal is to assess crossword effectiveness in a more natural setting; namely, whether students will benefit if a crossword learning task is added to their regular study routine. If so, it will show how much learning gain a person may achieve simply by spending time on a crossword and can use a different type of control condition. This assessment takes into account the idea that students are more likely to choose crosswords over word lists or self-learning modules, regardless of their actual benefits, because they are more game-like and less intimidating than the alternatives (Rambli et al., 2013).

To answer questions about the relative effectiveness of learning interventions such as crossword learning, two types of control conditions are typically used. Type A compares the use of a crossword to a no-crossword condition (Kruawong &

Soontornwipast, 2021; Yuriev et al., 2016). This is usually the case when a crossword is added as an additional learning activity to the regular study routine. Type B assesses the addition of a crossword over the assigned learning strategy (Abuelo, Castillo, & May, 2016; Whisenand & Dunphy, 2010). This comparison controls for previous learning experiences by presetting learning strategy to be the same for all students prior to the crossword intervention. To assess crossword's effectiveness in a naturalistic way (either Type A or B), the preferred setup is therefore a within-subjects design. Doing so allows each student to serve as their own control, and the difference between control and intervention conditions can be directly attributed to the crossword.

2 Learning Enhancement

This chapter discusses three key factors that have been shown to be beneficial to overall learning outcomes in various learning contexts. The factors I will discuss include constructive learning, collaborative learning, and elaboration. The chapter ends with suggestions on a subset of elaboration strategies that can potentially be adapted to and then included in crossword puzzles as additional learning tasks.

2.1 Operationalization of constructive learning

For its benefit to improve, a crossword puzzle may need to include constructive-learning elements. It should be noted here that the term *constructive*, in the broad literature of education, has a wide range of definitions. For example, constructive learning sometimes refers to a process through which the learner discovers knowledge that is new to either the whole of mankind or simply just himself or herself (Bruner, 1961). In most publications in educational research, constructive learning refers to either a constructivist perspective or a sociocultural perspective on learning theory (Cobb, 1994). From the constructivist perspective, the term is often taken to mean a learning process in which a learner, upon being exposed to new and unfamiliar information, gains knowledge through reconstruction of his or her existing knowledge (Piaget, 1977; Schifter & Simon, 1992). On the other hand, the term from a sociocultural perspective means a learning process in which the learner gains knowledge through interaction with another individual(s), society, or physical environment (Bailey & Pransky, 2005; Vygotsky, 1978).

The main difference between the two perspectives is the emphasis being placed on how the learning process occurs. While the constructivist perspective emphasizes that constructive learning is a result of cognitive development that occurs within an individual, the sociocultural perspective emphasizes external factors such as social interaction and culture, that is, knowledge is gained through the use of symbol systems such as language (Cobb, 1994; Windschitl, 2002). However, there are aspects that both perspectives share. In either case, constructive learning refers to a process of reconstruction of knowledge or, specifically, integration of new information into existing knowledge, resulting in new knowledge (Amineh & Asl, 2015). Also, learners gain new knowledge through active thinking as opposed to knowledge being fed by other individuals such as parents, teachers, and colleagues (Richardson, 1997).

However, this dissertation deliberately avoids the theoretical and application-based debates about how constructivism is defined or should be implemented. Instead, I will use the term **constructive learning** as operationalized and adopted by Chi and Wylie (2014) in their framework on four modes of cognitive engagement. The four modes include *Interactive*, *Constructive*, *Active*, and *Passive* (henceforth referred to as the ICAP framework). Because the ICAP framework focuses on classification of specific actions taken to process new knowledge, which corresponds with the dissertation's key concept of adding a learning task to the crossword puzzle. In this dissertation, constructive learning (or generative learning) refers to a process through which the learner produces an output including content that goes beyond what is provided in the original learning materials (Chi & Wylie, 2014; Fiorella & Mayer, 2016). For example, if learner writes a summary of learned knowledge from his or her own interpretation or draws a diagram that is not initially included in a problem set, such a learning process is considered constructive learning. Thus, although there is substantial overlap between the prescriptive notion of *constructivism* to describe a theoretical approach to education practice and the much more limited descriptive use of *constructive* within the ICAP framework, I will use constructive primarily in the same sense as it appears within the ICAP framework.

2.2 ICAP Framework

According to the ICAP framework, different learning activities involve different levels of student engagement that produce different levels of learning benefits (Chi & Wylie, 2014). The framework operationalizes the term *engagement* as “overt behaviors that students can undertake and teachers can see.” (Chi & Wylie, 2014, p. 220). That is, depending on how student overtly interacts with learning materials, such interaction (e.g., hand movement during the note-taking process) results in a certain level of learning outcome. The framework divides engagement into four different modes as follows.

- *Passive* – In this mode of engagement, students do not show any overt behavior related to the currently engaging learning activity while they are receiving information from learning materials. Some examples of *passive* learning activities include simply sitting still while listening to a lecture or watching a video. It is important to note here that passive learning does not necessarily mean a lack of student attention towards learning activity. In this mode, students can fully focus on listening and processing the materials even though they do not show any sign of doing so.
- *Active* – This mode of engagement involves students’ overt motoric action while receiving information as well as physical manipulation of learning materials.

Some examples of *active* learning activities include finger movement when students point at texts as they read a book (Alibali & DiRusso, 1999), hand movement when students underline or highlight a portion of texts on which they are currently focused (Katayama, Shambaugh, & Doctor, 2005), and hand and arm movements necessary in mixing two chemicals together (Yaron, Karabinos, Lange, Greeno, & Leinhardt, 2010). The important point here is that overt motoric behaviors must relate to currently engaging learning activity and cause focused attention at the object of the behavior (Chi & Wylie, 2014).

- *Constructive* – This mode of engagement focuses more on the learning output than overt behavior involved in a learning activity. In this mode, students "generate or produce additional externalized outputs or products beyond what was provided in the learning materials" (Chi & Wylie, 2014, p. 222). For example, summarize a text passage and rewrite it in one's own words as opposed to copying the passage as it is (which is categorized as *active*). Both activities involve hand motoric movement in writing texts, but the outputs are different. Since the focus in this categorization is placed on the generation of outputs whose content goes beyond what is provided in the original learning materials, *constructive* activities are sometimes referred to as generative learning (Fiorella & Mayer, 2016).
- *Interactive* – This mode of engagement basically covers *constructive* learning activities that are conducted by multiple students. While *interactive* mode refers to the interaction between students, the key points here are that each individual student involved in a learning activity must first take actions that are considered

constructive, and their interaction itself must contribute to the addition of knowledge to one another. For example, an interaction that occurs between two students as they are blindly copying each other's notebooks (which categorized as *active*) does not fall into this mode (Chi & Wylie, 2014). *Interactive* often involves constructive learning that is done in a group. For example, if the students look at each other's notebooks, summarize the content, and then rewrite it down in their own words, such interaction is then considered *interactive*.

Different modes of engagement do not involve the same set of underlying knowledge-change processes, and the level of learning benefits will also be different. Specifically, the ICAP framework predicts that learning activities that fall under the *interactive* mode of engagement will lead to the achievement of a higher level of learning and, as a result, a deeper understanding of learned materials than activities that fall under the c*onstructive* mode. Furthermore, c*onstructive* activities result in a greater level of such benefits than *active* activities, which in turn are greater than *passive* activities (Chi & Wylie, 2014). In short, *Interactive* > *Constructive* > *Active* > *Passive* in terms of level of learning. In addition to evaluating crossword's effectiveness to enhance memory retention of learned technical terms, the other purpose of this dissertation is to explore the potential of crosswords as a learning-aid tool that goes beyond learning vocabulary and associated facts. One might argue that filling blank boxes with letters is an act of constructive learning (because the letters are not originally provided and, therefore, must be generated). However, it can also be thought of as an act of turning those letters that are originally in the boxes visible, making it a simple active recall process. To ensure that a crossword can provide a legitimate constructive learning experience and produce learning outcome at a similar level, this learning-aid tool needs an additional task that falls under the *constructive* category.

2.3 Constructive learning: Going beyond factual knowledge

According to Bloom's taxonomy of educational objectives (Bloom, 1956; Wilson, 2016), high-level scientific reasoning, including analysis, evaluation, and creation of concepts, is regarded as a higher-level objective, whereas remembering and understanding facts and concepts are regarded as a lower-level objective. Since the nature of crossword puzzles involves recognizing the connection between words and their definition or associated facts which are provided as hints, its learning benefit is limited due to appropriate transfer, meaning that one can get better at recognizing word-definition pairs and not anything else beyond (Koedinger, Corbett, & Perfetti, 2012; Roediger & Karpicke, 2006). Thus, it is expected that the benefit of crosswords on learning gain will be limited to the lower-level objective (Whisenand & Dunphy, 2010). That is not to say, however, that crossword does not contribute at all to achieving the higher-level objective. In order to reach the higher level, one needs to first master the lower level, and excelling in vocabulary provides a strong foundation of knowledge as well as accelerates higher-level learning (Sweller, 1994). Therefore, it is necessary to search for possible modifications to upgrade crossword puzzles so that they can help a person achieve a higher-level objective.

Constructive or generative learning strategies involve generating something beyond what is already provided in learning materials. They have been shown to not only produce stronger memory traces of newly gained knowledge, but also develop a more in-depth understanding which goes beyond the content of that new knowledge (Chi et al., 1989; Chi & Wylie, 2014; Fiorella & Mayer, 2016; Levin, 1988). Elements of constructive learning can be observed being implemented in various learning environments both directly, such as an intelligent tutoring system (VanLehn, 2011), and indirectly, such as learning-oriented conversation within a group of students during a problem-based learning cycle (Yew & Schmidt, 2009). In the case of a crossword puzzle, it may be possible to include constructive learning elements in a form of learning task that asks learners for more information than just knowledge of the words and their definition or associated facts. Some of the tasks that promote constructive or generative learning and are suitable for crosswords include summarization, mapping knowledge, and self-explanation (Fiorella & Mayer, 2016). Using summarization, a person may digest information provided as word hints, selecting only their important parts, and writing them out. The task of generating paraphrased hints can be counted as summarization. In mapping knowledge, a person may use the connection between the learned word and its hint to generate other related words or concepts. The task of generating synonyms, antonyms, or other semantically related words can be counted as mapping (Carpenter, 2011). Self-explanation is similar to summarization, but the task also allows possible add-on of information. Through the act of explaining the learned knowledge to oneself, a person may think of some other related information that helps one remember and apply the knowledge better (Fiorella & Mayer, 2016).

2.4 Collaborative learning: Sharing benefits among multiple learners

Another factor whose contribution to achieving higher-level knowledge is on par with that of constructive learning is collaborative learning. It has become a norm that schools and workplaces to promote collaborative learning in some way such as group assignments and team projects (Wang, 2009). Collaborative learning occurs when multiple learners work together as a group toward the same learning objectives (Bonwell & Sutherland, 1996; Smith & MacGregor, 1992). Collaborative learning can provide constructive learning benefits to all learners in a group, given that each of them shares his or her own knowledge and ideas and coordinates with others to put that knowledge and ideas into practice (Johnson, Johnson, & Smith, 1991). Similarly, the ICAP framework states that only when learners constructively contribute their knowledge to the group will interaction (a crucial process in collaborative learning) between them be beneficial. Research in this area has resulted in the development and implementation of many practical strategies which promote collaborative learning; some of those strategies are small-group teaching (Sharan & Sharan, 1976), and jigsaw strategy (Aronson et al., 1978), and Teams-Games-Tournament (DeVries & Slavin, 1978).

For example, Teams-Games-Tournament, as one strategy that enforces collaborative learning, has been widely accepted and implemented in many practical classroom settings (DeVries & Slavin, 1978; Slavin, 1980). In this strategy, students are put together in a team of 4-5 members. They receive a lecture and then a worksheet covering the same academic materials, and their task is to study the worksheet together and help each other in mastering the materials. Days later, student's knowledge of learned materials is evaluated in a tournament format, in which each person representing their team competes against other students from different teams. The tournament is structured in such a way that students from each team with the same level of academic performance compete against each other. The knowledge test is a quiz game covering academic materials similar to the content in the lecture and worksheet.

The mechanics behind this strategy is that collaborative learning will have an effect favoring constructive learning for each student in a team. During the learning phase (team members studying a worksheet together), the learning task encourages students to generate discussion in order to arrive at a common understanding of the worksheet content. For that to happen, the exchange of knowledge must be done in a productive way. Strong students who quickly learn the materials help other members in their learning process. The evaluation phase is when the effect of collaborative learning becomes apparent. At this point, students no longer receive help from their team, instead, they contribute to the team by trying to score as many correct answers as possible. Thus, the more knowledge students gain from their team, the more points they can earn for the team. This strategy results in significantly higher academic performance in curriculum-specific tests and, in some cases, even in standardized tests (Slavin, 1980). Compared to the benefits of constructive learning by individual effort, collaborative learning results in even higher productivity and learning outcomes (Johnson & Johnson, 2009; Johnson et al., 1991).

2.5 Elaboration in human cognition

Among several learning strategies that promote constructive learning, this dissertation focuses on elaboration because it seems to go along well with crosswords as a follow-up task. The concept of elaboration and its effect on memory and learning have been studied for more than half a century (Tulving & Madigan, 1970). Back then, the term referred to deep-level information processing involving an effortful attempt to form connections between new information to what a person already knows (Craik & Lockhart, 1972; Craik & Tulving, 1975; Fisher & Craik, 1980; Greene, 1987). Ever since then, a large number of studies in the human memory literature have continued to explore the elaboration process in the human brain, attempting to get a better understanding of its underlying mechanisms and beneficial effects on memory as well as on learning.

What is known so far about encoding and information retrieval in the human memory system is as follows. When people intend to remember some new information, they take it in through their perceptual sense (e.g., eyes and ears), and the information gets held temporarily inside their working memory or WM (Unsworth & Engle, 2007). It is widely known that WM contains short-term storage of 3 - 5 chunks of information (Cowan, 2001; 2008). When the total amount of to-be-remembered information exceeds that capacity, chunks of information that are already inside WM must be processed first to be transferred to long-term memory (LTM), otherwise, those chunks may be lost (Cowan 2008; Unsworth & Engle, 2007). If information is successfully encoded and stored inside LTM, then it will be there permanently, and one can say that the information is learned. The issue here, though, is how to retrieve learned information stored in LTM when it is needed. A critical factor that determines the success rate of retrieval lies within the encoding process. Specifically, information needs to be encoded in such a way that leaves a strong enough memory trace for that particular information (Marschark et al., 1987). One solution to this is to apply elaboration techniques for encoding information as they leave an enriched memory trace and thus increase the odds of successful retrieval of information stored in LTM (Bailey, Dunlosky, & Kane, 2011; Galli, 2014; Marschark et al., 1987). The high degree of elaboration results in broader understanding as well as more durable retention of learned information (Craik & Lockhart, 1972).

It turns out that, even today, the essence of elaboration has not differed much from the past, as the process was still described in one of the latest studies as "enriching the memory representation of an item by activating many aspects of its meaning and by linking it into the pre-existing network of semantic associations" (Bartsch & Oberauer, 2021, p. 1). What this means is that when a person elaborates on an item (incoming information, to-be-learned knowledge, etc.), he or she explores it for features by mentally activating many aspects of the item's meaning. The person then forms connections between each of those features to his or her existing memories. Based on this explanation, elaboration seems to refer to any learning activity or technique that requires a person to analyze new knowledge for pieces that are meaningful, thus memorable, or those that fall within the scope of his or her prior knowledge.

2.6 Elaboration strategies: Add-ons to crossword task

Perhaps the fruitful result from on-going research on elaboration is not so much an expansion of its definition but rather strategies implemented to exercise it. This notion of integrating prior and new bodies of knowledge is highly associated with the concept of constructive learning (Glaser, 1991). Therefore, the aforementioned constructive learning tasks: mapping knowledge, self-explanation, and high-level summarization, if done right, can also be counted as elaboration strategies. Other studies have either devised and tested new strategies or applied and re-evaluated old ones (Bartsch & Oberauer, 2021; Richardson, 1998). As a result, there are now many elaboration strategies along with evidence of their positive effect on memory performance; among those strategies include verbal mediators (Wood & Bolt, 1968), imagery (Paivio & Yuille, 1969), keyword mnemonic (Fritz et al., 2007), elaborative retrieval (Carpenter, 2011), and sentence generation (Turley-Ames & Whitfield, 2003). As this dissertation focuses on improving the effectiveness of crossword puzzles in learning technical terms by adding an elaboration element to them, three elaboration strategies that seem to fit with crosswords have been considered.

2.6.1 Strategy 1: Summarization

The first elaboration strategy chosen for this study is summarizing knowledge of learned term-definition pair associates. Summarization, a widely known learning strategy, is mostly used to digest a content of information and extract the important portion of it. Since the process oftentimes results in generating the summarized text in one's own words, which is not part of the original content, it can be considered a legitimate constructive learning strategy, given done properly (Fiorella & Mayer, 2016). However, summarization might not be perceived as elaboration because it is not apparent that the process requires integration with prior knowledge (Hidi & Anderson, 1986). However, that is not necessarily the case. To provide a high-quality summary, it is crucial that learners need to know enough about the original information. After effortful learning attempt(s), original information then becomes old knowledge (prior to summarization), and learners must ensure a solid connection between summarized text (newly generated knowledge) and original information (Hidi & Anderson, 1986). Additionally, generating a high-quality summary requires deep processing of the original information to understand, recognize, and recall the important portion (Stein & Kirby, 1992). These reasons justify summarization being counted as an elaboration strategy. In this dissertation, the strategy will be applied in reading and comprehending words and their corresponding definitions (initially presented as crossword hints), after which a short summary will be generated for each word-definition pair. A portion of the generated summary can then be selected as a shorter cue for future recall. Each cue can also then be treated as a new hint for its corresponding word.

2.6.2 Strategy 2: Association Generation

The second elaboration strategy chosen for this study is generating words conceptually related to the target words. Other studies on learning and remembering words or concepts have tested the effectiveness of this strategy (Carpenter, 2011; Karpicke & Smith 2012; Lehman, Smith, & Karpicke, 2014). For example, Karpicke and Smith (2012) conducted a two-learning-phase experiment. During phase 1, all participants learned a list of word pairs, so that they supposedly shared the same base knowledge of the words before being split into several experimental conditions. During phase 2, participants within the elaboration condition restudied the same word pairs and generated additional words that could be associated with original words in the pairs. It was assumed that participants would encode both original words and newly generated words within the process and that remembering any of them would assist in correctly recalling target words. The finding was that participants within the elaboration condition significantly outperformed those in the control condition who only studied words during the first phase.

This suggests that one can improve learning by simply doing an additional elaborative learning task (i.e., generating related words) which supposedly consumes little extra time. With this evidence showing such benefit, the strategy seems to have the potential in improving crossword's effectiveness in learning word-definition pairs as well. For this dissertation, the strategy was slightly modified – generating three words or concepts related to only the learned technical terms, but not necessarily related to their corresponding definitions (provided as crossword hints).

2.6.3 Strategy 3: Sentence Generation

The third elaboration strategy chosen for this study is the generation of a sentence or story from target words. Other studies on learning and remembering words or concepts have tested the effectiveness of this strategy (Dunlosky & Kane, 2007; Hamilton, 1997; McNamara & Scott, 2001; Turley-Ames & Whitfield, 2003). In one study, for example, participants were given a task to remember words (Turley-Ames & Whitfield, 2003). Some of the participants were prompted to generate a sentence or story which contains to-be-remembered words. The finding was that participants who were prompted to utilize this elaboration strategy performed better in the working-memory task than those who received no such prompt.

While it has already been shown that this elaboration strategy has some benefits in remembering words, it is important to note some major differences between those studies and the ones in this dissertation. The other works studied the implementation and effectiveness of the strategy in maintaining target words in working memory (Dunlosky & Kane, 2007; McNamara & Scott, 2001; Turley-Ames & Whitfield, 2003). However, in this dissertation, the strategy would be used to increase the probability of successful retrieval of target words from long-term memory after a several-day delay. Additionally, in the previously mentioned study by Turley-Ames and Whitfield (2003), a generated sentence or story would contain all target words. This might cause participants to not remember exactly which words were target words or misremember filler words as target words. To mitigate the occurrence of such a problem, the elaboration strategy used in this dissertation would be to generate, for each of the learned words, a short sentence that portrays a brief story or scenario; in other words, generating one sentence per word.

3 Overview of Dissertation Experiments

In order to investigate the effectiveness of crossword puzzles on learning technical terms in psychology, I have conducted seven experiments. Specifically, throughout the series of seven experiments, my goal is to investigate the effect of a crossword puzzle with an add-on elaboration on students' ability to (a) retain memory of learned terms for short period (1-3 days) as well as a moderately long period (5-7 days); and (b) provide a more in-depth explanation for those terms.

In Experiments 1 and 2, I measured students' learning outcomes from solving crossword puzzles, both with and without an additional elaboration task of generating new crossword hints for each technical term they learned throughout the process (Hidi & Anderson, 1986). Using a within-subjects design, each student served as their own control. I compared their performance from how they normally learn new terms (e.g., attending lectures and reading textbooks) to how they learned with the crossword puzzle or a crossword puzzle with elaboration. I inserted a delay of at least 24 hours between the learning phase and the knowledge evaluation phase, representing a situation in which students study technical terms a day right before their exam.

In Experiment 3, I incorporated a control for time-on-task by providing the same study time with and without elaboration tasks. I did so by adding a task of solving the same crossword a second time and studying it to the word-game-only condition as well as putting an equal time-limit on both conditions. For the elaboration condition, I further assigned students with one of the two elaboration tasks: generating new crossword hints (same as in experiments 1 and 2) or generating connections between new terms and other terms students have already known (Carpenter, 2011). The purpose of this change was to test out the benefit of a new elaboration task on learning outcome as compared to that without elaboration. For the three experiments previously mentioned, I added a collaborative learning factor to the study. To have my experiments represent a practical classroom environment, I therefore designed my experiments such that participants are instructed to work together in solving the crossword puzzles.

For experiments 4, 5, and 6, I moved my experiment to an online environment due to the COVID-19 pandemic. This change resulted in a shift from collaborative learning to individual learning because the crossword sheets were electronically distributed to each participant individually. My objective stayed the same – to investigate the crossword's effect on abilities to (a) retain memory of learned terms for a moderately long period of time and (b) to provide a more in-depth explanation for those terms. To that end, I decided to keep the new elaboration task of generating term connections and removed the hint-generation task. I also modified the delay period by extending it to five days to explore the long-term benefit of learning terms with crosswords. Other procedures stayed the same.

Experiments 4 and 5 shared the exact same materials and procedure, but they were intentionally conducted at different times of the year to investigate two different teaching contexts. Specifically, Experiment 4 was conducted at the end of the semester, so the crossword served as a review tool, while Experiment 5 was conducted at the beginning of the semester, so the crosswords served as an introduction of new knowledge. Given that all other factors were exactly the same, comparing the results from these two experiments should give me a difference (if any) in the efficacy of crosswords as a review tool versus an introduction of new knowledge.

Experiment 6 was the same as Experiments 4 and 5, except that I changed the elaboration task from generating term connections to generating sentences from the learned terms (Turley-Ames & Whitfield, 2003). The purpose of this change was to allow students to practice putting the learned term into practical concepts based on their own knowledge and experience. This also allowed me to evaluate the effect of the new elaboration task when added to a crossword.

Experiment 7 compared the effectiveness of different elaboration tasks used in this dissertation. It was conducted to fill in the gap in experimental designs, since none of the previous six experiments directly compared different types of elaboration tasks. Through this experiment, I could determine which of the three elaboration strategies would be most compatible with crossword puzzles.

4 Preliminary Experiments Adding Elaboration to Crosswords (Experiments 1 and 2)

4.1 Introduction

The purpose of these first couple of experiments was to answer a simple question of whether crossword puzzles have any effect at all on learning and retention of technical terms. Therefore, in treatment conditions, participants would engage in solving a puzzle on top of what they have normally done inside and outside of class to study the terms. This was to establish if there is an effect at all in a naturalistic context, that is, whether crosswords serve as an additional and not alternative learning option. Experiment 2 served as a replication of Experiment 1, with one change in stimuli being a different set of technical terms.

4.2 Hypotheses

H1: Adding a crossword puzzle (both with and without add-on elaboration) will improve memory performance compared to the baseline.

H2: Adding a structured elaboration task to a crossword puzzle will improve memory performance compared to a crossword alone.

4.3 Methods

4.3.1 Participants

Experiment 1 consisted of 26 undergraduates and Experiment 2 consisted of 40 undergraduates recruited from *Introduction to Psychology* courses, including 15 males (58%) and 11 females (42%) for Experiment 1 and 28 males (70%) and 12 females (30%)

for Experiment 2. All participants are over 18 years old, and they received course credit for their voluntary participation. All experiments received IRB approval. Each person received a unique ID to complete the second part of the experiment.

4.3.2 Materials

Forty-two technical terms of similar difficulty were selected from three different chapters of the course textbook (Myers & DeWall, 2017). For Experiment 1, the selected chapters included nature and nurture, development through the lifespan, and sensation and perception. For Experiment 2, the selected chapters included consciousness, learning, and memory. Crossword puzzles for these scientific terms were generated using a crossword puzzle maker from *education.com*. Each puzzle consisted of 14 scientific terms from a single chapter (see APPENDIX A).

4.3.3 Experimental Design and Variables

The experimental design was a 1 x 3, within-subjects study to examine the effect of training condition on learning. Training condition order and chapter were counterbalanced using a within-subjects, Latin-squares design (see Table 2). The independent variable consisted of three training conditions: Word-Game + Elaboration (WGE), Word-Game-Only (WGO), and baseline without additional crossword training (control). In WGE condition, participants were given 35 minutes to solve a crossword puzzle and then elaborate by generating a new crossword hint for each term based on their own knowledge of the term. For example, presented with the term *withdrawal* along with a hint, “discomfort and distress that follow discontinuing an addictive drug or behavior,” participants might generate a new hint for the term such as, “abnormalities that occur to your body when it needs an addictive substance." Prompt for the elaboration task in WGE condition was as follows.

Elaboration Task – Instructions

Please type the crossword terms that you just solved and a new clue/hint for each of those terms.

Example

Term:	Withdrawal
Original Clue:	Discomfort and distress that follow discontinuing an addictive drug or behavior
Your Answer:	Abnormalities that occur to your body when it needs an addictive substance

Table 2. Latin-squares design to counterbalance three training conditions for all participants

Participants	Puzzle #1	Puzzle #2	No Puzzle
Group 1	Chapter A	Chapter B	Chapter C
Group 2	Chapter B	Chapter C	Chapter A
Group 3	Chapter C	Chapter A	Chapter B
Group 4	Chapter A	Chapter C	Chapter B
Group 5	Chapter B	Chapter A	Chapter C
Group 6	Chapter C	Chapter B	Chapter A

Note: Chapters A, B, and C represents 3 different sets of technical terms.

In the WGO condition, participants were given 25 minutes to solve a different crossword puzzle. The baseline condition involved no additional training other than what students did outside the experiment (e.g., attending class, reading textbook, etc.) and, thus, a crossword puzzle that fell under this condition would not be presented to participants. The dependent variable was recognition of memory for the learned terms, which was measured using a 42-item multiple-choice test covering all three chapters. This means that the test would measure participant's knowledge on learned terms with crossword puzzle interventions (i.e., book chapters that fell under WGO and WGE conditions), as well as those without an intervention (i.e., chapters that fell under baseline conditions).

The aforementioned design of knowledge test is practically the same as pretest-posttest. In many cases, pretests evaluate participant's knowledge before an intervention while posttests evaluate knowledge after intervention (Abuelo et al., 2016; Gaikwad & Tankhiwale, 2012; Gilani et al., 2020; Yuriev et al., 2016). If the actual time when those tests take place is to be disregarded, however, then pretest-posttest is essentially an evaluation of knowledge with and without intervention. As in Kruawong and Soontornwipast's (2021) study, their pretest was actually a final knowledge test covering materials learned from the regular classroom only, while the posttest was a knowledge test covering different materials from classroom learning with a crossword intervention. The majority of experiments described in this dissertation used one single test that covered learned technical terms both with and without crossword interventions, making the test highly similar to pretest-posttest. Also, since the experiments utilized within-subjects design, all participants were measured against themselves, meaning they served as their own control for knowledge with and without interventions.

Using this test design would also help prevent the testing effect which has been found to have strong influence on posttests in pretest-posttest design (Johnson & Mayer, 2009; Roediger & Karpicke, 2006). The testing effect causes a participant to perform better on a retention test after a practice test (e.g., pretest) and each subsequent posttest. It is possible to mitigate the effect by ensuring that test questions and covered materials are not the same for pretests and posttests (Abuelo et al., 2016). Nevertheless, by giving participants just one retention test that functions similarly to pretest-posttest, one can completely remove the testing effect from design as well as further reduce possibility for other confounds from doing multiple retention tests.

4.3.4 Procedure

Experiments 1 and 2 were conducted in the lab. Participants worked together in groups using the course textbook and made responses via a computer. Each experiment consisted of two phases: a 60-minute training phase covering both WGE and WGO training conditions (as explained in the previous section) and a 60-minute evaluation phase which was to be completed 24 hours later. During the training phase, participants were put together into small groups of 2-3 people upon entering an experiment site. Each group was provided a computer station, a copy of the course textbook, two different crossword puzzles, and writing utensils. Participants were instructed to utilize the textbook to help them find keywords based on the crossword hints. Each group was randomly assigned to one of the six configurations for a training condition order and completed all conditions in that order (see Table 2). For example, a group of three participants worked together to first solve a crossword puzzle consisting of technical terms from chapter B (for WGO condition); they then solved a crossword puzzle consisting of technical terms from chapter C and generated new crossword hints for each of those terms (for WGE condition). Training phase ended when participants submitted all of their responses including solved crossword puzzles and generated hints. During the evaluation phase, each participant received an access to an online form including a 42-item multiple-choice test. Participants could choose to complete the test at their preferred time and place. Participants were instructed to work alone and not to use any resources besides their own memory to help them complete the test.

4.4 Results

Figure 1A depicts the means and standard errors of participants' performance in the multiple-choice test in Experiment 1 while Figure 1B depicts those in Experiment 2. To evaluate H1, a 1 x 3 ANOVA was conducted using R statistical computing program, taking both training conditions (3 conditions) and crossword puzzles (3 different sets of technical terms) into account. The test indicated a significantly large effect of training condition on participant's *recognition performance* (multiple-choice test results): Experiment 1, $F(2, 48) = 35.41$, $p < .001$, and Experiment 2, $F(2, 76) = 53.97$, $p < .001$. In addition, planned comparisons using one-tailed, paired t-tests were conducted using pwr package in R. Aside from p-value denoting statistical significance, the following reports effect size denoted by Cohen's d (Cohen, 1988). Cohen's d indicates standardized difference of mean between two groups, which basically informs how large (or small) an effect is due to an intervention or change in treatment. As a general rule of thumb, d of

0.2 can be interpreted as small effect size, 0.5 as medium effect size, and 0.8 as large effect size (Cohen, 1988).

The paired t-tests showed statistically significant improvement between the WGO and Control conditions in Experiment 1, $t(25) = 5.22$, $p < .001$, $d = 1.01$, as well as in Experiment 2, $t(39) = 3.98$, $p < .001$, $d = 0.88$. Significant improvement between the WGE and Control conditions could also be found in Experiment 1, $t(25) = 8.12$, $p < .001$, $d = 1.76$, as well as in Experiment 2, $t(39) = 7.50$, $p < .001$, $d = 1.72$. Based on these results, H1 was supported, suggesting that solving a crossword for technical terms (with or without an add-on elaboration task) improved memory performance on the terms. To evaluate H2, one-tailed, paired t-tests showed significantly greater performance when comparing WGE condition to WGO condition in Experiment 1, $t(25) = 3.04$, $p = .003$, $d = 0.67$, as well as in Experiment 2, $t(39) = 3.37$, $p < .001$, $d = 0.73$. Based on these results, H2 was supported, suggesting that adding a structured elaboration task to a crossword puzzle improved resulting memory performance even further compared to a crossword without an elaboration task.

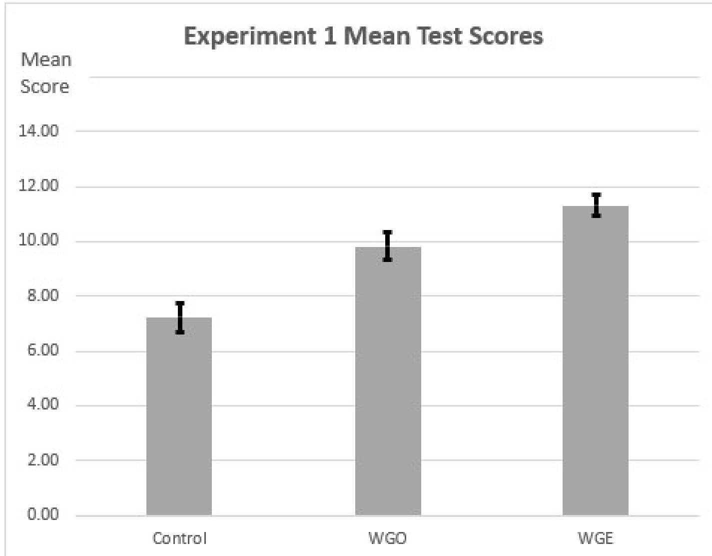

Figure 1A. Experiment 1 mean scores (out of 14 points) and standard error by training condition after 24 hours. The three conditions are Control, Word-Game-Only (WGO), and Word-Game + Elaboration (WGE).

Figure 1B. Experiment 2 mean scores (out of 14 points) and standard error by training condition after 24 hours. The three conditions are Control, Word-Game-Only (WGO), and Word-Game + Elaboration (WGE).

4.5 Discussion

These experiments showed that having students spend approximately half an hour solving crossword puzzles for technical terms in addition to studying for the terms by attending class, reading textbooks, and working on assignments greatly improved their memory retention of gained knowledge on the terms. Also, by simply adding an extra elaboration task to crossword puzzles, which took only 10 more minutes of student's time, the retention was further improved. These improvements can be attributed to two factors: elaboration and time-on-task. The elaboration task used in these experiments (i.e., generation of new hints based on knowledge on learned terms) shares the same nature as summarization, in which an individual selects and integrates incoming information to existing knowledge, producing a different version of the information (Fiorella & Mayer, 2015). Therefore, elaboration can be considered a form of constructive learning which results in stronger memory retention of learned terms (Chi & Wylie, 2014). However, it is important to note that a low level of summarization, such as simply applying synonyms to generate new hints, does not necessarily produce new information. While such work of elaboration improves memory retention of gained knowledge, it might not lead to a better understanding of that knowledge.

Regarding time-on-task, the more time spent on performing a learning task, the better the learning outcome is (Wagner, Schober, & Spiel, 2008). In a more practical circumstance, however, students may have a limited amount of time in a day they can spend studying (Romero & Barbera, 2011). In pursuing improved learning outcomes, adding more learning tasks, thus resulting in more study time, might not be a practical solution. In order to prove the elaboration task's benefits, it must be compared with non-elaboration conditions for the same amount of time spent. This issue led to new experiments that control for time-on-task for both WGO and WGE conditions.

5 Evaluation of Elaborative Summarization Strategy in the Crossword (Experiment 3)

5.1 Introduction

The purpose of this experiment was to take potential issues found in the previous experiments into account: unequal time on tasks across the two crossword conditions and simple paraphrasing for elaboration. In solving the time-on-task issue, I extended the task time in the WGO condition to be equal to that in WGE condition. Looking at participants' responses from Experiments 1 and 2, there seemed to be a considerable portion of generated hints that came from simple paraphrasing – a mere change of wording in the original hints by using synonyms. In order to unpack the relationship between the quality of the generated hints and participant learning gain, the hints were coded and analyzed degree of elaboration. Additionally, foreseeing that such elaboration might not be powerful enough, an alternative elaboration task – generating connections with conceptually related words, that relies more on a student's semantic knowledge was added. The experiment did not remove the hint-generation task from the design; instead, both tasks were included in the same WGE condition and counterbalanced. Hypotheses for Experiment 3 were the same as those for Experiments 1 and 2. The methods and procedure were the same as Experiments 1 and 2, but with the following changes: I reduced the number of technical terms (stimuli) presented to participants, used a combination of chapters from Experiments 1 and 2, controlled for time on task, and added the aforementioned alternative elaboration task.

5.2 Methods

5.2.1 Participants

Experiment 3 consisted of 24 undergraduates recruited from the *Introduction to Psychology* course. All participants were over 18 years old and received course credit for their voluntary participation.

5.2.2 Materials

Thirty technical terms were selected from the following three chapters: sensation and perception, learning, and memory. Each of three crossword puzzles consisted of 10 scientific terms from a single chapter. The coding scheme for hints participants generated consisted of three aspects: correctness, word count, and quality. Correctness was determined by whether or not the generated hint corresponded to its original definition (scored as 0 or 1). Word count evaluates length of generated hint (scored as number of words included in the hint). Finally, quality of the hint was adapted from Veinott and Whitaker's (2019) hint coding scheme and scored from 0 (no information) – 3 (adding new information).

Coding scheme for elaborated-hint data – Adapted from Veinott and Whitaker (2019)

Sample Term:	Weber (from Weber's law)
Original Hint:	_____'s Law states that two stimuli must differ by a constant percentage to be perceived as different

Scoring	**Code with definition and examples**
0 Point	**No information**: No hint or a hint with semantically unrelated information. Examples: "What do spiders make?" (*P*03 & *P*04).
1 Point	**Verbatim**: Hint contains the whole or a portion of original information, with minimum change in wording. Examples: "Percent difference" (*P*28 & *P*29).
2 Points	**Restatement:** Hint restates information, demonstrating some understanding of concepts. Examples: "Two things must be a different percentage to be different" (*P*20 & *P*21).
3 Points	**Elaboration**: Hint demonstrates a deeper understanding of the concept with possible add-on information from the original. Examples: "Idea that stimuli can be perceived as different if they vary by some minimum percentage" (*P*26 & *P*27).

5.2.3 Experimental Design and Variables

The overall experimental design was the same as in Experiments 1 and 2. The within-subjects ANOVA design examined 3 training conditions: Word-Game + Elaboration (WGE), Word-Game-Only (WGO), and baseline (control). In the WGE condition, participants first solved a crossword puzzle and then did an elaboration task. In the WGO condition, participants only solved a puzzle. The baseline condition involved no additional training than what students did outside the experiment. Training condition order and chapter were counterbalanced using a within-subjects, Latin-squares design.

For the dependent variables, memory retention of learned terms was measured using a 30-item multiple-choice test.

5.2.4 Procedure

Overall procedure was the same as that in Experiments 1 and 2. Participants worked together in a small group of 2-3 people to complete learning tasks for both WGO and WGE conditions within a 60-minute time period. After a delay of 24 hours, each participant individually completed an online evaluation test on the learned technical terms within a 60-minute time period. Regarding the elaboration task in the WGE condition, participants would, at random, do one of the two tasks: participants generated either (a) new crossword hints for the solved terms (the same task as that in Experiments 1 and 2) or (b) three words or phrases that are conceptually related to the solved terms. For example, presented with the term *implicit*, participants might generate “retrieved unconsciously, without thinking, learned skills” (*P*35 & *P*36). Participants were given 30 minutes to first solve the puzzle and then complete the randomly assigned elaboration task. In the WGO condition, after participants solved a crossword puzzle, they were provided another copy of the same puzzle for them to solve again. Participants were given 30 minutes to complete the task. If they finished before 30 minutes, they were instructed to spend the remaining time studying the solved puzzle.

5.3 Results

Figure 2 depicts the means and standard errors of participants’ performance in the multiple-choice test. To evaluate H1, a 1 x 3 ANOVA was conducted using R, taking both training conditions (3 conditions) and crossword puzzles (3 different sets of technical terms) into account. The test indicated a significantly large effect of training condition on participant's *recognition performance*, $F(2, 44) = 31.89$, $p < .001$. In addition, planned comparisons using one-tailed, paired t-tests were conducted using pwr package in R. The tests showed statistically significant improvements between the WGO and Control conditions, $t(23) = 6.63$, $p < .001$, $d = 1.50$, as well as between the WGE and Control conditions, $t(23) = 5.31$, $p < .001$, $d = 1.09$. Based on these results, H1 was supported, suggesting that solving a crossword for technical terms (with or without an add-on elaboration task) improved memory performance on the terms. To evaluate H2, one-tailed, paired t-tests showed non-significant difference between performance in WGE and WGO conditions, $t(23) = 1.49$, $p = .075$, $d = 0.24$. Based on these results, H2 was not supported, suggesting that adding a structured elaboration task to a crossword failed to improve the resulting memory performance even further compared to a crossword without an elaboration task.

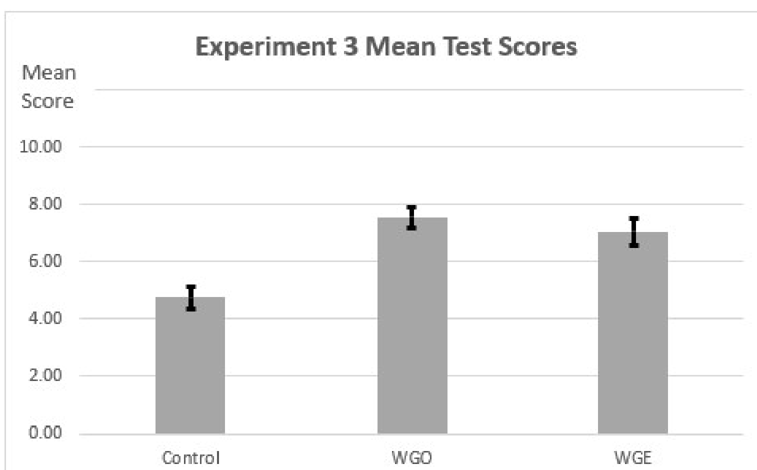

Figure 2. Experiment 3 mean scores (out of 10 points) and standard error by training condition after 24 hours. The three conditions are Control, Word-Game-Only (WGO), and Word-Game + Elaboration (WGE).

Additionally, I coded and analyzed elaborated hints which were generated by participants during the training phase. This coding applies to the hint generation task only (the same task as that in Experiments 1 and 2). Thus, only a portion of participants who did this particular task were included in this analysis ($N = 17$). Out of 90 generated responses, ten of them were selected as samples for testing inter-rater reliability (IRR). Two raters (one being the author) independently coded the selected responses based on the same coding scheme previously described under the *Materials* section. Correctness and quality of the generated responses (the latter is also referred to as *AddInfo*) are categorical variables. Thus, I chose Cohen’s kappa (Cohen, 1960; McHugh, 2012) as a measure for IRR because its calculation considers and then removes proportion of agreement in responses that is expected to occur by chance. Based on the coding scheme, the quality aspect or *AddInfo* has four dimensions which are ordinal ($0 < 1 < 2 < 3$), thus I calculated for weighted Cohen's kappa. Weighted Cohen's kappa takes into account the degree of disagreement between raters (Cohen, 1968). This is preferred over unweighted (classical) Cohen's kappa when ratings have more than two dimensions (e.g., 0 - 3 as opposed to 0 - 1 or agree/disagree) and when they are ordinal. Larger disagreement has more weight, that is, disagreement in ratings of 1 and 3 has more weight than disagreement of 2 and 3. In weighted Cohen's kappa, weight differences are included in calculations of proportion of observed agreement and proportion of expected chance agreement; both are key variables in a formula for calculating Cohen's kappa coefficient. As a result, weighted Cohen's kappa does not penalize the difference in ratings between 2 and 3 as hard as that between 0 and 1.For this experiment, weighted Cohen's kappa coefficients were measured to be 1.00 for correctness and 0.90 for *AddInfo*. For the collected data to be meaningful to research, it is important that the coefficient for IRR is greater than 0.8 (almost perfect agreement) for all dimensions (McHugh, 2012).

Once the reliability of the two raters was confirmed, participant's scores from three criteria: correctness, *AddInfo*, and word count, along with their corresponding test scores were analyzed using a logistic regression model. It appeared that 12 responses were left completely blanked by participants. After removing the blanked responses, I then applied ANOVA on the models. The analysis revealed that none of the three factors is considered statistically significant (Figure 3). This suggested that correctness, quality, and length of elaborated information on learned terms were not capable of predicting participant's test score. Additionally, Figures 4A and 4B depict scatter plots of participant's test score versus number of elaborated responses and *AddInfo* respectively. From these plots it can be inferred that a participant can achieve either high- or low-test score regardless of whether he or she generates all responses or leave some of them blank. Similarly, participants can achieve any score regardless of quality of their elaboration. These results support the findings above, explaining why participant's test performance after elaboration training (as in the WGE condition) was not better than the performance after solving the same crossword puzzle twice (as in the WGO condition).

```
            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          157     197.24
PID         16   39.256       141     157.98 0.0009989 ***
Correctness  1    0.145       140     157.84 0.7038104
AddInfo      1    1.785       139     156.05 0.1815439
WordCount    1    2.675       138     153.38 0.1019442
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3. Logistic regression analysis conducted in R on elaborated hints generated during training phase of Experiment 3. PID indicates individual difference among participants.

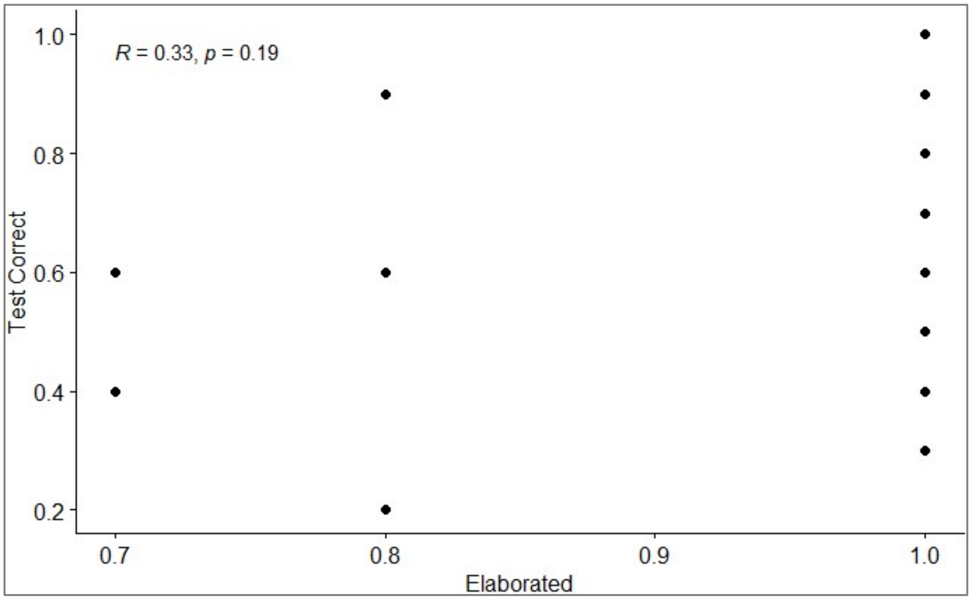

Figure 4A. Plot shows participant's test score based on how many elaborated responses they generated (as opposed to leaving blank).

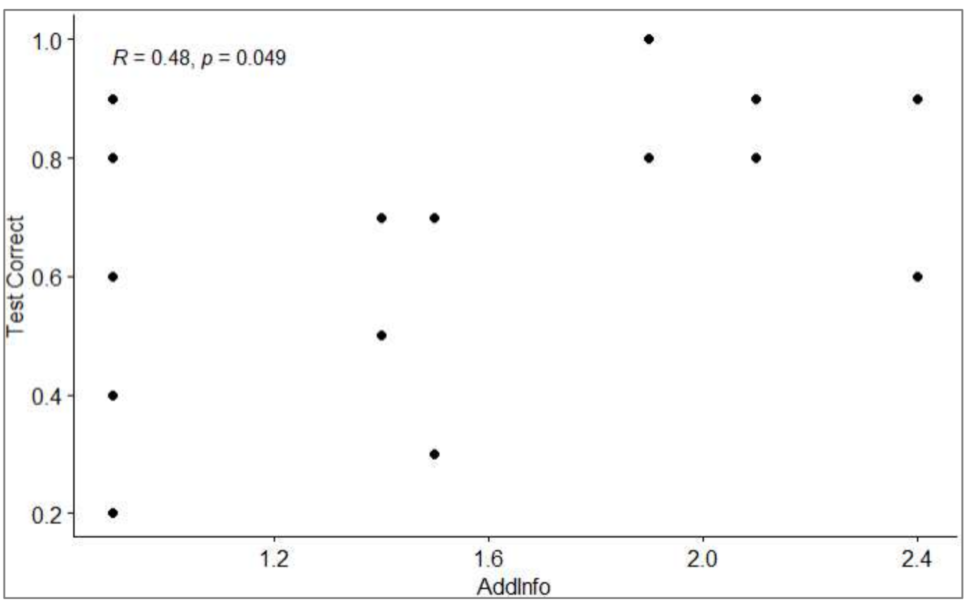

Figure 4B. Plot shows participant's test score based on quality of elaboration (*AddInfo*) in their generated responses.

5.4 Discussion

In this experiment, time on task was controlled, participants were instructed to solve the same puzzle twice. This was done mainly to make up for the difference in time spent within each crossword condition. Experiment 4 showed no difference between the WGO and WGE conditions. This suggests that the extension of time on task by adding another round of puzzle-solving could also have improved memory for the WGO through repetition on learning (Bromage & Mayer, 1986). Even though the repetition ended up adding some value to learning outcome, its effect was not expected to match that of elaboration. Yet, these results suggested that adding an elaboration task to a crossword puzzle produced nearly the same learning gain, as simply repeating the puzzle one more time.

While it is good to learn that a repetition effect is powerful enough that it rivals an add-on elaboration task to a crossword puzzle, it is also disappointing that a technically more complex elaboration task could not provide better learning outcome than just doing the same crossword again. Perhaps the problem lay within the elaboration task itself. Due to the mix-task design used in this experiment, it was not certain which task: hint generation or word-connection generation, that produced insufficient learning outcome. However, the literature suggests that the latter task (i.e., word-connection generation) risks causing cue overload, in which additionally generated words or concepts, along with target word, can all become cues and create even more cues for themselves (Unsworth & Engle, 2007). Cue overload is known to diminish the encoding performance of target words, resulting in insufficient learning outcome (Lehman et al., 2014). In order to find out whether the word-connection generation, as an add-on task to a crossword puzzle, was the true culprit that led to the aforementioned results, it was necessary to redesign the experiment so that the task could be directly compared to crossword alone. This led to new experiments whose WGE condition included only one type of elaboration task, that is, the word-connection generation (which will also be referred to as association generation).

6 Evaluation of Association-Generation Strategy in the Crossword (Experiments 4 and 5)

6.1 Introduction

The purpose of these two experiments was to directly investigate the efficacy of association generation – generation of conceptually related words to each of the learned technical terms, as an add-on elaboration task to the crossword puzzle. To do so, the WGE condition would include association generation as a sole type of elaboration task (the hint-generation task was completely removed from the design). Experiment 5 served as a replication of Experiment 4, with one difference being the time of the year the two experiments occurred. Experiment 4 was conducted near the end of a semester, so the crossword served as a review tool for technical terms which participants already learned in the classroom. On the other hand, Experiment 5 was conducted early in the semester, so participants had little exposure to the technical terms in the classroom. In this case, the crossword served as an introduction tool for the terms.

Hypotheses for these two experiments were the same as those for the previous three experiments. The design of Experiments 4 and 5 was similar to Experiment 3, with a few notable changes on the mode and the delay. As it turned out, these experiments were conducted during an academic year in which COVID-19 pandemic the took place. Conducting an experiment in-person in a physical lab room was not an option. As a result, the experimental was conducted online, and changed from group to individual participation. In order to investigate the effect of crosswords on memory, the length of retention the delay between training and evaluation phases was extended from 24 hours to at least 5 days.

6.2 Methods

6.2.1 Participants

Experiment 4 consisted of 14 undergraduates and Experiment 5 consisted of 30 undergraduates recruited from the *Introduction to Psychology* course. All participants were over 18 years old and received course credit for their voluntary participation. Participants were recruited at different time periods from different semesters. Specifically, Experiment 4 was conducted at the end of semester, while Experiment 5 was conducted at the beginning of the semester. The experiments received IRB approval.

6.2.2 Materials

These experimented used the same materials as Experiment 3, implemented in an online environment. Thirty technical terms were selected from the following three chapters: sensation and perception, learning, and memory. Each of three crossword puzzles consisted of 10 scientific terms from a single chapter. Online questionnaires were developed using *Qualtrics*.

To evaluate the quality of participants' elaboration during training phase, a slightly modified coding scheme was used. The coding scheme for connections between technical terms being used as stimuli and other terms participants knew, which were generated as part of elaboration task, consists of three criteria: correctness, added information, and word count. Correctness evaluates whether or not a connection between two words makes logical sense based on each word's original definition or provided hint (scored as 0 or 1). Added information evaluates whether or not a connection between two words comes directly from each word's original definition or provided hint (scored as 0 or 1). Word count evaluates number of items being generated (scored as 0.5 per related word and 0.5 per description of connection).

Elaboration Task – Instructions

Please type in at least 3 words/phrases that are conceptually related to crossword terms that you just solved and provide explanation for those relations.

Example

Term:	Habituation
Original Clue:	A decrease in responding with repeated stimulation
Your Response:	Attention - Its processing includes habituation
	Diminish over time - Characteristics of habituation
	White noise - stimulus that can cause habituation

In addition to the 30-item multiple-choice test (the same as those in Experiment 3), a new 6-item short-answer test was included. The short-answer test consisted of free-response type questions, and it was designed to measure participant's ability to provide a more in-depth understanding of the learned technical terms. Two out of ten terms from each of three chapter were chosen for a total of six questions. Below is a sample short-answer question:

Test Question – Instructions

- What is the difference threshold (in the context of Sensation & Perception)?

- Describe its connections with another two key terms/concepts.

- Provide an example of how it might affect your perception or discrimination of something when you are driving a car.

6.2.3 Experimental Design and Variables

Experiments 4 and 5 changed from an in-person to an online experiment, from being run with groups to individuals, extended the delay between the training and evaluation phases, and added a new short-answer test as a dependent variable (see *Materials* section). The overall experimental design was the same as that in the previous experiments. The within-subjects design examined three training conditions: Word-Game + Elaboration (WGE), Word-Game-Only (WGO), and baseline (control). In the WGE condition, participants first solved a crossword puzzle and then did an elaboration task. In the WGO condition, participants only solved a puzzle. The baseline condition involved no additional training. Training condition order and chapter were counterbalanced using a within-subjects, Latin-squares design. For the dependent variables, multiple-choice test measured memory retention, and short-answer test measured high-level understanding of learned terms.

6.2.4 Procedure

The overall procedure was the same as in Experiments 1, 2, and 3. After a participant signed up to take part in the experiment and signed an online written consent form, the participant received two links to the online questionnaire forms (one for WGE and one for WGO) to be completed individually. In WGE condition, participants first solved a crossword puzzle and then did an elaboration task. The elaboration task was to generate three words or phrases that were conceptually related to the solved crossword terms, with an additional prompt that asked participants to also provide written explanation on how they are related. Participants were given 30 minutes to first solve the puzzle and then complete the elaboration task. In WGO condition, participants had 30 minutes to solve a puzzle twice and study it (the same as Experiment 3).

The evaluation phase was set to begin five days after participants completed their individual training, during which they received a link to another online questionnaire form. Participants were instructed to access the link and complete the form as soon as possible. The average actual delay between phases was 7.3 days with a range of 4 - 19 days. The questionnaire form started with six short-answer test questions, each asking participants to provide (a) definition of the presented term, (b) two other words or phrases that are conceptually related to the term along with justification, and (c) application-based knowledge of the term. Following the short-answer test questions, 30 multiple-choice test questions were included in the same form. The multiple-choice test was presented after the short-answer test to prevent participants from being influenced by potential gain of information from test questions. Participants were given 60 minutes to complete all 36 test questions.

6.3 Results

For Experiment 4, Figure 5A depicts means and standard errors of participants' performance in the multiple-choice test while Figure 5B depicts those in the short-answer test. During training phase, participants in WGO condition successfully filled, on average, 9.5 out of 10 words in a crossword puzzle, with a range of 5 - 10 words. The same participants in WGE condition successfully filled, on average, 8.2 out of 10 words in a crossword puzzle, with a range between 0 - 10 words. To evaluate H1 for multiple-choice test, a 1 x 3 ANOVA was conducted using R, taking both training conditions (3 conditions) and crossword puzzles (3 different sets of technical terms) into account. The test indicated a marginal effect of training condition on participant's recognition performance, $F(2, 24) = 3.32$, $p = .053$. In addition, planned comparisons using one-tailed, paired t-tests were conducted using pwr package in R. The tests indicated a statistically significant improvement only between the WGO and Control conditions, $t(13) = 2.32$, $p = .019$, $d = 0.69$. There was no difference between the WGE and Control conditions, $t(13) = 1.00$, $p = .168$, $d = 0.30$. Based on these results, H1 was not supported due to that solving a crossword for technical terms with add-on elaboration task failed to improve memory performance on the terms. To evaluate H2, one-tailed, paired t-tests showed non-significant difference between performance in the WGE and WGO conditions, $t(13) = 1.30$, $p = .109$, $d = 0.36$. Based on these results, H2 was also not supported, suggesting that adding a structured elaboration task to a crossword failed to improve resulting memory performance even further compared to crossword without an elaboration task.

To evaluate H1 for the short-answer test, a 1 x 3 ANOVA indicated a significant effect of training condition on participant's in-depth understanding of learned technical terms, $F(2, 24) = 4.49$, $p = .022$. Planned one-tailed, paired t-tests showed, once again, a significant improvement only between WGO and Control, $t(13) = 2.90$, $p = .006$, $d = 0.55$. The other comparison was not significant: WGE and Control, $t(13) = 1.38$,

$p = .095$, $d = 0.37$. To evaluate H2, one-tailed, paired t-tests showed non-significant difference between performance in the WGE and WGO conditions, $t(13) = 0.82$, $p = .214$, $d = 0.18$. Based on these results, both H1 and H2, again, were not supported, suggesting that solving a crossword, with or without an add-on elaboration task, failed to improve elaborative thinking on the terms.

For Experiment 5, Figures 5C and 5D depict participants' performance in the multiple-choice test and short-answer test, respectively. During training phase, participants in WGO condition successfully filled, on average, 9.6 out of 10 words in a crossword puzzle, within the range between 5 - 10 words. The same participants in WGE condition successfully filled, on average, 7.5 out of 10 words in a crossword puzzle, within the range between 0 - 10 words. To evaluate H1 for multiple-choice test, a 1 x 3 ANOVA indicated a significantly large effect of training condition on participant's recognition performance, $F(2, 56) = 18.47$, $p < .001$. Planned comparisons to evaluate H1 using a one-tailed, paired t-tests showed a statistically significant improvement and a large effect size between the WGO and Control conditions, $t(29) = 5.88$, $p < .001$, $d = 1.22$, as well as between the WGE and Control conditions, $t(29) = 4.55$, $p < .001$, $d = 1.08$. Based on these results, H1 was supported, suggesting that solving a crossword for technical terms (with or without an add-on elaboration task) improved recognition performance on the terms. To evaluate H2, one-tailed, paired t-tests showed non-significant difference between performance in the WGE and WGO conditions, $t(29) = 1.05$, $p = .150$, $d = 0.27$. Based on these results, H2 was not supported, suggesting that adding a structured elaboration task to a crossword failed to improve the resulting recognition performance even further compared to a crossword without an elaboration task.

As for the short-answer test, a 1 x 3 ANOVA showed no significant difference across training conditions on participant's in-depth understanding of learned technical terms, $F(2, 56) = 1.78$, $p = .178$. However, planned one-tailed, paired t-tests showed a significant improvement only between WGO and Control, $t(29) = 1.83$, $p = .039$, $d = 0.27$. The other comparisons were not significant: WGE and Control, $t(29) = 1.67$, $p = .053$, $d = 0.27$; WGE and WGO, $t(29) = 0.04$, $p = .485$, $d = 0.01$. Based on these results, both H1 and H2 were not supported, suggesting that solving a crossword, with or without add-on elaboration task, failed to improve elaborative thinking on the terms.

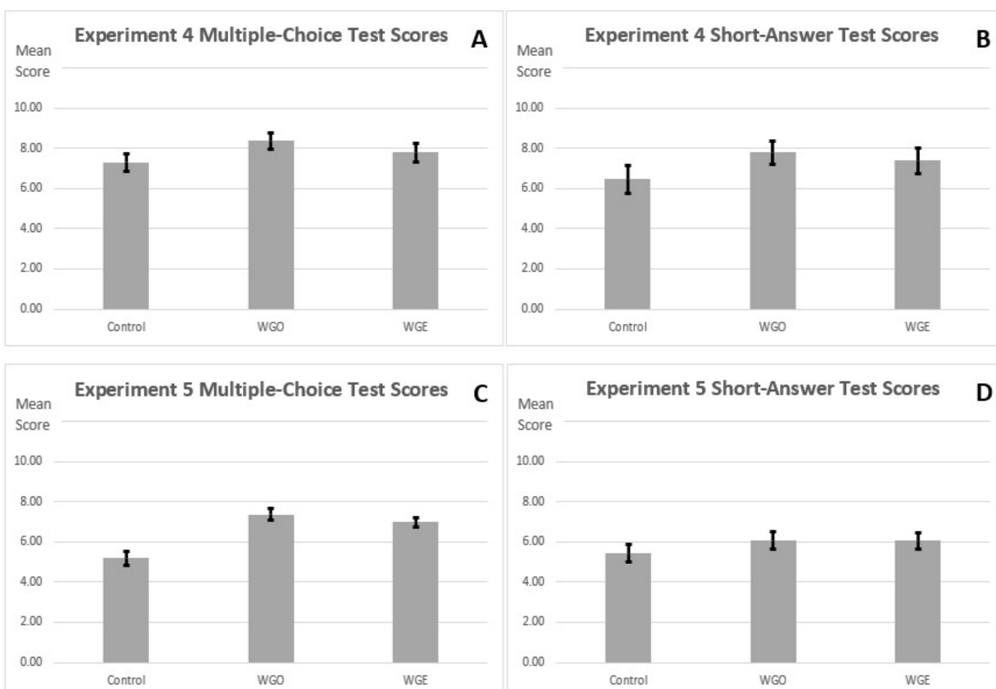

Figure 5. Experiments 4 and 5 mean scores (out of 10 points) and standard error by training condition for multiple-choice (left panels) and short-answer (right panels) after five days. The three conditions are Control, Word-Game-Only (WGO), and Word-Game + Elaboration (WGE).

Additionally, I coded and analyzed connections between technical terms generated by participants as part of the elaboration task (in WGE condition) during the training phase. Due to the fact that experiments were conducted online, it was difficult to monitor participants' progress in the task. Thus, only a portion of participants who properly did the task were included in this analysis ($N = 27$). Out of 270 generated responses, thirty of them were selected as samples for testing IRR. Two raters (one being the author) independently coded the selected responses based on the same coding scheme previously described under the *Materials* section. Correctness and added information of the generated responses (the latter is also referred to as *AddInfo*) are categorical variables. Thus, Cohen's kappa was again chosen as a measure for IRR. For these experiments, weighted Cohen's kappa coefficients were 1.00 (almost perfect agreement) for correctness and 0.59 (moderate) for *AddInfo*. In order to adjust their coding standard to become more agreeable, the two raters discussed with each other regarding coding schemes and scores they each gave. Afterwards, raters independently revisited the same set of selected responses and reevaluated them as they deemed necessary. Cohen's kappa analysis was run again, and finally $k \geq 0.8$ for both dimensions were achieved: 1.00 for correctness and 0.84 (substantial) for *AddInfo*.

First, I fit participant's scores from three criteria: correctness, *AddInfo*, and word count, along with their corresponding test scores in a logistic regression model. It appeared that 18 responses were left completely blanked by participants from Experiment 4 and 51 responses were left blanked by participants from Experiment 5. After removing the blanked responses, I then applied ANOVA on the models. The analysis revealed that none of the three factors is considered statistically significant (Figure 6). This suggested that connections between technical terms generated by participants did not correlate with their test scores in any ways. Additionally, Figures 7A and 7B depict scatter plots of participant's test score versus number of elaborated responses and *AddInfo* respectively. From these plots it can be inferred that a participant can achieve either high- or low-test score regardless of whether he or she generates all responses or leave some of them blank. Similarly, participants can achieve any score regardless of whether or not new information was added to the generated connections. These results support the findings above, explaining why participant's test performance after elaboration training (as in the WGE condition) was not better than the performance after solving the same crossword puzzle twice (as in the WGO condition).

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			200	220.98	
PID	25	18.4116	175	202.57	0.8244
Correctness	1	0.0302	174	202.54	0.8620
AddInfo	1	0.8104	173	201.73	0.3680
WordCount	1	0.0766	172	201.65	0.7819

Figure 6. Logistic regression analysis conducted in R on elaborated hints generated during the training phase of Experiments 4 & 5. PID indicates individual differences among participants.

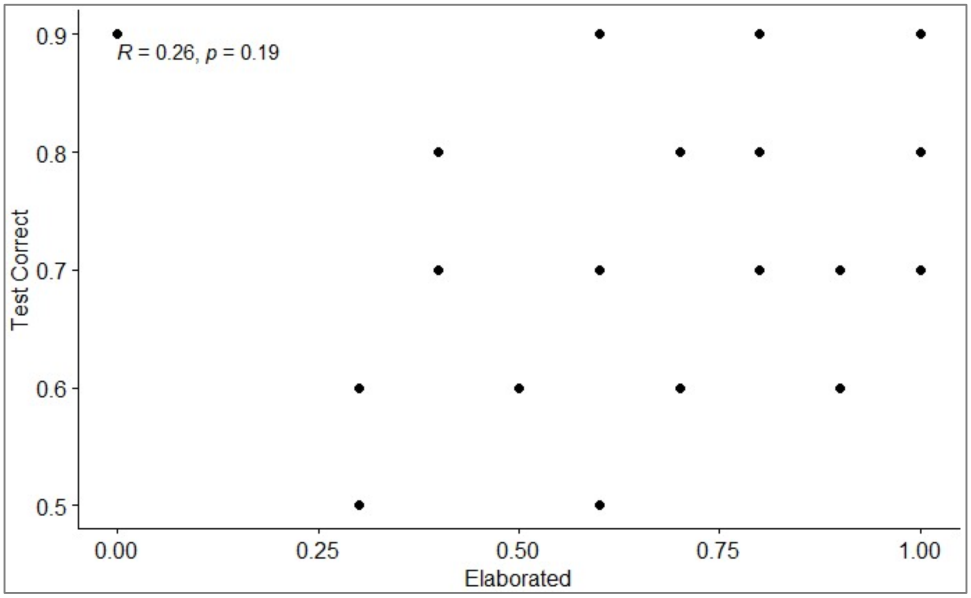

Figure 7A. Plot shows participant's test score based on how many elaborated responses they generated (as opposed to leaving blank).

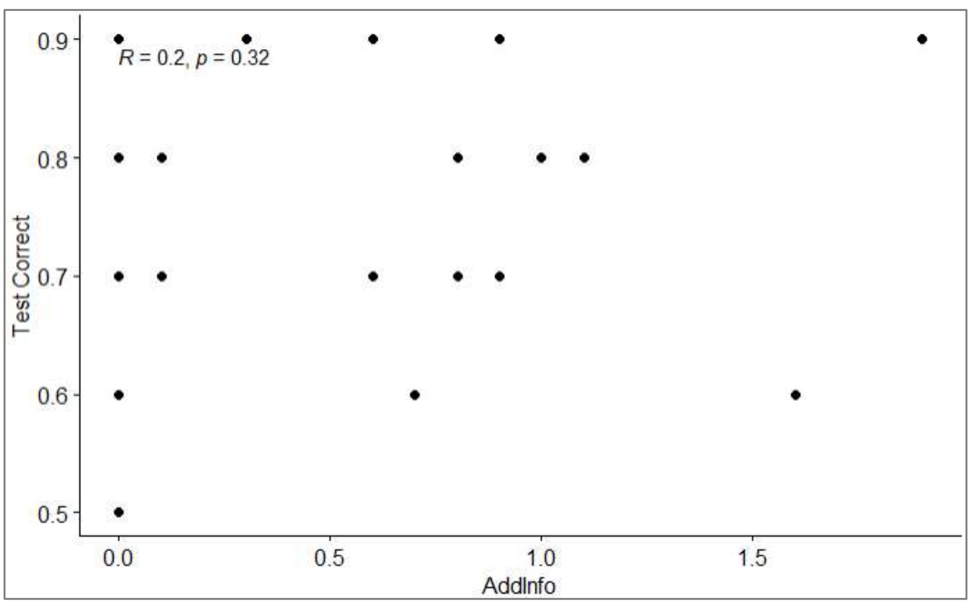

Figure 7B. Plot shows participant's test score based on added information (*AddInfo*) in their generated responses.

6.4 Discussion

Throughout these two experiments, I tested the efficacy of association generation. The add-on elaboration task involves (a) generating several technical words that are conceptually related to the target technical word learned during the experiments and (b) explaining their relations. While in Experiment 4, I showed the elaboration effect on crosswords which served as a review tool for already known words, in Experiment 5 the crossword served as an introduction tool for unfamiliar words. Within these experiments, the elaboration effect was tested both at the recognition level and at the more in-depth understanding level. Regardless of several differences in context, all four cases shared similar trends of results. That is, for all of the cases, spending approximately half an hour solving a crossword puzzle for technical terms and reviewing it one more time provides knowledge gain that is retained over five days; adding the elaboration task to a crossword puzzle, however, is as effective as solving the same puzzle again and does not provide further benefits to the knowledge gain. In Experiment 5, where student's prior knowledge of technical terms is comparatively less, the gain from add-on elaboration becomes apparent. When students are already familiar with technical terms, there seems to be no difference in knowledge gains from crosswords (both with and without the add-on elaboration) between recognition and high-level understanding. However, when students are unfamiliar with the technical terms, the crossword puzzle improves recognition more than understanding level.

Based on all the findings mentioned above, an add-on elaboration task fails to provide greater benefit than simply repeating a crossword twice. This is also the case in Experiment 3, in which a different elaboration strategy is compared against the repetition of crossword. Looking at the three experiments as a whole, however, there were many changes in transition from Experiment 3 to Experiments 4 and 5 including change in elaboration strategy, change in delay between training and evaluation phases, change from in-person to online participation, and change from teamwork to individual effort. Originally, Experiments 4 and 5 were conducted to test the effect of the association generation strategy for a more practical situation. Therefore, changes from Experiment 3 should have only been on elaboration strategy, delay period, and addition of short-answer test questions. Due to the COVID-19 pandemic, however, the whole experiment must be moved to an online environment. Corresponding to this shift, participation was also changed from group level to individual level because it was no longer feasible to monitor participant's behavior remotely. Moreover, most of the changes stated here are known to reduce learning gain. Change from group to individual participation removed the collaborative-learning factor, which negatively affects constructive learning (Johnson & Johnson, 2009; Johnson et al., 1991). It is also possible that changing from an in-person to an online environment may have negatively affected participants' level of engagement with the experiment (Baber, 2020; Morgan, 2020). Extension of delay between learning and evaluation from 24 hours to five days can lead to poorer overall performance (Mueller & Veinott, 2018). Nevertheless, based on the results from the three experiments, these changes did not seem to impact the trend between the two crossword conditions and the baseline, suggesting consistent validities of repetition of crossword and the not-so-promising add-on elaboration.

That association generation did not lead to the acquisition of higher-level scientific thinking can be explained by the Knowledge-Learning-Instruction (KLI)

Framework (Koedinger et al., 2012). From the framework's perspective of appropriate transfer of knowledge, if the ultimate goal is to achieve higher-level scientific thinking, the learning event (e.g., use of a learning-aid tool) needs to provide knowledge component that can take on variation of values. In the case of crossword puzzle, it can only enforce learning of paired associates, which is considered constant, as opposed to variable (Roediger & Karpicke, 2006). Hence, crossword puzzles need an add-on elaboration task that provides a variable knowledge component. However, since association generation also involves generation of paired associates, coupling it with crossword does not make any difference, and the end product is still constant knowledge component (Koedinger et al., 2012).

In the end, both elaboration strategies: summarization (Experiment 3) and association generation (Experiments 4 and 5) failed my expectation in producing better learning outcome than crossword alone. My next step then was to explore another elaboration strategy that has potential of providing variable knowledge component. This led to new experiments whose WGE condition utilize a different elaboration task, that is, the sentence generation.

7 Evaluation of Sentence-Generation Strategy in Crossword (Experiment 6)

7.1 Introduction

The purpose of this experiment was to investigate the efficacy of sentence generation as a new elaboration technique. This technique involves using knowledge of learned technical terms to generate a sentence portraying a scenario. The WGE condition included sentence generation as an add-on elaboration task to the crossword puzzle. Hypotheses for this experiment were the same as those for Experiments 1 and 2. The overall methods and procedure were identical to those in Experiments 4 and 5, with a sole difference being the elaboration task.

7.2 Methods

7.2.1 Participants

Experiment 6 consisted of 50 undergraduates recruited from *Introduction to Psychology* courses. All participants are over 18 years old, and they received course credit for their voluntary participation.

7.2.2 Materials

Same materials as those in Experiments 4 and 5 were implemented in an online environment. Online questionnaires were developed using *Qualtrics*. The instructions for the elaboration task and short-answer test question, which was modified to accommodate the new task, are as follows.

Elaboration Task – Instructions

Please type in a scenario (in 1-3 sentences) that includes the crossword terms that you just solved, either the word itself or context related to the word's definition.

Example

Term:	Habituation
Original Clue:	A decrease in responding with repeated stimulation
Your Response:	Because I work in a factory everyday, I have become habituated to the sound of machine humming.
	Because I work in a factory everyday, the sound of machine humming does not bother me anymore.

Test Question – Instructions

- What is the difference threshold (in the context of Sensation & Perception)?

- Provide an example of how it might affect your perception or discrimination of something when you are driving a car.

7.2.3 Experimental Design and Variables

Overall design was identical to that in Experiments 4 and 5. The within-subjects design examined 3 training conditions: Word-Game + Elaboration (WGE), Word-Game-Only (WGO), and baseline (control). In WGE condition, participants first solved a crossword puzzle and then did an elaboration task. In the WGO condition, participants only solved a puzzle. The baseline condition involved no additional training than what students did outside the experiment. Training condition order and chapter were counterbalanced using a within-subjects, Latin-squares design. For the dependent variables, multiple-choice test measured memory retention, and short-answer test measured high-level understanding of learned terms.

7.2.4 Procedure

Overall procedure was identical to that in Experiments 4 and 5. The only difference was the change of the elaboration task. In WGE condition, after participants solved a crossword puzzle, they used the solved terms to generate sentences. For example, presented with the term *flashbulb*, participants might generate a sentence with the term such as, “My friend’s car accident in 2011 would probably remain in my head for life as a flashbulb memory.” Alternatively, they might generate a sentence that is conceptually related to the term without using the term directly such as, “It has been many years since, but I can still clearly remember reading news about my friend’s car accident.” The evaluation phase was set to begin five days after participants completed their individual training. The average actual delay between phases turned out to be 6.7 days within the range between 2 - 11 days.

7.3 Results

Figures 8A and 8B depict participants’ performance in multiple-choice test and short-answer test, respectively. During training phase, participants in WGO condition successfully filled, on average, 8.7 out of 10 words in a crossword puzzle, within the range between 0 - 10 words. The same participants in WGE condition successfully filled, on average, 7.5 out of 10 words in a crossword puzzle, within the range between 0 - 10 words. To evaluate H1, a 1 x 3 ANOVA was conducted using R, taking both training conditions (3 conditions) and crossword puzzles (3 different sets of technical terms) into account. The test indicated a significantly large effect of training condition on participant's recognition performance, $F(2, 96) = 14.44$, $p < .001$. In addition, planned comparisons using one-tailed, paired t-tests were conducted using pwr package in R. The tests showed statistically significant improvements between the WGO and Control conditions, $t(49) = 2.21$, $p = .016$, $d = 0.37$, as well as between the WGE and Control conditions, $t(49) = 5.17$, $p < .001$, $d = 0.81$. Based on these results, H1 was supported, suggesting that solving a crossword for technical terms (with or without add-on elaboration task) improved recognition performance on the terms. To evaluate H2, one-tailed, paired t-tests showed significantly greater performance when comparing WGE condition to WGO condition, $t(49) = 2.98$, $p = .002$, $d = 0.46$. Based on these results, H2 was supported, suggesting that adding a structured elaboration task to a crossword improved resulting recognition performance even further compared to crossword without an elaboration task.

As for the short-answer test, a 1 x 3 ANOVA showed no significant difference across training conditions on participant's in-depth understanding of learned technical terms, $F(2, 96) = 1.35$, $p = .264$. Planned one-tailed, paired t-tests also showed no significant difference between any of the two conditions: WGO and Control, $t(49) = 1.51$, $p = .069$, $d = 0.26$; WGE and Control, $t(49) = 0.88$, $p = .193$, $d = 0.15$; WGE and WGO, $t(49) = 0.69$, $p = .246$, $d = 0.12$. Based on these results, both H1 and H2 were not supported, suggesting that solving a crossword, with or without add-on elaboration task, failed to improve elaborative thinking on the terms.

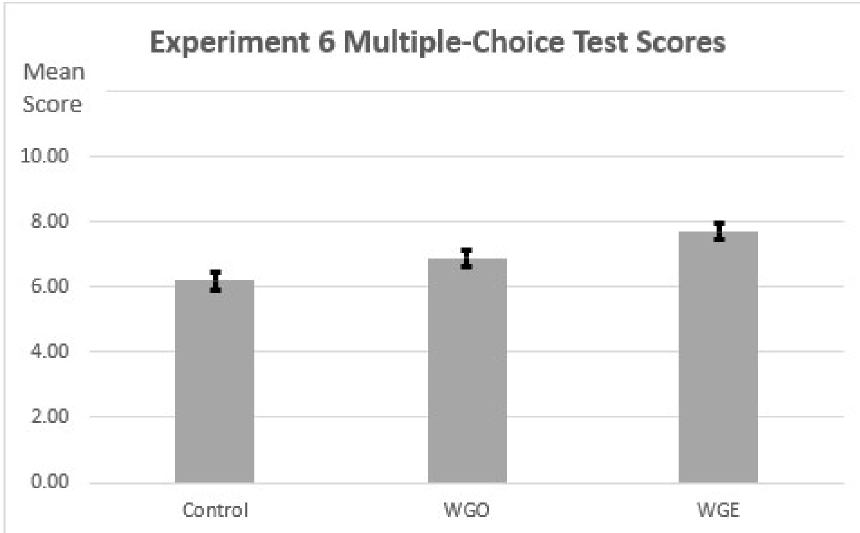

Figure 8A. Experiment 6 mean scores (out of 10 points) and standard error by training condition for multiple-choice after five days. The three conditions are Control, Word-Game-Only (WGO), and Word-Game + Elaboration (WGE).

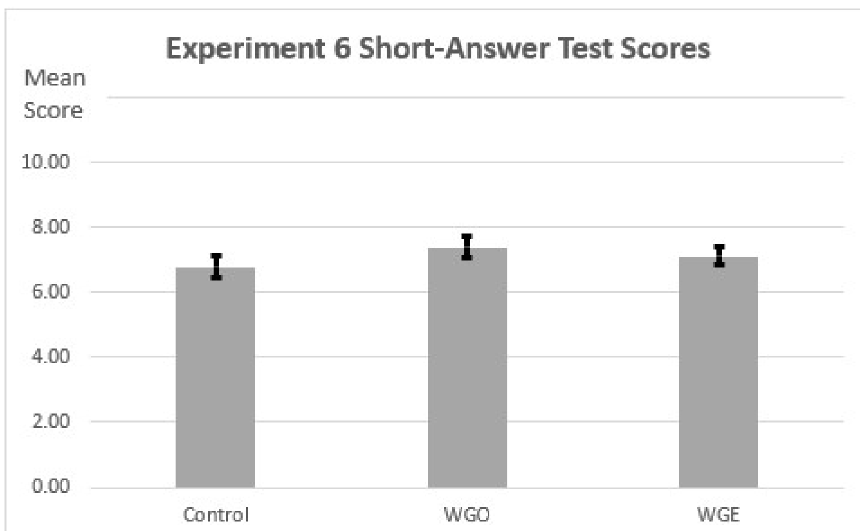

Figure 8B. Experiment 6 mean scores (out of 10 points) and standard error by training condition for short-answer after five days. The three conditions are Control, Word-Game-Only (WGO), and Word-Game + Elaboration (WGE).

7.4 Discussion

These results suggest that a single change in the elaboration task greatly improved its effect on recognition as measured by the multiple-choice test. Compared to the task of generating conceptually related words used in the previous experiments, this task did not rely as much on the broad knowledge around the learned terms, but rather the understanding of the terms themselves. It required students to know enough of the terms for them to be able to generate something more than the definitions, specifically sentences portraying scenarios that make practical sense. Additionally, students were able to generate self-related scenarios, making them more personal, and therefore more memorable. This flexibility in student's response to this elaboration task makes the task's learning output variable, and therefore it should lead to achievement of higher-level scientific thinking (Koedinger et al., 2012). These benefits, however, failed to appear from the short-answer generation test (similar to the one used in Experiments 4 and 5). One possible reason was that the test was not modified to match exactly with the new task, thus elaboration had little to no effect (Morris, Bransford, & Franks, 1977; Tulving, 1979). While the task during training asked students to generate their own scenarios, the test asked them to utilize the knowledge on the terms under provided specific scenarios.

8 Comparison of Efficacy Between Three Elaboration Strategies (Experiment 7)

8.1 Introduction

Throughout the first six experiments, none directly compared different types of elaboration tasks to determine their effect on memory or deeper understanding. Across these experiments, three types of elaboration tasks were used: (1) generation of new hints based on summarized knowledge on learned technical terms; (2) generation of words or phrases conceptually related to the terms; and (3) generation of sentences portraying practical use of the terms. These produced mixed results regarding effectiveness of the tasks compared to the non-elaboration conditions, and these different elaboration methods were not compared directly. Consequently, this final experiment directly compared the effectiveness of different elaboration tasks used in this dissertation. Furthermore, because different forms of elaboration might produce benefits for different aspects of knowledge, this experiment involved four distinct means of assessing knowledge retention—the standard multiple-choice response, as well as three dependent measures linked to specific elaboration approaches. That is, an elaboration involving generating new hints will benefit assessments of knowledge involving similar hint-generation techniques, word generation elaboration will differentially benefit word-generation tests, and sentence-generation elaboration will benefit practical use of terms the most.

8.2 Hypotheses

The experiment was designed to test the following hypotheses:

H1: There will be a difference in degree of effectiveness on recognition of learned terms among three types of elaboration training tasks.

H2: There will be a difference in degree of effectiveness on higher level understanding of learned terms among three types of elaboration training tasks.

H3: Specific elaboration training will produce the best knowledge transfer for dependent measures aligned with the mental processes involved in the elaboration.

8.3 Methods

8.3.1 Participants

Experiment 7 consisted of 86 undergraduates recruited from *Introduction to Psychology* courses, including 55 males (64%) and 31 females (36%). All participants were over 18 years old, and they received course credit for their voluntary participation. The experiment received IRB approval.

8.3.2 Materials

This experiment used one crossword puzzle containing 12 technical terms selected from three chapters focusing on (1) sensation and perception; (2) learning; and (3) memory. Follow-up evaluation test included 6 short-answer questions and 12 multiple-choice questions. The short-answer questions were designed to accommodate all three types of elaboration tasks, with examples of each type shown in Figure 9.

Instruction *- Tell us 1 thing you know about the term gestalt (in the context of perception).* *- Give us 2 other key terms/concepts that are related to the term and briefly explain the relationships.* *- Provide us 1 scenario depicting how the term might play a role in a person's life.*

Figure 9. Example short-answer questions used in Experiment 7, highlighting the three kinds of elaborative assessments.

8.3.3 Experimental Design and Variables

The between-subjects design examined three different types of elaboration tasks used during training phase as the independent variable. Three conditions were Type A - generation of summarized information of learned technical terms which can then be used as their crossword hints, Type B - generation of words or phrases conceptually related to the terms along with explanations for how they are related, and Type C - generation of sentences portraying practical use of the terms. Dependent variables were recognition performance assessed by multiple-choice test questions and effectiveness of elaboration assessed by short-answer test questions shown in Figure 9.

8.3.4 Procedure

Participants were first randomly assigned into one of the three conditions. This experiment consisted of two phases: 60-minute training phase and 60-minute evaluation phase which was to be completed at least 5 days later. possible. The average actual delay between phases was 5.7 days with a range between 5 - 8 days. The training phase was conducted in the lab, while the evaluation phase was conducted online. During the training phase, participants were given a crossword puzzle to solve on paper. They were provided with the course textbook that they can use to search for information related to the terms that were included in the puzzle. The time to complete this task was limited to 25 minutes. Then, participants proceeded to complete the elaboration task for the remaining 35 minutes. The task (as explained above) was different for each participant depending on their assigned condition. During the evaluation phase, participants completed a test which took the form of an online questionnaire. The test started with short-answer questions, asking participants to elaborate on a portion of the learned terms using all three types of elaboration techniques consistent with three types of elaboration tasks. Participants had 50 minutes to complete this first part of the test. The second part involved multiple-choice test questions to be done within 10 minutes.

8.4 Results

Figure 10 depicts participants' performance in multiple-choice test, and Figure 11 plots the performance alongside baseline performances from previous six experiments to show that crosswords with add-on elaboration of any type produce learning gain over baseline controls. During training phase, participants successfully filled, on average, 8.5 out of 12 words in a crossword puzzle, within the range between 4 - 12 words. To evaluate H1, a 1 x 3 ANOVA was conducted using R. The test indicated no significant main effect for training condition in participant's multiple-choice test score, $F(2, 80) = 1.24$, $p = .295$. Based on these results, H1 was not supported, suggesting that solving a crossword for technical terms with an add-on elaboration training task equally improved recognition performance on the terms regardless of the type of training task.

To evaluate H2 and H3, participant's performance in short-answer test was analyzed. Each test question included three test prompts (see *Materials* section) which evaluated cognitive ability to recall information related to learned technical terms and elaborate on those terms in two different ways. For each of those prompts, participants under three different training conditions were compared against one another on each of the three elaborative dependent variables. Figure 12 depicts participants' performance in recall prompt, association-generation prompt, and sentence-generation prompt respectively in Panels A through C. Additionally, Figure 13A compiles results from three test prompts together to make it easier to look at interaction between training condition and type of test prompt, which tests hypothesis H3. Figure 13B shows a graph of z scores for the same thing. The car package in R was used to run a 2 x 3 ANOVA on linear mixed effect regression model. The test indicated no significant main effect for training conditions, $F(2, 249) = 0.91$, $p = .406$, but found significant main effect for types of test prompt, $F(2, 249) = 22.62$, $p < .001$, which simply indicates that some kinds of elaborative testing higher scores than others. The test also indicated non-significant interaction between training condition and type of test prompt, $F(4, 249) = 0.39$, $p = .815$, which failed to support hypothesis H3. Looking at just student's performance in each individual test prompt, 1 x 3 ANOVA F-tests showed no significant difference among training conditions for any of the three dependent measures: recall prompt, $F(2, 83) = 0.89$, $p = .413$, association-generation prompt, $F(2, 83) = 0.38$, $p = .683$, sentence-generation prompt, $F(2, 83) = 0.42$, $p = .656$. Based on these results, H2 was not supported, suggesting that solving a crossword for technical terms with an add-on elaboration training task equally improved overall high-level understanding on the terms regardless of the type of training task.

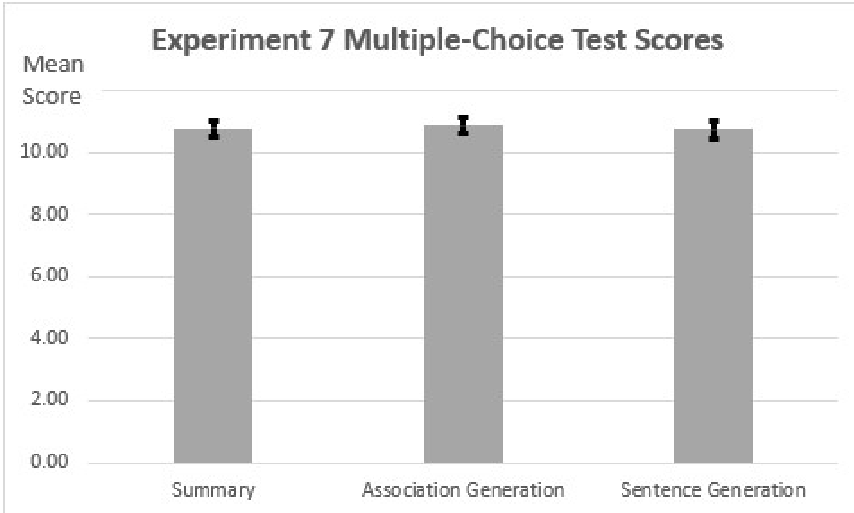

Figure 10. Experiment 7 mean scores (out of 12 points) and standard error by training condition for multiple-choice test after five days. The three conditions are Type A: train summary, Type B: train association generation, and Type C: train sentence generation.

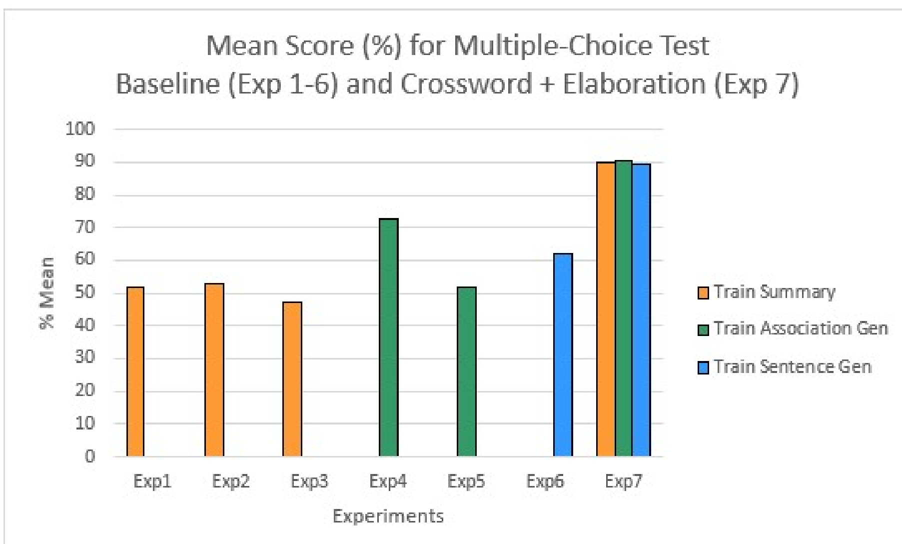

Figure 11. Mean scores on multiple-choice test across 7 experiments. Experiments 1 - 6 show baseline performance, whereas Experiment 7 shows performance after learning technical terms from a crossword with add-on elaboration. Color indicates type of elaboration task.

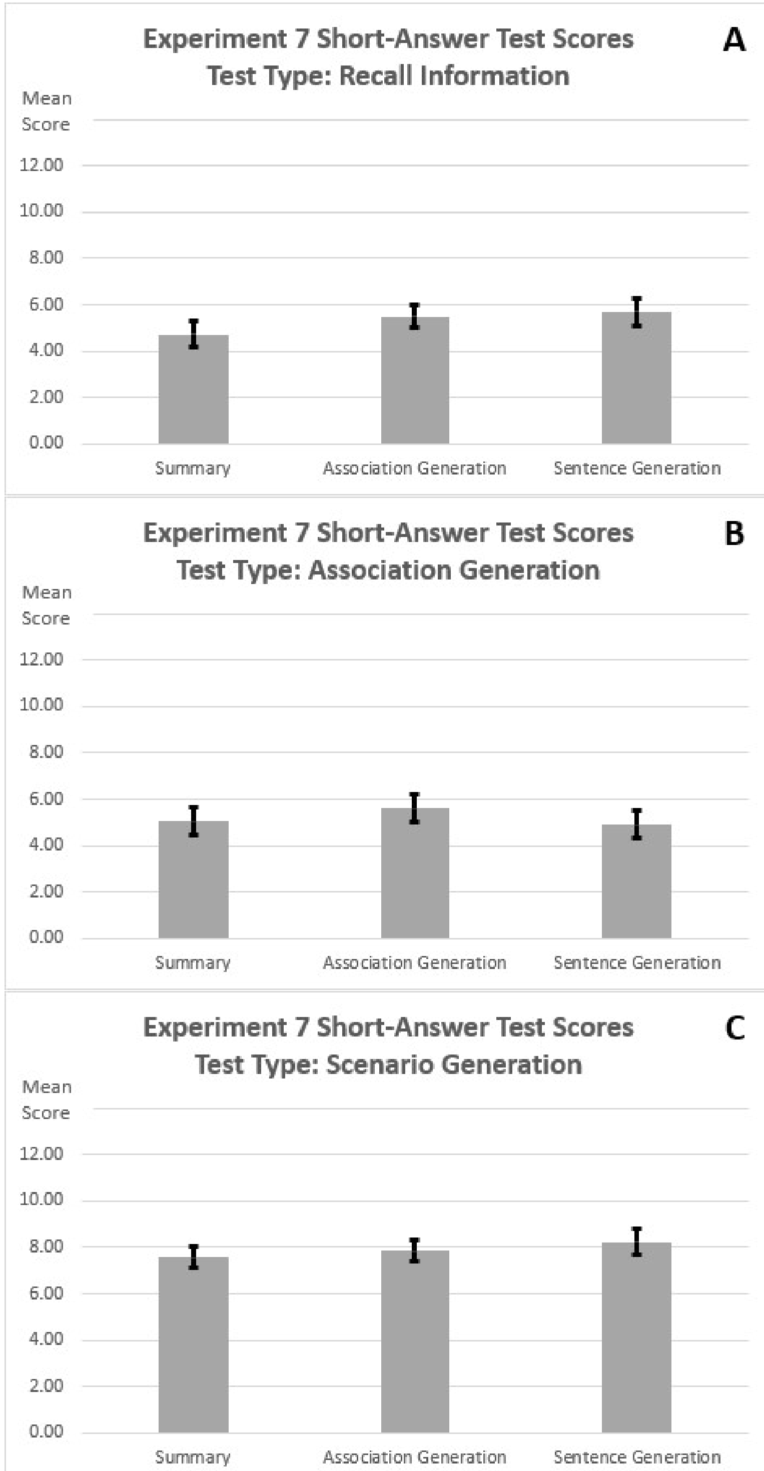

Figure 12. Experiment 7 mean scores (out of 12 points) and standard error by training condition for short-answer test after five days. The three conditions are Type A: train summary, Type B: train association generation, and Type C: train sentence generation.

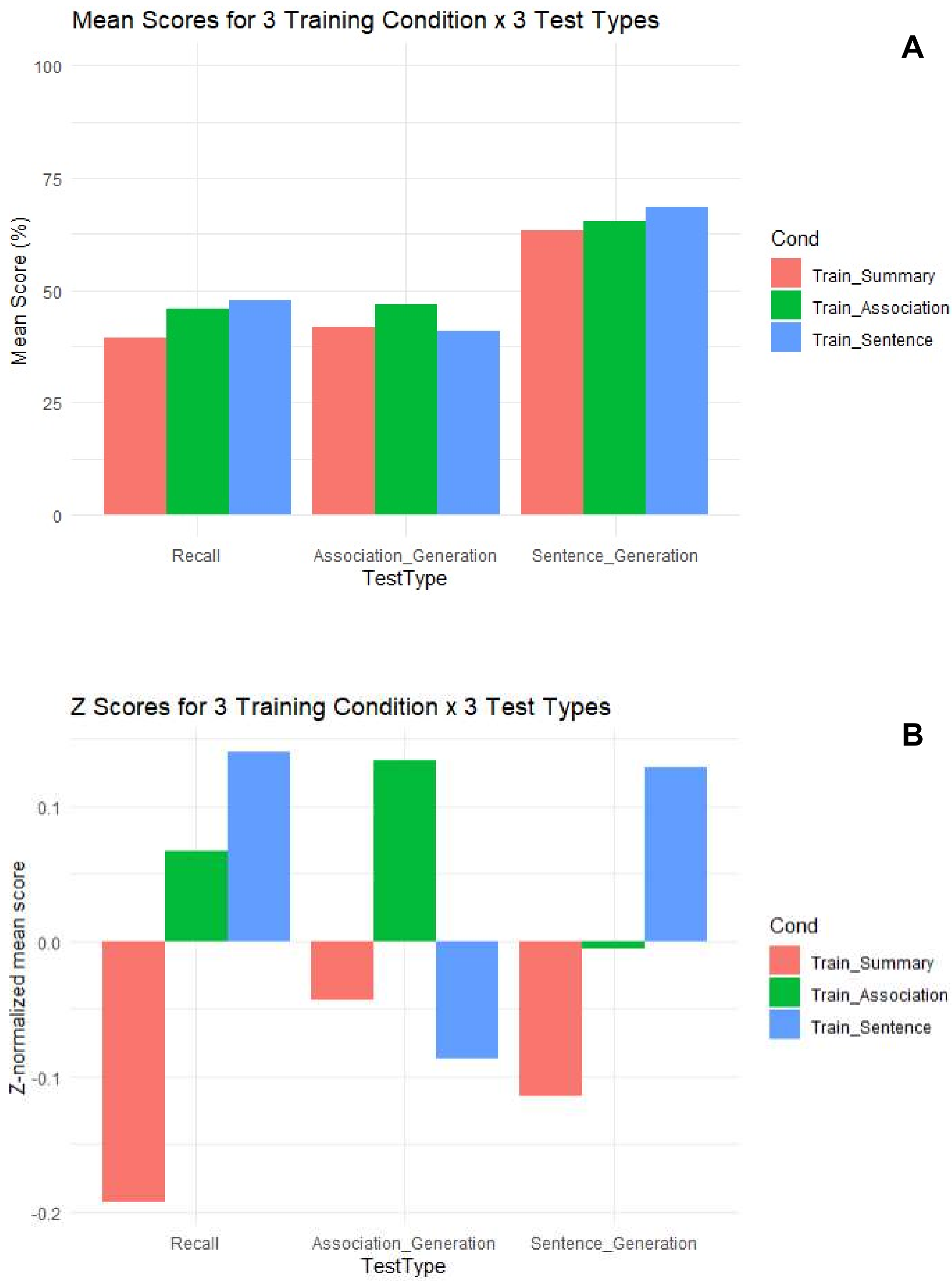

Figure 13. Experiment 7 mean scores for short-answer test in raw score and z-normalized score for 3 training conditions by 3 test types.

8.5 Discussion

In this experiment, I directly compared the efficacy of three elaboration strategies: summarization, association generation, and sentence generation as an add-on task to a crossword puzzle. Two knowledge tests were used to evaluate the efficacy of those strategies. A multiple-choice test measured memory performance at the recognition level. A short-answer test measured participants' ability to elaborate on learned technical terms which requires student's higher-level understanding of those terms. For recognition performance, the results suggested that if students learn technical terms by solving a crossword puzzle followed by completing one of the three elaboration tasks, they will do equally well in remembering the terms when presented with the term's associated facts, regardless of which type of task they completed. This finding agrees with results from other experiments conducted for this dissertation. Looking at recognition performance in WGE condition from Experiments 3, 4, 5, and 6, the performance was about the same level (within 10% range of test score).

Two possible explanations might account for this finding. The first explanation is that these elaboration strategies have about the same effect as one another on learning, and therefore produce equal amount of learning gain. If this is true, it appears to contradict the predictions made from the depth-of-processing theory (Craik & Lockhart, 1972; Craik & Tulving, 1975). The theory predicts that deeper processing involved in encoding learned information results in stronger memory trace. In this experiment, association generation strategy normally involves deeper processing than writing a summary from word-definition pairs because it relies more on semantic network to generate semantically related words from target words (Brown & Perry Jr, 1991).

Likewise, sentence generation strategy has been shown to involve deeper processing than elaboration from just word's definition (Coomber, Ramstad, & Sheets, 1986). As such, one might expect the three elaboration strategies used in this dissertation to produce different benefits on learning.

A second possibility is that the three elaboration strategies in fact have unequal influence on learning in isolation, but these differences are negated by combining them with a crossword. The previous discussion about difference in level of processing involved in different elaboration strategies only concerns with elaboration as a stand-alone task in a learning activity. Most published studies in the literature of vocabulary learning concern the effect of either a crossword or constructive learning strategy on learning, but not a combination of both. Therefore, it is difficult to disprove this conjecture. However, predictions made by the KLI Framework (Koedinger et al., 2012) offer some support for this notion of crossword negating elaboration. The framework suggests that students benefit most in learning constant knowledge component (e.g., recognition of paired associates) from instructional event (or learning activity) which enforces application of constant knowledge component (e.g., crossword puzzle in learning word-definition pairs). This notion can be used to infer that a crossword will have the strongest effect on recognition of paired associates, to the point that differences in effectiveness among types of add-on elaboration tasks are minimal. Future research should evaluate this idea.

As for students' performance on higher-level understanding of learned technical terms, the findings are similar with those for recognition performance previously discussed. That is, with respect to test type, different elaboration strategies provide nearly the same level of benefits on learning. This is rather puzzling, however, given that three test prompts were designed to correspond with three different elaboration strategies. Therefore, one would either expect to see performance in the matching training condition (elaboration strategy) significantly greater than performance in the other two non-match conditions for any particular test type (i.e., transfer-appropriate learning; Morris et al., 1977; Tulving, 1979), or see one learning mode dominate the others in all measures if it were simply more effective. And since two out of three test types (besides recall of definition) required students to generate elaborative responses that go beyond word-definition pairs, it suggests that the baseline knowledge gained through crossword practice is sufficient to support the delayed elaborative assessments used in this experiment.

In the other six experiments conducted for this dissertation, it was difficult to attribute success or failure of elaboration to the add-on task due to inconsistency of other factors involved, i.e., control for time-on-task, collaborative-learning factor, and nature of evaluation test. Therefore, this experiment was conducted to keep all other factors consistent in order to put a focus on the difference (if any) in efficacy of elaboration strategy. If there was a significant difference in main effect, it might be attributed directly to the add-on elaboration task. As it happened, however, such a difference was not found. Furthermore, it is important to note that it is unlikely that the design and treatment failed to find an effect because of small sample size or lack of statistical power. Future research is needed to identify the true cause of lack of difference in effect of elaboration as an add-on task to crossword puzzle.

9 General Discussion

9.1 Lessons learned from this study

This dissertation compiles a series of studies on integrating elaboration learning task to crossword puzzle in order to extend its learning benefits to go beyond learning and retention of semantic information of technical vocabulary. Crossword puzzle, when used as a learning-aid tool, can only provide learners paired associate of technical terms including the terms and their corresponding definitions or factual information. As such, its benefits are limited to acquisition of lower-level, constant knowledge component which does not get transferred to higher-order scientific thinking such as reasoning, analysis, and concept creation (Koedinger et al., 2012; Whisenand & Dunphy, 2010). In order to break through the limitation, crossword puzzles need to include a constructive learning element. Specifically, it needs a task prompt that provides opportunity for learners to generate something beyond its original content of paired associates (Chi & Wylie, 2014; Fiorella & Mayer, 2016). Across seven experiments, three different elaboration strategies: summarization (Fiorella & Mayer, 2016; Stein & Kirby, 1992), association generation (Karpicke & Smith 2012; Lehman et al., 2014), and sentence generation (Coomber et al., 1986; Turley-Ames & Whitfield, 2003) were tested for their efficacy when integrated to a crossword as an add-on learning task. The main hypotheses for these experiments were (a) crosswords (both with and without add-on elaboration) would improve memory performance compared to learner's baseline, and (b) adding a structured elaboration task to a crossword would improve memory performance compared to a crossword alone.

Based on findings from the first two experiments, spending nearly half an hour solving crossword puzzles for technical terms greatly improved memory retention of gained knowledge on the learned terms. Furthermore, adding just ten more minutes of study time to complete the add-on elaborative summarization task results in even greater retention. This is due to that summarization involves learner selecting and integrating new information to his or her existing knowledge. It requires learners to understand the learning materials to some great extend beyond the original content in order for them to be able to generate a different version of the same information (Fiorella & Mayer, 2015). In such manner, summarization that is done properly can be considered a form of constructive learning which results in stronger memory retention of learned information (Chi & Wylie, 2014).

However, the positive findings from the two experiments can also be due to unequal time-on-task between the crossword alone and crossword with the add-on elaboration. Because the latter ends up consuming more time to complete, it is within expectation that it will result in better learning outcome (Wagner et al., 2008). In order to prove that main benefit on learning comes from the add-on elaboration task and not from extra time-on-task, direct comparison must be done between the two crossword conditions (with and without the add-on elaboration) when both have the same time limit. Experiment 3 tested the efficacy of crossword with add-on elaboration against crossword-only condition which involved solving the same crossword puzzle twice to control for time-on-task. The findings showed that a crossword with an add-on elaboration task failed to provide greater benefits than a repeated crossword. This suggested that time-on-task, rather than summarization, was the main contributor to the benefits found in the first two experiments.

One possible explanation to why time-on-task beat out elaboration in this case was that the extension of time to allow the 2nd round completion of the same crossword puzzle ended up introducing repetition effect to the crossword-only condition. The repetition effect is known to add value to learning outcome (Bromage & Mayer, 1986). If this was the case here, then it suggests that even though summarization did in fact have some effect on memory retention of learned technical terms, its effect was weaker compared to the repetition effect. This raises the need to find new elaboration strategies with stronger effect.

Hence, Experiments 4 and 5 were conducted to test the efficacy of a different elaboration strategy, namely, association generation. However, the same result trend as that in Experiment 3 was found – crossword with add-on association generation task also failed to provide greater benefits than repetitive crossword. In the cases of Experiments 4 and 5, the result trend indicating weaker elaboration effect applied to both memory retention and high-level understanding of the learned terms. This finding that association generation does not lead to acquisition of higher-level scientific thinking can be explained by the KLI Framework (Koedinger et al., 2012). From the framework's perspective of appropriate transfer of knowledge, if the ultimate goal is to achieve higher-level scientific thinking, the learning event must provide variable knowledge component. In the case of crossword puzzle, it can only enforce learning of paired associates, which is considered constant, as opposed to variable (Roediger & Karpicke, 2006). Hence, crosswords need an add-on elaboration task that provides a variable knowledge component. However, the chosen elaboration task, namely association generation, was not found to be an appropriate choice since it also involved generating a paired associate, and the end product was still constant knowledge component (Koedinger et al., 2012).

Another possible explanation is that association generation risks causing cue overload, in which additionally generated words or concepts, along with target word, can all become cues and create even more cues for themselves (Unsworth & Engle, 2007). Cue overload is known to diminish encoding performance of target words, resulting in insufficient learning outcome (Lehman et al., 2014). It can also be the case that the task is too complex since it requires generation of semantic connections between target word and additional words. It is complex because learners must first become knowledgeable of target word and additional words along with their corresponding definitions or associated facts; only then they can make semantically plausible connections for those words (it is prerequisite that learners have already mastered necessary skills for generating word connection) (Sweller & Chandler, 1994). As such, high degree of interaction among several knowledge components contributes to complexity of the learning task, and this interactivity is what introduces high intrinsic cognitive load (Sweller, 1994). Therefore, it is reasonable to believe that association generation comes with high cognitive load which hinders the task's benefit in learning (Sweller & Chandler, 1994). This explains why this elaboration strategy does not lead to acquisition of higher-level scientific.

Two elaboration strategies: summarization and association generation failed my expectation in producing better learning outcome than repetitive crossword. This led to Experiment 6 whose goal was to explore the efficacy of sentence generation strategy which involved using knowledge of learned technical terms to generate, for each of those terms, a sentence portraying a scenario. The elaboration strategy was shown to have stronger effect than some of the other elaboration strategies, e.g., learning from examples (Coomber et al., 1986). The finding from this experiment suggested that sentence generation, as an add-on elaboration task to crossword, had inconsistent effect strength. For multiple-choice test, which assessed learning gain at memory level, the add-on elaboration resulted in higher performance than repetitive crossword. This did not happen, however, for short-answer test that was designed to evaluate learner's performances in self-explanation and application of the learned terms in other contexts, entailing in-depth understanding of the terms that goes beyond paired associates. One possible reason was that the test was not modified to match exactly with the new task, thus elaboration had little to no effect (Morris et al., 1977; Tulving, 1979). While the task during training asked students to generate their own scenarios, the test asked them to utilize the knowledge on the terms under provided specific scenarios.

Thus far, three elaboration strategies: summarization, association generation, and sentence generation were investigated within their own experiments, and there were mixed results regarding their efficacy as an add-on learning task to crossword. So, which task is the most suitable? To answer this question, Experiment 7 was conducted, in which all three strategies were directly compared against one another. At memory level, students performed equally well regardless of which type of task they completed. This finding agrees with results from other experiments conducted for this dissertation. Looking at recognition performance after students study technical terms using crossword with add-on elaboration from Experiments 3, 4, 5, and 6, the performances were about the same level (within 10% range of test score). Similar result trend also appeared in student's performance on higher-level understanding of the terms. With respect to test type, different elaboration strategies provide nearly the same level of benefits on overall learning.

Based on these results, one might assume that the three elaboration strategies have about the same level of effect on learning and, therefore, they produce equal amount of learning gain. However, this is most likely not the case because other studies have shown that different elaboration techniques involved different levels of thought processing (Brown & Perry Jr, 1991; Coomber et al., 1986). Another possible explanation is that if the three elaboration strategies in fact have unequal influence on learning gain, then the difference is most likely negated by combining them with crossword. This explanation is supported by the KLI Framework (Koedinger et al., 2012), suggesting that students benefit most in learning constant knowledge component from learning activity which enforces application of constant knowledge component. In this case, using crossword puzzle to learn word-definition pairs (constant) will highly benefit recognition of paired associates (constant). Hence, it can be inferred that crosswords have the strongest effect on recognition of paired associates, to the point that differences in the effect strength among types of add-on elaboration tasks is practically ignored.

Aside from add-on elaboration, there are some other interesting factors with potential effect on learning experience and the resulting knowledge gain that should be considered, namely collaborative learning, learning environment (i.e., in-person vs. online), and length of delay period between treatment and evaluation phases. These factors might have come into play during the transition from Experiment 3 to Experiment 4, where several changes in experimental design occurred. The design changes included change from in-person (presented in Experiments 1 - 3) to online participation (presented in Experiments 4 - 7), change from teamwork (presented in Experiments 1 - 3) to individual effort (presented in Experiments 4 - 7), and extension of delay from 24 hours (presented in Experiments 1 - 3) to five days (presented in Experiments 4 - 7). These three changes are known to negatively affect learning experience and, as a result, reduce learning gain. Change from group to individual participation removed the collaborative-learning factor, which negatively affects learning (Johnson & Johnson, 2009; Johnson et al., 1991). From the ICAP Framework's (Chi & Wylie, 2014) perspective, this change resulted in stepping down from *Interaction* mode to *Constructive* mode, reducing potential learning gain to be achieved. Change from in-person to an online environment can possibly negatively affect a student's level of engagement towards the experiment (Baber, 2020; Morgan, 2020). According to Mueller and Veinott (2018), while the effect of crossword still appears even after adding a short delay between training and testing phases, the effect drops tremendously compared to immediate-test results. This suggested that as delay period grows longer, the benefit is also diminished at larger rate.

However, overall findings across seven experiments indirectly suggested that these changes and their negative influence had very little impact here. Looking at results from Experiment 3 and from Experiments 4 and 5, it appeared to have very small drop in performance in either control or two crossword conditions, and the trends are highly similar. One possible explanation for the minimum reduction (if any) caused by removal of collaborative-learning factor from experimental design has to do with the content of an interaction. Chi and Wylie (2014) stated that interaction produced the highest learning, even more than constructive learning, but they also pointed out that not all interactions do so. The key points here are that each individual student involved in a learning activity must first take actions that are considered constructive, and then their interaction must contribute to adding knowledge to one another (Chi & Wylie, 2014). Based on this notion, if removal of collaboration between students does not result in declined performance, it suggests that students fail to engage in a meaningful interaction with each other. Given that none of the seven experiments conducted for this dissertation looked at direct comparison for these changes, the presence of negative impact caused by design change is uncertain. Future research is needed to investigate these changes under the context of crossword with add-on elaboration.

9.2 Practical implementation of this study

This section provides several recommendations for educators who are considering adding crossword puzzle to their course curriculum. First, there is a benefit in having students solve the same crossword puzzle multiple times. Comparing performance results from Experiments 1 and 2 with those from Experiments 3, 4, 5, and 6 reveals that, in most of the cases, solving crossword twice produced higher learning gain than doing it just once. The increase in learning gain comes from the repetition effect, which has been shown to have consistent benefit on learning (Bromage & Mayer, 1986; Mueller & Veinott, 2018). Furthermore, the benefit is positively correlated with number of exposures. So, in general, repeating a learning activity again is better than doing it once, and thrice is better than twice. However, note that there is a limit for number of repetitions before the effect starts to weaken and eventually fade away. For a crossword task in learning vocabulary, the critical point in applying the repetition effect correctly is to first take away a crossword sheet filled by students during their first run and then provide them a new copy. Students must rely on their own memory of either answers themselves or source of information where they find the answers for the first time. Even if students might struggle to solve the same crossword in the second round, it should take them tremendously less time than in the first round (assuming that students manage to fill all the puzzle boxes correctly during the first round). On the other hand, students who have excellent memory might finish the second round so quickly that they have time to slack off. In such cases, educators should encourage their students to spend the remaining time simply reviewing the information in the crossword puzzle.

Even though students were informed of the benefit of repetition effect, it is difficult to deny the possibility that they will feel demotivated in doing the task. Some students might find repeating the same crossword puzzle boring or, worse, pointless and will tackle the task half-heartedly. In such case, rather than force them to do it anyway, it is better to apply a new strategy. My second recommendation is to replace the second round of solving the same crossword with an elaboration task. Across four experiments described in this dissertation, specifically Experiments 3, 4, 5, and 6, three elaboration techniques were investigated for their efficacy as an add-on task to crossword. The results from those experiments showed that adding an elaboration task to crossword for learning vocabulary make the crossword as good as repeating it twice in terms of the produced learning gain. While the downside of crossword with add-on elaboration is that it generally takes more time to complete (depending on how complex the add-on task is), the advantage is that students will more likely not perceive the activity as meaningless and might feel motivated in engaging the task. As for which of the three techniques should educators add to crossword, Experiment 7 shows that they all provide nearly the same level of benefits on overall learning. Therefore, it is best to inform students of available options for elaboration technique and let them choose the one they are most interested in doing. Note that there are other elaboration techniques not investigated in this dissertation. Future research should look into those. In the meantime, I encourage educators to try those techniques out if they feel like experimenting with them, for I believe it should not result in severely less learning gain.

My final recommendation has to do with a factor not directly investigated in this dissertation, namely collaborative learning. Comparing performance results from

Experiments 1, 2, and 3, in which students work together in solving crossword puzzles, with those from Experiments 4, 5, and 6, in which students work on the puzzles individually, reveals that the performances from the two groups do not differ much. Chi and Wylie (2014) mentioned that there are interactions between students that are productive and those that contribute nothing to students' learning gain. In the context of crossword for learning vocabulary, productive collaboration occurs when students assist each other in finding correct words and get into a discussion on which word is the most correct (when they find more than one potential correct answers). Additionally, they could also inform each other of information associated with words they are working on, which many times go beyond the scope of the word's definition. On the other hand, non-productive collaboration is when students try to find correct words on their own, and when one student succeeds, the other simply agrees with and then they both move on without further discussion. Thus, the benefit of collaboration depends heavily on students themselves. My suggestion here is to encourage students to collaborate in solving crossword puzzle. Educators can give their students a training on how to collaborate productively prior to engaging the crossword. Note that, however, even without the training, students might naturally work together productively; and even if that is not the case, non-productive collaboration only adds no value to learning and does not negatively impact the learning gain. Nevertheless, diligent students will be more likely to motivate each other in completing the learning activity.

Mastering technical vocabulary is critical for students who choose to pursue careers related to the STEM fields. Therefore, educators should consider implementing any learning-aid tools that have been scientifically proven to work. Variation of methods in adding elaboration technique to crossword puzzle, as described in this study, can easily be implemented with minimal cost, yet considerably improving student's learning performance. The method also applies to both in papers and online versions, making the learning of technical terms equally available to a large population even during pandemic. I believe that this dissertation serves as a contribution toward the literature of elaboration as one form of constructive learning. It also serves as another step toward the best practices to implement word games in learning technical terms.

10 Reference List

Abuelo, A., Castillo, C., & May, S. A. (2016). Usefulness of crossword puzzles in helping first-year BVSc students learn veterinary terminology. *Journal of veterinary medical education, 43*(3), 255-262.

Alibali, M., & DiRusso, A. (1999). The function of gesture in learning to count: More than keeping track. *Cognitive Development, 14*(1), 37-56.

Amineh, R. J., & Asl, H. D. (2015). Review of constructivism and social constructivism. *Journal of Social Sciences, Literature and Languages, 1*(1), 9-16.

Aronson, E., Blaney, N., Stephan, C., Sikes, J., & Snapp, M. (1978). *The jigsaw classroom*. Beverly Hills, CA: Sage.

Baber, H. (2020). Determinants of students' perceived learning outcome and satisfaction in online learning during the pandemic of COVID-19. *Journal of Education and e-Learning Research,7*(3), 285-292.

Bailey, F., & Pransky, K. (2005). Are" Other People's Children" constructivist learners too?. *Theory into Practice, 44*(1), 19-26.

Bailey, H., Dunlosky, J., & Kane, M. J. (2011). Contribution of strategy use to performance on complex and simple span tasks. *Memory & cognition, 39*(3), 447-461.

Bartsch, L. M., & Oberauer, K. (2021). The effects of elaboration on working memory and long term memory across age. *Journal of Memory and Language, 118*(104215), 1-16.

Berry, D. C., & Miller, M. G. (2008b). Crossword puzzles as a tool to enhance athletic Training student learning: Part 2. *Athletic Therapy Today, 13*(1), 32.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Handbook 1: Cognitive domain. New York: Longmans.

Bonwell, C. C., & Sutherland, T. E. (1996). The active learning continuum: Choosing activities to engage students in the classroom. *New Directions for Teaching and Learning, 1996*(67), 3-16.

Bromage, B. K., & Mayer, R. E. (1986). Quantitative and qualitative effects of repetition on learning from technical text. *Journal of Educational Psychology, 78*(4), 271-278.

Brown, T. S., & Perry Jr, F. L. (1991). A comparison of three learning strategies for ESL vocabulary acquisition. *Tesol Quarterly, 25*(4), 655-670.

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31*, 21-32

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough,&J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society*. 56–64. New York: Worth.

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.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(2), 145-182.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist, 49*(4), 219-243.

Cobb, P. (1994). Where is the mind? Constructivist and sociocultural perspectives on mathematical development. *Educational researcher, 23*(7), 13-20.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(4), 213-220.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Coomber, J. E., Ramstad, D. A., & Sheets, D. R. (1986). Elaboration in vocabulary learning: A comparison of three rehearsal methods. *Research in the Teaching of English, 20*(3), 281-293.

Coticone, S. R. (2013). Utility of self-made crossword puzzles as an active learning method to study biochemistry in undergraduate education. *Journal of College Science Teaching, 42*(4), 33-37.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences, 24*(1), 87-114.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory?. *Progress in brain research, 169*, 323-338.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671-684.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268-294.

Crossman, E. K., & Crossman, S. M. (1983). The crossword puzzle as a teaching tool. *Teaching of Psychology, 10*(2), 98-99.

Davis, T. M., Shepherd, B., & Zwiefelhofer, T. (2009). Reviewing for exams: Do crossword puzzles help in the success of student learning? *The Journal of Effective Teaching, 9*(3), 3-9.

DeVries, D. L., & Slavin, R. E. (1978). Teams-Games-Tournament: A research review. *Journal of Research and Development in Education, 12*, 28-38.

Dunlosky, J., & Kane, M. J. (2007). The contributions of strategy use to working memory span: A comparison of strategy assessment methods. *The Quarterly Journal of Experimental Psychology, 60*(9), 1227-1245.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review, 28*(4), 717-741.

Fisher, R. P., & Craik, F. I. (1980). The effects of elaboration on recognition memory. *Memory & Cognition, 8*(5), 400-404.

Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology, 21*(4), 499-526.

Gaikwad, N., & Tankhiwale, S. (2012). Crossword puzzles: Self-learning tool in pharmacology. *Perspectives on medical education, 1*(5), 237-248.

Galli, G. (2014). What makes deeply encoded items memorable? Insights into the levels of processing framework from neuroimaging and neuromodulation. *Frontiers in Psychiatry, 5*(61), 1-8.

Gilani, R., Niranjane, P., Daigavane, P., Bajaj, P., Mankar, N., & Vishnani, R. (2020). Crossword puzzle: An effective self-learning modality for dental undergraduates. *Journal of Datta Meghe Institute of Medical Sciences University, 15*(3), 397.

Glaser, R. (1991). The maturing of the relationship between the science of learning and cognition and educational practice. *Learning and Instruction, 1*(2), 129-144.

Greene, R. L. (1987). Effects of maintenance rehearsal on human memory. *Psychological Bulletin, 102*(3), 403-413.

Hamilton, R. J. (1997). Effects of three types of elaboration on learning concepts from text. *Contemporary Educational Psychology, 22*(3), 299-318.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of educational research, 56*(4), 473-493.

Ismayenti, I., Anwar, S., & Khairani, K. (2018, April). Improvement of activity and social science learning outcomes through Teams Game Tournament using crossword puzzle. In *International Conferences on Educational, Social Sciences and Technology* (pp. 807-815). Fakultas Ilmu Pendidikan UNP.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*(3), 621.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Journal of Educational Researcher, 38*(5), 365-379.

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1991). *Cooperative Learning: Increasing College Faculty Instructional Productivity*. ASHE-ERIC Higher Education Report No. 4. George Washington University, One Dupont Circle, Suite 630, Washington, DC 20036-1183.

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.

Katayama, A. D., Shambaugh, R. N., & Doctor, T. (2005). Promoting knowledge transfer with electronic note taking. *Teaching of Psychology, 32*(2), 129-131.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge - Learning – Instruction framework: Bridging the science - practice chasm to enhance robust student learning. *Cognitive science, 36*(5), 757-798.

Kruawong, T., & Soontornwipast, K. (2021). The development of Thai EFL secondary school students' English science vocabulary knowledge through Science Vocabulary Crossword Puzzle (SVCP) practices: Action research. *Shanlax International Journal of Education, 9*(4), 44-54.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1787-1794.

Levin, J. R. (1988). Elaboration-based learning strategies: Powerful theory = powerful application. *Contemporary Educational Psychology, 13*(3), 191-205.

Marschark, M., Richman, C. L., Yuille, J. C., & Hunt, R. R. (1987). The role of imagery in memory: On shared and distinctive information. *Psychological Bulletin, 102*(1), 28-41.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica, 22*(3), 276-282.

McNamara, D. S., & Scott, J. L. (2001). Working memory capacity and strategy use. *Memory and Cognition, 29*(1), 10-17.

Morgan, H. (2020). Best practices for implementing remote learning during a pandemic. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 93*(3), 135-141.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of verbal learning and verbal behavior, 16*(5), 519-533.

Mueller, S., & Veinott, E. (2018). Testing the effectiveness of crossword games on immediate and delayed memory for scientific vocabulary and concepts. doi:10.13140/RG.2.2.34982.63047.

Paivio, A., & Yuille, J. C. (1969). Changes in associative strategies and paired-associate learning over trials as a function of word imagery and type of learning set. *Journal of Experimental Psychology, 79*(3, Pt.1), 458-463.

Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures. (Trans A. Rosin)*. New York: The Viking Press.

Pratiwi, A. (2021). The effect of crossword puzzle learning strategies on fourth grade students' learning outcomes on science learning contents at Gunung Sari 1 Elementary School Rappocini District Makassar City. *International Journal of Elementary School Teacher, 1*(1), 1-12.

Pruitt, C. M. (1927). Activities of chemistry clubs. *Journal of Chemical Education, 4*(8), 1037-1042.

Rambli, D., Matcha, W., & Sulaiman, S. (2013). Fun learning with AR alphabet book for preschool children. *Procedia Computer Science, 25*, 211-219.

Richardson, J. T. E. (1998). The availability and effectiveness of reported mediators in associative learning: A historical review and an experimental investigation. *Psychonomic Bulletin & Review, 5*(4), 597-614.

Richardson, V. (1997). Constructivist teaching and teacher education: Theory and practice. *Constructivist teacher education: Building a world of new understandings*, 3-14.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210.

Romero, M., & Barbera, E. (2011). Quality of learners' time and learning performance beyond quantitative time-on-task. *International Review of Research in Open and Distributed Learning, 12*(5), 125-137.

Rosenthal, J. S. (1995). Active learning strategies in advanced mathematics classes. *Studies in Higher Education, 20*, 223-228.

Rowles, C. J. (2013). Strategies to promote critical thinking and active learning. *Teaching in nursing: A guide for faculty*, 258-279.

Rubin, J., & Herbert, M. (1998). Peer teaching-model for active learning. *College Teaching, 48*, 26-30.

Ruron, M. N., & Hastuti, S. P. (2020). The development of crossword game media on virus material. *EDUSAINS, 12*(2), 188-195.

Saran, R., & Kumar, S. (2015). Use of crossword puzzle as a teaching aid to facilitate active learning in dental materials. *Indian Journal of Applied Research, 5*(4), 456-457.

Saxena, A., Nesbitt, R., Pahwa, P., & Mills, S. (2009). Crossword puzzles: Active learning in undergraduate pathology and medical education. *Archives of pathology & laboratory medicine, 133*(9), 1457-1462.

Schifter, D., & Simon, M. A. (1992). Assessing teachers' development of a constructivist view of mathematics learning. *Teaching and Teacher education, 8*(2), 187-197.

Sharan, S., & Sharan, Y. (1976). *Small-Group Teaching*. Englewood Cliffs, NJ: Educational Technology Publications.

Shawahna, R., & Jaber, M. (2020). Crossword puzzles improve learning of Palestinian nursing students about pharmacology of epilepsy: Results of a randomized controlled study. *Epilepsy & Behavior, 106*, 1-10.

Shetgar, M. P. S., & Thalange, M. A. V. (2018). Crossword puzzle: An active learning strategy. *IJRAR-International Journal of Research and Analytical Reviews, 5*(2), 1317-1322.

Slavin, R. E. (1980). Cooperative learning. *Review of educational research, 50*(2), 315-342.

Smith, B. L., & MacGregor, J. T. (1992). What is collaborative learning. *Towards the Virtual University: International Online Learning Perspectives*, 217-232.

Stein, B. L., & Kirby, J. R. (1992). The effects of text absent and text present conditions on summarization and recall of text. *Journal of Reading Behavior, 24*(2), 217-232.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295-312.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and instruction, 12*(3), 185-233.

Turley-Ames, K. J., & Whitfield, M. M. (2003). Strategy training and working memory task performance. *Journal of Memory and Language, 49*(4), 446-468.

Tulving, E. (1979). Relation between encoding specificity and levels of processing. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory*. Hillsdale, N.J: Erlbaum.

Tulving, E., & Madigan, S. A. (1970). Memory and verbal learning. *Annual review of psychology, 21*(1), 437-484.

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin, 133*(6), 1038-1066.

Van Vleet, R. C. (1925). Crossword puzzles in Hollywood High School. *Journal of Chemical Education, 2*(4), 292-294.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197-221.

Veinott, E. S., & Whitaker, E. (2019, July). Leaving hints: Using player in-game hints to measure and improve learning. In *International Conference on Human-Computer Interaction* (pp. 222-230). Springer, Cham.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge: Harvard University Press.

Wagner, P., Schober, B., & Spiel, C. (2008). Time students spend working at home for school. *Learning and Instruction, 18*(4), 309-320.

Wang, Q. (2009). Design and evaluation of a collaborative learning environment. *Computers & Education, 53*(4), 1138-1146.

Weisskirch, R. S. (2006). An analysis of instructorcreated crossword puzzles for student review. *College Teaching, 54*(1), 198-201.

Whisenand, T. G., & Dunphy, S. M. (2010). Accelerating student learning of technology terms: The crossword puzzle exercise. *Journal of Information Systems Education, 21*(2), 141-148.

Wilson, L. O. (2016). Anderson and Krathwohl – Bloom's taxonomy revised. *Understanding the New Version of Bloom's Taxonomy*.

Windschitl, M. (2002). Framing constructivism in practice as the negotiation of dilemmas: An analysis of the conceptual, pedagogical, cultural, and political challenges facing teachers. *Review of educational research, 72*(2), 131-175.

Wong, C. L., Chu, H. E., & Yap, K. C. (2014). Developing a framework for analyzing definitions: A study of the Feynman Lectures. *International Journal of Science Education, 36*(15), 2481-2513.

Wood, G., & Bolt, M. (1968). Mediation and mediation time in paired-associate learning. *Journal of experimental psychology, 78*(1), 15.

Yager, R. E. (1983). The importance of terminology in teaching K-12 science. *Journal of Research in Science Teaching, 20*(6), 577-588.

Yaron, D., Karabinos, M., Lange, D., Greeno, J. G., & Leinhardt, G. (2010). The ChemCollective—virtual labs for introductory chemistry courses. *Science, 328*(5978), 584-585.

Yew, E. H., & Schmidt, H. G. (2009). Evidence for constructive, self-regulatory, and collaborative processes in problem-based learning. *Advances in Health Sciences Education, 14*(2), 251-273.

Yuriev, E., Capuano, B., & Short, J. L. (2016). Crossword puzzles for chemistry education: Learning goals beyond vocabulary. *Chemistry Education Research and Practice, 17*(3), 532-554.

A Appendix

Table A1. List of technical terms selected as stimuli for Experiment 1

Selected Terms	Chapter 5	Chapter 6	Chapter 7
Term 1	chromosomes	accommodation	absolute threshold
Term 2	darwin	adolescence	adaptation
Term 3	epigenetics	assimilation	visual cliff
Term 4	evolutionary	cognition	cones
Term 5	genome	egocentrism	difference threshold
Term 6	heredity	habituation	disparity
Term 7	heritability	imprinting	fovea
Term 8	interaction	intimacy	gestalt
Term 9	monozygotic	maturation	helmholtz
Term 10	mutation	object permanence	hue
Term 11	natural selection	piaget	priming
Term 12	nurture	schema	subliminal
Term 13	social script	teratogens	transduction
Term 14	temperament	zygote	weber

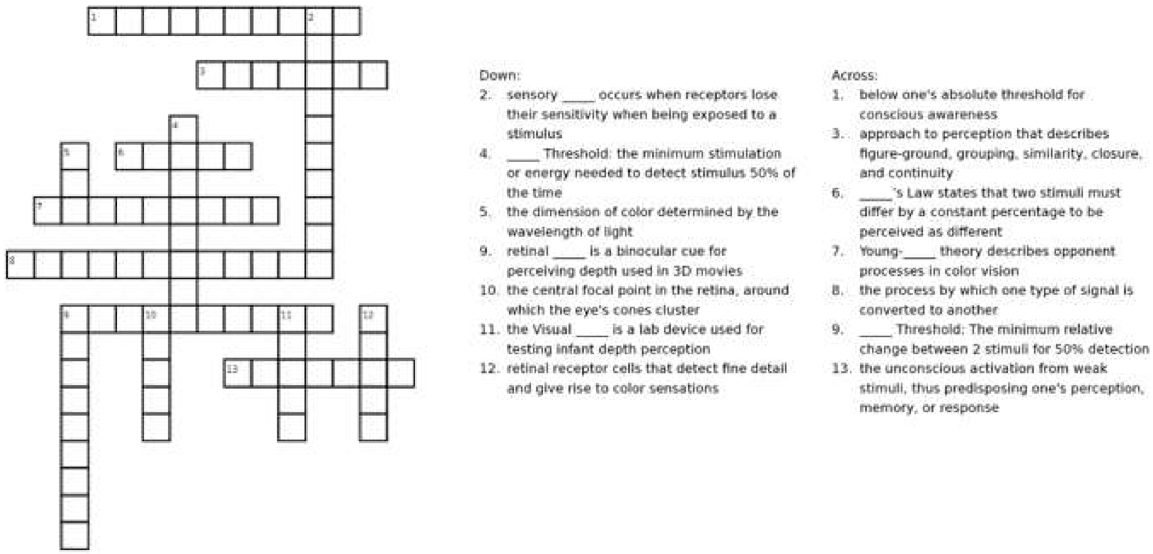

Figure A1. Sample crossword puzzle with corresponding hints used in Experiment 1.

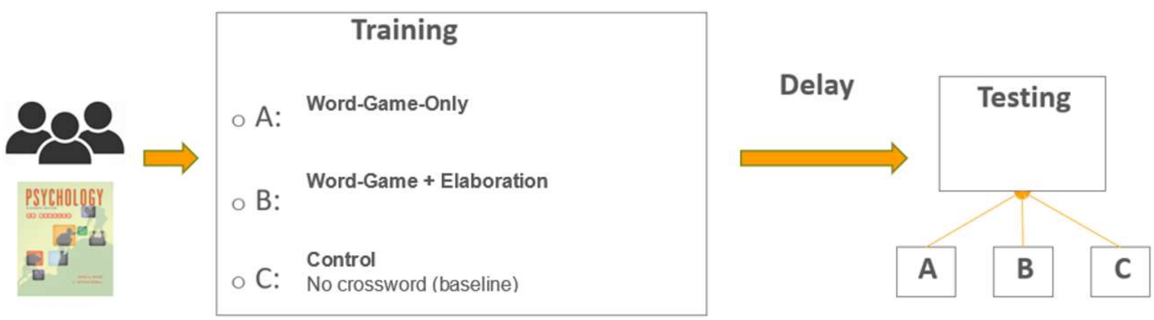

Figure A2. Diagram depicting procedures for Experiments 1 & 2.