ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 14, No. 7, pp. 2002-2012, July 2024
DOI: https://doi.org/10.17507/tpls.1407.06

Exploring the Relationship Between Working Memory Capacity and L2 Oral Fluency

Danh C. Vu*
Faculty of Foreign Languages, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Thanh V. Nguyen
Faculty of Foreign Languages, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Nakhon Kitjaroonchai
Faculty of Arts and Humanities, Asia-Pacific International University, Muaklek Saraburi, Thailand

***Abstract*—This study investigated the relationship between working memory capacity (WMC) and L2 oral fluency and explored the maintenance of newly presented information in L2 speaking production. When performing speech production tasks, knowledge from long-term memory is at frequent use and hence involves the operation of the episodic buffer, one component of the working memory system, where information is retrieved from long-term memory and stored in chunks as we speak in chunks. In this study, we measured WMC by assessing chunk capacity and chunk size held in and retrieved from the episodic buffer. Chunk capacity was measured by Pair-word Speaking Task; chunk size was measured by Procedure Description Task; and oral fluency was measured by speech rate and mean length of run using an IELTS speaking task. Twenty-nine English-major students participated in this study. The results showed a strong positive correlation between chunk capacity and the two measures of fluency while chunk size negatively correlated with fluency. Data from the recall interviews revealed that participants employed various strategies for the maintenance of the presented information which involved different types of information binding.**

***Index Terms*—working memory capacity, episodic buffer, chunk capacity, chunk size, L2 oral fluency**

I. Introduction

It is understandable that anyone who learns English as a second or foreign language has a desire to speak English with fluency and that people are much different from each other regarding this aspect of oral performance (Daneman, 1991; Mostafa et al., 2020). However, the factors contributing to the desired fluency that make a fluent speaker different from a regular one have remained the subject of an ongoing debate in L2 research (Aslan & Şahin, 2020; Ngoc & Dung, 2020). Since the introduction of the multi-component model of working memory (Baddeley & Hitch, 1974), the construct has been widely exploited in different areas of research. In L2 research the construct of working memory has been used to explain learner differences in language comprehension and performance. One of the common claims about working memory in previous L2 studies which employed the dual-tasks for the assessment of working memory capacity, i.e. the Speaking Span Test & Reading Span Test (Daneman & Green, 1986), is that its capacity is the predictor of success in L2 oral production (Bergsleithner, 2011; Finardi, 2006; Rezai & Okhovat, 2016; Vieira, 2017; Weissheimer & Mota, 2009). The results of those previous studies showed a positive correlation between the storage capacity and oral fluency. However, a limitation of measuring working memory capacity using dual-tasks is its exclusive emphasis on the maintenance of individual words that are not related to each other. This may not accurately capture the nature of the functional working memory in authentic speaking tasks, as these tasks often involve the maintenance of related information. Moreover, the maintenance of information in the functional working memory during oral production has been under-researched, and thus, much is unknown about how an L2 speaker manages to maintain the information to produce speech. The present study first sought to validate the claim about WMC as a predictor of L2 oral fluency. It also explored the mechanisms involved in retaining newly introduced information during L2 oral production, referred to hereafter as "the input." Furthermore, new assessment tools were developed to accurately measure working memory capacity based on real-life speaking events.

II. Literature Review

A. Models of Working Memory

Working memory is referred to as a limited capacity "brain system" that temporarily stores and manipulates necessary information for the performance of such complex cognitive activities as comprehension, learning, and

* Corresponding Author (https://orcid.org/0009-0007-8518-6818; Email: vucongdanh@tdtu.edu.vn)

reasoning (Baddeley, 1992, p. 281). With that conceptualization, Baddeley and Hitch (1974) originally proposed the model of working memory that comprises the three-component system and that simultaneously stores and processes information. The three components are (1) the Central executive, which is capable of focusing attention, storing information, and making decisions, and two active slave systems, namely (2) the Visuospatial Sketchpad, which holds and manipulates visuospatial information, and (3) the Phonological Loop, which keeps speech-based information. The original model of working memory by Baddeley and Hitch (1974) received a great deal of attention from other researchers in the field as it primarily emphasizes the active manipulation of information rather than passive storage, which distinguishes it from earlier short-term memory models (Atkinson & Shiffrin, 1971; Miller, 1956). This model was adopted by many researchers studying span and loop capacity and its influence on language (Daneman & Carpenter, 1980; Daneman & Green, 1986; Just & Carpenter, 1992; Just et al., 1996). Different from other researchers, Engle and others (Conway & Engle, 1996; Towell et al., 1996; Engle et al., 1999) argued that working memory is a unitary, domain-free construct that is strongly related to general fluid intelligence. They highlighted the importance of the focus of attention rather than domain-specific processes. Unlike Baddeley's model, Cowan (1999, 2005) proposed an embedded-processes model of working memory which emphasizes the connection between long-term memory and attention. Cowan's (1999) model comprises four elements including central executive, long-term memory, activated memory, and the focus of attention, the core of working memory. Cowan's model tackles the issue of access to long-term memory, which Baddeley's original model fails to explain why complex cognitive tasks can be done without accessing the resources from long-term knowledge.

It was not until 2000 that Baddeley's model was substantially revised, with the proposal of a new component of the working memory system, the 'episodic buffer'. In this revised multi-component model of working memory, the episodic buffer is assumed to hold integrated episodes or chunks in multidimensional codes. It, therefore, acts as a buffer store between the components of working memory, and it links working memory to perception and long-term memory. However, it is noted that the episodic buffer is capacity-limited and that the retrieval from the buffer occurs through conscious awareness (Baddeley, 2012). Baddeley's (2000) influential model of working memory is believed to be easier to get hold of and to help explain many important phenomena, e.g. individuals with short term memory deficits are capable of storing and manipulating complex information in the limited-capacity slave systems. Cowan (2005) also considered this model as a basis on which "many predictions have been made and tested" (p. 21). Hence, this model serves as a theoretical framework for the present study.

B. Measuring Working Memory Capacity

Over the past two decades, a plethora of L2 research has been conducted to investigate the relationship between working memory capacity and L2 speech production (e.g. Ahmadian, 2012; Finardi, 2008; Finardi & Prebianca, 2006; Mizera, 2006; Mota, 2003; Prebianca et al., 2014; Rezai & Okhovat, 2016; Tavares, 2016) mainly by using the dual-tasks which are assumed to simultaneously measure the processing and storage functions of working memory. Among the most frequently-employed measures are the Speaking Span Test (Daneman & Green, 1986) and the Reading Span Test (Daneman & Green, 1986). These kinds of tests assess the capacity of working memory by counting the number of sentences produced by the speakers after they are presented with sets of individual words. In doing this, the tasks require the coordination between the storage and processing functions of working memory (Daneman & Green, 1986). A speaker, therefore, needs to plan what to speak and temporarily store lexical input until readily articulate them in sentences. Researchers argued that L2 learners with larger working memory capacity tended to outperform the ones with lower spans in speaking tasks due to their greater efficiency in spoken production processes (Finardi & Prebianca, 2006), whereas Daneman and Tardif (1987) posited that the greater a learner's efficiency in processing information, the greater the capacity left available for storage of the products of the processing and of material retrieved from the long-term memory.

The present study supports the view that working memory is a task-specific capacity (Daneman & Green, 1986; Daneman, 1991; Fortkamp, 1999) which means WMC in speech production varies across tasks and differs from other cognitive tasks such as language comprehension, learning, and reasoning. Furthermore, in performing spoken tasks, knowledge from long-term memory seems to be at frequent use and hence involves the function of the episodic buffer where information is retrieved from long-term memory and stored in chunks as we speak in chunks (Brown, 1974). Therefore, it is argued that the capacity of the episodic buffer (EBC) is an indicator of WMC in oral production. In the present study, we assessed WMC by measuring the capacity of the episodic buffer or the number of chunks as well as the size of the chunk maintained and retrieved during L2 speech production.

C. L2 Oral Fluency

The definition of L2 fluency may vary across different studies. It can encompass various aspects such as vocabulary range, grammatical correctness, pronunciation, idiomaticness, appropriateness, and relevance of speech, or alternatively, or alternatively it may refer only to the speed of oral production without hesitations (Lennon, 1990). In this study, we focus on two categories proposed by Segalowitz (2010): cognitive fluency and utterance fluency. Cognitive fluency refers to an individual's ability to plan and deliver their speech while utterance fluency pertains to the actual performance during speaking (Fortkamp, 1999). Perceived fluency is not relevant for this study as it reflects the listener's judgment rather than cognitive abilities.

Previous research has examined utterance fluency by analyzing various speech characteristics. For instance, a study conducted by Bui and Huang (2016) investigated different aspects of L2 oral fluency using 18 measures related to speed, stretch, voicing, pauses within clauses (mid-clause pauses), independent clause pauses, dependent clause pauses, filled pauses and repairs. In this study, the researchers adapted the speed-related measures from Bui and Huang's work and recall interviews for assessing cognitive and utterance fluency during an L2 speaking task. The purpose was to explore if there is a correlation between working memory capacity and L2 oral fluency while also examining how input is maintained in L2 speech production. To address these objectives, the following research questions were posed:

1) Are there any relationships between WMC in a speaking task as indicated by the episodic buffer capacity and L2 oral fluency as measured by the speech rate and mean length of run?

2) How do the participants manage to maintain the input for L2 speech production?

III. Methodology

A. Setting and Participants

The research took place at Asia-Pacific International University, a private university situated in Saraburi province, Thailand. The study involved 29 Thai undergraduates majoring in English, aged between 19 and 23 years old. Their English proficiency levels ranged from low intermediate to advanced (A2 to C1 on the CEFR scale). Out of these participants, there were 25 females and five males. All participants were informed of the purpose of the study and provided verbal consent to participate voluntarily. Measures were taken to protect participants' privacy and anonymity in data analysis and reporting. Data collection was carried out during a term break and after final exams had been completed, ensuring that exam results did not influence the findings of the study.

B. Research Instruments

The study employed three different instruments: (1) an IELTS Speaking Part 2 Task, (2) a Pair-word Speaking Task (PST), and (3) a Procedure Description Task (PDT). In the IELTS speaking task, participants were given a topic and allotted one minute to prepare for a two-minute monologue aimed at evaluating their proficiency in L2 oral fluency. To assess the capacity of the episodic buffer in working memory, both PST and PDT tasks were utilized. The former measured chunk capacity whereas the latter measured chunk size that could be stored in the episodic buffer.

The PST was developed based on the Speaking Span Test proposed by Daneman and Green (1986), with a significant modification in terms of using paired words to measure chunks. In the PST, participants were presented with sets of seven pairs of words, intentionally selected to have semantic or phonological associations. The word pairs were simultaneously displayed both visually and audibly on a computer screen for two seconds each. After each set, participants were asked to produce one grammatically correct and meaningful sentence for each pair of words they could recall. Oral production of a sentence using the words in the same pair can be regarded as the evidence for the binding processes. In this study, these include both semantic and phonological bindings.

The PDT was adapted from the story recall task developed by Kapikian and Briscoe (2012). In the PDT, participants were presented with five steps of a procedure in which each step was presented in a phrase, one step at a time. Each step consistently consisted of five words, comprising three content words and two function words. The steps were presented in both written and spoken forms. After viewing the steps, participants were asked to retell the procedure while maintaining as many original words as possible. The syntactic units in the steps produced should be grammatically correct and semantically acceptable. The assumption is that reproducing the procedural steps may indicate an above-sentence level of binding process beyond just having pair-words construct sentences in PST.

For the designs of PST and PDT, a list of words was compiled with the following criteria: all of the words were taken from the Cambridge KET Vocabulary list (CEFR A2), they were one-syllable words and they were all content words.

To further investigate the binding process as possibly observed in the participants’ performance in the PST and PDT, we employed recall interviews. While PST and PDT can show what participants held in their working memory, they do not offer much insight into how the participants stored this information. Recall interviews serve as a valuable method for exploring this matter for two reasons. Firstly, by conducting interviews, we can identify any words or information that participants remembered beyond what was assessed in the tests. Secondly, we can gain a better understanding of how specific words or information were either forgotten or retained during and after the speaking tasks These insightful data can only be gathered from the perspectives of the participants themselves and may not necessarily reflect actual processes occurring within their working memory; nonetheless, they have potential to illuminate the strategies employed by participants to maintain the input.

C. Data Collection Procedure

The data collection took place in two days. The first task, PST, took approximately 25 minutes and the second task, PDT, took another 10 minutes for each participant. The recall interviews were conducted immediately after the participants finished doing each word set in the PST and at the end of the PDT. Generally, the participants were interviewed about: (1) What word(s) they remembered but could not manage to make a sentence and (2) how they managed to remember the presented pairs of words. One of the techniques we used in the recall interviews was to tell the participants a random word from the pairs that they failed to make a sentence. This was done to assess their ability to remember the other word in the pair. Then we proceeded with questioning them about the ways that they could recall the other word of the pairs, otherwise, asking them why they could not. It is noted that while participants had the option of being interviewed in Thai or English, only two of them preferred conducting their interviews in Thai. The IELTS speaking task was subsequently administered after the interviews.

D. Data Analysis

Data obtained in the IELTS speaking task were transcribed verbatim and calculated for (1) the speed rate (SR) which is the total number of words per minute after the deletion of vocal fillers, incomplete words, and repeated words and (2) the mean length of run (MLR) which is the average number of words before the participants encountered any vocal filler or incomplete word or repetition or unnatural pause of more than 2 seconds. These measures of L2 oral fluency were adapted from Bui and Huang (2016). For the calculation of SR and MLR, refer to the formula below.

$$SR = \frac{(total\ words\ produced\ -\ vocal\ fillers\ -\ incomplete\ words - repeated\ words)x\ 60}{total\ time\ of\ speech\ (in\ second)}$$

$$MLR = \frac{\sum words\ in\ all\ stretches}{number\ of\ stretches}$$

In the PST, every word reproduced by the participants was given strict or lenient scores. Strict scores were given when the two words of each pair were reproduced exactly as originally presented to them. However, lenient scores will be given if the reproduced words are synonyms or derivational forms of the words presented in the word set. Additionally, in all circumstances, the words should be used in (1) grammatically correct sentences (if not, the score is reduced half) and (2) semantically acceptable (according to native speaker's norms).

Regarding the PDT, the measure was the chunk size (CS) which is the total number of content words in a step. Content words can be reproduced in a derived form, however, the syntactic units in the steps reproduced should be (1) grammatically correct (if not, the score is reduced half) and (2) semantically acceptable.

The interviews were transcribed and analyzed for the contents by two researchers (Guba & Lincoln, 1994). The coding process was carried out to reduce data into easily locatable segments. Three researchers performed the cross-checking of data coding to increase the level of inter-rater reliability and agreement on data coding was reached among the researchers before any further categorization of the data. The analysis aimed to explore the participants' memory strategies used during the performance of PST and PDT in order to seek explanations for the participant's performance in the tasks as well as to explore the quality of the maintenance of the input in the speaking tasks. The analysis was guided by the two following questions:

How did the participant manage to remember the words for speech production?

What are the words they could remember but failed to produce when doing the tasks?

IV. RESULTS

Twenty-nine English majored students performed in three different tasks and the results of which were rated manually by the three researchers and then processed by SPSS for further statistical analysis. The first task PST measured the chunk capacity as indicated by the strict scores (PSTS) and lenient scores (PSTL). The second task PDT measured the chunk size (CS). The IELTS speaking task measured L2 oral fluency as indicated by the speech rate (SR) and the mean length of run (MLR). Table 1 shows the means and standard deviations of each measure in the three tasks.

TABLE 1
DESCRIPTIVE STATISTICS OF PST, PDT AND IELTS SPEAKING TASK (N = 29)

Task	N	Min.	Max.	Mean	Std. Deviation
PSTS	29	1.00	18.0	8.4138	4.11024
PSTL	29	2.00	9.00	5.2241	1.65608
CS	29	0.20	3.00	1.3517	0.63898
SR	29	5.00	.90	13.5704	19.27698
MLR	29	.18	.33	9.5214	11.66742

Note. PSTS = pair-word speaking task strict scores; PSTL = pair-word speaking task lenient scores; CS = chunk

As shown in Table 1, the strict scores of PST range from 1.00 to 18.00 (*M* = 8.41, *SD* = 4.11) and the lenient scores range from 2.00 to 9.00 (*M* = 5.22, *SD* = 1.65). The chunk sizes (CS) range from 0.20 to 3.00 (*M* = 1.35, *SD* = 0.64). The speech rates range from 75.00 to 151.90 words per minute (*M* = 113.57, *SD* = 19.28), whereas the means length of run range from 7.18 to 50.33 words per stretch (*M* = 19.52, *SD* = 11.66).

To address our research question 1: Are there any correlations between the measures of working memory capacity (WMC) and the measures of L2 oral fluency? Spearman's correlation coefficient (ρ) was employed. Table 2 shows how the WMC measures correlated with the L2 oral fluency.

Regarding the relationship between chunk capacity and oral fluency measures, it was found that the lenient measurement of chunk capacity (PSTL) positively correlated with both measures of oral fluency, evident by a strong correlation between PSTL and speech rate (ρ=.732, P=.000) and a moderate correlation between PSTL and mean length of run (ρ=.506, P=.005). In contrast, the strict measurement of chunk capacity negatively corresponded with the speech rate, despite its relatively low correspondence (ρ=-.449, P=.015).

Interestingly, the statistical analysis of the measurement of chunk size and speech rate also revealed a low, negative correlation (ρ=-.411, P=.027) but showed no significant relationship with the length of run. Thus, it appeared that the size of the maintained chunk tended to negatively respond to the number of words the participants could produce per minute.

TABLE 2
CORRELATIONS BETWEEN THE MEASURES OF WMC AND THE MEASURES OF L2 ORAL FLUENCY (N = 29)

Spearman's rho		SR	MLR
PSTS	ρ	-.449*	-.180
	Sig. (2-tailed)	.015	.350
	N	29	29
PSTL	ρ	.732**	.506**
	Sig. (2-tailed)	.000	.005
	N	29	29
S	ρ	-.411*	-.074
	Sig. (2-tailed)	.027	.705
	N	29	29

Note. PSTS = pair-word speaking task strict scores; PSTL = pair-word speaking task lenient scores; CS= chunk size; SR = speech rate; MLR = mean length of run. ***P < .01, significant correlation. *P < .05, significant correlation*

In Table 3, simple linear regression showed a significant relationship between PSTL and speech rate (F(1,27) =20.324, P < 0.001). The slope coefficient for PSTL was 7.628. The R^2 *value* was .429, so it could be suggested that 42.9% of the speech rate can be explained by the model of PST with lenient measurement, and 57.1% could be explained by other factors that contributed positively to the speech rate. This indicated that PSTL can be a good predictor of L2 oral fluency in terms of speech rate.

TABLE 3
REGRESSION RESULTS

Variable	B	SE	β	t	Sig.
Predictors (constant)	**73.720**	9.258		7.962	**.000**
PSTL	**7.628**	1.692	.655	4.508	**.000**
R^2	.429				
F	20.324				

Note. Dependent variable: Speech rate

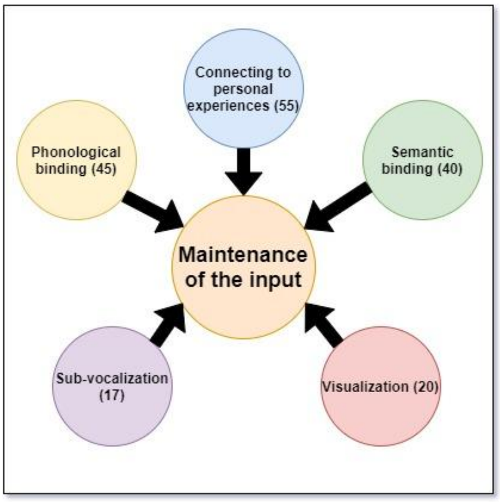

Figure 1. Strategies Used for Maintaining the Input in the PST

To respond to Research Question 2: *How do L2 speakers maintain the input for speech production?*, which examines how L2 speakers retain information for speech production, immediate recall interviews were carried out immediately after participants finished each set of pair-word speaking tasks and procedure description tasks. The data was transcribed and analyzed following the methods described in the previous section. Figure 1 illustrates the five primary strategies used by participants to maintain the information presented during the first task, the PST. The frequency count of mentions for each strategy is provided in brackets.

A. Connecting to Personal Experiences

As found in the analysis of the recall interviews, there were 55 times the participants mentioned that they benefited from their own experiences. This means that for maintaining the information presented to them in a very short time many of the participants tried to link these pieces of information with their personal experiences. The following excerpts illustrate this most frequently mentioned strategy. The two participants explained how they managed to maintain the pair-words *guest – room* and *field – trip* during the task:

> I can remember the last pair well because I used to work in the university guestroom before. (E13-S6)
>
> During I studied here in the first and second year we went on a field trip, so I remember that word. (E22-S5)

B. Phonological Binding

The second frequent strategy was to maintain the pair-words by creating a connection based on the phonological similarity of the words rather than meaning. The following excerpts demonstrate some mentions of this strategy:

> I can remember these pairs because they are similar in sound like cheap and chips even though I'm not quite sure at first about the meaning. (E10-S3)
>
> The ship and the sheep is like almost the same. It sounds the same. (E20-S5)

C. Semantic Binding

Alternatively, many participants described a strategy that helped connect two words by their meaning rather than their phonological features. These words are often co-occurring items in discourse. The following excerpt shows a typical mention of this strategy: “When I saw the words in pairs and they go together like the sun set, or the boys go with toys, the road goes with the map and they are easy to remember” (E20-S7).

D. Visualization

Some participants reported that they visualized the presented verbal information in order to hold it for later spoken production. The following excerpts illustrate these instances.

> I remember these words because I pictured it in my mind, like a town people feed a cow. (E09-S2)
>
> I think I make a picture in my mind about the beach and when activities there always in the beach is sports. (E22-S1)

E. Sub-Vocalization

The least mentioned strategy was the use of sub-vocalization of the input. To sub-vocalize the input, the participants simply repeated the presented information in mind. Below are some participants’ descriptions of this strategy.

> I try to repeat. I have to spell the word like ‘red’ and ‘green’ in mind but without noise. (E28-S5)
>
> Normally I choose the words that are short and easy or the ones that I already know the meaning to repeat in my mind. When I do this I can make a sentence easily. (E26-S6)

The procedure description task (PDT) required the subjects to describe five steps of a procedure of planning for a summer vacation by using five phrases that were previously presented in sequential order. From the analysis, it was found that the participants also employed some of the strategies they used for the pair-word speaking task when performing PDT. Figure 2 illustrates the four underlying strategies the participants used to complete the task.

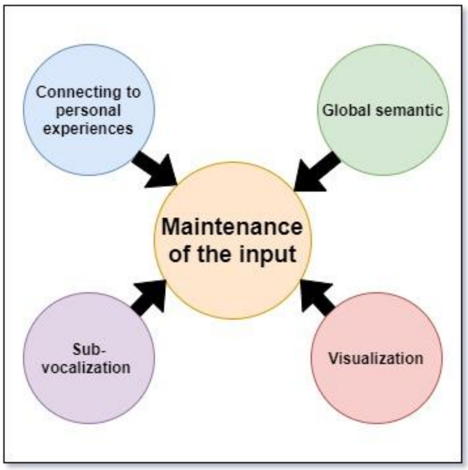

Figure 2. Strategies for Maintaining the Input in the PDT

Like the PST, the participants revealed that they made use of their personal experience connection, visualization, and sub-vocalization strategies. For example, some of the participants explained that the task was somewhat related to their past experiences when they had to go on a tourism class field trip in which all the students in the class were required to collaboratively plan for the trip and activities. In the following excerpt, a participant described the use of this strategy:

> I tried to recall these phrases by closing my eyes and think about my task in my tourism class this semester. I think to link it with my tourism class field trip when we were asked to plan for a class field trip together. We

need to go through many steps in making our plan which is similar to the task here. (E05-T2)

At least four of the participants stated that they picturized the verbal input and organized them in a situation or a story in order to assist them in maintaining the information. This excerpt depicts such strategy use: "I tried to picture things like I used to do and when I saw these phrases I tried to make picture of the situation and put it as a story" (E10-T2).

F. Global Semantic Structuring

The most common strategy used in the PDT was global semantic structuring, a derived strategy of semantic binding. The assumption of global semantic structure is that language comprehension processes are hierarchically constructed, that is to say, words comprising sentences are bound together and an overarching structure binds together groups of sentences in order to create global meaning in relation to propositional content (Kapikian & Brisco, 2012). Sentences or phrases that have causal event structure or presented in intact order can be better recalled or bound than the same constituent sentences that are presented in a different order (Christoffels, 2006). In the PDT the participants explained that they could retrieve the presented information because those phrases came in sequential order. The following excerpts demonstrate such strategy use.

> They already list the steps for us. I just see the main point of the sentence and I will put it together. I notice these phrases are in order and after you know the first one, you will know the second one and the third one. (E23-T2)
>
> I can remember these three phrases because they are the process of going for a trip and when we plan a trip we need to follow the steps. (E01-T2)

In brief, the results from the immediate recall interviews suggested that the participants in this study used a range of strategies to enhance the retention of information, some of which were common across both tasks. Furthermore, it is argued that these strategies may not be exclusively used at once but can be simultaneously exploited for maintaining the input. In the next part of the paper, we will provide some explanations for the results that have been discussed in this part.

V. Discussion

A. Chunk Capacity and L2 Oral Fluency

The concept of chunking in this study refers to the cognitive process of binding individual pieces of information together and storing them as a cohesive unit. This chunking process is an integral part of the comprehension process in which meaning is built by filtering and binding the extracted pieces of information so that they fit with the mental structure (Baddeley et al., 2009; Gernsbacher, 1990). What is different from the structure-building model of comprehension (Gernsbacher, 1990) is that chunking information in our study does not necessarily imply only semantic binding but also involves all types of binding, ranging from semantic, phonological, visual, spatial, experiential, perceptual et cetera bindings. According to Baddeley's (2000) model of working memory, these bound pieces of information, or chunks, are stored in multidimensional codes in the episodic buffer which prepares the resources for speech production. Our results showed that there was a positive close correspondence between the participants' chunk capacity, measured by the lenient scores, and how fast they can produce L2 speech, measured by the speech rate. There was also a moderate positive correlation between the number of chunks the participants could retain and the average number of words they could produce before they encountered any hesitation in speech production. One explanation for the observed relationships is that chunking can reduce the processing demands and promote the overall fluency because chunks are easier to remember and faster to retrieve and produce as compared to recalling individual pieces of information separately (Newell, 1990; Taguchi, 2008). As mentioned earlier, binding the individual units of information together and maintaining these units as chunks can allow more information to be stored, which caters more ready-to-use resources for oral production (Taguchi, 2008). As Pawley and Syder (1983) suggested, fluent speakers make use of 'ready-made memorized chunks' because they help to lower the load of information processing in speaking and thus would allow more space for other processes. Many researchers also supported the idea that improving chunk capacity and efficiency can improve L2 oral fluency (Amir Mahdavi & Zafarghandi, 2015; Schmidt, 1992; Shen, 2015; Towell et al., 1996). Our study seems to support that view. As evidenced from the regression statistical analysis we believe that chunk capacity, or the number of chunks stored in the episodic buffer, can facilitate L2 oral fluency and it can be a reliable predictor of L2 oral fluency.

Interestingly, however, in this study we measured the chunk capacity in a speech production task in two distinct ways, reflecting the strict and lenient ways of maintaining the chunks, and only the lenient way of maintaining the chunks positively correlated with the measures of speech fluency while the other, the stricter way of maintaining the chunks, negatively corresponded with the participants' oral fluency. This means that maintaining the bindings of the input in a strict, accurate manner may be responding to less fluent speech production. Our assumption is that maintaining the bindings in such strict and accurate way may require more attentional resources (Baddeley, 2012) and thus may leave less capacity for other cognitive processes of speech production, often referred to as the 'trade-off effects' (Daneman, 1991; Finardi & Prebianca, 2006; Fortkamp, 1999; Skehan, 1998). This trade-off effect is based on the view that working memory is a dual function system, consisting of processing and storage. Since the two functions compete for the limited capacity of working memory, a greater effort dedicated to the strict maintenance of bindings means that less space is spared to the processing function. Therefore, trading off the capacity for maintenance may lead to a decrease in processing capacity, meaning a slowdown in speech production.

B. Chunk Size and L2 Oral Fluency

In the present study, we examined not only the chunk capacity but also the chunk size. In the PDT, we aimed to measure the size of the chunk and we were particularly interested in finding out if the chunk size had any relationship with the measures of oral fluency. The results revealed that chunk size had a low correlation with the speech rate but what was surprising was that it was a negative correlation, meaning that maintaining a large size of a chunk is not a good thing to oral fluency. The observed phenomenon can be explained by Newport's (1990) 'less is more hypothesis'. Newport (1990) claimed that the difference in language acquisition between children and adults lies in the differences in their working memory capacity. Children's low capacity leads them to processing language input in its minimal components thus allowing them to easier maintain more accurate and meaningful chunks rather than larger but misleading chunks. This hypothesis is supported by several studies such as Chin and Kersten (2010), Jones (2012), Kersten and Earles (2001), and Smalle et al. (2016), to name just a few. Our findings also suggested that the larger the size of the chunk to be maintained tended to be responding to the less fluency in L2 speech production, which fits well with Newport's (1990) less-is-more theory.

C. The Maintenance of the Input in L2 Speech Production

Regarding the research question about the maintenance of the input in L2 speech production, the data from the interviews provided evidence for the chunking process which assisted the participants in the maintenance of the input. As discussed in the results section, participants reported the various types of binding that they used to maintain the information when they performed in the speaking tasks. These included phonological, semantic, and visual binding as well as using personal experiences and repeating the pairs of words. These bindings of information preparing necessary resources for later speech production can be explained with reference to the different functions of working memory's components.

Firstly, phonological binding was most frequently used in the first task, the PST, in which the maintenance of the word-pairs was achieved by binding the two words based on their phonological similarity. As Baddeley (2012) stated in his multi-component model of working memory, phonological similarity is one of the fundamental characteristics of the phonological loop. Likewise, participants also retained the information by chunking it based on the semantic similarity of the input. The results from the PDT suggested that participants could benefit even greater in the maintenance of the input when the information was presented semantically connected with a theme and in a meaningful order. This finding was in line with Kapikian and Briscoe (2012) whose study concluded that semantic binding could generate stronger memory representations than words or texts that lack global semantic structure.

Participants also maintained the input by connecting it to their personal experiences. This was done by accessing the long-term memory where personal knowledge and experiences are stored. As noted in the model of working memory, this access to long-term memory can be made possible through the episodic buffer which links perception, knowledge, and personal experience with the working memory system. The binding itself may not take place in the buffer, however, the bound pieces of information or the resulted chunks, are stored there (Baddeley, 2012).

Some participants reported the use of visualization in the maintenance of the input. As mentioned earlier, the visual system is responsible for processing and storing visual and spatial information (Baddeley, 2000). In this case, it is interesting that the presented verbal information was processed and converted into visual codes and bound together. This was probably done by the visual system before being sent to and stored in the episodic buffer.

Other participants told us that they repeated some pair-words in mind to maintain them for later speech production. The mechanism underpins this strategy of information maintenance largely relies on the function of the phonological loop. As mentioned earlier in the model (Baddeley, 2000), the phonological loop has two sub-systems, which are the phonological store and the articulatory control process. While the former is responsible for storing information in speech-based codes, the latter is responsible for rehearsing information from the phonological store by subvocalizing the articulatory codes. These two sub-systems work together as the inner ear and inner voice to maintain the subvocalized information for speech production.

In addition, one interesting observation from the interviews was that some participants could recall some of the pair-words that they failed to do earlier in the task. We found out that when they were given the first word in a pair, in many cases, they could tell us the remaining word in that pair which they insisted to be forgotten during the task performance. One possibility is that the maintenance of the bound information rather than the binding process requires additional attention from the central executive (Baddeley, 2012), and hence maintaining the information and retrieving it from the episodic buffer is restricted to the central executive control. In other words, the participants did store the bound information in their episodic buffer, however, because the central executive did not allocate attentional resources for the retrieval of the stored chunks but for other cognitive processes for speech production, which resulted in the failure to retrieve the maintained chunks. Until later, when the burden of performing the task was over, the central executive reconnected with the episodic buffer and the retrieval of the maintained chunks was made possible again.

VI. Conclusion

The results of this study have reinforced the claim about the close relationship between working memory capacity and L2 oral fluency. Moreover, it was found that working memory capacity indicated by different chunk measures exhibited various types of relationship with L2 oral fluency. Firstly, since chunk capacity closely relates to speech fluency, the learning of chunks should be promoted in L2 speaking classrooms. Instead of introducing a list of individual, unrelated words, it is suggested that teachers should present the input in ways that facilitate the students' chunking process as this will benefit both the maintenance of the presented resources and speech production. What is more, since the chunking process can happen not only when there are semantic similarities but phonological and visual resemblances or experiential connections as well, the practice of chunking all possibly noticeable features of the presented information should be advocated. Secondly, however, it is noteworthy that because only the lenient way of maintaining the chunks can predict the fluency, using chunks in a flexible, creative, personalized way should be encouraged among L2 students. Especially in fluency-based activities, the importance of maintaining the input in a flexible way needs to be highlighted. More importantly, our study revealed that although learning many chunks is necessary, maintaining large-sized chunks might not be a good thing for oral performance. It is, therefore, whilst we recommend the practice of chunking in L2 speaking, presenting long and complicated chunks to students might be avoided, especially for less proficient students, as it can add an extra burden on the maintenance of the input in their oral production.

One limitation of this study is that L2 oral fluency was only measured in terms of quantity, i.e. speech rate and length of run, but the language quality of the speech was completely neglected. It is thus recommended that future studies may include accuracy measures of the speech so that it can reflect fluency in a more meaningful way. In addition, the group of participants in the study was homogeneous in terms of their academic discipline and English proficiency, consisting of only English majored students. Future research can also cover a wider range of participants with various academic disciplines and levels of English proficiency, for example, recruiting non-English major students. Another limitation of the study is that the chunking effect was only validated by the data from the retrospective interviews. In fact, in addition to the experimental condition, we originally intended to have a control condition for the two tasks measuring WMC, which could assist to claim the chunking effect. However, we failed to have enough participants for the two conditions. The control condition for the PST can be done by rearranging the words in each word set so that the words in the same pair do not share semantic or phonological similarity any longer. Likewise, the steps in the PDT can be rearranged to eliminate the effect of the sequential, logical order of the procedure. It is highly recommended that future research may need to include a control condition for the validation of the chunking effect as a useful way to back up the qualitative data. Finally, the present study is a correlational study in which we found a prediction, hence, future researchers can conduct an experimental study to explore the causal relationship between WMC and L2 oral fluency so that more pedagogical actions can be suggested and implemented.

References

[1] Ahmadian, M. J. (2012). The relationship between working memory capacity and L2 oral performance under task-based careful online planning condition. *TESOL Quarterly*, *46*(1), 165-175.

[2] Aslan, R., & Şahin, M. (2020). 'I feel like I go blank': Identifying the factors affecting classroom participation in an oral communication course. *TEFLIN Journal - A Publication on the Teaching and Learning of English*, *31*(1), 19. https://doi.org/10.15639/teflinjournal.v31i1/19-43

[3] Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American*, *225*(2), 82-91.

[4] Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 8, pp. 47-89). New York, NY: Academic press.

[5] Baddeley, A. D. (1992). Working memory: The interface between memory and cognition. *Journal of cognitive neuroscience, 4*(3), 281-288.

[6] Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, *4*(11), 417-423.

[7] Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2009). Working memory and binding in sentence recall. *Journal of Memory and Language*, *61*, 438-456.

[8] Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual review of psychology, 63*, 1-29.

[9] Bergsleithner, J. M. (2011). The role of noticing and working memory capacity in L2 oral performance. *Organon*, *26*(51). https://doi.org/10.22456/2238-8915.28841

[10] Brown, D. F. (1974). Advanced vocabulary teaching: The problem of collocation. *RELC Journal*, *5*(2), 1-11.

[11] Bui, G., & Huang, Z. (2016). L2 fluency as influenced by content familiarity and planning: Performance, measurement, and pedagogy. *Language Teaching Research*, *22*(1), 94-114.

[12] Chin, S. L., & Kersten, A.W. (2010). The application of the less is more hypothesis in foreign language learning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 150-155). Austin, TX: Cognitive Science Society.

[13] Conway, A. R. A. & Engle, R.W. (1996). Individual differences in working memory capacity: More evidence for a general capacity theory. *Memory, 4*, 577-590.

[14] Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62-101). New York, NY: Cambridge University Press.

[15] Cowan, N. (2005). *Working memory capacity*. East Sussex, UK: Psychology Press.

[16] Daneman, M., & Carpenter, P.A. (1980). Individual differences in working memory and reading. *Journal of Learning and Verbal Behavior*, *19*, 450-466.
[17] Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language*, *25*, 1-18.
[18] Daneman, M., & Tardif, T. (1987). Working memory and reading skill re-examined. In M. Coltheart (Ed.), *Attention and performance XII* (pp. 491-508). London: Erlbaum.
[19] Daneman, M. (1991). Working memory as a predictor of verbal fluency. *Journal of Psycholinguistic Research*, *20*, 445-464.
[20] Engle, R. W., Kane, M.J., & Tuholski, S.W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In: A. Miyake & P. Shah (Eds.), *Models of working memory Mechanisms of active maintenance and executive control* (pp. 102-134). Cambridge, UK: Cambridge University Press.
[21] Finardi, K. (2008). Working memory capacity and L2 speech production in a picture description task with repetition. *Revista de Estudos Da Linguagem*, *16*(2). https://doi.org/10.17851/2237-2083.16.2.129-144
[22] Finardi, K., & Prebianca, G. V. V. (2006). Working memory capacity and speech production in L2: evidence from a picture description task. *Revista de Estudos da Linguagem*, *14*(1), 231- 260.
[23] Fortkamp, M. B. M. (1999). Working memory capacity and aspects of L2 speech production. *Communication & Cognition*, *32*, 259-296.
[24] Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Lawrence Erlbaum Associates, Inc.
[25] Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. *Handbook of qualitative research*, *2*(163-194), 105.
[26] Jones G. (2012). Why Chunking Should be Considered as an Explanation for Developmental Change before Short-Term Memory Capacity and Processing Speed. *Frontiers in psychology*, *3*, 167. https://doi.org/10.3389/fpsyg.2012.00167
[27] Just, M. A., & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122-149.
[28] Just, M. A., Carpenter, P.A., & Keller, T.A. (1996). The capacity theory of comprehension: New frontiers of evidence and arguments. *Psychological Review*, *103*(4), 773-780.
[29] Kapikian, A., & Briscoe, J. (2012). Semantic binding, not attentional control, generates coherent global structure in children's narrative memory. *Journal of Cognitive Psychology*, *24*(6), 751-764.
[30] Kersten, A. W., & Earles, J. L. (2001). Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language*, *44*(2), 250-273.
[31] Lennon, F. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*, 387-417.
[32] Mahdavi-Zafarghandi, A., Tahriri, A., & Bandari, M.D. (2015). The impact of teaching chunks on speaking fluency of Iranian EFL learners. *Iranian Journal of English for Academic Purposes*, *1*(4), 36-47.
[33] Miller, G. A. (1956). The magical number seven plus-or-minus two: Some limits of our capacity for processing information. *Psychological Review*, *63*, 81-97.
[34] Mizera, G. J. (2006). *Working memory and L2 oral fluency*. Unpublished doctoral dissertation, University of Pittsburgh. Retrieved December 05, 2023, from http://d-scholarship.pitt.edu/7655/1/Mizera.Dissertation.pdf
[35] Mota, M. B. (2003). Working memory capacity and fluency, accuracy, complexity, and lexical density in L2 speech production. *Fragmentos: Revista de L ńgua e Literatura Estrangeiras*, *24*, 69-104.
[36] Mostafa, T., Kim, Y., & Friginal, E. (2020). Examining a developmentally based measure of L2 oral performances: Does it predict L2 learners' oral proficiency? *System*, *89*, 102197. https://doi.org/10.1016/J.SYSTEM.2019.102197
[37] Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
[38] Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28.
[39] Ngoc, D. T. B., & Dung, T. T. (2020). Key factors influencing learners' oral fluency in English speaking classes: a case at a public university in Viet Nam. *VNU Journal of Foreign Studies*, *36*(6), 93. https://doi.org/10.25073/2525-2445/vnufs.4631
[40] Pawley, A., & Syder, H. (1983). Two puzzles for linguistic theory: Native-like selection and native-like fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191-226). London: Longman.
[41] Prebianca, G. V. V., Finardi, K. R., & Weissheimer, J. (2014). Working memory capacity across L2 speech proficiency levels. *Revista Brasileira de Lingu śtica Aplicada*, *14*(2), 441–462. https://doi.org/10.1590/S1984-63982014005000009
[42] Rezai, M. J., & Okhovat, B. (2016). The Effect of Working Memory on EFL Learners' Oral Fluency. *International Journal of English Linguistics*, *6*(5), 74. https://doi.org/10.5539/ijel.v6n5p74
[43] Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition*, *14*, 357-385.
[44] Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York, NY: Routledge.
[45] Shen, L. (2015). Impacts of chunks input on Chinese English major students' oral production. *Theory and Practice in Language Studies*, *5*(10), 2138-2143.
[46] Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.
[47] Smalle, E. H. M., Bogaerts, L., Simonis, M., Duyck, W., Page, M. P. A., Edwards, M. G., & Szmalec, A. (2016). Can chunk size differences explain developmental changes in lexical learning? *Frontiers in Psychology*, *6*(JAN), [1925]. DOI: 10.3389/fpsyg.2015.01925
[48] Taguchi, N. (2008). Building language blocks in L2 Japanese: Chunk learning and the development of complexity and fluency in spoken production. *Foreign Language Annals*, *40*(1), 132-156.
[49] Tavares, M. D. G. G. (2016). Learners' processes during pre-task planning and working memory capacity. *Ilha do Desterro*, *69*(1), 79-94.
[50] Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*, 84-119.
[51] Vieira, G. V. (2017). Lexical Access in L2 Speech Production: a controlled serial search task. *Ilha Do Desterro A Journal of*

English Language, Literatures in English and Cultural Studies, *70*(3), 245–264. https://doi.org/10.5007/2175-8026.2017v70n3p245
[52] Weissheimer, J., & Mota, M. B. (2009). Individual Differences in Working Memory Capacity and the Development of L2 Speech Production. *Issues in Applied Linguistics*, *17*(2), 93-112. https://doi.org/10.5070/L4172005115

Danh C. Vu is a lecturer in applied linguistics at the Faculty of Foreign Languages, Ton Duc Thang University, Vietnam. He is interested in exploring the potential of technological affordances and how to use interdisciplinary knowledge to make practical applications for second language education. His research interests include technology-based approaches to second language learning and teaching, personalization in language learning, second language listening and speaking, and corpus linguistics. His recent research involves developing a technology-based learning platform for L2 listening comprehension. (Email: vucongdanh@tdtu.edu.vn)

Thanh Van Nguyen is an English lecturer at the Faculty of Foreign Languages, Ton Duc Thang University, Vietnam. He completed his PhD at Suranaree University of Technology, Thailand. His current research interests include the learning theory of Connectivism, Technology Enhanced Teaching and Learning, Innovation in English Teaching and Learning, Online Teaching and Learning, applications of Instructional Design in ELT, and student engagement in online learning environments.

Nakhon Kitjaroonchai is a lecturer at the Faculty of Arts and Humanities, Asia-Pacific International University, Thailand. His research interests include second language acquisition, learning motivation, and L2 learning achievement.