THE EFFECTS OF WORKING MEMORY AND DIRECTIONALITY ON SENTENCE BUILD-UP DRILL PERFORMANCE

BY

YOU LI

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in East Asian Languages and Cultures
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Adviser:

Professor Jerome Packard

ABSTRACT

This study investigated the role of working memory (WM) and input direction in the use of sentence build-up drills (SBD). Thirty Chinese L2 speakers participated in the experiment, which consisted of a SBD test, a R-span test, and a language background questionnaire. The results indicate that the backward input direction has several advantages over the forward direction. First, the backward direction elicited repetitions with fewer total and whole-sentence errors. Second, it also extended the tipping point, which was defined as the serial position of the word at which participants started making their first errors. Finally, there was an interaction effect between WM and input direction, indicating that participants with different WM levels performed equally well in the backward direction, whereas the low-WM group made significantly more errors in the forward direction, compared to their high-WM counterparts. Based upon these results, we argue that the backward build-up is more effective, and that it constitutes a better way to conduct sentence build-up repetition drills in the classroom.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Despite the current popularity of communicative and task-based second language (L2) teaching methodologies, sentence repetition tasks, which are commonly considered to be associated with the audio-lingual method, still plays an important role in language teaching, learning and testing (Dekeyser, 2007; Ellis et al., 2009; Larsen-Freeman, 2012; van Moere, 2012). When sentences are too long to be remembered at once, the sentence build-up drill (SBD), is recommended to improve participants' performance. SBD requires teachers to parse sentences into smaller meaningful chunks, and then present them incrementally to the students for verbatim repetition, until the whole sentence is repeated.

In language classrooms, SBD can be implemented in two different directions: forward build-up drill (FBD) starts from the first word and builds the whole sentence forward, whereas backward build-up drill (BBD) starts from the last word and moves toward the front. Currently, the choice between these two directions is a matter of teachers' preference. On one hand, forward direction is more intuitive for many teachers, since sentences in natural language always unfold in a forward direction. On the other hand, however, some proponents of the audio-lingual method assert that it is important to start from the last word and work backwards in order to elicit more accurate repetition (O'Connor, 1960; Benwell, Mathieu, & Holton, 1961; Dauer & Browne, 1992; Anderson-Hsieh & Dauer, 1997; Larsen-Freeman & Anderson, 2013). Because of this belief, BBD is widely practiced in language classrooms, despite the perception that its unfolding direction, both in terms of form and in terms of information, is non-intuitive. However, this claim about the efficacy of BBD over FBD in facilitating sentence repetition has never been put to an empirical test. Therefore, one purpose of the current study is to investigate the directionality effect on SBDs.

Another related issue is the effect of working memory capacity (WMC) on sentence repetition tasks. There are two ways in which WM can affect learners' performance. Frist, it is obvious that to complete a repetition task, participants need to hold the input in their WM while listening to the sentences, and later retrieve the stored information for their own production. What is not clear is whether the WMC influences performance on SBDs. While sentence repetition tasks have been widely criticized for their lack of higher cognitive processing and for relying merely on rote memorization, some recent studies (Okura & Lonsdale, 2012; Kim, Tracy-Ventura, & Jung, 2016; Yan, Maeda, Lv & Ginther, 2016) report that WMC may not be an indicator of sentence repetition performance. It has been claimed that repetition of long sentences is reconstructive rather than literal in nature, because the information retrieved from WM is not the verbatim form of the input, rather, it is the meaning, together with the recently activated lexical items that are retrieved and used to reconstruct the sentences. As a result, the performance of the task is more accurately predicted by participants' implicit language knowledge (Ellis, 2005; Ellis et al, 2009) than their WMCs.

In addition, previous literature has shown that WM can interact with other factors to influence learners' performance (see Wen, Mota, & McNeill, 2015, and Juffs & Harrington, 2011, for review). For example, Santamaria and Sunderman (2015) found that, compared to low-WMC participants, high-WMC participants had a better and longer-lasting learning effect in production tasks, but not in comprehension tasks. Ahmadian (2015) investigated the relationship between WMC and self-repair behaviors while conducting online output planning. He found that WMC correlated positively with the number of self-repairs on phonological, lexical and syntactic errors, but correlated negatively with D-repairs, in which speakers encoded new information. The results suggest that as WMC increases, participants make more form-related but fewer information-related self-corrections. Ahmadian argued that speakers with different levels of WMC allocate their attentional resources to different aspects of L2 speech, that is, high-WMC speakers devote more processing resources to the pre-articulatory and external loops of monitoring. These two studies explore the interaction between WM and different learning factors, with the first suggesting that WM effect is different across modalities with a larger effect on production, and the second suggesting that speakers with different levels of WMC might allocate their attention differently. Since the major task of the present experiment is oral production BBD in which participants might pay more attention to form, it is reasonable to predict that WM may play a role in SBDs. The second goal of the current study is to investigate whether or not there is empirical evidence for this prediction.

To sum up, the present experiment is designed to investigate the effect of directionality, WMC, and their potential interaction on the performance of SBDs by Chinese L2 speakers.

In the remaining sections of this chapter, there will be a selective review of related literature. Section 1.1 introduces audio-lingual claims about BBDs, which is followed by section 1.2 that discusses the nature of the sentence repetition task and its functions in SLA. Section 1.3, reviews the research on the effect of WMC on sentence repetition tasks. Finally, in the 1.4 section, research questions and predictions of the current study are formulated.

1.1 Backward Build-up Drill

Repetition tasks were especially favored by the audio-lingual method, which considered language learning to be "basically a mechanical system of habit formation, strengthened by reinforcement of the correct response." (Paulston, 1971, p.3) Hence, both the action and the accuracy of the repetition were important. To help students to articulate a more accurate response, when a sentence was too long to be remembered all together, the BBD technique was recommended, since it would facilitate production by breaking down long sentences into small chunks and expanding the repetition, part by part, from the last words to the beginning words of a sentence.

Advocators of BBD claimed that expanding a sentence backward had several advantages over FBD. O'Connor (1960) proposed that repetition practice should be used to introduce new phrases or sentences, through which the students can be forced to remember the phrase accent, and the melody of the new sentences. She further recommended that when sentences are too long, they must be parsed by meaningful parts and be built-up from the end. In this way, sentences could be presented with correct parsing, and the normal intonation at the end of a sentence would also be preserved throughout the building-up practice. Benwell et al. (1961) also argued that a repetition task might be very memory demanding. Participants not only need to imitate the sounds they hear, but also to memorize the sequence of words. Therefore, BBD should be used in order to ease the memory load, and to ensure that the melody of each unit is preserved as well as that of the complete sentence. Besides the function of preserving the intonation of the cue sentences, BBD is also considered to be useful as a pronunciation practice. Dauer and Browne (1992) emphasized that English learners should learn to link words together in pronouncing their sentences to improve the overall pronunciation and intelligibility. The authors presented that backward build-up was a technique that could help for this purpose. Further, Anderson-Hsieh and Dauer's (1997) presentation argued that slow-down speech could be used for teaching both listening comprehension and pronunciation to L2 English learners. Like Dauer and Browne, they also suggest that backward build-up would be a helpful tool when the students could not follow the teacher to pronounce phrases accurately. Although all these suggestions in pedagogical literature linked BBD with a more accurate pronunciation, these arguments had never been empirically tested.

Larsen-Freeman and Anderson (2011) also included BBD as a teaching technique of the audio-lingual method. They described a hypothetical observation of BBD when students stumbled over a sentence in their repetition. The authors explained that BBD were recommended in such cases, because it could “direct more student attention to the end of the sentence, where new information typically occurred,” (p. 47) and consequently, could elicit more accurate responses. This explanation took a further step from previous assertions and implied that BBD could also facilitate the memorization of the wording and information of sentences.

For the current study, the major goal is to test the efficacy of BBD, in comparison to FBD, on the broader implication, which is suggested by Larsen-Freeman and Anderson. Therefore, the experiment is designed to test whether BBD elicits more accurate memorization and articulation of the syllables in the cue sentences, and the responses are not evaluated for their pronunciation accuracy, rather, they are graded at the syllabic level, which also overlaps with the morphemic boundaries in Chinese. The rationale is that if the BBD leads to better memorization of the sequences of words, there should fewer errors at the syllabic level in the repetition.

1.2 Sentence Repetition: Its Nature and Functions

On the surface, sentence repetition drills entail a “complete control of the response, (and) there is only one correct way of responding,” which fits the definition of mechanical drills as defined by Paulston and Bruder (1976, p.4). Although it has been largely criticized for neglecting the meaningful and communicative aspects of language using, there is evidence in previous literature showing that repetition is not merely parroting, rather, it can be reconstructive, and may require participants' implicit language knowledge.

Potter and Lombardi (1990) tested their hypothesis that immediate sentence repetition could be a reconstruction by the speakers generated from the conceptual representation of the cue sentences with words that had been recently activated. In their experiments, when a synonym of a word in the sentence was presented before or after the stimuli, for both adults and 4-year-old children, this synonymic word intruded frequently in the repetition, indicating that the repetition was not parroting, and the participants reconstructed the sentence with activated words in the cache. Other researchers observed similar spontaneous word changing in the repetition of ungrammatical sentences. For example, Erlam (2006) gave participants both grammatical and ungrammatical sentences as stimuli, and reported that, although participants were not explicitly informed that there were ungrammatical sentences, native speakers automatically corrected 91%, and L2 learners corrected 35% of the ungrammatical sentences. (see also, Munnich, Flynn & Martohardjono, 1994; and Hamayan, Saegert & Larudee, 1997) Both studies showed that sentence repetition could be reconstructive in nature. In other words, repetition of sentences relied not simply on rote memorization, and participants who already had some knowledge of the target language could use this knowledge, spontaneously, to reconstruct the sentences.

Another line of research on the nature of sentence repetition examines the use of the repetition task, also termed as "elicited imitation (EI)," as a measurement of participants' implicit knowledge of the target language. Ellis (2005) used principal component analysis to investigate what was really measured by five different proficiency tests, including: an EI test, an oral narrative test, a timed grammaticality judgement test (GJT), an untimed GJT, and a metalinguistic knowledge test. The results revealed that these five tests loaded on two different underlying components. Ellis argued that the first three tests tapped into learners' implicit knowledge, while the other two tests measured explicit knowledge of the target language. Similarly, Erlam (2006) replicated these results with a different set of sentences in English. Bowles (2011) replicated these results with Spanish L2 and heritage learners.

This reconstructive view of sentence repetition tasks challenges the claims of those who consider repetition to be mechanical and irrelevant to language proficiency improvement. Larsen-Freeman (2012) pointed out three major roles of repetition drills in language teaching and learning, including: rote learning, enhancing working memory and automaticity access. He explained that, from a connectionist view of learning, repetition can result in strengthening the weight of connection in relevant neural networks, and consequently would increase automaticity. In a similar vein, van Moere (2012) argued that repetition drills provide frequent opportunities to connect components of utterances, and it would eventually help learners to memorize production chunks, which are considered to be "the building-blocks of fluent spoken discourse." (Ellis, 2001, p.45). Ellis also claimed that, when syllables are connected into chunks, they are believed to be stored and retrieved as entireties. As a result, WMC would also be increased since the size of each stored item was enlarged.

Previous research on repetition tasks have two implications for the current study. First, since sentence repetition is a worthwhile practice that leads to fluency and automaticity in L2, this study, which aims to find the optimal direction for BBD, has pedagogical value in teaching. Second, if repetition of sentences is reconstructive in nature, it puts the intuitive connection of WMC and repetition performance into question. As matter of fact, whether or not WM has an effect on sentence repetition, and, if so, how it influences performance is still a contentious question. Hence, investigating the WM effect and its interaction with input direction on BBDs will provide new evidence for this ongoing debate.

1.3 Working Memory Effect on Sentence Repetition Tasks

Individual differences, such as language proficiency level and WMC, can influence L2 production greatly. For repetition tasks, however, as discussed in the previous section, due to the reconstructive nature of the task, the intuitive link between WM and repetition performance might not hold true. Since no experiment has been conducted to test the WM effect on SBDs, the studies on sentence repetition tasks will be reviewed here.

Okura and Lonsdale (2012) gave English L2 learners two tests: a WMC test and an English sentence repetition test, which was termed as EI in their paper. The results showed no significant correlation between the WM scores and the repetition scores, indicating that WM had only a minor involvement in repetition. It is also noteworthy that this study employed a new paradigm to measure WMC. Participants were instructed to listen to sequences of nonce syllables and then to repeat what they heard. The verbal sequences varied in length from 1 to 10 syllables, and scores were given to each correctly repeated syllable string. Results of this paradigm showed that, participants could repeat sequences with 3 or less syllables very well, whereas the average accuracy across participants dropped below 80% when there were four syllables, and fell below 50% at 5 syllables. These results were in line with Cowan's (2001) claim, which indicated that WMC should be of "4±1" items.

Kim et al. (2016) replicated the results in Okura and Lonsdale's experiment with a more widely-accepted task, the forward digit span task, in which participants were asked to listen to, and repeat, series of random digits in the presented order. They also found that the repetition scores correlated with WM scores weakly, which indicated a nonsignificant modulating effect of WM on repetition performance.

On the one hand, these two studies provided empirical evidence showing that WMC had a limited effect on repetition performance. On the other hand, it is not clear whether the lack of correlation was due to problems in the measurement of WMC. Both experiments employed simple span tasks, which are considered to assess the capacity limitation in terms of the amount of information actively held in WM (Wen, 2016). However, there is evidence in previous literature (e.g., Perfetti & Lesgold, 1977; Daneman & Carpenter, 1980; Turner & Engle, 1989; Engle, Tuholski, Laughlin & Conway, 1999; Juffs & Harrington, 2011) suggesting that it is the complex span tasks, such as reading-span and listening-span, that involve both storage and processing aspects of WM. Hence, WM scores, measured by complex span tasks, were found to correlate with the performance of language processing tasks (e.g., reading comprehension task) better than simple span tasks. In Okura and Lonsdale (2012) and Kim et al.'s (2016) studies, the true relationship of the WMC and the repetition scores might have been masked by the task effect. It would be stronger evidence for their claim if their results could be replicated with WMC measured by a reading-span (R-span) task.

The experiment presented here aims to investigate the effect of WMC on SBDs. The reviewed literature on sentence repetition tasks may shed some light on the issue. First, there is a great chance that no significant effect of WM will be found on SBDs, even when WMC is measured by the R-span task, because SBD is a variant of the sentence repetition task. Second, Okura and Lonsdale observed dramatically decreased accuracy when the number of nonce syllables exceeded five. Consistent with Cowan's "4±1 items" WMC hypothesis, these results implied that there was a tipping point, where participants were most likely to start making errors, and the tipping point was around 4 items. Regarding the current study, questions arise over whether the tipping point, defined as where participants make their first errors, will be influenced by the input direction; and whether the "4±1" capacity hypothesis holds for SBDs, when items are defined as parsed chunks; and, finally, whether the tipping point is significantly larger for participants with larger WMC.

1.4 The Current Study: Research Questions and Predictions

Three research questions are addressed with this experiment. First, compared with FBD, does BBD elicit more accurate repetitions? Second, does WMC have a significant effect on performance? Third, is there any interaction between WMC and directionality?

Regarding the effect of directionality, no a priori prediction is being made. Any significant effect will suggest the optimal direction of SBDs. However, a null result might suggest that FBD is the better option, since BBD is less intuitive and perhaps more demanding compared to FBD. As for WM, basing on the results from previous literature, an effect of WM on error count is not expected. However, WM may influence the tipping point, because it is reasonable to assume that the performance of high-WMC participants may be better than their low-WMC counterparts. Finally, there is no prediction on the interaction between the directionality and the WM effects.

CHAPTER 2: METHODS

2.1 Participants

Thirty Chinese L2 learners between the ages of 18 to 30 participated in the experiment. Data from one participant was discarded due to technical difficulties resulting in an incomplete WMC test, yielding a final total of twenty-nine participants. All participants were registered students at Beijing Language and Culture University. Results from a language background questionnaire showed that participants varied in their L1 background and all of them had some experience in learning other foreign languages.

Each participant's Chinese proficiency level was also self-reported in the questionnaire. An HSK-5 level was coded as high proficiency (HP), whereas an HSK-4 level was coded as intermediate proficiency (IP). In addition, the proficiency level of participants who had no HSK scores was coded based on their course registration status. That is, since all these students were registered for courses to prepare for the HSK-4 test, this last group was coded as low proficiency (LP). Thus, all twenty-nine participants were divided into three subgroups basing on their proficiency levels, with six in the HP group, thirteen in the IP group, and ten in the LP group. Informed consent was obtained from all participants prior to the experiment, and all participants were paid for their participation.

2.2 Stimuli for Sentence Build-up Drills

Twenty cue sentences were constructed using only the core vocabulary from Chinese Link, Level 1, which is a popular first year Chinese textbook in the US. Sentences all began with a complex NP as the subject, which was followed by a VP as the predicate, after which another complex NP occurred as the object. Note that the complex NPs were all subject-gap relative clauses, regardless of their positions in the main clauses. Four sample sentences of this construction can be seen in Table 1.

The length of verbal sequences plays a critical role in repetition performance (Yan et al, 2016), and should be strictly controlled across build-up conditions. In this experiment, distinctions were made between: the length of sentences (SL), which referred to the number of syllables contained in each cue sentence, the total repeated length (RL), which referred to the integrated number of syllables repeated through the building-up procedure, and the length of chunks (CL), which referred to the number of syllables contained in each chunk. The SL varied from 18 to 20 syllables, whereas the RL varied from 72 to 80 syllables. Each sentence was parsed into seven chunks, guided by two general principles: first, all chunks should be meaningful; second, chunks should not be too long or too short. Hence, the chunking criteria were as follows: (1) monosyllabic words should be grouped with the following syllables to avoid chunks containing only one syllable; (2) disyllabic and three-syllable words may be separated alone or grouped with other words before or after, as long as the chunk size did not exceed four syllables. As indicated by the pipes “|”, in Table 1, all chunks were meaningful parts, which varied from 2 to 4 syllables. Sentences were presented to participants from the first (FBD) or last (BBD) chunk, adding one chunk at a time.

<table>
<tr><th>Cue sentences</th><th>Sent. L (SL)</th><th>Tota. L (TL)</th></tr>
<tr><td>1. 常喝|冰红茶的|爸爸|有位|不去|图书馆的|朋友。
often drink | black ice-tea | dad | have one | not go | library | friend.
The dad who often drinks black ice-tea has a friend who does not go to library.
<b>Probe word:</b> 有位 have a. (Answer: “yes”)
<b>Probe statement:</b> The friend does not go to the library. (Answer: “yes”)</td><td>18</td><td>72</td></tr>
<tr><td>2. 爱吃|日本菜的|妈妈|有位|不去|体育馆的|室友。
love eat | Japanese food | mom | have one | not go | gym | roommate.
The mom who loves Japanese food has a roommate who does not go to gym.
<b>Probe word:</b> 有位 have one. (Answer: “yes”)
<b>Probe statement:</b> Mom loves Japanese food. (Answer: “yes”)</td><td>18</td><td>72</td></tr>
<tr><td>3. 留学生|常去的|书店|前边有|老师|跑步的|健身房。
foreign student | often go | book store | front have | teacher | run | fitness room.
In front of the book store where foreign students often go, there is the fitness room where the teacher goes to jog.
<b>Probe word:</b> 前面有 there is (Answer: “no”)
<b>Probe statement:</b> Foreign students go to fitness room. (Answer: “no”)</td><td>19</td><td>76</td></tr>
<tr><td>4. 女朋友|买来的|桌子|后边有|妹妹|要骑的|自行车。
girlfriend | buy | desk | back have | sister | go to ride | bike.
Behind the desk which was bought by the girl friend, there is the bike that the sister is going to ride.
<b>Probe word:</b> 后面有 in front have. (Answer: “no”)
<b>Probe statement:</b> The girlfriend rides the bike. (Answer: “no”)</td><td>19</td><td>76</td></tr>
</table>

Table 1. Samples of cue sentences. SL: the number of syllables contained in each cue sentence. RL: to the integrated number of syllables repeated through SBDs.

In order to compare the number of errors in different build-up directions, sentences were created in comparable pairs, within which the SL, the RL, as well as the CL of the two sentences were all matched. Taking sentence 1 and 2 in Table 1 as an example, these two sentences were same in their SL and RL. In addition, the CL at a given serial position was also matched. For instances, in both sentence 1 and 2, the third chunk contained two syllables, whereas the sixth chunk contained four syllables. Moreover, for all cue sentences, the RL was kept constant in both build-up directions. In this way, when the two sentences within a pair were assigned randomly into FBD and BBD, there was no difference in the total number of syllables that needed to be repeated, and consequently, the number of errors was comparable within the pair. In contrast, sentences in different pairs were not matched on their lexical categories, length, and structure. When analyzing the data, each pair was coded as one item, which was fitted into the statistical models as a random effect.

Two practice sentences, and twenty filler sentences were created in a similar pattern, yielding a total number of forty testing stimuli, 20 for FBD, and 20 for BBD. All sentences were prerecorded and programed for a self-paced audio presentation. Critical sentences in each pair were randomly assigned to different build-up directions, and were separated by at least 19 testing sentences in order to avoid a learning effect. Throughout the stimuli list, FBD and BBD were arranged alternatively, one after another.

To test memorization and comprehension of the cue sentences, the repetition task was followed by a probe word, and then, a probe statement. The probe word was employed to check the memorization of the exact wording in the cue sentences. There is empirical evidence showing that in serial recall tasks, memorization is worst for items that occur in the middle of the sequences. (Murdock, 1962) Therefore, all probe words in critical sentences were designed to test the memorization of the fourth chunk in each sentence, which was the one that occurred in the middle position. Half of the probe words were the same as those used in the sentences, while the other half were synonyms of the original words, but with only slight differences in form. For example, the probe word 前面 *in front*, has the same meaning to the original word 前边 *in front*, and differs from it only with the second syllable. Probe words in filler sentences varied in their positions, in order to distract participants' attention from the critical region. As for the probe statements, they were used to assess participants' comprehension of the stimuli. The "yes" and "no" answers were balanced in their counts, with 50% of each across all stimuli sentences. But the answers were kept the same within critical pairs.

2.3 Measurement of Working Memory

Participants' WMC was measured by the Reading Span (R-span) test, developed by Daneman and Carpenter, (1980, as modified by Unsworth et al. 2009). Daneman and Carpenter reported in their paper that, unlike the digit span test, R-span correlated very well with reading comprehension performance. They also argued that WM measurement must involve processing tasks in order to predict processing related performance. As it was discussed in Section 1. 3, previous studies (Okura & Lonsdale, 2012; Kim et al., 2016) using simple span tasks did not find significant correlation between WM and sentence repetition scores. It is possible that the lack of correlation was due to imprecise measurement of WM. Therefore, the current experiment employed the R-span test as a more appropriate assessment tool for WMC.

The Chinese R-span test was created using only vocabulary from *Chinese Link*, Level 1. Participants were required to read increasingly longer sets of sentences, with the set sizes varied from 2 to 8 sentences. Of each size two trials were presented. There were also two practice trials of size-two before the critical sentences were presented, yielding a total number of 74 sentences. These sentences were presented visually, each followed by a random alphabetic letter. After reading the sentence, as well as the following letter, a probe statement appeared on the screen. Participants were asked to judge whether the statement was true based on their comprehension of the cue sentence. After all sentences in a given set were presented, participants were asked to recall the letters following each sentence in the presented order. This Chinese R-span test was programed and implemented using *E-prime 2.0 Professional*.

The scoring procedure was developed by Packard and Qian (2015), which added the greatest number of correctly recalled letters in a set with the proportion of correct recalls from the next two sets. For example, if the letters in the two sets of size-four were all correctly recalled, and the participant started making errors in size-five sets, ending up with 6 letters correctly recalled, then the WMC score for this participant would be 4+6/10=0.6. Participants were then divided into high- and low-WM groups by calculating the mean and median scores. Since there were 29 participants, and the median score was below the mean, 15 participants were assigned to the low-WM (LWM) group and 14 participants were assigned to the high-WM (HWM) group.

2.4 Procedures

The present experiment consists of three parts: a SBD test, a R-span test, and a language background questionnaire. For the SBD test, participants were instructed to start to repeat immediately after the audio files were played. Although the presentation was self-paced, the experimenter supervised the procedure to make sure that there was no pause between the stimuli and the repetition. After repeating each cue sentence, participants were required to respond to the probe word and probe statement by hitting a button on the keyboard, after which a cue line appeared on the screen indicating the start of a new trial. There was a ten-minute break between the SBD and the R-span tests, during which participants filled out the language background questionnaire. The complete session took about one hour and forty minutes on average.

CHAPTER 3: RESULTS

3.1 Transcription and Error Analysis

Data of six critical sentences were removed from the analysis due to technical difficulties resulting in incomplete sentence repetitions. Including data from the discarded participant, there was a total loss of 4.5% of the data. Repetition responses to all critical sentences were transcribed into text for error analysis. Since the main goal of this experiment was not on the pronunciation factor, but rather on the wording and meaning aspects of the repetitions, all errors were identified at the syllabic level. Scoring criteria can be seen below:

1. Deletion: count each deleted syllable as one error;
2. Addition: count each inserted syllable as one error; inserted pauses with meaningless sounds, such as “blah-blah-blah,” do not count as errors;
3. Substitution: count the deleted syllables or the syllables used for substitution, whichever is larger, as the number of errors;
4. Shifting + deletion: when a syllable is moved to another position to substitute for the original word, count as substitution only;
5. Shifting and exchanging syllables: count each moved syllable as one error;
6. Self-repairing: correctly self-repaired syllables do not count as errors.

Two types of error count were calculated for each participant. The total error count (totE) represented the integrated error count throughout the SBD, whereas the whole-sentence error count (wholE) indicated how many errors occurred when the complete sentence was repeated. Notice that, the totE should not exceed the RL, whereas the wholE should not exceed the SL of the sentence. In addition, the tipping points (TiP), defined as the serial position numbers of the chunk, where a participant made the first mistake, were also collected.

	Forward			Backward			Overall		
	M	*SD*	Error Rate	*M*	*SD*	Error Rate	*M*	*SD*	Error Rate
totE	13.22	10.33	.17	10.17	8.48	.13	11.70	9.57	.15
wholE	4.67	3.82	.25	3.95	3.48	.21	3.00	3.67	.23
TiP	4.61	1.42	----	4.98	1.46	----	4.79	1.45	----

Table 2. Descriptive results of error analysis. “totE” is the integrated error counts of SBDs; “wholE” is the error count in repeat the complete sentence; “TiP” is the tipping point.

A summary of the descriptive results of error analysis are presented in Table 2. The results showed that the means of error count, as measured by the totE and the wholE, were lower for BBD, compared with FBD. Further, the mean tipping point was also larger with backward build-up, indicating that, on average, the repetition in backward build-up remained intact longer. Of particular note, the means of the tipping points for the FBD (*M* = 4.61, *SD* = 1.42), the BBD (*M* = 4.98, *SD* = 1.46), as well as for all the sentences together (*M* = 4.79, *SD* = 1.45), were in support of Cowan’s “4±1 items” WMC hypothesis, when each chunk was considered as one “item.”

Although the descriptive results suggested a numerical tendency, indicating the higher effectiveness of the BBDs over the FBDs, whether or not the differences were statistically significant still needs further testing.

3.2 Working Memory Scores

The R-span test yielded two scores for each participant: a WM score and an accuracy rate for the probe statements. The average accuracy rate (*M* = .89, *SD* = .05) was about 90% for both WM groups, indicating that participants were actively involved in the comprehension task, and thus, the WM scores yielded from the R-span test should be valid. Descriptive statistics of WM scores are reported in Table 3.

	Low WM		High WM	
	M	*SD*	*M*	*SD*
WM score	4.29	1.22	6.96	.89
Accuracy	.88	.05	.91	.05

Table 3. Descriptive results of the WM scores.

Independent samples *t* tests conducted on the means of the two WM groups showed that the HWM group had significantly higher WM scores ($t = 30.22$, $p < .000$) and accuracy rates ($t = 7.98$, $p < .000$).

3.3 Results

The data from the error analysis were analyzed in generalized linear mixed effects models using the *lme4 package*, version 1.1-12 (Bates, Maechler, Bolker, & Walker, 2016) of R, version 3.1.2 (R Core Team, 2016). Three separate poisson mixed effects models were applied to the data of the total errors, the whole-sentence errors, and the tipping points. Two other separate logistic mixed effects models were applied to analyze the results of the word verification task and the statement judgement task.

For all the analyses, the build-up direction (two levels), the WMC (two levels), the Chinese proficiency level (two levels), and the trial number were included as fixed effects. Subjects and items were included in these models as random effects, which were fitted using a "maximal" random effect structure supported by the data (Barr, Scheepers, & Tily, 2013). This results in random intercepts for subjects and items. In addition, two-way interactions between the build-up direction and the WMC, and between the build-up direction and the proficiency level were also included.

3.3.1 Total errors

As illustrated in Table 4, there was a significant main effect of direction ($p < .000$), indicating that there were fewer errors occurred in BBDs in compare to FBDs. The trial number was also significant ($p < .000$), indicating that the participants made fewer errors as the experiment went on. The main effects of proficiency level ($p = .11$) and WMC ($p = .81$) were not significant. In addition, the interaction between the build-up direction and the WMC was also significant (*p* = .04). Follow-up t-tests indicated that the effect of WMC was only significant for FBD (*t* = -2.57, *p* = 0.01), not for BBD (*t* = -0.86, *p* = 0.39).

Parameters	Estimate	SE	Pr(>|z|)
Direction: F	0.29	0.04	< 0.000 ***
WMC: H	0.02	0.16	0.91
Proficiency: IP	-0.16	0.19	0.39
Proficiency: HP	-0.66	0.23	< 0.00 **
Trial	-0.03	0.00	< 0.000 ***
Direction: F × Proficiency: IP	-0.00	0.05	0.93
Direction: F × Proficiency: HP	0.17	0.08	0.03 *
Direction: F × WMC: H	-0.14	0.05	< 0.00 **

Table 4. Results for the total errors.

Notes. F = forward. H = high. IP= intermediate proficiency. HP = high proficiency.
***p < .001, **p < .01, *p < .05

3.3.2 Whole-sentence errors

Results of the whole-sentence errors showed the same pattern as those of the total errors. As presented in Table 5, a significant main effect was found for direction (*p* < .000), suggesting that there were fewer whole-sentence errors in BBDs than in FBDs. Again, the trial number was also significant (*p* < .000). However, the main effects of proficiency level (*p* = .10) and WMC (*p* = .96) did not reach significance. The two-way interaction between the build-up direction and the WMC was marginally significant (*p* = .07). Follow-up t-tests showed that the effect of WMC was only significant for FBD (*t* = -2.44, *p* = 0.02), not for BBD (*t* = -0.50, *p* = 0.62).

Parameters	Estimate	SE	Pr(>|z|)
Direction: F	0.29	0.07	< 0.000 ***
WMC: H	-0.01	0.19	0.96
Proficiency: IP	-0.23	0.21	0.29
Proficiency: HP	-0.73	0.26	0.01 **
Trial	-0.03	0.00	< 0.000 ***
Direction: F × Proficiency: IP	-0.12	0.09	0.17
Direction: F × Proficiency: HP	0.06	0.12	0.65
Direction: F × WMC: H	-0.16	0.08	0.04 *

Table 5. Results for whole-sentence errors.

Notes. F = forward. H = high. IP= intermediate proficiency. HP = high proficiency.
***p < .001, **p < .01, *p < .05

3.3.3 Tipping points

When proficiency levels were included in the analysis, the model failed to converge. Since this factor was never significant in the previous analyses, it was removed from the model for the analysis of the tipping points. The model without the proficiency factor converged and yielded results as shown in Table 6. The direction effect was significant (*p* < .000) showing that FBDs had a smaller tipping point. In other words, the repetition in

BBDs remained intact longer than that in FBDs. the trial number was again significant, but in a reverse direction, that is, as the trial number increased, the performance became worse, as indicated by a smaller tipping point.

Parameters	Estimate	SE	Pr(>|z|)
Direction: F	-0.11	0.05	0.05 *
WMC: H	-0.02	0.06	0.73
Trial	-0.01	0.00	0.02 *
Direction: F × WMC: H	-0.06	0.08	0.44

Table 6. Results for tipping points.

Notes. F = forward. H = high. ***p < .001, **p < .01, *p < .05

3.3.4 Word verification and statement judgement

Proficiency was also removed from the analysis due to converging problem. As for the word verification task, results showed a significant effect of the trial number (*p* = .03), and a marginally significant main effect of WM (*p* = .10), suggesting that participants with larger WMC tended to remember the wording of the sentences better. Regarding the statement judgement task, the only significant factor was the trial number (*p* = .02), indicating that participants' performance improved over time for the comprehension task.

CHAPTER 4: DISCUSSION

This study aimed to investigate how directionality and WM affect the efficacy of SBDs. The major findings are summarized as follows. The numerical tendency in the error analysis suggested that BBDs elicited fewer errors and a larger tipping point in comparison to FBDs. This tendency was proved by the statistical analysis, which showed significant main effects of the sentence build-up direction on total errors ($p < .000$), whole-sentence errors ($p < .000$), as well as on the tipping point ($p < .000$) of the repetition performance. Furthermore, the main effects of WM were not significant for all the analysis, except that a marginal significant effect was found in the word verification task, indicating that participants in different WM groups did not differ in their repetition accuracy, but those who had larger WMC tended to have better memorization on the exact words that were used in the cue sentences. However, WM indeed played a role in the repetition task, having a significant interaction effect with the build-up direction on total errors and whole-sentence errors. Follow-up *t*-tests suggested that WM effect was only significant in the forward direction. In other words, participants from different working memory groups did not differ in their performance with the BBD, but the high-WM group did significantly better with the FBD.

4.1 Effects of Directionality

The facilitative effects of the backward direction were robust throughout all of the analyses, which supported the suggestion of the audiolingual method. Findings of the present experiment showed several advantages of BBD. First, with BBD, participants made fewer errors throughout the build-up practice. And when they repeated the sentences as a whole, which was the very goal of the SBDs, BBD also resulted in more accurate responses. Second, with BBD, participants could repeat more chunks before they made their first errors. Finally, while the low-WM group did worse in FBD, compared to the high-WM group, participants of both groups did equally well in BBD. In other words, the backward direction appeared to be easier for the low-WM participants.

The cognitive mechanism underlying these advantages of BBD is far from clear. Larsen-Freeman and Marti Anderson (2011) implied that the effect may relate to the information structures within sentences, that is, new information usually occurs towards the end of a sentence. However, this explanation cannot explain the results in the current experiment, since the sentences were controlled to ensure that the beginning of the sentences did not differ from the end in terms of given and new information.

One explanation could be that it was the serial position, where new chunks were added, that caused the effect. Previous studies showed different serial position effects for unorganized versus sequentially organized verbal materials (Deese & Kaufman, 1957; Murdock, 1962). In unorganized materials where adjacent words have no sequential association, a recency effect occurred such that the first items were moderately well recalled, whereas the last items were most frequently recalled. In contrast, for the organized materials, like passages of connected discourse, the order of recall was the same as the order with which the material was presented. It is obvious that words in phrases and sentences are sequentially organized. Therefore, the memorization for the first items should be better than the last items. In the present experiment, the FBDs had each newly added chunk located at the end of the verbal sequence, whereas in BBDs, newly added chunks always occurred at the beginning of the sequences, which would be the most frequently recalled position for organized verbal strings. Thus, participants' performance with the newly added chunks seemed to be facilitated by their position in the verbal sequence.

4.2 Effects of WM

Another finding of this experiment was that WM levels did not predict the performance on SBDs. This overall result supported the findings in previous studies on WM and the sentence repetition tasks (Okura & Lonsdale, 2012; Kim et al., 2016) Kim et al. (2016), in which no significant correlation was found between these two variables. One improvement of the current study was that a complex span task was employed to measure participants' WM. Since complex span tasks are reported to correlate better with the sentence processing tasks (Juffs & Harrington, 2011), the results from the current experiment provide stronger evidence for the claim that repetition of sentences does not rely merely on WM.

Nevertheless, the interaction effect between WM and build-up direction suggest that WM influences repetition in the forward direction, but not in the backward direction, implying that forward build-up requires more memory resources. Therefore, it is possible that in a regular sentence repetition task, without the build-up procedures, WM, as measured by the R-span test, may influence participants' performance on their repetition. A separate experiment is needed to test this possibility.

4.3 Pedagogical Implications

The results of this experiment suggest that BBDs are a better choice for SBDs for two reasons. First, BBDs appear to elicit more accurate repetition. As was found in this experiment, both the integrated error count and the number of whole-sentence errors were smaller in the backward direction, in which the tipping point was larger. Second, BBD can be especially facilitative for low WM participants, in that participants in this experiment performed as well as their high WM counterparts in BBDs, but not in FBDs.

Some people may argue that repetition tasks are now out of fashion. However, many scholars hold the position that repetition can proceduralize the process of formulating oral productions through which fluency can be enhanced. (Larsen-Freeman, 2012; Muranoi, 2007; DeKeyser, 2007). Furthermore, repetition is not parroting. In our experiment, there were several cases where participants substituted words with synonyms. This observation was similar to the phenomenon reported in the Potter and Lombardi (1990) paper, where the authors found that synonymous words intruded frequently in the repetitions. Therefore, it was clear that speakers were not simply parroting what they heard, since they made efforts to integrate the new part into the previous structure and tried to understand it before they gave utterance.

To sum up, sentence repetition tasks can be meaningful. When a sentence is too long to be remembered at once, sentence build-up drills can be used in the backward direction to facilitate more accurate repetitions, especially for students with lower WMC.

4.4 Limitations

Surprisingly, the experiment showed no main effect of proficiency. However, it was also suggested in the results that, although proficiency was not significant as a factor, the contrast between the high- and low-levels was significant, that is, the accuracy of high proficiency level was significantly better than that of the low proficiency level. These results could be partially explained by the way in which proficiency levels were coded in this experiment. Participants with an HSK 5 level were coded as high proficiency. Those who had an HSK 4 level were coded as intermediate proficiency. Finally, the low proficiency group included the participants who were preparing for the HSK 4 test. As several participants in the last group indicated in their language background questionnaires, they were going to take the test in one to a few weeks. Therefore, it was possible that some participants in this group had already achieved the intermediate level at the time of the experiment. This could explain why the repetition accuracy was not significantly different between the low proficiency and intermediate proficiency levels. Since these two groups constituted the major portion (23/29) of the participants, the lack of variance between these two groups might lead to the lack of an effect for proficiency. However, since there was evidence showing that high proficiency participants did perform better in the SBDs compared with the low proficiency group, it is reasonable to argue that proficiency, had it been measured better, would play a role.

Another limitation of this experiment is that there is a possibility that the effects are structure specific. All sentences in this experiment were constructed using the same pattern, with relative clauses occurring at both the matrix subject and matrix object positions. Since the processing of relative clauses itself is very complex, the possibility exists that this type of structure is more direction-sensitive. If this were the case, it would indicate that the findings of this experiment cannot be applied to all sentence structures. Therefore, replication studies on other sentence patterns are needed for a stronger claim about the directionality effect.

CONCLUSION

The reported experiment was conducted to investigate the effects of working memory and input direction on the use of sentence build-up drills. The results were straightforward, indicating that the backward input direction has several advantages over the forward direction. First, the backward direction elicited repetitions with fewer total and whole-sentence errors. Second, it also extended the tipping point, so that participants' performance was kept intact longer in this direction. Finally, participants with different WM levels performed equally well in the backward direction, whereas the low-WM group made significantly more errors in the forward direction, compared with their high-WM counterparts. Furthermore, the main effect of WM was not significant, which was consistent with several previous studies. In particular, since the more appropriate R-span test was used to measure participants' WM, the results from this experiment provide stronger evidence for the claim that WM has no significant effect on repetition performance.

Based upon these results, it is clear that backward build-up is more effective and that it is therefore a demonstrably better way to conduct sentence build-up repetition drills in the classroom.

REFERENCES

Ahmadian, M. J. (2015). Working memory, online planning, and L2 self-repair behavior. *Working memory in second language acquisition and processing*, *87*, 160.

Anderson-Hsieh, J., & Dauer, R. M. (1997). Slowed-Down Speech: A Teaching Tool for Listening/Pronunciation. Paper presented at the 31st Annual Meeting of the Teachers of English to Speakers of Other Languages, Orlando, FL

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255-278.

Bates D., Maechler, M., Bolker, B., & Walker, S. (2016). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-12,

Benwell, F. P., Mathieu, G., & Holton, J. S. (1961). Suggestions for Teaching Foreign Languages by the Audio-Lingual Method: A Manual for Teachers.

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition*, *33*(02), 247-271.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral & Brain Sciences*, 24(1), 87-185.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, *19*(4), 450-466.

Dauer, R. M., & Browne, S. C. (1992). Teaching the Pronunciation of Connected Speech. Paper presented at the 26th Annual Meeting of the Teachers of English to Speakers of Other Languagesm, Vancouver, British Columbia, Canada

DeKeyser, R. (2007). Introduction: Situating the concept of practice. *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*, 1-18. Cambridge University Press.

Deese, J., & Kaufman, R. A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of experimental psychology*, *54*(3), 180.

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33-68). Cambridge: Cambridge University Press.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in second language acquisition*,*27*(02), 141-172.

Ellis, R., Loewen, S. D., Elder, C., Erlam, R., Philp, J., & Reinders, H. (2009). Implicit and Explicit Knowledge in Second Language Learning, Testing and Teaching.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of experimental psychology: General*, *128*(3), 309.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied linguistics*, *27*(3), 464-491.

Hamayan, E., Saegert, J., & Larudee, P. (1977). Elicited imitation in second language learners. *Language and Speech*, *20*(1), 86-97.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, *44*(02), 137-166.

Kim, Y., Tracy-Ventura, N., & Jung, Y. (2016). A Measure of Proficiency or Short-Term Memory? Validation of an Elicited Imitation Test for SLA Research. *The Modern Language Journal*, *100*(3), 655-673.

Larsen-Freeman, D. (2012). On the roles of repetition in language teaching and learning. Applied Linguistics Review, 3(2), 195-210.

Larsen-Freeman, D., & Anderson, M. (2011). *Techniques and Principles in Language Teaching 3rd edition*. Oxford university press.

Munnich, E., Flynn, S., & Martohardjono, G. (1994). Elicited imitation and grammaticality judgment tasks: What they measure and how they relate to each other. *Research methodology in second language acquisition*, 227-243.

Muranoi, H. (2007). Output practice in the L2 classroom. *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*, 51-84. Cambridge University Press.

Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of experimental psychology*, *64*(5), 482.

O'Connor, P. (1960). Modern Foreign Languages in High School: Pre-Reading Instruction. Bulletin, 1960, No. 9. OE-27000. *Office of Education, US Department of Health, Education, and Welfare*.

Okura, E., & Lonsdale, D. (2012). Working memory's meager involvement in sentence repetition tests. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 2132-2137). Austin, TX: Cognitive Science Society.

Packard, J. L., & Qian, Z. (2016). A working memory explanation for recency effects in Mandarin second-language sentence processing. *Chinese Teaching in the World*, *30*(1), 75-100.

Paulston, C. B. (1971). The sequencing of structural pattern drills. *TESOL quarterly*, 197-208.

Paulston, C. B., & Bruder, M. N. (1976). Teaching English as a Second Language. Techniques and Procedures.

Perfetti, C. A., & Lesgold, A. M. (1977). Discourse comprehension and sources of individual differences.

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*(6), 633-654.

R Core Team. (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Santamaria, K., & Sunderman, G. (2015). 12 Working Memory in Processing Instruction: The Acquisition of L2 French Clitics. *Working memory in second language acquisition and processing*, *87*, 205.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent?. *Journal of memory and language*, *28*(2), 127-154.

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory*, *17*(6), 635-654.

van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, 29(3), 325-344.

Wen, Z. E. (2016). *Working memory and second language learning: Towards an integrated approach* (Vol. 100). Multilingual Matters.

Wen, Z., Mota, M. B., & McNeill, A. (Eds.). (2015). *Working memory in second language acquisition and processing* (Vol. 87). Multilingual Matters.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*(4), 497-528.