

*Polish Listening SPAN:  
A new tool for measuring verbal working memory*

**Katarzyna Zychowicz**

Pomeranian University, Słupsk, Poland

[katarzyna.zychowicz@apsl.edu.pl](mailto:katarzyna.zychowicz@apsl.edu.pl)

**Adriana Biedroń**

Pomeranian University, Słupsk, Poland

[adriana.biedron@apsl.edu.pl](mailto:adriana.biedron@apsl.edu.pl)

**Mirośław Pawlak**

Adam Mickiewicz University, Kalisz, Poland

State University of Applied Sciences, Konin, Poland

[pawlakmi@amu.edu.pl](mailto:pawlakmi@amu.edu.pl)

**Abstract**

Individual differences in second language acquisition (SLA) encompass differences in working memory capacity, which is believed to be one of the most crucial factors influencing language learning. However, in Poland research on the role of working memory in SLA is scarce due to a lack of proper Polish instruments for measuring this construct. The purpose of this paper is to discuss the process of construction and validation of the Polish Listening Span (PLSPAN) as a tool intended to measure verbal working memory of adults. The article presents the requisite theoretical background as well as the information about the PLSPAN, that is, the structure of the test, the scoring procedures and the steps taken with the aim of validating it.

*Keywords:* working memory; central executive; listening span

## 1. Introduction

Working memory (WM) is a term adapted from cognitive psychology, which generally refers to our ability to maintain and operate on a limited amount of information when doing some mentally demanding tasks (Baddeley, 2015). There is much evidence that WM storage and executive components are involved in foreign or second language (L2) learning and processing (Linck, Osthus, Koeth, & Bunting, 2014; Wen, 2015, 2016); however, this relationship is difficult to pinpoint due to various methodological problems, the method of measurement being one of the most important issues. In order to examine the relationship between WM and second language acquisition (SLA), valid and reliable tools are needed. One of the prerequisites of the reliability of cognitive tests is the use of the participants' native language. Therefore we decided to construct two Polish tools for measuring WM capacity: a listening span, which is a measure of the central executive (CE), and a nonword list, which is a measure of the phonological loop (PL). This article describes the process of construction of the first one, that is the Polish Listening Span Test (PLSPAN). The PLSPAN, based on Daneman and Carpenter's (1980) listening span and Polish Reading Span (Biedroń & Szczepaniak, 2012a, 2012b), is a tool employed to assess the CE. The instrument is designed for testing adult native speakers of the Polish language. At first we present the theoretical background to our study: the concept of WM, together with its two most important components, that is the PL and the CE as well as methods of their measurement. Then, we describe the newly developed tool and the procedures implemented in the construction process. Finally, we offer some conclusions and suggestions for further research.

## 2. Working memory

WM (Baddeley, 2003, 2015; Baddeley, Gathercole, & Papagno, 1998; Baddeley & Hitch, 1974) has recently been high on the agenda of SLA researchers as a significant factor determining the outcomes of L2 learning (Biedroń & Pawlak, 2016; Biedroń & Szczepaniak, 2012a; DeKeyser & Juffs, 2005; DeKeyser & Koeth, 2011; Doughty, Campbell, Mislevy, Bunting, Bowles, & Koeth, 2010; Doughty, 2013; Juffs & Harrington, 2011; Mackey, Philip, Egi, Fujii, & Tatsumi, 2002; Miyake & Friedman, 1998; Papagno & Vallar, 1995; Pawlak, 2017; Robinson, 2003; Sawyer & Ranta, 2001; Skehan, 2012; Wen & Skehan, 2011; Wen, Mota, & McNeill, 2015; Wen, 2016; Williams, 2012). Recently, there have been some suggestions that WM can be another foreign language aptitude (Wen & Skehan, 2011; Wen, 2015, 2016).

Baddeley and Hitch (1974) proposed the multicomponent WM model that comprises two storage systems, that is a *phonological loop* (PL) and a *visuospatial*

*sketchpad*, regulated by a *supervisory attention-limited control system* (CE). Later, they extended the original tripartite model by adding a fourth component, the *episodic buffer*, which stores information (Baddeley, 2000). The most relevant for language learning are the PL and the CE. The PL temporarily stores sound-based information through an articulatory rehearsal process. The PL, viewed as equivalent to a language acquisition device (Baddeley et al., 1998), plays a crucial role in learning novel phonological forms of new words. The CE is responsible for executive functions, such as controlling, allocating and inhibiting attentional resources in higher-level cognitive processes.

Besides the modular model of WM proposed by Baddeley (2003), there are other models emphasizing the factor of executive attention as central to the WM system. The most popular are two, namely the *embedded process model* (Cowan, 2005) and the *attentional control model* (Engle, Kane, & Tuholsky, 1999a), in which WM is an activated subset of long-term memory (LTM). In these models, attention capability accounts for the predictive validity of WM span tests and underlies other cognitive abilities, including fluid intelligence. Consensual theories of WM, which aim at unifying discrepancies (e.g., WM as a gateway to LTM; Baddeley, 2012; Conway et al., 2008; Cowan, 2014) have significant implications for research on the effects of WM on human cognition. A more unitary approach to WM theory has been proposed by Wen (2016, p. 24), who states that “WM is best conceived as a primary memory system (as opposed to LTM as secondary) for learning that functions as an interface between STM components . . . and LTM . . . , which in turn affects real-world actions.”

Research in WM has provided ample evidence that it plays an important role in a number of complex cognitive abilities, such as first language (L1) acquisition and L2 learning, reasoning, comprehension and cognitive control. It is relevant to many everyday tasks, such as reading, making sense of spoken discourse, problem-solving and mental arithmetic. Moreover, WM measures overlap with fluid intelligence test results (Conway, Macnamara, & Engel de Abreu, 2013; Engle, Laughlin, Tuholski, & Conway, 1999b; Kane, Conway, Hambrick, & Engle, 2008). It is quite likely that WM, with its origin in and dependence on rapid developments in modern cognitive science, may hold the *very* key to elaborating the concept of foreign language aptitude (Chan, Skehan, & Gong, 2011; DeKeyser & Koeth, 2011; Miyake & Friedman, 1998; Sawyer & Ranta, 2001; Wen, 2016; Wen, Biedroń, & Skehan, 2016). There is much evidence for this suggestion. First of all, there are clear individual differences among L2 learners, both in relation to their phonological component and executive functions (Wen, 2015, 2016; Williams, 2012). For example, L2 learners have displayed individual variation in their PL, as measured by the simple version of memory span task, and their CE, as indexed by the complex version of memory span task (Linck et

al., 2014). Moreover, a great number of empirical studies in cognitive psychology and SLA (see Wen, 2016, for a review) have provided ample evidence that both the PL and the CE exert consistent and distinctive influences on various aspects of L2 acquisition and processing, and that their relevance varies according to the proficiency level. The PL has been shown to be most important for the acquisition and development of vocabulary, formulaic sequences and grammar (Ellis, 2012; Martin & Ellis, 2012), mostly in L2 beginners. The CE has been demonstrated to be involved mainly in noticing, monitoring, and self-repair in language comprehension and production in intermediate L2 learners (Linck et al., 2014). Results and findings from WM-SLA studies are summarized in Table 1.

Table 1 Results and findings from WM-SLA studies (adapted from Wen, 2016)

SLA domains and activities	PL	CE	Major SLA studies
L2 vocabulary acquisition and development	Instrumental in storing and acquiring novel phonological forms	Not yet clear	Bolibaugh and Foster (2013); Cheung (1996); Ellis and Sinclair (1996); Foster, Bolibaugh and Kotula (2014); French (2006); French and O'Brien (2008); Service (1992); Speciale, Ellis and Bywater (2004)
Acquisition and development of L2 grammar and/or morpho-syntactic constructions	Facilitates the storage and chunking of morpho-syntactic constructions	Not yet clear	Martin and Ellis (2012); Williams and Lovatt (2003)
L2 language comprehension (listening and reading)	Used to maintain a phonological record that can be consulted during offline language processing	Facilitates processing syntactic and semantic information	Alptekin and Erçetin (2011); Berquist (1997); Harrington and Sawyer (1992); Havik et al. (2009); Leeser (2007); Miyake and Friedman (1998)
Language production (speaking and writing)	Predicts narrative vocabulary at early stage; predicts grammatical accuracy at later stage	Is related to performance measures of L2 speech (e.g., accuracy)	Abu-Rabia (2003); Ahmadian (2012); Bergsleithner (2010); Fortkamp (1999, 2003); Guará-Tavares (2008); O'Brien Segalowitz, Collentine and Freed (2006, 2007); Payne and Whitney (2002)

Still, despite all the promising evidence, there is much controversy surrounding WM and the results are often contradictory or ambiguous. One such ambiguity relates to grammar learning. A few studies (e.g., Fortkamp, 2003; Linck et al., 2014; Martin & Ellis, 2012; Williams & Lovatt, 2003) provide evidence for a

complex relationship between WM and grammar learning. Fortkamp (2003) examined the relationship between the CE component of WM, operationalized as a speaking span in an L2, and speech production during a picture description and a narrative. Her investigation revealed that WM positively correlates with fluency, accuracy and structural complexity, which led her to conclude that grammatical encoding in L2 speech production depends on the regulation of attention and control, which are seen as key elements of the CE component of WM. Williams and Lovatt (2003) conducted two experiments targeted at relating PL and grammar learning. They found an important link between PL and grammar rule learning in a semiartificial language; however, the link only partially explained the variance in the acquisition of grammar. Therefore, they concluded that for a fuller understanding of the process of grammar learning research should include tests of both PL and CE. O'Brien et al.'s (2006) research concentrated on the role of phonological short-term memory, that is the PL, as measured by serial nonword recognition, in speech production focusing on lexical, grammatical and narrative abilities of adults. The results of their study clearly indicate that PL plays an important role in the grammatical proficiency of L2 students at later stages of L2 development. Kormos and Sáfár (2008) studied the relationship between PL, measured by a nonword repetition test, and CE, measured by a backward digit span test, and performance in the L2 in an intensive language program, and found a high positive correlation between FCE Use of English, Reading, Listening and Speaking parts and both PL and CE. However, since FCE Use of English measures both grammar and vocabulary at the same time, it is difficult to draw conclusions concerning exclusively grammar results. Martin and Ellis (2012) investigated the influence of PL, operationalized as a nonword repetition span and a nonword recognition span, and CE, operationalized as a listening span test capacities on the learning of vocabulary and grammar in an artificial language, and documented separate effects of PL and CE on grammar learning, either direct or mediated by vocabulary. The CE component of WM turned out to be a stronger predictor of learning outcomes, with CE explaining 14% and PL explaining 10% of the variance in production, and 11% and 17%, respectively, in comprehension. Summing up, research on the relationship between WM and the knowledge of grammar is relatively scarce and the results are inconclusive; however, the CE subsystem seems to be definitely more strongly implicated in grammar production than the PL (see Linck et al., 2014).

### 3. Working memory measurement

The definition and structure of WM as well as its variable impact on different aspects of SLA and processing affect the construction of tasks employed in its

measurement. The construct of WM is widely operationalized to refer to the total resources that are available to an individual for simultaneous processing and storage. According to Just and Carpenter (1992), any individual possesses finite resources that are consumed by both the processing and storage of information. This means that the processing and storage demands of a task can be traded off against each other. For example, in an easy task processing demands will be low and so storage capacity will be relatively high. In this view, measuring the storage capacity of the individual without reference to a particular processing task does not seem to make sense and therefore WM tests should involve storage and processing of information simultaneously.

In line with this view, Daneman and Carpenter created the first test measuring WM capacity, namely the Reading SPAN Task (RST). In the original RST (Daneman & Carpenter, 1980), participants were instructed to read series of sentences aloud, while remembering the final word of each sentence in a particular series. In addition, Daneman and Carpenter (1980) developed a listening version of the RST. The listening span also required the retention of sentence-final words, but the participants listened to, rather than read, lists of sentences. In order to ensure subjects' focus on both processing and remembering information, Daneman and Carpenter added a true/false component to the test, where subjects decided if a sentence they listened to was true or false within 1.5 seconds from hearing it; however, they did not monitor the accuracy of the answers. Engle et al. (1999a) decided to alter this procedure for their reading span and asked their subjects to verify the correctness of the presented sentences, excluding all subjects with processing scores below 80% from analysis, which helped ensure that attention was paid to the processing task. In what follows, we discuss the construction, scoring procedures and validation of the PLSPAN.

## 4. The study

### 4.1. Aims

The aim of the study, which took place at Pomeranian University in Słupsk, Poland, in May 2015, was to design a valid and reliable tool for measuring WM capacity in Polish. The PLSPAN test is based on the same principle as that followed by Daneman and Carpenter (1980), and Engle et al. (1999a), but the language of the input is Polish. It has often been stressed (e.g., Linck et al., 2014) that cognitive tests, including WM tests, should be conducted in participants' native language, as tasks performed in the L2 would indicate not only WM capacity but also L2 proficiency. This would negatively influence any analysis of the results, especially if the study was to be held in the field of SLA and later correlated with any linguistic outcome.

## 4.2. Participants

Fifty eight first- and second-year English majors enrolled in a BA program agreed to take part in the study. The sample consisted of 36 females and 22 males, aged 19-23, with the mean age of 21.6. They were monolingual Polish learners of English as a foreign language whose proficiency level was intermediate (B1/B2 in terms of the *Common European Framework of Reference*). They had been studying English for 3-11 years, with the mean length of about 9 years, either at school or in additional courses or private tutoring. In the BA program they attended classes in English, including the four skills, namely speaking, listening, reading and writing, as well as classes dealing with grammar and pronunciation. They also participated in a number of content classes, such as introduction to linguistics, strategic training, introduction to literary studies and varieties of English, all of which were taught in the target language.

## 4.3. The test

The test consists of 9 sets of sentences of growing sizes, from 2 sentences in Set 1 to 10 in Set 9, producing a total of 54 sentences. The sets were recorded using Audacity software, with 1.5-second gaps between sentences. The length and complexity of the items was controlled for. Each is a grammatically correct complex sentence, approximately 8 words in length and, when recorded, lasts from 2.77 seconds to 3.56 seconds with the average length of 3.06 seconds. 50% of the sentences were altered lexically so that some of them do and some of them do not make sense in everyday life. For example, the sentence: *Marek jest po egzaminach, więc wyjeżdża na biwak* 'Mark has already taken his exams, so he is going camping' makes sense. On the other hand, the sentence: *Koza szybko powiedziała, że na pewno woli mikrofon* 'The goat quickly said that it surely preferred the microphone' is senseless as goats do not speak. The altered words are nouns, verbs and adjectives placed in any but final position in a sentence. The participants' task is to determine whether or not each sentence makes sense to ensure the processing of the input, and, at the same time, remember the last word of each sentence for subsequent recollection. Each sentence-final word is a common noun in the nominative case to avoid confusion with word endings. Test reliability and validity were verified in two ways: The material was first evaluated by judges and later a pilot study was conducted.

## 4.4. Administration

As with most tests in the field of cognitive science, subjects take the test individually, which allows them to focus on both tasks that they are requested to

perform. Additionally, it gives the researcher an opportunity to observe the subjects and ensure that they focus on both processing and storage. The administration of the test takes about 10 minutes. Before they begin the test, they are informed of its content and the tasks they are supposed to perform. During the listening to the sentences they are to judge whether each sentence makes sense and mark all those that do on the answer sheet, ignore the senseless sentences, and remember all the sentence-final words. After each set there is a pause during which participants are supposed to recollect all the words they remember from the set. The order of recall is free, that is, they can list the words in any order, not necessarily in the order the sentences were presented. The actual test is preceded by two trial sets in order to make sure that subjects understand both tasks, learn to judge sentence sensibility and practice focusing on two things at the same time. One trial set is presented below:

*Posialiśmy już marchewkę i pietruszkę, został jeszcze seler* 'We have already planted carrots and parsleys; all we are left to do are celeries.'

*Nie mam czasu, niech pomoże ci drewniane krzesło* 'I do not have time, our wooden chair can help you.'

*Karolina jest już dorosła, może posmarować na wybory* 'Caroline is already an adult, she can butter to the election.'

#### 4.5. Scoring and analysis

In Daneman and Carpenter's traditional test, each subject was assigned an absolute span score. The test started with a 2-element item and continued until the subject failed to retrieve an item. The test ended at that time, and the last item size (e.g., 4 or 5) recalled was the span score. However, absolute spans have several shortcomings (Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle, 2005; Linck et al., 2014). First of all, such scores take on one of very few values, usually from 2 to 6, thus limiting the sensitivity of the measure and disallowing diversification of results. Secondly, by just estimating the item size for a participant and then discontinuing the test, data on all other trials are ignored. Moreover, the difficulty of a span item may vary on many dimensions, thereby threatening span reliability (Conway et al., 2005, p. 774). In summary, absolute span measures cannot be applied to research on individual differences. Instead, the use of scoring procedures exhausting the information collected is advised, such as the partial scoring procedure, where correct responses to individual elements within an item are assigned 1 point, and all other responses are assigned 0 points, with no attempt to classify the type of error (Conway et al., 2005).



Given the above, the result of the PLSPAN is a partial score, that is the number of correctly remembered words in all the sets. It allows for greater diversification of the results as well as preventing the floor and the ceiling effects (Conway et al., 2005). Furthermore, points are assigned to all elements recalled, irrespective of the correctness on the processing component. The outcome of the processing task, that is, the judgments concerning the logic of the sentences, serves only as a distractor precluding subjects from mental rehearsal and is usually close to the ceiling. However, it is taken into consideration while calculating the score, as results with the score below 80% of correct answers in the processing task are excluded from the sample, the reason being the lack of ample concentration on the task.

## 4.5. Results

### 4.5.1. Reliability

Reading span, operation span and listening span have been used in hundreds of independent studies involving thousands of subjects. According to Conway et al. (2005, p. 776):

One conclusion that can be drawn from this body of research is that measures obtained from these tasks (span scores) have adequate reliability . . . For example, estimates of reliability based on internal consistency, such as coefficient alphas and split-half correlations, which reflect the consistency of participants' responses across a test's items at one point in time, are typically in the range of .70-.90 for span scores.

WM span tests seem to be reliable across time as well. Typical test-retest results correlate in the range of .70-.90.

In order to verify the reliability of the PLSPAN, the test-retest method was applied. The correlation between the initial test and the retest which took place 3 weeks later was .91, which indicates a high reliability of the test. The Kuder Richardson Alpha for internal consistency reliability for the test was .76. Split-half reliability was estimated at .78, which allows a conclusion that the test is a reliable measure of CE.

### 4.5.2. Validity

#### 4.5.2.1. Construct validity

The test can be said to possess high construct validity as it was constructed following leading experts in the field of cognitive neuroscience who verified their

tools in numerous empirical studies. The results of their research indicate that the construct measured by WM span tests is the ability to control attention and thought. Measures of WM capacity reflect individual differences in the aforementioned ability. Also, as described in the first part of this paper, results of WM span tests correlate with numerous tests of higher-order cognition, including intelligence, thus demonstrating high predictive validity. Construct validity also refers to convergent and discriminant validity. WM span tasks correlate extremely well with each other and, at the same time, correlate mildly with more traditional simple span tasks.

In order to measure the convergent validity of the PLSPAN, we correlated the results of our test with the results of the Polish reading span by Biedroń and Szczepaniak (2012a, 2012b), which is supposed to tap the same construct, and a nonword repetition test, which is to measure only storage capacity. The results we obtained are as follows: For the Polish reading span and the PLSPAN Pearson coefficient  $r$  was  $.77$ ,  $p = .000$ , which is a high or very high correlation. For the PLSPAN and the nonword repetition test Pearson coefficient  $r$  was  $.33$ ,  $p = .011$ , which is a low moderate correlation.

Such results allow us to conclude that although all the three tests measure one concept, that is memory, which is visible in the positive correlations between them, the PLSPAN and the Polish reading span measure a different aspect of it, namely the CE component of WM whereas the nonword repetition measures only its phonological aspect. Even though it would seem that the two verbal memory tests using the same modality, that is, aural reception, would correlate better than those using two different modalities, the results of the analysis clearly show that the effect of modality is far weaker than could have been expected.

#### 4.5.2.2. Content validity

Content validity of the test was assessed by five competent judges, four linguists and a psychologist. The judges were familiarized with the concept of WM and the purpose of the test. Next, they were asked to evaluate all the test tasks on a 5-point Likert scale, where 1 indicated *total disagreement* and 5 *total agreement*. After reading each sentence they answered three questions:

- Is the sentence comprehensible?
- Is it possible to immediately decide whether the sentence is acceptable in everyday speech?
- Does the sentence make sense?

After reading all the sentences in a given set the judges were asked two additional questions:

- Are the sentences in the set thematically connected?

- Are the words at the end of the sentences thematically connected?

The judges were also asked whether the test as a whole measures WM. The answers of the judges were analyzed, and all the sentences with mean values below 4.5 were replaced with new ones, which were also evaluated. Kendall's coefficient of concordance for all the sets was above .9, with the value of .94 for the entire test. The high concordance among the judges indicates that the test is valid.

#### 4.5.2.3. Face validity

The next step in verifying test validity was the face validity check. For this purpose, a group of ten university students was chosen since young adults and adults were the targets of the test. They were asked to listen to the entirety of the test and decide whether the gaps between the sentences of 1 second were long enough to judge sensibility. Later, they evaluated the test according to the same criteria as the competent judges, but they listened to the sentences instead of reading them. Again, the analysis of their answers indicated that the test is valid, with Kendall's coefficient of concordance for all the sets equaling .91.

The evaluation of the test by the students was followed by a focus session, in which the students expressed their opinions about the content and the form of the test. Their opinions were very positive. They said they had fun judging the sensibility of the sentences, as lexical changes made in senseless sentences created funny images of, for example, singing tattoos or writing buckets. According to one respondent, "*it was funny . . . and strange. I'm not used to doing two things at the same time, so it was also challenging and very difficult.*" They also believed that the test would measure memory, as well as intelligence and concentration, as mentioned by another respondent: "*I think it will measure memory and concentration, and I think . . . intelligence, too.*" That was a surprising finding since they could not have known that the original version of the listening as well as the reading span correlated well with results of IQ tests. However, all of them agreed that the pace of the presentation of the sentences was too high, that is, 1 second was not enough to decide if a sentence makes sense or not. One respondent even said: "*It was too difficult for me. Maybe because it was so fast.*" On the basis of their opinions the gaps were lengthened to 1.5 seconds, which was the original timing in Daneman and Carpenter's (1980) test.

#### 4.5.3. Processing task

As expected, the processing task turned out to be a very simple one, thus allowing the subjects to achieve very high results, often even 100%. However, one person refused to finish the test as he "*couldn't concentrate on remembering*

*while thinking.*" Another person achieved a 57% level of correctness and was also excluded from the analysis.

The only factor influencing sensibility judgment was the grammatical category of the word altered, which we chose to be nouns, verbs or adjectives. Noun alternations achieved over 99% correctness, verbs seemed to cause some initial confusion and achieved almost 97% correctness, with the first two sentences achieving only 86% and the rest of the sentences close to 99%. The sentences with adjective alternations seemed to be the most difficult to process, since they achieved only 83% correctness and one sentence, that is, *Wiał tak zielony wiatr, że potamał ogromne drzewo* 'The wind was so green that it broke a huge tree,' reached only 57% correctness.

#### 4.5.4. Storage task

The mean result of the test was 26.52, which is almost half of the 54 elements of the test. The minimum score was 8 points and the maximum was 41 points, which shows that the sensitivity of the measure is considerable. Besides no floor or ceiling effect was observed, which shows that the span of the test is accurate. All the measures of test reliability show that the test is a reliable measure; however, the discriminating power of several items within the test is still not satisfactory, possibly due to the very strong primacy and recency effects observed during the analysis.

### 5. Discussion

The analysis of the processing task revealed several interesting findings. As mentioned above, a strong ceiling effect was observed, which had been expected, and which indicates that participants had few problems with judging sentence sensibility. We had expected that any problems connected with this task might result from the position of the word altered, namely that the later the alternation appeared in the sentence, the more difficult it would turn out to be to evaluate. Yet, no such effect was observed in the analysis, which allows a conclusion that the position of the senselessness in a sentence has no influence on the sensibility judgment. Another presupposition we had was that any problems appearing while judging sensibility might result from the grammatical category of the semantic alternation. This proved to be right, and the results show that while altering nouns and verbs poses no difficulty, changing adjectives seems to mislead some subjects.

The storage task brought findings we had expected. The high sensitivity of the test, its accuracy, reliability and validity appear to indicate that the test is a fine measure of the CE. The only limitation of the PLSPAN is the low discriminating

power of several positions, which we attributed to the primacy and recency effects. This is consistent with the results obtained by other researchers (Murdock, 1962; Unsworth & Engle, 2007).

## 6. Conclusions

The study reported in the present paper aimed to design an instrument that could be used to examine, in the Polish educational context, the subcomponent of WM which is the most relevant for SLA research, that is the CE. In line with the theoretical suggestions, we constructed the PLSPAN, which is a complex span test intended to measure the CE. The test is designed for adults and young adults. It is based on classical tests of WM, that is, the reading span and listening span. The procedures applied to assess test reliability and validity proved that the test is a good measure of the CE component of WM. Our study suffers from a number of limitations that can mainly be attributed to highly individualized cognitive abilities of the participants. A problem that is very difficult to solve is the primacy and recency effects. Another is the grammatical category of the altered words.

There are a number of methodological issues that should be addressed in further research. One such problem is domain specificity versus domain generality of tasks. In view of lack of any reliable criterion, the choice of a task depends on the researcher, and this can significantly affect the results of a particular study. We agree with Wen (2016) that future research should specify the consequences of using the two different types of measures. Moreover, the relationship between WM components and aspects of L2 learning is far more complicated and nuanced than the relationship that can be revealed through simple correlation analysis. Wen suggests that the measures of WM should be functionally oriented by targeting specific functions, such as, for example, information updating. In this way, an integrated WM profile that comprises all individual WM components or functions can be obtained. A precise multi-span profile will allow for individualization of the learning process and compensation for weaker areas.

Summing up, in the process of test construction, the theoretical conceptualizations of WM have been complemented by established assessment procedures to examine the CE, which paves the way for further studies in the field of SLA. We are hopeful that, as a result of the construction and validation of the PLSPAN, the explanatory power of WM as foreign language aptitude in L2 learning will be greatly enhanced.

## Acknowledgements

The study reported in this paper represents a contribution to the research project no. 2015/17/B/HS2/01704 (2016-2018) funded by the National Science Centre, Poland.

## References

- Abu-Rabia, S. (2003). The influence of working memory on reading and creative writing processes in a second language. *Educational Psychology, 23*, 209-222.
- Ahmadian, M. J. (2012). The relationship between working memory and oral production under task based careful online planning condition. *TESOL Quarterly, 46*(1), 165-175.
- Alptekin, C., & Erçetin, G. (2011). The effects of working memory capacity and content familiarity on literal and inferential comprehension in L2 reading. *TESOL Quarterly, 45*, 235-266.
- Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*, 417-123.
- Baddeley, A. D. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*, 189-208.
- Baddeley, A. D. (2012). Working memory: Theories, models and controversies. *Annual Review of Psychology, 63*, 1-30.
- Baddeley, A. D. (2015). Working memory in second language learning. In Z. Wen, M. B. Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 17-28). Bristol: Multilingual Matters.
- Baddeley, A. D., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language acquisition device. *Psychological Review, 105*, 158-173.
- Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47-90). New York, NY: Academic Press.
- Bergsleithner, J. M. (2010). Working memory capacity and L2 writing performance. *Ciências & Cognição, 15*(2), 2-20.
- Berquist, B. (1997). Individual differences in working memory span and L2 proficiency: Capacity or processing efficiency? In A. Sorace, C. Heccock, & R. Shillcock (Eds.), *Proceedings of the GALA' 1997 Conference on language acquisition* (pp. 468-473). Edinburgh: The University of Edinburgh.
- Biedroń, A., & Pawlak, M. (2016). The interface between research on individual difference variables and teaching practice: The case of cognitive factors and personality. *Studies in Second Language Learning and Teaching, 6*(3), 395-422. doi: 10.14746/sslit.2016.6.3.3
- Biedroń, A., & Szczepaniak, A. (2012a). Polish reading span test – an instrument for measuring verbal working memory capacity. In J. Badio & J. Kosecki (Eds.), *Cognitive processes in language*. (pp. 29-37). Frankfurt am Main: Peter Lang.
- Biedroń, A., & Szczepaniak, A. (2012b). Working-memory and short-term memory abilities in accomplished multilinguals. *Modern Language Journal, 96*, 290-306.
- Bolibaugh, C., & Foster, P. (2013). Memory-based aptitude for nativelike selection: The role of phonological short-term memory. In G. Granena & M. H.

- Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 203-228). Amsterdam: John Benjamins.
- Chan, E., Skehan, P., & Gong, G. (2011). Working memory, phonemic coding ability and foreign language aptitude: Potential for construction of specific language aptitude tests: The case of Cantonese. *Ilha Do Desterro: A Journal of English Language, Literatures and Cultural Studies*, 60, 45-73.
- Cheung, H. (1996) Nonword span as a unique predictor of second-language vocabulary learning. *Developmental Psychology*, 32 (5), 867-873.
- Conway, A. R. A., Jarrold, Ch., Kane, M. J., Miyake, A., & Towse, J. N. (2008). Variation in working memory. An introduction. In A. R. A. Conway, Ch. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 3-17). Oxford: Oxford University Press.
- Conway, A., Kane, M., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, 12, 769-786.
- Conway, A., Macnamara, B., & Engel de Abreu, P. (2013). Working memory and intelligence: An overview. In T. P. Alloway & R. G. Alloway (Eds.), *Working memory: The new intelligence* (pp. 13-36). New York, NY: Psychology Press.
- Cowan, N. (2005). *Working memory capacity*. New York, NY: Psychology Press.
- Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, 26(2), 197-223.
- Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450-466.
- DeKeyser, R. M., & Juffs, A. (2005). Cognitive considerations in L2 learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 437-454). Mahwah, NJ: Lawrence Erlbaum.
- DeKeyser, R. M., & Koeth, J. (2011). Cognitive aptitudes for second language learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 395-407). New York, NY: Routledge.
- Doughty, C. J. (2013). Optimizing post-critical-period language learning. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 153-175). Amsterdam: John Benjamins.
- Doughty, C. J., Campbell, S. G., Mislavy, M. A., Bunting, M. F., Bowles, A. R., & Koeth, J. T. (2010). Predicting near-native ability: The factor structure and reliability of Hi-LAB. In M. T. Prior, Y. Watanabe, & S-K. Lee (Eds.), *Selected proceedings of the 2008 Second Language Research Forum* (pp. 10-31). Somerville, MA: Cascadilla Proceedings Project. Retrieved from [www.lingref.com](http://www.lingref.com), document #2382
- Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal Teddy Bear. *Annual Review of Applied Linguistics*, 32, 17-44.

- Ellis, N. C., & Sinclair, S. G. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. *The Quarterly Journal of Experimental Psychology*, 49A(1), 234-250.
- Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999a). 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). London: Cambridge Press.
- Engle, R. W., Laughlin, J. E., Tuholski, S. W., & Conway, A. R. A. (1999b). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, 128(3), 309-331.
- Fortkamp, M. B. M. (1999). Working memory capacity and aspects of L2 speech production. *Communication and Cognition*, 32, 259-296.
- Fortkamp, M. B. M. (2003). Working memory capacity and fluency, accuracy, complexity and lexical density in L2 speech production. *Fragmentos*, 24, 69-104.
- Foster, P., Bolibaug, C., & Kotula, A. (2014). Knowledge of nativelike selections in an L2: The influence of exposure, memory, age of onset and motivation in foreign language and immersion settings. *Studies in Second Language Acquisition*, 36(1), 101-132.
- French, L. M. (2006). *Phonological working memory and second language acquisition: A developmental study of francophone children learning English in Quebec*. New York, NY: Edwin Mellen.
- French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics*, 29, 463-487.
- Guará-Tavares, M. G. (2008). *Pre-task planning, working memory capacity and L2 speech performance* (Unpublished doctoral dissertation). Universidade Federal de Santa Catarina, Florianópolis, Brazil.
- Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition*, 14, 25-38.
- Havik, E., Robert, E., van Hout, R., Schreuder, R., & Haverkort, M. (2009). Processing subject-object ambiguities in the L2: A self-paced reading study with German L2 learners of Dutch. *Language Learning*, 59, 73-112.
- Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, 44, 137-166.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 98, 122-149.
- Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2008). Variation in working memory capacity as variation in executive attention and control. In A. R. A. Conway, Ch. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 21-49). Oxford: Oxford University Press.



- Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, 11(2), 261-271.
- Leeser, M. (2007). Learner-based factors in L2 reading comprehension and processing grammatical form: Topic familiarity and working memory. *Language Learning*, 57, 229-270.
- Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, 21(4), 861-883.
- Mackey, A., Philip, J., Egi, T., Fujii, A., & Tatsumi, T. (2002). Individual differences in working memory, noticing interactional feedback and L2 development. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 181-209). Philadelphia, PA: John Benjamins.
- Martin, K. I., & Ellis, N. C. (2012). The roles of phonological STM and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition*, 34(3), 379-413.
- Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. Healy & L. Bourne (Eds.), *Foreign language learning* (pp. 339-364). Mahwah, NJ: Lawrence Erlbaum.
- Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5), 482-488.
- O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics*, 27, 377-402.
- O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition*, 29, 557-582.
- Papagno, C., & Vallar, G. (1995). Verbal short-term memory and vocabulary learning in polyglots. *Quarterly Journal of Experimental Psychology*, 38A, 98-107.
- Pawlak, M. (2017). Overview of learner individual differences and their mediating effects on the process and outcome of interaction. In L. Gurzynski-Weiss (Ed.), *Expanding individual difference research in the interaction approach: Investigating learners, instructors, and other interlocutors* (pp. 19-40). Amsterdam: John Benjamins.
- Payne, J. S., & Whitney, P. J. (2002). Developing L2 oral proficiency through synchronous CMC: Output, working memory, and interlanguage development. *CALICO Journal*, 20, 7-32.
- Robinson, P. (2003). Attention and memory during SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 631-679). Oxford: Blackwell.

- Sawyer, M., & Ranta, L. (2001). Aptitude, individual differences, and instructional design. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 319-354). Cambridge: Cambridge University Press.
- Service, E. (1992). Phonology, working memory and foreign-language learning. *Quarterly Journal of Experimental Psychology*, 5, 21-50.
- Skehan, P. (2012). Language aptitude. In S. Gass & A. Mackey (Eds.), *Routledge handbook of second language acquisition* (pp. 381-395). New York, NY: Routledge.
- Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, 25, 293-321.
- Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, 114, 104-132.
- Wen, E. Z. (2015). Working memory in second language acquisition and processing: The phonological/executive model. In E. Z. Wen, M. B. Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 41-62). Bristol: Multilingual Matters.
- Wen, E. Z. (2016). *Working memory and second language learning: Towards an integrated approach*. Bristol: Multilingual Matters.
- Wen, E. Z., & Skehan, P. (2011). A new perspective on foreign language aptitude: Building and supporting a case for "working memory as language aptitude". *Ilha Do Desterro: A Journal of English Language, Literatures and Cultural Studies*, 60, 15-44.
- Wen, E. Z., Biedroń, A., & Skehan, P. (2016). Foreign language aptitude theory: Yesterday, today and tomorrow. *Language Teaching*, 50(1), 1-31.
- Wen, E. Z., Mota, M. B., & McNeill, A. (Eds.). (2015). *Working memory in second language acquisition and processing*. Bristol: Multilingual Matters.
- Williams, J. N. (2012). Working memory and SLA. In S. Gass & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 427-441). Oxford: Routledge/Taylor & Francis.
- Williams, J. N., & Lovatt, P. (2003). Phonological memory and rule learning. *Language Learning*, 53, 67-121.