8-4-2023 10:00 AM

Learning Verb-Noun Collocations Through Multiple-Choice Exercises: Do Distractors Benefit or Hinder Later Recall?

Mengxue (Alyssa) Li,

Supervisor: Boers, Frank, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Arts degree in Education

Abstract

This study aimed to investigate under what conditions multiple-choice exercises benefit second language learners' acquisition of lexical phrases. Of particular interest was the question whether the distractors in the multiple-choice items create interference when learners' later try to recall the lexical phrases. Twenty advanced ESL (English as a Second Language) learners were given 20 multiple-choice items on verb-noun collocations (e.g., *run a business, take a toll, speak volumes*) followed by feedback. They were then tested on the same collocations two weeks later by means of gap-fill items. The participants were invited to verbalize their thoughts during the exercise and the test. They were subsequently shown the multiple-choice exercise again and asked if they could remember how they had responded to each item and what the correct response option was.

This mixed-methods study revealed poor effectiveness of the multiple-choice exercises overall. When learners failed to produce the correct response in the post-test, this was either because they could not remember which of the response options in the multiple-choice item turned out to be the correct one or because they simply had not recollection of the exercise item. The likelihood of producing a correct response in the post-test increased when learners (a) chose the correct response option in the multiple-choice item, (b) remembered the multiple-choice item, and (c) accurately recalled the feedback received on the item. Individual learner differences and item characteristics also influenced the effectiveness of multiple-choice exercises for learning collocations. The findings suggest that, for multiple-choice exercises to be relatively beneficial for collocation learning, they need to be designed and implemented in a way that ensures a high accuracy rate at the exercise stage.

Keywords

Multiple-choice, Selected-response exercise, Think-aloud, Stimulated recall, Collocation

Summary for Lay Audience

Selected-response items such as multiple-choice questions are commonly used for learning and testing purposes. However, their effectiveness in introducing new language knowledge is debatable. Selected-response exercises require learners to identify the correct answer among distractors, and exposure to incorrect information during a learning phase can be detrimental to later recollection (Remmers & Remmers, 1926). Some studies in the domain of language learning (e.g., Boers et al., 2014; Boers et al., 2017) have reported that such exercises lead to only small learning gains according to tests administered a couple of weeks later, because learners often produce one of the incorrect response options that they were asked to consider in the exercises. However, it remains unclear whether these wrong test responses are due to interference from the distractors in the exercises. After all, it is also possible that learners have simply forgotten all about an exercise item when they are tested after a time lapse. Besides, other studies have shown that making mistakes can be beneficial for learning, provided one recalls the mistake (and that it was a mistake), because remembering the mistake can help one to recall what the correct response turned out to be (Metcalfe & Huelser, 2020). However, this effect of learning from errors has been demonstrated for mistakes learners generate themselves. This may be different in the case of multiple-choice exercises where learners are asked to make a choice between response options some of which might not have spontaneously occurred to them in the first place.

In this mixed-methods study, we examined the impact of multiple-choice distractors when introducing new verb-noun collocations (e.g., *pose a threat, run a business, set the tone*). Twenty advanced ESL learners participated in the study and were invited to learn 20 verb-noun collocations through multiple-choice exercises, after which they received feedback. Two weeks later, they were tested on the same items using a gap-fill format. They were also provided with the original multiple-choice exercises and asked to recall how they tackled the exercises two weeks previously. The learners were encouraged to verbalize their thoughts during the exercises and the two-week delayed post-test.

The results indicated that some of the failed responses in the post-test could be attributed to interference from the exercise items while in other cases the learners could not recall the exercise item. Unsurprisingly, learners who accurately remembered how they had tacked a

given exercise item and what feedback they had received stood the best chance of providing the correct response in the post-test. Interestingly, some learners who vividly recalled tackling certain exercise items were either not certain which response option turned out to be correct or misremembered the feedback they received. The likelihood of correct post-test responses was the highest when learners had chosen correctly at the exercise stage and had accurate recollections of the feedback they had received. Additionally, individual learner differences and item characteristics influenced the effectiveness of multiple-choice exercises for learning collocations.

Should teachers or course designers wish to keep using such exercises, then they are advised to implement them in a way that ensures a high accuracy rate at the exercise stage so as to reduce the risk of competition in memory between correct and wrong responses.

Acknowledgments

This dissertation marks the culmination of my M.A. journey. I would like to take this opportunity to express my gratitude to the people whose support has been indispensable in completing this journey.

First and foremost, I would like to express my heartfelt appreciation to my supervisor, Dr. Frank Boers. I thank you for encouraging me to explore my research interests, lending a patient ear to my concerns, providing me with constructive feedback, and offering me valuable opportunities for my research career. Your guidance throughout this process has been truly invaluable. I would also like to extend my appreciation to Dr. Ruslan Suvorov for his insightful comments on my research proposal, which offered a fresh perspective from the field of language assessment.

A big thank-you goes to my fellows in the Applied Linguistics group at the Faculty of Education. I thank you for the suggestions on the statistical analyses and your comforting words during my moments of self-doubt. I also enjoyed the jokes, the drinks, and the get-togethers – those non-academic happy moments have brought joy to this journey as well. Additionally, I would like to express my gratitude to the participants of this study; your active involvement has been crucial to its completion.

To my family members who have been there for me with their understanding and support, this journey would not have been possible without your backing. To Bubble, my little four-legged fluffy one, who is turning three as I write this, happy birthday. Your constant company has kept me busier but also healthier and happier throughout these years.

Lastly, I acknowledge SSHRC (Social Sciences and Humanities Research Council) for providing funding support for this research project.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1 Introduction

Second language (L2) learners find it challenging to master verb-noun collocations (see Boers et al., 2014; Stengers & Boers, 2015, for review), and contemporary language courses include exercises intended to help learners to overcome this challenge. Textbook analyses (e.g., Boers et al., 2014; Boers et al., 2017) have revealed that collocation exercises are in fact a common means to *introduce* new collocations to learners (rather than serving the purpose of practicing previously studied collocations). These textbook analyses also reveal that collocation exercises are commonly of one or the other *selected-response* format (e.g., multiple-choice exercises), where learners need to distinguish the correct target from a set of plausible distractors. The effect of the distractors that present learners with erroneous language instances at the learning stage is the topic of interest in the research project reported here.

Erroneous information in exercises can be harmful. A considerable number of studies have indicated that when participants take selected-response tests (i.e., essentially variants of multiple-choice tests), they may later mistakenly recollect distractors as the correct 'knowledge' (e.g., Brown et al., 1999; Fazio et al., 2010; Marsh et al., 2009; Roediger & Marsh, 2005; Toppino & Luipersbeck, 1993). To avoid this, selected-response tasks should be followed by *feedback* to serve as learning material (Butler & Roediger, 2008).

In the domain of language learning, however, research has revealed that **corrective feedback may not prevent interference from selected-response tasks**. Boers and colleagues' studies (Boers et al., 2014; Boers et al., 2017) indicated that selected-response exercises followed by feedback yielded little gain in collocation learning. In these studies, learners sometimes reproduced their exercise errors in the post-test or replaced their correct exercise answer by one of the other response options they had seen in the exercise, although they were provided with post-exercise feedback. This led the authors to speculate that the poor learning gains attested in their studies were due to interference from incorrect response options which the learners had considered in the exercises. However, the reason for these post-test errors is not entirely clear. While they

may be due to interference from the distractors in the exercise, it is also possible that the re-emergence of exercise distractors among post-test responses is coincidental. Perhaps learners have no recollection of the exercise item in the first place, resort to the same intuition or guessing strategy in the post-test, and thus reproduce the same error.

However, **there is also evidence that making errors can be beneficial**, because one can learn from feedback received on errors (see Metcalfe, 2017, for a review). It is important to note that the evidence in support of this view concerns mostly errors generated by participants themselves (i.e., in constructed-response tasks, not selected-response tasks). Metcalfe and Huelser (2020) attributed such error-generation benefit to *episodic memory*. In their study, the learners who could recall their exercise errors outperformed those who failed to recall their errors when they took the post-test. Whether this beneficial effect of episodic memory is transferable to selected-response exercises is not clear—and this is one of the questions addressed in this study.

This study aims to investigate the effect of exercise distractors with a focus on L2 English collocation learning. Twenty advanced English L2 learners were invited to an exercise session and a post-test session. The exercises used a selected-response format (multiple-choice, where the participants need to select the right verb from three options to complete a blank before a noun phrase) (e.g., run a business; pose a threat). The post-test session used a constructed-response format (gapped sentences, where the participant needs to supply the missing verb). After the post-test, the multiple-choice exercises were presented to the participants again as a prompt for stimulated recall, to see if the learners had any episodic memories of how they had tackled certain exercise items and of the feedback received. The participants met the researcher individually for both sessions. They were asked to 'think aloud' while they did the multiple-choice exercises and took the tests. The think-aloud data from the exercise session helped to gauge per exercise item what distractors (if any) were explicitly considered by the participant. The participant's post-test responses were then interpreted with reference to this information. The think-aloud data from the post-test session helped to gauge whether the participant remembered the exercise items, their responses, and the feedback received.

The study thus adopted a mixed-methods design to investigate if multiple-choice distractors linger in learners' memory, and, if so, if this is harmful or beneficial for later recall of the correct verb-noun combinations. The purpose is to relate the participants' performance in the post-test to how they processed the multiple-choice exercise items, and to suggest steps to make the latter more effective. This investigation of the effects of exercise distractors will help to evaluate the merits and drawbacks of multiple-choice exercises for collocation learning, and provide implications for language learning, general learning, and memory.

Chapter 2 Background

As mentioned in the previous chapter, ESL textbook analyses (Boers et al., 2014; Boers et al., 2017) have revealed that collocations exercises are commonly of selected-response formats and used to introduce new language instances. However, using selected-response exercises to introduce collocations to learners could be problematic because of potential interference in memory by the incorrect response options (i.e., the distractors) that the learners are invited to consider.

This chapter aims to provide an overview of the problems of distractors in selected-response exercises. Do distractors hinder or benefit later recall? Supportive evidence of both sides can be found in the available literature. This chapter will review those theories and empirical studies, highlight some key concepts, and state the research gaps this study was meant to address.

2.1 Multiple-Choice and Selected-Response Exercises

Selected-response and *constructed-response* are two test formats that are commonly used in the domain of assessment and testing (see Downing, 2009, for review). Selected-response items are ones where test-takers are presented with a list of plausible answers and expected to select the correct one(s) from the list. Multiple-choice questions are the prototypical selected-response format. It is worth noting that diverse selected-response formats (e.g., matching, true-false, choosing the right word from two or more options) are all essentially multiple-choice (or binary-choice) tasks where test-takers need to distinguish between correct and incorrect options. Constructed-response formats are those where test-takers produce their own responses to a prompt, such as a gap-fill exercise without a bank of options to choose from, short-answer questions, and free-response tasks.

Testing tools are oftentimes used as learning tools. This may firstly be because the learning materials are designed to prepare learners for high-stakes tests in some contexts. Practicing an exercise format which is basically the same as the test format may efficiently improve learners' later test performance. Additionally, research has shown that

testing rather than simply re-presenting information to learners for review boosts long-term retention. This is called the retrieval effect or *Testing Effect* (see, Roediger & Karpicke, 2006a, 2006b, for review). However, using selected-response exercises for learning **new** language items may be questionable. The effect of distractors that present learners with erroneous language instances at the learning stage is the topic of interest in this study.

2.2 Harmful Exercise Distractors

A core feature of selected-response items is that there are one or more distractors (or lures) that are mixed with the correct response. Research has indicated that erroneous information in exercises can cause confusion. According to the *Negative Suggestion Effect* (Remmers & Remmers, 1926) in Memory Research, exposure to incorrect information will increase the chance of later false recollection. That is, people might recall erroneous information and consider it correct at a later retrieval stage. In the domain of learning, a considerable number of studies are consistent with the Negative Suggestion Effect, indicating that in *selected-response* tasks such as multiple-choice, the misinformation (i.e., distractors) may later be mistakenly recollected as correct 'knowledge' (e.g., Brown et al., 1999; Fazio et al., 2010; Roediger & Marsh, 2005; Toppino & Luipersbeck, 1993). The extent of such negative effect may depend on the test-taker's initial test performance. Marsh et al. (2009) found that test-takers who scored better on the initial multiple-choice test showed less negative effect and a larger positive testing effect than the more poorly achieving test-takers.

Corrrective feedback is effective to reduce this negative effect and thus to enhance the positive effect of multiple-choice testing, although it does not appear to totally prevent interference (Butler & Roediger, 2008). Pairing multiple-choice tasks with corrective feedback is how multiple-choice *exercises* (instead of tests) are normally implemented, and so this should reduce the negative effects of exposure to distractors (e.g., Marsh et al., 2009).

However, this positive effect of feedback is not guaranteed for selected-response exercises. In the domain of language learning, Boers (2021) has suggested that selected-

response exercises, where learners are presented with erroneous language uses (as distractors), risk leaving erroneous impressions on learners' memory and causing interference at a later retrieval stage. It is worth noting that the distractors in language exercises are often made up by the material designers rather than having been generated by the learners themselves. It may be different to give feedback on material-presented errors rather than on learners' self-generated errors. In fact, some empirical studies gauging different formats of collocation exercises revealed that, even with post-exercise corrective feedback, the learners sometimes reproduced the errors they made in the exercise or produced another distractor that they were presented with in the exercise (Boers et al., 2014; Boers et al., 2017).

The reproduction of errors and emergence of distractors in post-tests might be manifestations of unconscious *false memories*. According to the *False Fame Paradigm* (Jacoby et al., 1989), people may mistakenly judge misinformation that they were told was false as true at a later recollection stage, simply because the information (which was initially new to them) now looks familiar. The findings of Boers and colleagues are consistent with the False Fame Paradigm. It is possible that learners have no explicit recollection of the exercise item but consider an incorrect verb-noun collocation to be correct simply because it looks familiar after having seen it in an exercise.

According to the above account, the distractors in selected-response exercises might later be treated as correct knowledge by learners because of the phenomenon of false memory. Alternatively, learners may be **conscious** of what choices they were exposed to in an exercise and yet fail to remember what the correct option was. This seems likely especially in the case when distractors look as plausible as the correct response option does, which may hold true for learning targets such as verb-noun collocations. Selected-response exercises on verb-noun collocations oftentimes require learners to compare semantically vague verbs (e.g., *Select the correct verb to complete the phrase: _____ assignment. 1, Do; 2, Make; 3, Take*), which all look plausible (after all, one can *make a presentation* and one can *take a test*; so why should it be *do an assignment*?) Similarly, exercise designers often use near-synonyms of the correct choice as distractors (e.g., *Select the correct verb to complete the phrase: _____ the truth. 1, Say; 2, Tell; 3, Speak*).

According to the *Noticing Hypothesis* (Schmidt, 1990), giving attention to new language elements is the initial and necessary step in learning them. In a selected-response exercise, unless the target is already well known, the learner will attend to plausible-looking response options to try to determine which is the most plausible one. This is, in essence, why the incorrect options in multiple-choice tests are called "distractors" in the first place—they invite attention. **As a result, incorrect options may be as likely as the correct option to be noticed** (i.e., held in working memory), and then possibly stored in long-term memory. Later, at the retrieval stage, despite having received corrective feedback, learners may find it difficult to recall which combination is correct. The verb-noun combinations appear to be **arbitrary**, which may make selected-response exercises on verb-noun collocations different from those where learners can refer to some sort of logic or generalized pattern or "rule" (e.g., as in the case of grammar exercises).

According to the above theories, the emergence in post-tests of distractors that learners were exposed to at the exercise stage could be manifestations of mistaken memories. It is worth noting that they can also be manifestations of **no memory**, in which case distractors do not linger in learners' memory. That is, learners may simply have totally forgotten the exercise items and the feedback received by the time they are tested, and at that testing stage use the same strategy (e.g., guessing) as before, in which case the errors are unrelated to the earlier exposure to the exercise items.

2.3 Beneficial Exercise Errors

Despite the above research providing evidence that exposure to errors at the exercise stage could be harmful by leaving erroneous impressions and hindering later recollection of the correct response, some research has indicated that exercise errors may in fact be beneficial, provided learners remember their errors (and that they were errors). Put differently, errors during learning, followed by feedback, may be beneficial if *episodic memory* occurs.

Episodic memory was first coined by Tulving (1972), who distinguished it from semantic memory. Episodic memory refers to people's recollection of the personal experience of a learning event, whereas semantic memory refers to people's recollection of general

knowledge, divorced from memories of when and how that knowledge was acquired (see, Tulving, 1993, for further clarification and review). In both cases, one may be able to verbalize one's knowledge when prompted to do so. For example, a proficient learner of English will know that English is an SVO language and may even be able to explicitly verbalize this knowledge ('SVO represents subject-verb-object order in affirmative sentences') but may not recall when they learned this. If so, this knowledge belongs to semantic memory. Episodic memory is, for instance, 'I remember learning the SVO concept in a high-school English class from my favorite teacher'.

A crucial function of episodic memory is correcting erroneous information (Baddeley & Wilson, 1994). The theory of *Recursive Reminding*, which states that people's recall of information is more accurate when they remember the related event that included this information (see, Jacoby & Wahlheim, 2013; Wahlheim & Jacoby, 2013, for review), is consistent with episodic memory — one learns from mistakes by remembering the circumstances in which the mistakes were made and how they were rectified.

Learners' memory is enhanced when an error is generated and then corrected (i.e., trial-and-error followed by feedback), as compared to when learners are only presented with the correct target (i.e., study-only). This phenomenon is referred to as *error-generated benefit*. Metcalfe and Huelser (2020) indicated that the error-generation benefit should be attributed to episodic rather than semantic memory. In their study, it was only when learners could recollect the errors which they had made that there was error-generation benefit; when learners failed to recollect the errors they had generated, there was no such benefit, as compared to the study-only condition. The learners who successfully recalled their exercise errors (and who by that token had episodic memories of making the errors) outperformed those who failed to recall their initial errors in a post-test. It is worth noting, however, that this line of research on 'learning from one's mistakes' has so far concerned the benefits of error generation (i.e., learners' own errors) rather than selected-response exercises. This is relevant because, as previously mentioned, selected-response exercises are very common in materials for L2 collocation learning. One of the questions addressed in the study reported here is therefore whether *episodic memories* of one's responses in multiple-choice exercises reduce the risk of later interference not only from

the incorrect response selected in the exercise but also from the other distractors in such exercises.

2.4 Summary and Research Gaps

Do the distractors in a multiple-choice exercise on collocations benefit, hinder, or have no effect on later retrieval? The answer is not clear. On the one hand, the theories of *False Memory* and *Negative Suggestion Effect* in cognitive psychology suggest that the distractors may leave erroneous memories and thus hinder later retrieval. Some empirical studies are consistent with these theories. However, in the domain of language learning, we would need further evidence to determine if post-test errors that duplicate distractors from multiple-choice exercises are to be attributed to false memories (whereby the verb-noun combination feels vaguely familiar because it was one of the distractors), or if such duplicates occur even though the learner remembers the exercise item (but is confused about which option turned out to be correct), or the learner remembers neither the exercise item or the feedback.

On the other hand, multiple-choice exercises may in fact be beneficial if the distractors leave *episodic memories* (i.e., memories of evaluating the distractors in the exercise, choosing them or dismissing them, and then finding out from the feedback what the correct response was), thus providing extra recollection material for learners at a later stage. If so, it is theoretically conceivable that material-presented errors benefit later recollection of the target just as well as self-generated errors do.

Summing up this chapter, more research is needed to investigate the effects of multiple-choice exercises to inform educational practice, in this case regarding collocation learning. This study aims to contribute to the below RGs (research gaps).

RG1. Memory or coincidence? Research has shown that when selected-response exercises are used for learning collocations, learners at times reproduce their exercise errors in the post-test, despite the provision of post-exercise corrective feedback. However, the sources of error-reproduction are not clear. Do learners reproduce errors due to interference (where the exposure to exercise distractors hinders learners' later

recall of the correct response) or coincidence (where the learners simply have no recollection of the exercise item)? I have found no study that directly answered this question.

RG2. Does episodic memory help? Research has provided evidence that, when learners benefit from generating errors, this is attributed to episodic memory of the errors that they made in the exercise. In other words, learners who can recollect their earlier exercise errors tend to outperform those who do not in a post-test. However, it is not clear whether episodic memory is as beneficial when it comes to material-designed errors in selected-response exercises (instead of learner's self-generated errors in constructed-response exercises).

RG3. What else predicts the relative effectiveness of selected-response exercises? Research has provided indirect evidence that using selected-response exercises to introduce new language items may bring about little learning gain, despite post-exercise corrective feedback. However, this might be related to the learners' exercise performance. For example, if the success rate at the exercise stage is poor, then learners will need to process a lot of corrective feedback to replace incorrect responses in memory. By contrast, if learners make few errors at the exercise stage, then it may be better manageable to remember the corrective feedback on those errors. In addition, the precise nature of the exercise items may matter, because some distractors may be easier to dismiss than others. It is conceivable that it is especially items which trigger hesitation between two plausible looking response options that leave an episodic memory. This is also likely for items where learners are surprised by the corrective feedback. To evaluate the role of these and aforementioned factors, this study endeavored to relate individual learners' post-test performance to how they experienced the exercises. As explained further below, this was done through analyses of "think-aloud" and "stimulated-recall" data.

Chapter 3 Research Questions

Previous studies have reported disappointing learning gains from exercises on verb-noun collocations that used a selected response format and that were used in a spirit of trial and error. Boers et al. (2014), Stengers & Boers (2015), and Boers et al. (2017) report a total of 20 trials of diverse exercise formats where learners were required to distinguish correct from incorrect response options. The average learning gain (calculated by comparing post-test scores and pre-test scores) was only 16%. It was speculated that the effectiveness of such exercises was so poor because the wrong response options that the learners were invited to consider interfered with their recall of the correct responses. The main purpose of the present study is to further evaluate the plausibility of this. More specifically, this study aims to investigate the effects of distractors in multiple-choice collocation exercises. The general research question of the study is whether multiple-choice distractors hinder, benefit, or have no effect on later recall of the correct response, with a focus on L2 English collocation learning. In addition, the aim is to examine the conditions under which distractors are likely to hinder or benefit learning, with reference to the theories reviewed earlier, especially regarding the role of episodic memory. Derived from the general question, below are the RQs (research questions) the study aims to address:

RQ1. When learners produce a response in a post-test that corresponds to a distractor in a multiple-choice exercise they did previously, then is this because they remember seeing this response option or is it a coincidence?

RQ2. Does episodic memory of how they tackled specific multiple-choice exercise items help learners to recall the correct response?

RQ3. Does a relatively high success rate at the exercise stage help learners to benefit from (corrective) feedback?

RQ1 is related to RG1 (see RGs in Background above), investigating if duplicates and emergences of distractor are related to ‘false memory’ (erroneous impressions lingering in learners’ memory) or ‘no memory’ (learners use the same strategy such as guessing).

RQ2 is related to RG2, examining if episodic memory of the exercise items benefits retrieval performance in the post-test (given that research has revealed that episodic memory of self-generated errors benefits later retrieval). RQ3 is related to RG3, aiming to see if implementing multiple-choice exercises in ways that reduce the error rate might reduce their side effects (if any).

Chapter 4 Methodology

This study adopted a mixed-methods, within-participant design. Twenty advanced ESL learners attended two sessions: (a) exercise + feedback and (b) two-week delayed post-test + stimulated recall. The target learning items were 20 verb-noun collocations. Participants were invited to 'think aloud' (i.e., verbalize their thoughts) while doing the exercise and post-test.

4.1 Participants

Twenty advanced ESL learners participated in this study. At the time of data collection, all participants were students enrolled in a TESOL program at a university in Canada. The TESOL program requires a minimum IELTS score 6.5 overall with no individual score less than 6.0 for admission. Of all participants, nineteen are L1 Chinese Mandarin speakers, one is a L1 Vietnamese speaker. The participants were informed that they would receive an Amazon gift card ($20) after each session. Their participation in this study was voluntary and they were assured that it would not have any impact on their grades in the TESOL program.

4.2 Procedure and Instruments

There was no pre-test in the study because pre-testing itself amounts to a guessing event (without feedback) with wrong responses potentially interfering later, making it hard to attribute findings solely to the multiple-choice exercise used in the experiment. Instead, participants' prior knowledge of the target collocations was assessed by the combination of learners' verbalization data (e.g., "I already know this one") and their correct exercise responses.

All participants read the Letter of Information (see Appendix A) and signed the Consent Form (see Appendix B) before the study. There were two sessions. The first was the exercise + feedback session. Each participant attended the first session to learn 20 English verb-noun collocations presented to them in a multiple-choice format. Corrective feedback was provided to learners right after the exercise. The second was the posttest + stimulated recall session. This second session followed two weeks after the learning

session. The participants were tested on the same 20 verb-noun collocations by means of a gap-fill format (supplying the missing verbs). Then, the same multiple-choice worksheet as in the former session was presented to the learners as a prompt for stimulated recall.

In both sessions, the participants met individually and in-person with the researcher, and they were asked to 'think-aloud' as they did the exercise and took the tests. Instructions for thinking aloud were kept minimal, so as not to change the way a participant would normally go about the exercise in a language class or as homework. The think-aloud data of the exercise session were meant to inform the researcher of which collocations a participant was already familiar with, and of which distractors the participant gave special attention to. As most participants and the researcher of this study are L1 Chinese speakers, learners were welcome to express their thoughts in Chinese if they preferred to. The verbalization data were recorded through a digital voice recorder.

In the post-test session, the instructions for thinking aloud were more specific. Per item in the gap-fill test, the participants were asked to express what (if anything) they could remember about the exercise they tackled the previous week. Next, per item in the multiple-choice test (which was identical to the multiple-choice exercise), they were asked not only if they could remember the correct response but also if they could remember what option they had selected the previous week and if it was right or wrong. This second post-test thus serves as a prompt for stimulated recall. The instruments of each section are introduced below.

4.2.1 Session 1: Multiple-Choice Exercise and Feedback

All 20 exercise items are of a multiple-choice format (i.e., selected-response format) with three options in each item. Per target collocation the participants were presented with a sentence where the verb of the verb-noun collocation was missing. The task was to choose the appropriate verb to complete the blank. The complete exercise worksheet and answer keys can be found in Appendix C. Figure 1 gives an example:

Figure 1

Example Item of Multiple-Choice Exercise

Choose the correct answer to complete the blank in each of the following sentences.

1. Alcohol addiction can ____ a heavy toll on your health and can shorten your life by many years.

A. have
B. take
C. pose

The exercise was delivered in a paper-and-pen format, allowing participants to jot down any notes on the worksheet during the exercise. Before taking the exercise, participants were told that they were expected to 'think-aloud' (see the script in Appendix E, for reference). Participants were encouraged to hazard a guess in the case of items they felt unsure about. There was no time limit, but most participants were able to finish all items in 20 minutes.

After finishing the exercise, all participants were given the answer keys verbally. Participants were asked to make corrections on their worksheets. After this, they handed in the worksheets. They were asked to come back for the second session two weeks later. This second session was introduced to them as another exercise session, even though the data served as post-test data for the purpose of the research study.

4.2.2 Session 2: Gap-Fill Delayed-Posttest and Stimulated-Recall

The post-tests took place two weeks after the exercise session. First, knowledge of the 20 collocations was assessed by means of a sentence-level gap-filling format again, but this time without providing response options to choose from (that is, the post-test used a constructed-response format). The sentences differed from those used in the exercise session, as did the order of presentation of the collocations. The post-test worksheet and answer keys can be seen in Appendix D.

Figure 2

Example Item of Gap-Fill Post-Test

Complete the blank of each of the following sentences with a verb that partners with the noun that follows.

1. The daycare is recruiting certified staff to _____ a watch on the kids after school.

Before taking the test, participants were told that they were expected to 'think-aloud' again (see the script in Appendix E, for reference). A constructed-response rather than a selected-response format was used first because it mimics real-life language use better. In real life, language users do not select from a set of presented options; instead, they retrieve appropriate language from their memory. Additionally, a constructed-response format allows participants to respond freely, whereas the selected-response format forces participants to select an available option. The constructed-response format thus leaves the possibility of participants supplying a response which corresponds to none of the options they met in the multiple-choice exercise the previous week, which would indicate that the distractors had *no* effect.

After completing the gap-fill test, the participants repeated the multiple-choice exercise they had tackled two weeks previously while saying aloud what they could remember about their own earlier choices and the feedback received. This elicited additional evidence of the extent to which the exercise items left reliable episodic memories.

The verbalization data of the exercise session revealed cases where (a) the target item was already known, and (b) the participant hesitated between certain response options. The verbalization data of the post-test sessions revealed whether the participant had episodic memories of the exercise, their own selected response, and the feedback received.

This study was approved by The Western University Non-Medical Research Ethics Board (NMREB). See Appendix F for the letter of approval.

4.3 Data Coding

The author revisited the recorded think-aloud data and the worksheets after both sessions and coded the data. Responses at the exercise stage (multiple-choice) were coded as (a) collocation already known/not yet known, (b) correct/incorrect response, and (c) verbalized hesitation/no hesitation regarding distractors.

Post-test responses in the gap-fill test were first coded as correct/incorrect. The correct test responses were coded as (a) claimed known and correct in the prior multiple-choice exercise, (b) claimed unknown or incorrect in the exercise but correct in posttest. This coding helped to exclude from the analysis any correct post-test responses on items that the participant was already familiar with.

The incorrect responses were coded as (c) same response as in the exercise item, i.e., error duplication, (d) different response than in the exercise item but corresponding to a distractor explicitly mentioned by the participant at the time, (e) different response than in the exercise but corresponding to a distractor **not** explicitly mentioned by the participant at the time, and (f) different response than in the exercise and corresponding to neither of the distractors of the exercise item (i.e., a new error).

Additionally, all the post-test responses were coded for presence/absence of evidence of episodic memory. This was operationalized by the participants' verbalizing their recollection of how they tackled the exercise item the previous week and the feedback received.

The stimulated-recall data collected while the participants were re-presented with the multiple-choice exercise items were coded as no memory/false memory/accurate memory of (a) their exercise response, (b) whether the response was correct or wrong, and (c) the feedback received.

Given the open-ended nature of think-aloud data, the author also took extra notes while revisiting the recorded files, in addition to the above coding. The extra notes were to qualitatively explore learners' experience during the exercise and recalls. The notes include learners' response strategies, evaluative comments about the exercises, etc.

4.4 Data Analysis

Jamovi (2022) was used for data entry and analysis. Mixed-effects modeling was used to find out what factors help to explain the variance in post-test performance (gap-fill) and stimulated recall (multiple-choice). The random effects are items (collocations) and participants. Fixed factors for the logistic regression modelling were chosen based on inspection of the descriptive statistics and preliminary chi-square tests (in this case, Fisher Exact Probability Tests) which suggested these factors were potentially associated with post-test and stimulated-recall performance. The fixed effects for **gap-fill accuracy** include (a) correct vs. incorrect exercise response, (b) hesitation vs. no hesitation about distractors at the exercise stage, (c) episodic memory vs. no episodic memory during gap-fill exercise, and (d) episodic memory of multiple-choice exercise at the stimulated-recall stage.

The models logically focus on collocations which were not yet known by participants before the experiment. Thus, the cases where participants said they already knew the collocation and then also got it right in the multiple-choice exercise were excluded from the analyses. To examine if the extent of prior knowledge influenced learning of the unknown collocations, the correlation coefficient was calculated for the association between the number of collocations that participants already knew and the proportion (%) of unknown collocations that were learned according to the gap-fill post-test.

Descriptive statistics were used to report the proportion of the post-test errors corresponding to the above error categories (e.g., what proportion are duplicates of exercise errors).

Chapter 5 Results

As mentioned in Data Analysis, the focus of this study is using multiple-choice exercises (henceforth, MC) in a trial-and-error fashion (i.e., to learn unknown collocations), and so the items that participants said they already knew and gave correct responses on in the MC exercise (n = 109) are excluded. This by no means indicates they were also correct in the gap-fill (henceforth, GF) post-test, which did not provide response options to choose from and is therefore a more challenging format—recall being harder than recognition. There were 26 instances where the participants failed to supply the correct verb of the collocation in the post-test GF, even though they reported that they knew this collocation and got it right in the exercise session. Of these instances, six GF responses corresponded to MC exercise distractors.

Of the total 400 cases (20 participants * 20 collocations), 291 were 'unknown' at the exercise stage. Table 1 indicates the GF response type in comparison with the MC response. Of the unknown items, 79 (27.15%) were correct in the GF, 210 were incorrect, and two blanks were left empty. Of the incorrect responses, 23.33% corresponded to one of the two distractors presented in the related MC exercise item.

5.1 Associations with Gap-Fill Post-Test Performance

Let us first focus on the 291 unknown items (of which 289 were answered and 2 were left blank in GF), and what predicts their recall in the GF test.

Choosing the correct option in the MC exercise two weeks previously is more strongly associated with successful recall than choosing a wrong option in the MC exercise, according to the Fisher Exact probability test ($p < .001$). As can be seen in Table 2, 41.38% (n = 48) of correct MC responses were followed by correct recall. Only 17.92% (n = 31) of wrong MC responses were followed by correct recall. So, it seems better to get it right from the start instead of relying on corrective feedback.

Table 1

Frequencies of GP Post-Test Response Types for Previously Unknown Items

GP Post-Test Response	*n*	% of Total
Correct	79	27.15%
Incorrect		
Same Distractor as in MC	25	8.59%
Another Distractor – Mentioned in MC	9	3.09%
Another Distractor – Not Mentioned in MC	15	5.15%
Non-Distractor Error	161	55.33%
No response supplied	2	0.69%
Total	291	100.00%

Table 2

Association of MC Exercise Accuracy with GF Accuracy

MC Exercise	GF		
	Incorrect	Correct	Total
Incorrect	142 (82.08%)	31 (17.92%)	173 (100.00%)
Correct	68 (58.62%)	48 (41.38%)	116 (100.00%)
Total	210 (72.66%)	79 (27.34%)	289 (100.00%)

Episodic memory of the MC item while doing the GF is associated with correct recall in the GF. As Table 3 indicates, when there was evidence of episodic memory, the success rate in the GF was 36.26%. This compares to only 23.23% when there was no evidence of episodic memory. The difference is significant: Fisher Exact probability test yields $p = .024$. As to error type, there was a slightly higher chance of supplying a distractor from the MC exercise when participants reported episodic memories of the MC items (see Table 4), but this difference is not significant ($p = .368$).

Table 3

Association of Episodic Memory of MC during GF with GF Accuracy

Memory during GF	GF		
	Incorrect	Correct	Total
Absence	152 (76.77%)	46 (23.23%)	198 (100.00%)
Presence	58 (63.74%)	33 (36.26%)	91 (100.00%)
Total	210 (72.66%)	79 (27.34%)	289 (100.00%)

Table 4

Association of Episodic Memory of MC during GF with GF Error Type

Memory during GF	GF Error Type		
	Distractor	Non-Distractor	Total
Absence	33 (21.71%)	119 (78.29%)	152 (100.00%)
Presence	16 (27.59%)	42 (72.41%)	58 (100.00%)
Total	49 (23.33%)	161 (76.67%)	210 (100.00%)

Do **hesitations between response options during the MC exercise** help to predict GF performance? As can be seen in Table 5, of 109 such documented hesitations regarding unknown items, 26.61% were followed by correct recall in the GF. This is very similar to the overall accuracy rate in the GF test: 27.34%. Are the wrong verbs supplied in the gap-fill often duplicates of the distractors after hesitations? As can be seen in Table 6, this was the case for 23.75% of the hesitation episodes, which is not different from the overall proportion of duplicates (23.33%).

Table 5

Association of Reported Hesitations during MC with GF Accuracy

Hesitation during MC	GF		
	Incorrect	Correct	Total
Not Reported	130 (72.22%)	50 (27.78%)	180 (100.00%)
Reported	80 (73.39%)	29 (26.61%)	109 (100.00%)
Total	210 (72.66%)	79 (27.34%)	289 (100.00%)

Table 6

Association of Reported Hesitations during MC with GF Error Type

Hesitation during MC	GF Error Type		
	Distractor	Non-Distractor	Total
Not Reported	30 (20.08%)	100 (76.92%)	130 (100.00%)
Reported	19 (23.75%)	61 (76.25%)	80 (100.00%)
Total	49 (23.33%)	161 (76.67%)	210 (100.00%)

The above associations concern the participants' verbal reports as they tackled the MC exercise in the exercise session and subsequently did the GF post-test two weeks later. Let us now turn to associations with their recollections of the MC exercise items when they were subsequently shown the MC exercise again as a prompt for stimulated recall. There are three components of **stimulated-recall data**: participants' memory of (a) their initial MC response, (b) whether that response was correct (henceforth MC result), and (c) feedback received (i.e., what response options was correct). As shown in Table 7, accurate memories of the MC feedback are more strongly associated with successful gap-fill recall (43.18%) than false or no memories of the feedback (20.40%), which is not surprising. Comparing correct episodic memories of feedback to false memories/absence of episodic memories by the Fisher Exact probability test confirms that correct episodic memories are more closely associated with correct recall: $p < .001$.

Accurate memory of MC response appears associated with a higher chance of successful GF recall (32.59%) than no/false memory (22.73%), and accurate memory of MC result seems to improve the chances of successful GF recall (34.00%) than no/false memory (23.81%). However, these differences fall short of significance ($p = .065$; $p = .072$, respectively).

The results of the above preliminary analyses are confirmed by the mixed-effects logistic regression model (Table 8): significant predictors of GF post-test accuracy include MC exercise accuracy (χ^2 (1) = 8.907, $p = .003$), verbalized memory of MC during GF (χ^2 (1) = 6.411, $p = .011$), and recall of MC Feedback during stimulated recall (χ^2 (1) = 7.523, $p = .006$). The total amount of variance explained by the model was 35.0% (conditional $R^2 = .350$), with 15.8% of the variance (marginal $R^2 = .158$) attributed to the fixed effects and leaving 19.2% of the variance associated with participants and items (collocations).

Table 7

Associations of Stimulated-Recall Responses with GF Accuracy

Stimulated-Recall	GF		
	Incorrect	Correct	Total
	MC Response		
No/False Memory	119 (77.27%)	35 (22.73%)	154 (100.00%)
Accurate Memory	91 (67.41%)	44 (32.59%)	135 (100.00%)
	MC Result		
No/False Memory	144 (76.19%)	45 (23.81%)	189 (100.00%)
Accurate Memory	66 (66.00%)	34 (34.00%)	100 (100.00%)
	MC Feedback		
No/False Memory	160 (79.60 %)	41 (20.40%)	201 (100.00%)
Accurate Memory	50 (56.82%)	38 (43.18%)	88 (100.00%)

Table 8

Mixed-Effects Logistic Regression Model for Post-Test GF (n = 289) Accuracy

Random Effects	SD	Variance
Participants	0.532	0.283
Items	0.828	0.686

Fixed Effects	X^2	df	p
MC Exercise Accuracy	8.907	1.000	0.003*
Memory of MC during GF	6.411	1.000	0.011*
Hesitation during MC	0.449	1.000	0.503
Episodic Memory of MC			
Recall of MC Response	0.176	1.000	0.674
Recall of MC Result	0.087	1.000	0.768
Recall of MC Feedback	7.523	1.000	0.006*

(continued)

Fixed Effects Parameter	Effect	Estimate	SE	OR	95% OR CI		z	p
					Lower	Upper		
(Intercept)	(Intercept)	-0.689	0.281	0.502	0.290	0.870	-2.455	0.014*
MC Exercise Accuracy	Correct - Incorrect	1.039	0.348	2.827	1.429	5.595	2.984	0.003*
Memory of MC during GF	Presence - Absence	0.980	0.387	2.666	1.248	5.694	2.532	0.011*
Hesitation during MC	Reported – Not Reported	-0.230	0.343	0.794	0.405	1.557	-0.670	0.503
Episodic Memory of MC								
Recall of MC Result	No/False - Accurate	0.126	0.427	1.134	0.491	2.621	0.295	0.768
Recall of MC Response	No/False - Accurate	-0.148	0.352	0.862	0.432	1.720	-0.420	0.674
Recall of MC Feedback	No/False - Accurate	-1.105	0.403	0.331	0.150	0.730	-2.743	0.006*

Note. Number of Obs: 289. OR = Odd Ratios. Asterisk (*) indicates statistical significance ($p < .05$).

The above analysis concerns all the post-test GF responses (for items that were not yet known by participants). This includes the many incorrect responses which did not correspond to the distractors in the MC exercise items. Let us now focus exclusively on post-test GF responses which correspond to MC options (either the correct option or the distractors in the MC items). There are 128 cases where learners produced one or another MC verb in the GF post-test, of which 61.72% ($n = 79$) were correct and 38.28% ($n = 49$) corresponded to a distractor.

Correct MC choices were more likely to be followed by correct GF responses than wrong MC choices, according to the Fisher Exact probability test ($p < .001$). 80.00% ($n = 48$) of correct MC responses were followed by correct recall. Only 45.59% ($n = 31$) of wrong MC responses were followed by correct recall. So, it is better to get it right from the start instead of relying on corrective feedback.

Do **episodic memories during the GF of the MC exercise** help to predict GF performance? While 67.35% of responses were correct when learners mentioned the MC item during GF, 58.23% of responses were correct when learners did not. The difference does **not** reach statistical significance according to the Fisher Exact Probability Test ($p = .30$). This appears not to be consistent with the prior analysis including all 289 GF responses, where episodic memory of MC items was significantly and positively associated with GF accuracy. However, the sample size is reduced in the present analysis, and this makes it less likely to reach statistical significance.

Do **hesitations between response options during the MC exercise** help to predict GF performance? There is **no** significant association ($p = .81$): 60.42% of responses are correct after hesitations; 62.5% of responses are correct after no reported hesitations.

Correct memories of the feedback were more strongly associated with successful recall (48.10% of correct GFs) than false memories (11.39%) or no memories (40.51%). The difference between accurate memory and no memory is significant: $p = .002$. The difference between accurate memory and false memory is significant: $p < .001$. However, the difference between no memory and false memory falls short of significance ($p = .32$).

This is probably because there are not enough cases of the former: 9 false memories associated with correct GFs and 13 false memories associated with wrong GFs.

The above preliminary analyses (focusing on the 128 GF responses where learners produced verbs corresponding to previously seen MC options) are confirmed by the mixed-effects logistic regression model (Table 9): significant predictors of GF post-test accuracy include MC exercise accuracy (χ^2 (1) = 12.009, $p < .001$) and recall of MC feedback (χ^2 (1) = 9.703, $p = .002$). The total amount of variance explained by the model was 41.0% (conditional $R^2 = .410$), with 30.2% of the variance (marginal $R^2 = .302$) attributed to the fixed effects and leaving 10.8% of the variance associated with participants and items (collocations).

Table 9

Mixed-Effects Logistic Regression Model for Post-Test GF (n = 128) Accuracy

Random Effects	SD	Variance
Participants	0.3801	0.1445
Items	0.6784	0.4602

Fixed Effects	X^2	df	p
MC Exercise Accuracy	12.009	1	< .001*
Hesitation during MC	0.659	1	0.417
Memory of MC during GF	1.029	1	0.310
Recall of MC Feedback	9.703	1	0.002*

(continued)

Fixed Effects Parameter	Effect	Estimate	SE	exp(B)	95% OR CI		z	p
					Lower	Upper		
(Intercept)	(Intercept)	1.089	0.360	2.971	1.467	6.017	3.024	0.002*
MC Exercise Accuracy	Correct - Incorrect	1.711	0.494	5.533	2.103	14.560	3.466	< .001*
Hesitation during MC	Reported – Not Reported	-0.391	0.482	0.676	0.263	1.739	-0.812	0.417
Memory of MC during GF	Presence - Absence	0.511	0.503	1.666	0.622	4.467	1.015	0.310
Episodic Memory of MC								
Recall of MC Feedback	No/False - Accurate	-1.754	0.563	0.173	0.057	0.522	-3.115	0.002*

Note. Number of Obs: 128. OR = Odd Ratios. Asterisk (*) indicates statistical significance (p < .05).

5.2 Multiple-Choice Stimulated-Recall Performance

Above, we reported on the 291 unknown items and factors that influence their recall in the GF test. We now turn more specifically to the second presentation of the MC exercise to the participants, which served as a prompt for stimulated recall but which at the same time provides an additional means to evaluate how much the participants learned from doing the MC exercises and receiving feedback two weeks previously.

Table 10 presents the data regarding learners' verbalized recollections of their MC performance. There were 88 instances where learners accurately recalled the feedback they received on their MC answers, but there were also 48 instances where learners were mistaken about the feedback they received, and so they considered one of the distractors to be the correct answer. In another 155 instances, they did not remember the feedback they received and so they still felt unsure which of the three response options was the correct one. This suggests that of the 291 previously unknown items, 29. Put differently, we may feel confident about new collocations having been learned through the MC exercise plus feedback for only 30% of the items, at least when it comes to being able to choose the correct verb from three options (that is, knowledge at the level of recognition, which does not imply the ability to produce the collocations accurately in discourse).

Interestingly, while the participants accurately recalled 46.74% of their own MC responses, they recalled whether their response had turned out to be correct or wrong only 34.46% of the time. Their recollection of the feedback (which informed them which response option was correct) was even poorer: 30.24%. This suggests that, even if learners can recall their original MC exercise choice, this by no means guarantees that they will also remember if their choice was correct or wrong, let alone which option turned out to be the correct one.

We already reported above that, when participants verbalized their recollections of the MC exercise, some of those recollections were inaccurate. For example, of the 136 instances where learners stated they remembered the feedback they received on a MC exercise item, 20.34% (n = 48) of these recollections were wrong. Altogether, there were

479 instances where learners mentioned things that they remembered regarding specific items on the MC exercise, but 32.36% (n = 155) of these memories were unreliable. In other words, doing the MC exercise created episodic memories regarding a fair number of exercise items, but these memories were often false.

Table 10

Frequencies of Stimulated-Recall Performance

Stimulated-Recall Performance	Counts	% of Total 291 Unknown Items
Recall of MC Response		
Accurate Memory	136	46.74%
False Memory	58	19.93%
No Memory	97	33.33%
Recall of MC Result		
Accurate Memory	100	34.46%
False Memory	49	16.84%
No Memory	142	48.80%
Recall of MC Feedback		
Accurate Memory	88	30.24%
False Memory	48	16.49%
No Memory	155	53.26%

5.3 Exploring Additional Factors

It seems plausible that the amount of attention given to feedback is positively related to the amount of thought a learner invests in a specific MC exercise item. When learners reported while thinking aloud that they were hesitating between two response options in an exercise item, then this could be taken as a manifestation of some cognitive effort. At the very least, it may be a sign that learners were becoming curious to find out which was the correct option. For these reasons, one would expect these learners to attend to the feedback and thus retain better episodic memory of the feedback received on the MC item. Table 11 seems to support this possibility. It is notable, however, that even in this scenario, episodic memory of the feedback does not at all guarantee that the memory is in fact accurate: In 14 out of the 55 instances where participants said they could recall the feedback they had received, they misremembered what was the right response option.

It is possible that the learners did not benefit much from the feedback because it concerned a large number of items which they were not yet familiar with. Of the 20 verb-noun collocations, roughly 15 on average were new to these participants. That said, some participants had better prior knowledge than others, which permits calculating the correlation between the number of items already known and the proportion (in %) of the unknown items that were learned according to the GF post-test. This yields a positive, but non-significant correlation between the extent of prior knowledge of the set of target collocations and these relative learning gains: $r_s = .245$ ($t = 1.07$; $p = .299$). This supports the idea that learners who face the challenge of taking in a lot of new information through feedback benefit less from it than learners for whom the feedback concerns a comparatively small number of items. The correlation is far from strong, however, suggesting that the role of prior knowledge should not be overestimated when it comes to the type of MC exercise examined here.

Table 11

Association of Hesitations during MC with Accurate Recall of Feedback

Hesitation during MC	Recall of MC Feedback			
	Accurate Memory	False Memory	No Memory	Total
Not Reported	47 (25.97%)	34 (19.54%)	110 (60.77%)	181 (100.00%)
Reported	41 (37.27%)	14 (11.97%)	45 (40.91%)	110 (100.00%)
Total	88 (30.24%)	48 (16.49%)	155 (53.26%)	291 (100.00%)

5.4 Qualitative Exploration of Learners' Experience

The above sections quantitatively reported the associations of GF post-test and MC stimulated-recall performance. This section will qualitatively further report the effects of the two random factors: participants and items, by exploring learners' experience while doing the exercise and recall, evidenced by the 'think-aloud' and 'stimulated-recall' data. Despite of small sample size (20 participants and 20 collocations), we found some individual differences and common tendencies of the participants and items that are worth reporting.

For starters, there was variation among participants in the extent to which they engaged with the MC exercise items and how they reacted to the feedback received on items. For example, some participants (e.g., #0306001, #0306004) expressed surprise and even disbelief when they were told what the correct response was, and they reported in session two that they had looked up the collocation in a dictionary after the exercise session to make sure. Some participants (e.g., #0307002) stated they were less interested in the feedback when they had made a blind guess, while others (e.g., #0307001) claimed to attend to the feedback for all unknown collocations.

The participants' guessing strategies when they tackled the MC exercise and/or the GF test are worth reporting as well. According to the think-aloud data, participants adopted certain strategies when they were not sure about the item they encountered.

One common strategy was to draw on L1 knowledge, where participants thought of the L1 translation of the English collocations. There were four participants (#0307001, #0308004, #0309002, and #0309004) whose think-aloud data indicate that they translated the collocation into their L1 (i.e., Chinese), referring to the context, and then supplied the English verb that is the closest equivalent to the L1 verb. This strategy sometimes worked, but sometimes did not. For example, using L1 translation strategy led to correct responses for the item *break the silence* [打破沉默]; however, the same strategy led to wrong responses for the item **do a prayer / *make a prayer (say a prayer)* [做祷告].

Another strategy was to use analogies with collocations that participants already knew, and which bear a semantic resemblance to the target collocations. For example, there were four participants (#0308001, #0308004, #0309003, #0309004) who supplied *keep a watch on […]* correctly, and who referred to *keep your eyes on […]* to explain their choice of verb. Inevitably, this strategy sometimes failed, too. For example, participant #0307003 supplied **give tribute* (instead of *pay tribute*) referring to *give respect*. Sometimes the semantic relation between the known collocation and the target was not so obvious. For instance, participant #0309001 supplied *pay tribute to* by drawing an analogy with *pay attention to.*

Some participants made guesses in the MC exercise that were led by semantic or experiential associations between the noun and one of the response options. For example, two participants (e.g., #0307003 and #0308002) opted for **run a sweat (*instead of *break a sweat)* because '*run* is the most dynamic word among the options' (and dynamic activities such as running make you sweat).

One more strategy discerned in the data is a simple test-taking strategy, where participants would supply the same high-frequency verb in all the blanks in the GF for items they could not remember. Two participants (#0306001, #0309002) stated that they would supply *take* for all the items where they did /not have a better alternative; one

participant (#0307005) preferred the verb *cause*, and one participant (#0309004) preferred *give* for any collocations they could not remember. Unsurprisingly, this 'using one word for all' strategy sometimes worked (because the verb in many verb-noun collocations is indeed a highly frequent lemma and so there is a statistical chance that such a highly frequent verb fits a new collocation too) but was inevitably also unsuccessful in many cases. Participant #0306001 supplied the word *take* in five GF post-test items regarding previously unknown collocations. Two *take*s turned out lucky guesses, but the other three *take*s turned out failed guesses. It is worth noting that this strategy included the high-frequency verbs which the participants had seen in the MC exercise. So, when a learner produced a GF post-test response that corresponded to an MC exercise distractor, this was not necessarily evidence of interference from the exercise item, but rather a result of a broader guessing and test-taking strategy.

Previous quantitative analysis confirmed that items (i.e., collocations), as well as participants, is a random predictor of later recall performance. As mentioned above, one and the same strategy can lead to a correct response to one item but a wrong response to another item. For example, 'L1 translation' strategy was effective for the collocation *break the silence* but not for *say a prayer.* According to the think-aloud data, some items were 'easier' than others, while some items appear particularly 'confusing'. For example, participant #0307005 stated in the stimulated recall session:

> *"哦对……是 speak！我发现有些就特别难记，特别容易混淆。像这个 speak my mind，当时（两周前做练习时）我就特别纠结，因为每一个（选项）都像是一个意思，当时好像还有 say、tell、还是 talk？就算告诉了我答案我也不会记得，感觉每一个都是对的。但是像 pay tribute 这种，一旦你记住了就不会搞错了，因为每个选项都很不同。" ["Oh right…It's 'speak'! I found that some are more difficult to memorize and easier to mix up. Like 'speak my mind', I hesitated that time (two weeks ago, when doing the MC exercise) because every option seemed like the same meaning. I remember there was 'say', 'tell', and 'talk'? Even though you told me what the correct answer was, I find it*

> *hard to remember. I felt like every option could be correct. But with something like 'pay tribute', once you remember you won't get confused because every option was distinctive.]*

This example illustrates that the extent of interference from distractors that a learner could experience varies from item to item. The difficulty level of the exercise items could affect learner's performance on the post-test and stimulated recall. According to the think-aloud data, many learners (n = 12) reported hesitation in the MC exercise when tackling the item for the collocation ***speak*** *my mind,* where the response options were all semantically similar. Of the 20 GF post-test responses for *speak my mind,* only 4 responses were correct, 12 were non-distractor wrong responses, and 4 were MC distractors. In contrast, only one of the 20 learners reported hesitation for the MC item concerning ***take*** *chances,* where the response options were semantically distinct. Of 19 GF post-test responses (one was missing) for *take chances*, 12 were correct and 7 were non-distractor wrong responses. In this scenario, the multiple-choice item for *speak my mind* is harder than that for *take chances* because the response options in the former were near synonyms (e.g., speak, tell, say).

Chapter 6 Discussion

In this chapter, we will discuss the results that are reported in the previous chapter. We will discuss the general findings, followed by the findings concerning the three RG/RQs.

The participants' GF post-test responses were compared with their MC exercise responses. In this study, there were 26 cases where learners gave an erroneous response in the post-test, although they claimed to know the collocation and gave a correct response in the MC exercise. Of the 291 previously unknown collocations, there were 49 cases where learners duplicated their exercise error or supplied another exercise distractor in the post-test. These findings are consistent with prior studies of Boers and colleagues (i.e., Boers et al., 2014; Boers et al., 2017: Stengers & Boers, 2015) where learners sometimes reproduced or produced exercise distractors even though they had received post-exercise corrective feedback.

As for the stimulated recall where participants were shown the MC exercise again and asked to recall their response and the feedback received two weeks previously, 32.36% of students' recollections were false. This finding is consistent with the conflicts between recollection and familiarity in the False Fame Paradigm: learners might judge something mistakenly as correct because it has become 'familiar' to them.

A strong take-away message of this study is that MC exercises on verb-noun collocations, when implemented in a trial-and-error fashion, bring about very little learning gain. This is owing to two complementary reasons: either the exercise items leave no durable memories, or they risk leaving memories that are unreliable. This leads to the first research question we asked: Is the poor effectiveness of such MC exercises owing to false memories or simply to the absence of memories?

6.1 Memory and Coincidence

A question we asked in relation to the previous research findings (e.g., Boers et al., 2014) was whether the re-emergence of distractors in a post-test was due to interference from the exercise, or simply a coincidence because students may have no recollection of the exercise. The answer to this question according to the data in this study is: **both**. In some

cases, the re-emergence of distractors can be attributed to interference (false memories) and in other cases it is just because the students have already forgotten the exercise (no memories).

Interestingly, the risk of interference is the strongest when students choose one of the distractors in the MC exercise. This is because they do not always remember whether their choice turned out to be right or wrong. If so, the question then becomes how we can guide students to choose correctly when they do the MC exercise—if (and it is a big IF) teachers and materials designers really insist on using such selected-response exercise format. We will discuss some of the predictors of successful recall further in section 6.3.

6.2 Limited Help of Episodic Memory on Later Recall

Another question we asked, in relation to the previous research findings (e.g., Metcalfe & Huelser, 2020), was whether episodic memory of the exercise items reduces the risk of interference. We collected evidence of episodic memory of the MC exercise during the GF post-test and in the stimulated recall phase where we showed the participants the MC exercises again. We were particularly interested in the participants' recollections of (a) their initial MC exercise response, (b) the result (i.e., right or wrong), and (c) the feedback received (i.e., which response option turned out to be the correct one).

Unsurprisingly, learners who had accurate episodic memories of the feedback received on their MC exercise responses stood a very good chance of supplying the correct response in the post-test. This finding is consistent with the theory of *Recursive Reminding* (as reviewed in section 2.3): remembering a learning episode benefits later recall of what was learned. The descriptive statistics suggested that learners who recalled their own initial responses on the MC exercise items also produced more correct responses in the post-test. This finding is consistent with Metcalfe and Huelser (2020), who found that participants who can recollect their initial error tend to outperform those who cannot in a post-test. In the present study, however, episodic memory of erroneous exercise responses was not a statistically significant predictor of post-test success. This finding answers the question we raised from RG2 in section 2.4: remembering the mistakes one made in an exercise is not as beneficial when it comes to material-designed errors in

selected-response exercises. Remembering that the chosen response option in a MC exercise was wrong does by no means guarantee accurate recall of the response option that was right.

The general answer to RQ2, 'Does episodic memory of how they tackled specific multiple-choice exercise items help learners to recall the correct response?', according to the data, is that the episodic memories resulting from the MC exercise session were not always reliable, and therefore not always helpful. It stands to reason that *reliable* memories of feedback are beneficial—and this is indeed what the analyses show. The problem is that, even though learners may distinctly remember their choice of response option in the exercise, they appear to find it harder to recall (a) if their choice was correct and (b) what was revealed to be the correct option. Although episodic memory, to some extent, helped learners to recall the correct response in the post-test, it was not significantly associated with the error type (distractor vs. non distractor, as can be seen in Table 4). This is possibly because episodic memory is sometimes not reliable, according to the MC stimulated-recall data.

In sum, the beneficial function of episodic memory to reduce the risk of interfere is limited. Accurate memory of the exercise, of course, helps later recall. However, memory is not always reliable. This finding suggests that episodic memory may not be as beneficial for learning from one's mistakes in the case of material-designed errors in selected-response exercises as it is in the case of learners' self-generated errors.

6.3 The Predictors of Successful MC Exercise

The third question that we asked regarded the predictors of successful later recall, when using MC as an initial exercise for learning new collocations.

Does a relatively high success rate at the exercise stage help learners to benefit from (corrective) feedback? The answer to this question is: yes. Gauging the success of MC exercise by both GF post-test accuracy and MC stimulated recall of feedback, both logistic regression models reported in section 5.1 (Tables 8 & 9) indicated that choosing the correct answer at the exercise stage is a significant predictor of correct later recall.

This finding accords with previous research (Marsh et al., 2009) which found that the students who were higher-achieving in the initial MC test showed less negative suggestion effect and a larger positive testing effect in a later cued-test. This finding confirms the assumption that we made in section 2.4, notably that when learners make relatively few errors at the exercise stage, they find it easier to process the corrective feedback and replace incorrect responses by the correct ones in their memory.

In addition to correct exercise response, episodic memory is another significant predictor of later successful recall. However, as discussed in section 6.2, because memory is not always reliable—a problem that is possibly exacerbated by the nature of material-designed errors in selected-response exercises—the benefits of episodic memory should not be overestimated.

The qualitative data allowed us to explore learners' experience during the sessions. In addition to the above predictors of post-test success (i.e., correct exercise response and episodic memories), differences among participants and among exercise items clearly influenced performance on the GF post-test and MC stimulated recall. For example, some participants engaged more with certain exercise items (and the feedback) than others, the corrective feedback on some items was experienced as more surprising than for other items (which in turn influenced the likelihood of episodic memories), and MC items where the response options bear strong semantic resemblance (i.e., *say, speak, talk, tell*) appear particularly likely to cause competition in memory between such options. The latter finding suggests that the Semantic Mediation Hypothesis, which holds that recall of a certain target item is facilitated by one's recall of a semantically associated item, does not apply very well here: semantic relatedness is not the locus of the 'learning from one's error' benefit in the case of multiple-choice exercises on collocations.

Chapter 7 Conclusions and Implications

The above chapter discussed the findings responding to the research gaps and questions of this study. In this chapter, we will conclude the findings and draw some pedagogical implications based on the findings.

It needs to be acknowledged that the title of this research oversimplifies matters. The proper question is probably not whether distractors hinder or benefit later recall, but rather under what conditions they either hinder or benefit later recall. It stands to reason, for example, that if a learner has a reliable (episodic) memory of tackling a multiple-choice item AND the feedback received on it, then the distractors in that exercise item should not hinder recall of the correct response. To the contrary, a distractor that turned out to be wrong may serve as mediator to re-collect the correct response ("I remember X was wrong, so I should avoid X and use Y instead"). In that case, recalling the distractor flagged as "don't say this" in memory may perhaps slow down the learner's retrieval of the correct language item, but the latter will be achieved. The question remains, however, how easy it is for learners to separate wrong options from correct ones in memory when it comes to learning verb-noun collocations which are experienced as arbitrary (leading to "I remember that one of these was right and the other was wrong, but I'm not sure anymore which was the wrong one"), and where several response options look plausible (as in the case of near-synonyms).

A general finding of this study is that verb-noun collocations MC exercises bring about limited learning gains when they are implemented in a trial-and-error fashion (that is, when they are used to introduce collocations that are new to the learner). Two scenarios help to account for this limited effectiveness. In one scenario, the exercises are experienced as too shallow to create robust memories, and so learners quickly forget the exercise items. If in a later test (or follow-up exercise) provide a wrong response that corresponds to one of the response options in the MC exercise, then this is just a coincidence (or the outcome of the same guessing strategy as before). In the other scenario, the exercise items do foster episodic memories, but, owing to the nature of the selected-response exercise, the learner retains different response options and forgets which one is correct. In this scenario, when learners later provide a wrong response that

corresponds to a distractor seen in the MC exercise, this can be attributed to interference from this distractor. The data collected for this study furnish evidence for both scenarios. It cannot be denied that episodic memory does to some extent help to recall the correct response: The learners who mentioned the relevant exercise items while they were taking the post-test and who reported they remembered the MC exercise items in the stimulated-recall activity tended to outperform those who appeared to have forgotten all about the exercise items. However, these memories were not always precise enough. Learners sometimes remembered their original response in the exercise, but then failed to recall if it was correct or not. Sometimes learners did remember clearly that their initial response was wrong, but then failed to remember which of the remaining two options turned out to be correct. In the worst case, learners had inaccurate memories of the feedback and were at the time of the stimulated recall convinced a certain response option was correct while it was not.

Differences between individuals and items also play a role. Learners adopted various strategies to the exercise and test items. A single strategy could lead to a correct response on one item but a wrong response on another item. The qualitative exploration of think-aloud data indicates that the MC items where options bear semantic resemblance (e.g., *speak, say, tell*) are more 'difficult', compared with the items where options are more distinct (e.g., *give, hold, pay)*.

There are several implications that can be drawn from the above findings. In the domain of language testing, MC and other selected-response formats may be an efficient choice, which is probably why such formats are commonly used in many high-stakes language tests tried by a large number of test takers. However, this study indicates that selected-response formats such as MC can be problematic as a learning tool. We suggest that language teachers and material designers exercise caution when using selected-response formats as a tool to introduce new language items to learners, especially in the case of items such as collocations that are prone to confusion.

Although using selected-response exercises for learning new items is not advisable, there are some suggestions to enhance the effectiveness of such exercises if practitioners

nonetheless wish to use them. First, given that correct exercise responses predict learners' later recollection of the target item, it is important to help learners to get the correct answer from the start. This can be done by choosing distractors that are not as plausible as the distractors one would create for the purpose of a MC *test*. Second, teachers may need to take additional steps to ensure their students process the feedback they receive on such exercises. Simply informing the students of the correct response option may not suffice. Occasionally, the teacher may be able to point something out about the correct collocation that makes it more memorable. For example, it is not unusual for collocations to be made up for words that alliterate (in the present study, for example, *run* + *risk* and *take* + *toll*). Minimally, teachers could ask students to repeat the whole correct collocation orally. Ideally, additional retrieval practice should follow the initial exercise for learners to review the collocations and solidify the newly acquired knowledge. Relying on just one exercise (perhaps especially a MC exercise) for robust learning to happen is probably naïve.

It was already mentioned previously that the designers of selected-response exercises on collocations are advised to create distractors that are semantically and formally sufficiently distinct from the correct response. If not for testing, 'distractors' should not be too distracting. That said, it is also conceivable that learners remember little of such decontextualized exercises after two weeks or so. If so, one would intuitively expect very little interference from the distractors in such exercises. But still, as discussed in the Background section in relation to the False Fame Paradigm (Jacoby et al., 1989), just being presented with verb-noun combinations which might otherwise not even have occurred to the learners in an exercise can render them vaguely familiar and by that token seemingly acceptable, even if the learners have no conscious recollection of when and where they encountered the combination. Potential interference remains, and thus teachers and learners are still advised to exercise caution when it comes to this type of language practice.

Chapter 8 Limitations and Suggestions for Future Research

There are inevitably important limitations to this study that need to be acknowledged. One is the relatively small samples of target collocations (n = 20) and participants (n = 20). This is because the data was collected in one-on-one sessions and because of the time-consuming nature of the data coding and analysis. Future research may consider involving larger sample sizes if conditions permit.

Another limitation is that think-aloud data and stimulated-recall data tap into the participants' processing at the level of awareness (or consciousness) but may reveal little about processing below that level. Attention may be unconscious, too. Another potential problem of think-aloud and stimulated-recall data regards individual differences. The differences include personality: while some participants are more talkative or outgoing, some may be more reserved and may not say much about what they are thinking. This is especially unpredictable at the exercise stage where instructions of 'think-aloud' were not specific (to avoid interfering with the learners' habitual processing). It is worth mentioning that personality differences also include (self-)confidence. For example, some participants were more confident than others that they "knew" certain collocations and more confident than others about what they could remember about the MC exercise. Below is an example illustrating how a learner's personality influenced her stimulated-recall data. This participant (#0307003) reported very few episodic memories of the MC exercise, despite her high GF post-test accuracy (75%). When asked to confirm that she had no memory of the MC feedback even though she had supplied the correct verb in the GF, she said:

> *"是这样，我是一个比较缺乏自信……或者说比较保守的人。除非我100%确定我记得，否则我不会说我记得的，就算是99%确定，我可能也会说我不记得。" ["Here is the case, I am a person who is short of self-confidence…or, conservative. I would not say I remember unless I am 100% sure. Even though I am 99% sure, I might not say I remember."]*

This example illustrates that personality differences can influence personality can influence think-aloud and stimulated-recall performance. To make up for these

limitations of think-aloud and stimulated-recall data, future research could adopt eye-tracking as an additional method. Eye-tracking data could help, for example, to investigate if one distractor in a multiple-choice item attracts more attention than another distractor does. Eye-tracking data could also serve as a prompt for think aloud and stimulated recall.

A downside of think-aloud procedures is also that they may alter the way participants would habitually go about an exercise. In this study, for example, asking participants to think aloud as they tackled the exercise may promote episodic memory. This is why the instructions for thinking aloud were left relatively unspecific at the exercise stage, while still—hopefully—being able to capture participants' prior knowledge of certain collocations and their hesitations between response options.

Perhaps the greatest limitation is that this study examined only one type of exercise on verb-noun collocations. In the absence of a comparison with other types (e.g., constructed-response tasks), it is impossible to judge the relative usefulness of multiple-choice exercises (or other selected-response tasks) for collocation learning. What the findings can indicate is what factors matter for such exercises to "work", but future research will need to evaluate just how well they work (e.g., the overall learning gains, the effect of episodic memory, and the chances of error duplicates) compared to other types of language practice. Future research can also compare different sub-formats of selected-response exercise. For example, in the present study the students were presented with three response options per exercise item, but it may be worth investigating if reducing this to two options might be a way of improving the effectiveness of the exercise.

Last, but not least, this study only examined the efficacy of selected-response exercises for collocation learning. Future research could investigate the comparative effectiveness of such exercise formats for item learning (e.g., knowledge that is associated with vocabulary, including words and lexical phrases such as collocations) and pattern (or 'rule') learning (e.g., knowledge associated with grammar such as tense-aspect patterns). The distinction between item and pattern learning may matter for the effectiveness of

exercise types. For example, learners might feel it easier to dismiss a distractor in a selected-response grammar exercise once they have discovered a systematic ‘rule’ that can be extended to new instances. In that case, each new exercise item provides an opportunity to put acquired knowledge into practice. This is different from collocations learning, where students can rely much less on generalizable ‘rules’ (apart from applying a few analogy-based heuristics such as *commit* + nouns denoting crimes, *conduct* + nouns denoting research, etc.).

References

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia*, *32*(1), 53–68. https://doi.org/10.1016/0028-3932(94)90068-X

Boers, F. (2021). *Evaluating second language vocabulary and grammar instruction: A synthesis of the research on teaching words, phrases, and patterns* (first). Routledge.

Boers, F., Dang, T. C. T., & Strong, B. (2017). Comparing the effectiveness of phrase-focused exercises: A partial replication of Boers, Demecheleer, Coxhead, and Webb (2014). *Language Teaching Research*, *21*(3), 362–380. https://doi.org/10.1177/1362168816651464

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb–noun collocations. *Language Teaching Research*, *18*(1), 54–74. https://doi.org/10.1177/1362168813505389

Brown, A. S., Schilling, H. E. H., & Hockensmith, M. L. (1999). The negative suggestion effect: Pondering incorrect alternatives may be hazardous to your knowledge. *Journal of Educational Psychology*, *91*(4), 756–764.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

Downing, S. M. (2009). Written Tests: Constructed-response and selected-response formats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in Health Professions Education* (pp. 149–184). Routledge. https://doi.org/10.4324/9780203880135-14

Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition*, *38*(4), 407–418. https://doi.org/10.3758/MC.38.4.407

Jacoby, L. L., Kelley, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology*, *56*(3), 326–338. https://doi.org/10.1037/0022-3514.56.3.326

Jacoby, L. L., & Wahlheim, C. N. (2013). On the importance of looking back: The role of recursive remindings in recency judgments and cued recall. *Memory & Cognition*, *41*(5), 625–637. https://doi.org/10.3758/s13421-013-0298-5

Marsh, E. J., Agarwal, P. K., & Roediger, H. L. (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applied*, *15*(1), 1–11. https://doi.org/10.1037/a0014721

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*(1), 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Metcalfe, J., & Huelser, B. J. (2020). Learning from errors is attributable to episodic recollection rather than semantic mediation. *Neuropsychologia*, *138*, 107296. https://doi.org/10.1016/j.neuropsychologia.2019.107296

Remmers, H. H., & Remmers, E. M. (1926). The negative suggestion effect of true-false examination questions. *Journal of Educational Psychology*, *17*(1), 52–56. https://doi.org/10.1037/h0070067

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

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Stengers, H., & Boers, F. (2015). Exercises on collocations: A comparison of trial-and-error and exemplar-guided procedures. *Journal of Spanish Language Teaching*, *2*(2), 152–164. https://doi.org/10.1080/23247797.2015.1104030

The Jamovi project (2022). Jamovi (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org

Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *The Journal of Educational Research*, *86*(6), 357–362. https://doi.org/10.1080/00220671.1993.9941229

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory* (pp. 381–403). Academic Press.

Tulving, E. (1993). What is episodic memory? *Current Directions in Psychological Science*, *2*(3), 67–70. https://doi.org/10.1111/1467-8721.ep10770899

Wahlheim, C. N., & Jacoby, L. L. (2013). Remembering change: The critical role of recursive remindings in proactive effects of memory. *Memory & Cognition*, *41*(1), 1–15. https://doi.org/10.3758/s13421-012-0246-9

Appendices

Appendix A: Letter of Information-Student

Project Title: Distinguishing Right from Wrong in Language Exercises

Principal Investigator
Prof. Frank Boers
Faculty of Education
The University of Western Ontario, London, Canada
Student Investigator
Mengxue Li
Faculty of Education
The University of Western Ontario, London, Canada

Thank you for being interested in this research project. Please read the following Letter of Information and decide whether you would like to participate in this project or not. If you decide to participate in this study, we will sincerely appreciate your help. If you decide not to take part in the study, we will also be thankful for your interest.

Invitation to the study project

The aim of the study is to evaluate the usefulness of exercises for collocation learning that are common in EFL textbooks. Collocations are conventional word partnerships that differ from one language to another (e.g., in English we say *have a dream*, but the counterpart in French is "make a dream"; in English, we say *make an effort*, but the counterpart in Dutch is "do an effort"). This makes learning collocations in another language challenging. EFL textbooks use various types of exercises to help students learn collocations. We would like to explore how well these exercises work. You are invited to participate in this study because you are a language learner and a language teacher yourself. Should you agree to participate, you will be asked to do a set of exercises on English collocations you may not yet know, react to the exercises while you are doing them, and help us to evaluate how efficient they are. Users of English as their first language are not invited to participate, because they will already be familiar with the collocations.

The rationale of the study

With the growing recognition of the importance of phraseology for language learners, contemporary language courses feature exercises on collocations (word partnerships). These exercises come in diverse forms (e.g., matching, multiple-choice, and gap-fill exercises), but hardly any research exists to inform teachers and textbook writers whether some exercise types are more helpful than others. In this project, we address this gap in the research by collecting learner-participants' reactions to such exercises and by assessing how well the exercises serve their intended purpose. The practical aim is to inform teachers and course designers of the pros and cons of the diverse exercise types, and to give guidance as to whether some exercise types should be given precedence over others.

The procedures of the study

If you agree to participate in the study, you will be invited to two meetings (one week apart) with the Student Investigator (Mengxue Li) and to do a series of exercises with a focus on English collocations, some of which you may not yet know and will therefore learn in these meetings. You will be asked to say what goes through your mind (i.e., to think aloud) while you are doing the exercises, and your reactions will be recorded on a digital voice recorder. This is necessary because the researcher will not be able to take notes of your reactions fast enough and so will need to relisten to what you said. You will be given feedback on your exercise responses after completing the exercises. The exercises will be of diverse types, so we can compare your reactions to them. Each meeting will take approximately 40 minutes and will take place in one of the small meeting rooms here at the Faculty of Education.

The risks and harms of participating in the study

We do not anticipate any risks or discomfort related to participating in this study project. The two meetings with the Student Investigator will be held in a room at the Faculty of Education at a time that is convenient to you (you'll be asked to choose a day/time from a proposed schedule).

The benefits of participating in the study project

This study will be beneficial for you as a language teacher because it presents a valuable opportunity for reflective practice about commonly used language exercises. It will be useful for you also as a language learner because you will probably extend your knowledge of English collocations. Preliminary research findings (based on early data analysis) will be communicated in one of your TESOL classes in spring. If you wish to be informed of the final findings, we will email a summary to you when data analysis is completed.

The option of leaving the study

You can leave the project at any time. We can then also remove any information already collected from you if you would like us to. You can simply send the Student Investigator (Mengxue Li) an email to let her know of your decision. Should you wish to withdraw your data, then do notify the Student Investigator within a week after your second meeting with her, before she starts processing the data.

Data privacy

All the data collected from you will be kept confidential. The results of the research project will be reported in a dissertation, in an article submitted for publication in a scholarly journal, and in a conference presentation. No names or any other information enabling others to identify you will be included in these reports. Your identity will be known only by the Student Investigator. She will be the only one present during your meetings. She will collect the consent forms and will keep these until after your graduation. You will not need to write your name on any exercise worksheets nor mention your name while you are thinking aloud about the exercises. The Student Investigator will use a code on worksheets and the audio-file. She will be the only person

to have a "master list" with your name and email address next to the code. Your verbal comments while doing the exercises (i.e., your "thinking aloud") will be audio-recorded and transcribed by the Student Investigator. She will delete the audio-files after she's completed the transcription. Should your name be used in the recording, then this will not be included in the transcript. The anonymized data (worksheets, transcripts, and spreadsheets with coded responses) will be stored only on a secure server at Western University and will be retained for 7 years by the Principal Investigator, after which the data will be deleted. It is important to note that a record of your participation must remain with the study, and as such, the researchers may not be able to destroy your signed letter of information and consent, or your name on the master list. However, any data may be withdrawn upon your request.

Delegated institutional representatives of Western University and its Non-Medical Research Ethics Board may require access to your study-related records to monitor the conduct of the research in accordance with regulatory requirements.

Compensation

You will receive an Amazon e-gift card worth $20 (Canadian) at the end of each meeting in return for your time and participation.

The rights of participants

Your participation in this study is voluntary. You may decide not to be in this study. Even if you consent to participate you have the right to not answer individual questions or to withdraw from the study at any time. If you choose not to participate or to leave the study, it will have no effect on your grades. You do not waive any legal right by consenting to this study. We will give you any new information that may affect your decision to stay in this study.

Contact for questions

If you have more questions about this study, contact Prof. Frank Boers. If you wish to withdraw, contact the student research assistant, Mengxue Li.

If you have any questions about your rights as a research participant or the conduct of this study, contact the Office of Human Ethics. This office oversees the ethical conduct of research studies and is not part of the study team. Everything that you discuss will be kept confidential.

This letter is yours to keep for your future reference.

Appendix B: Consent Form

Project Title: Distinguishing right from wrong in language exercises: The case of collocations

Principal Investigator: Prof. Frank Boers

Student Investigator: Mengxue Li

For participants

- I have read and understood the Letter of Information.
- All relevant questions have been explained satisfactorily by the investigators.
- I will keep a copy of the Letter of Information and this consent once I have signed.

_____	_____	_____
Print Name of Participant	*Signature*	*Date (DD-MM-YYYY)*

For person obtaining consent

My signature means that I have explained the study to the participant named above. I have answered all questions.

_____	_____	_____
Print Name of Participant	*Signature*	*Date (DD-MM-YYYY)*

About the results of study:

Would like to receive a written summary of the research findings via email? Check the appropriate box.

YES ☐ NO ☐

Appendix C: Exercise Session 1 Worksheet

Choose the correct answer to complete the blank in each of the following sentences.

1. Alcohol addiction can ____ a heavy toll on your health and can shorten your life by many years.
    A. have
    B. take
    C. pose

2. Few people believed Peggy had the right qualities to ____ a business, but she proved them wrong. She's now head of one of the most profitable firms in the country.
    A. do
    B. hold
    C. run

3. There's something you need to be aware of, Jimmy: If you start dating Sarah, you ____ the risk of losing Jenny's friendship because Sarah is her worst enemy.
    A. have
    B. run
    C. pose

4. Can you ____ a watch on the children while I go and buy a coffee?
    A. hold
    B. keep
    C. take

5. Neither of her parents can ____ a tune, and yet their daughter is a famous singer.
    A. carry
    B. hold
    C. keep

6. The new legislation will ____ effect on June 1st.
    A. have
    B. make
    C. take

7. The nurse asked me to ____ my breath while the X-ray of my chest was taken.
    A. save
    B. hold
    C. keep

8. My son is very fit. He can cycle all the way to school, and he doesn't even ____ a sweat, when he does this.
    A. break
    B. draw
    C. run

9. Hello, David. If you're not busy, could we ____ a word?
    A. have
    B. hold
    C. speak

10. Many Americans still ____ a prayer before their meal. They call this 'grace'.
    A. make
    B. say
    C. tell

11. I've decided to ____ a totally new approach to this problem.
    A. give
    B. make
    C. take

12. No-one said anything for a minute, but then Frank ____ the silence with a humorous remark.
    A. broke
    B. cut
    C. killed

13. I told Bill to move the barbecue away from the hedge. I was afraid it would ____ fire.
    A. attract
    B. catch
    C. take

14. Wildfires ____ a serious threat to endangered species.
    A. give
    B. pose
    C. set

15. Once a year we ____ tribute to the soldiers who sacrificed their lives in the war.
    A. give
    B. hold
    C. pay

16. I can sense that you have reservations about this proposal. Please ____ your mind.
    A. say
    B. speak
    C. tell

17. The announcement that global warming was happening much faster than expected _____ the tone for the conference on climate change.
    A. set
    B. settled
    C. struck

18. After failing to establish peace in Afghanistan, the Americans _____ their losses and left.
    A. dropped
    B. cut
    C. settled

19. Hi Angela, can you _____ a minute? I'd like to ask you a couple of quick questions.
    A. give
    B. save
    C. spare

20. I'm going to _____ no chances, and make sure I have good travel insurance this time.
    A. draw
    B. run
    C. take

Exercise Answer Keys:

1. B
2. C
3. B
4. B
5. A
6. C
7. B
8. A
9. A
10. B
11. C
12. A
13. B
14. B
15. C
16. B
17. A
18. B
19. C
20. C

Appendix D: Exercise (Test) Session 2 Worksheet

Complete the blank of each of the following sentences with a verb that partners with the noun that follows.

1. The daycare is recruiting certified staff to _____ a watch on the kids after school.

2. I will _____ a word with the disruptive student after class.

3. I don't consider myself a religious person, but when I go skydiving, I nonetheless _____ a prayer before jumping out of the airplane.

4. For several minutes nobody said anything, but then Bill decided to _____ the silence and inquired if anyone had been to the movies lately.

5. If you keep turning up late for work, you _____ the risk of losing your job.

6. Move those candles away from the curtains, please. We don't want them to _____ fire.

7. A stressful job can _____ a heavy toll on one's health and family life.

8. Peterson had _____ the business with great success for ten years, but then he was unexpectedly asked to step down by the new board of shareholders.

9. I was wondering if you could _____ a minute of your time to answer a couple of quick questions.

10. Mrs. Hamilton opened the meeting by cracking a couple of jokes. That _____ the tone for the whole of the meeting, which was relaxed and friendly.

11. Opening our restaurant in this neighborhood was a bad idea. No-one here seems interested in French cuisine. Let's _____ our losses and try again in a different part of town.

12. I would like to _____ tribute to all the brave fire fighters who put their own lives at risk to save others immediately after the disaster.

13. I'm still in pretty good shape. I walked all the way to campus in just 30 minutes, and I didn't even _____ a sweat.

14. It can take a little while for medicine to _____ effect, but I'm sure you'll soon feel better.

15. My parents encouraged me to think critically and to _____ my mind when I disagree with someone.

16. Don't expect me to sing in the choir. I've never been able to _____ a tune.

17. Unexpectedly, the government changed its health guidelines and decided to _____ a new approach to the pandemic.

18. Rising temperatures _____ a serious threat to numerous animal species, including polar bears.

19. Please don't _____ any chances and be well prepared for each of the exams.

20. How long can you _____ your breath under water?

Test Answer Keys:

1. keep
2. have
3. say
4. break
5. run
6. catch
7. take
8. run
9. spare
10. set
11. cut
12. pay
13. break
14. take
15. speak
16. carry
17. take
18. pose
19. take
20. hold

Appendix E: Scripts of Instructions for the Think-Aloud Sessions

Script Session 1

Thank you again for helping us with our research.

I'm going to give you a worksheet with an exercise on verb-noun collocations. It imitates the kind of exercises that are commonly used in EFL textbooks, notably multiple-choice exercises. We would like to find out what kind of learning strategies and processes are prompted by these types of exercises, and whether they are always helpful.

I would like you to do the exercises and to "think aloud" while you're doing them. For example, I would like to find out how you decide which verb is the correct one in the sentences. I would also like to know how sure you feel about your choices. You may feel totally confident, for example, because you know the collocation, but in other cases you may hesitate between possible answers because, and in yet other cases you may simply be guessing.

[If you want to express your thoughts in Chinese, then you're welcome to do so—it's my first language too.]

Feel free to ask me for clarifications of any words in the sentences that you are not familiar with. Also let me know if you need a break at any point.

At the end of the exercise, I will of course tell you what the correct answers are. Also at that stage, I would like you to freely express your thoughts.

Do you have any questions about what I'm expecting you to do? Okay, let's get started!

[…]

The answer keys for the exercise are: …

Thank you! See you two weeks later.

Script Session 2

Today we're going to do another type of exercise on verb-noun collocations. The exercise focuses on the same collocations that you practiced last week. But instead of using a multiple-choice format or a find-and-correct-the-errors format, the worksheet just has gapped sentences, with the verbs missing.

Like last week, I'd like you to think aloud as you do the exercise. For example, you may want to say that you have seen this collocation on some occasions, or that you cannot remember which verb was correct, or that you are just guessing. Please say if you are hesitating between two or more possible verbs. [Feel free to use Chinese to express your thoughts if this makes the think-aloud task easier.]

I'll give you the right answers when you've completed the whole exercise.

Okay? Let's get started.

[…]

Let me show you the multiple-choice exercise worksheets that we used two weeks ago.For each multiple-choice item I will ask you three questions: (a) what option did you choose two weeks ago, (b) was your choice right or wrong, and (c) what was the answer key that I gave you. You don't have to guess if you do not recall any of them, just say 'I don't remember'.

Now, I'm giving you the answer keys for the gap-fill items: …

Now that you have a sense of how well you remembered the collocations from last week's exercises, how do you feel about the usefulness of those exercise types? Have you sometimes hesitated between verbs in today's exercise because last week's exercises required you to consider different possibilities, including wrong ones? Did you remember the different options in last week's exercises, but were not sure anymore which option turned out to be correct in the end?

This is in what our research intends to find out: Do multiple-choice exercises and find-the-error exercises create memories of wrong options that interfere with learners' recall of the correct choices at a later stage? Is this a phenomenon you have experienced before, perhaps?

My supervisor, Frank Boers, will tell you more about this line of research in his course on Teaching and Learning Grammar next term. By then, I will also have completed most of the data analysis for this specific project and will be able to share the findings with you.

We want to evaluate the usefulness of these exercise types because they are very common in language courses, while no research is available to show they are effective. So, we need empirical data to assess whether these exercise types cause confusion—and what can be done to reduce the risk of confusion. Without the participation of volunteers like you, collecting such empirical data would not be possible. So, thanks again for your help!

Appendix F: NMREB Approval Letter

Western Research

Date: 21 December 2022

To: Professor Frank Boers

Project ID: 121836

Study Title: Distinguishing right from wrong in language exercises: Do wrong choices linger in memory?

Short Title: Distinguishing right from wrong in language exercises

Application Type: NMREB Initial Application

Review Type: Delegated

Full Board Reporting Date: 13/Jan/2023

Date Approval Issued: 21/Dec/2022 14:55

REB Approval Expiry Date: 21/Dec/2023

Dear Professor Frank Boers

The Western University Non-Medical Research Ethics Board (NMREB) has reviewed and approved the WREM application form for the above mentioned study, as of the date noted above. NMREB approval for this study remains valid until the expiry date noted above, conditional to timely submission and acceptance of NMREB Continuing Ethics Review.

This research study is to be conducted by the investigator noted above. **All other required institutional approvals and mandated training must also be obtained prior to the conduct of the study.**

Documents Approved:

Document Name	Document Type	Document Date	Document Version
Scripts for think aloud instructions_v2	Interview Guide	21/Dec/2022	2
Exercises_v2	Other Data Collection Instruments	21/Dec/2022	2
Test_v2	Other Data Collection Instruments	21/Dec/2022	2
verbal recruitment script_v2_clean	Oral Script	21/Dec/2022	2
Letter of Information and Consent Forms (LOI_C)_collocation exercises_v2_clean	Written Consent/Assent	21/Dec/2022	2

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario. Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB. The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

Please do not hesitate to contact us if you have any questions.

Sincerely,

Ms. Zoë Levi, Research Ethics Officer on behalf of Dr. Randal Graham, NMREB Chair

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).

Curriculum Vitae

Name:	Mengxue (Alyssa) Li
Post-secondary Education and Degrees:	Shanghai Normal University Shanghai, China 2013-2017 B.A. The University of Western Ontario London, Ontario, Canada 2018-2019 M.P.Ed The University of Western Ontario London, Ontario, Canada 2021-2023 M.A.
Honours and Awards:	Academic Scholarship 2015-2016 Entrance Scholarship 2022-2023
Related Work Experience	Teacher (Chinese) Shanghai Zhongguo High School Shanghai, China 2016-2018 Research Assistant The University of Western Ontario London, Ontario, Canada 2021-2023