



ORIGINAL ARTICLE

Reliability and validity assessment of working memory measurements

Kexin Liu  and Remi Murao 

Graduate School of Humanities, Nagoya University, Nagoya, Japan

Corresponding author: Kexin Liu; Email: liu.kexin.a4@s.mail.nagoya-u.ac.jp

(Received 13 May 2024; revised 1 December 2024; accepted 8 January 2025)

Abstract

This research aims to identify a reliable method for measuring working memory (WM) within the context of second language learning. The goal of the study is to address the persistent problem of determining the most appropriate method for measuring WM. To achieve this objective, various WM measurement tasks, including the Digit Span Task, Listening Span Task, Sentence Recall Task (SRT) (both written and spoken), and Sentence Recognition Task, were administered to 39 participants. The experiments were conducted twice to assess the consistency and reliability of these measurement methods.

Through statistical analyses of results, this study endeavors to elucidate the relationship between diverse WM measurement tasks and English listening proficiency. The results of the test-retest correlation, Cronbach's alpha coefficient, and Rasch reliability indicate that SRT (written mode) exhibited the highest reliability while other measurements also demonstrated decent reliability. Additionally, the SRT showed the strongest correlation with the TOEIC Listening Test, administered to test criterion-related validity. This research has the potential to provide valuable insights into the role of WM in second language acquisition and may serve as a methodological guide for future studies in this field.

Keywords: L2 listening; Sentence Recall Task; working memory measurements

Introduction

Working memory (WM) refers to the “temporary storage and manipulation of information that is assumed to be necessary for a wide range of complex cognitive activities” (Baddeley, 2003, p.189). As a fundamental cognitive function, WM is essential for daily functioning and significantly impacts language development. For many cognitive tasks, such as comprehension and problem-solving, the capacity to store and manipulate information in our mental workspace is crucial. In second

language (L2) processing, learners often struggle to retain longer sequences of words in WM, which hinders their ability to effectively integrate sentence meanings.

Recent models of WM emphasize the interaction between WM and the language knowledge stored in long-term memory (LTM), exploring how previously acquired knowledge facilitates encoding in WM and how information retained in WM is subsequently transferred to LTM (Baddeley, 2000; Cowan, 1999; Ericsson & Kintsch, 1995; Schwering & MacDonald, 2020; Wen, 2016). Among these models, Schwering and MacDonald (2020) conceptualize WM as an activated portion of LTM, emerging from knowledge of the statistical regularities of language. According to their view, verbal WM tasks do not measure a separate memory capacity; instead, they assess language skills required to encode, maintain, and order verbal information temporarily, supported by long-term representations of language sequences. Consequently, this study adopts the term *WM efficiency* rather than *WM capacity* to reflect what is actually measured by WM tasks.

However, current methods for measuring WM efficiency, such as the Reading Span Task (RST) and Listening Span Task (LST) (Daneman & Carpenter, 1980), fall short in capturing L2 learners' long-term linguistic sequence knowledge and the chunking processes occurring in WM. Moreover, the reliability of RST/LST in measuring L2 WM is questionable, as learners often employ varying strategies, focusing either on word storage or sentence processing. In this context, the Sentence Recall Task (SRT) (Alloway et al., 2004; Baddeley et al., 2009; Jefferies et al., 2004) has emerged as a promising tool for assessing the integration of long-term memory and the fluid linguistic information processed in the episodic buffer (Baddeley, 2000).

The present study underscores the need to address the limitations of standard WM measures, which often lack adequate verification and validation. Consequently, these measurements cannot be confidently used to evaluate the efficiency of L2 working memory. The primary goal of this study is to close this gap by demonstrating the reliability and validity of methods for assessing L2 WM efficiency. By addressing concerns related to the reliability and validity of these evaluation approaches, the study contributes to the advancement of research in second language acquisition.

Specifically, this study aims to determine the reliability and validity of various WM assessment methods and identify which methods provide the most dependable measures of WM. While the study does not delve into how these methods capture the intricate workings of WM during second language acquisition, it focuses on developing instruments that evaluate and explore the interactions between WM and language learning. Through this effort, the research seeks to deepen our understanding of the cognitive processes underpinning language acquisition and support the creation of effective assessment tools for both research and educational purposes.

Literature review

Theoretical framework of working memory

A significant theoretical framework for understanding WM is provided by Baddeley's Multi-Component Model of Working Memory (2000). This model is composed of three components: the Articulatory Loop, responsible for storing

sound-based data; the Visual-Spatial Sketchpad, which processes visual and spatial information; and the Central Executive, which coordinates and manipulates information in working memory. In an updated version of the model, Baddeley introduces the concept of the Episodic Buffer as a fourth component. This buffer serves as a storage unit aiding the Central Executive in integrating information from WM and long-term memory, thus facilitating a smoother flow of information. This element emphasizes the complex relationship between the immediate, WM processes and the larger, more lasting aspects of long-term memory, emphasizing a dynamic interaction essential for comprehending cognitive functioning.

The executive control model of working memory, proposed by Wen et al. (2015), highlights the intricate interplay between WM and attentional processes. It suggests that WM encompasses the ability to efficiently manage and guide attentional resources, guaranteeing that information relevant to the task is stored and available while inhibiting irrelevant stimuli. This model presents WM as a dynamic cognitive system closely connected to broader cognitive functions, including those regulated by long-term memory.

Wen (2016) introduced a unified framework that integrates insights from both WM research and second language acquisition, particularly focusing on key executive functions such as information updating, shifting, and inhibition. It acknowledges the complex relationship between WM and long-term memory resources, including knowledge of linguistic sequences at multiple grain sizes in both the first and languages. This shift in focus reflects an acknowledgment of the interconnected nature of these memory systems within the broader cognitive framework.

These insights collectively underline the importance of understanding the relationship between working memory, attentional processes, and long-term memory. Ericsson and Kintsch (1995) proposed the Long-Term Working Memory Model, which offers a unique perspective on how WM interacts with long-term memory. It suggests that skilled individuals can effectively use their long-term memory as a component of working memory, achieved through retrieval structures that allow rapid access to relevant information stored in long-term memory. This concept aligns well with the insights provided by the previously discussed models.

More recently, Schwering and MacDonald (2020) proposed an emergent model of working memory, where they view WM as an activated portion of long-term memory. They suggest that verbal WM is the skill of maintaining and ordering linguistic materials, arising from long-term memory and shaped by experience. These models and frameworks are particularly relevant in the context of second language acquisition, where the ability to process and retain linguistic information is crucial. The theories and frameworks discussed provide a foundational understanding of how WM interacts with and is supported by long-term memory structures, offering valuable perspectives for cognitive psychology and language studies.

Domain-specific and domain-general functions of working memory

As mentioned above, WM as a sophisticated cognitive system encompassing the temporary retention and manipulation of information, can be broadly categorized into two distinct types: domain-general and domain-specific working memory. In Camos's (2017) study, she also concluded that there is clear evidence supporting the

existence of both domain-specific contributions and domain-general components in working memory.

The domain-general component of WM is recognized for its versatility and general-purpose functionality. It is essential to many different cognitive tasks because it is not limited to processing information from a single area. Its capability as a general processor allows it to handle diverse types of information, underscoring its importance in various cognitive activities. Researchers have referred to the term “working memory capacity” (Conway, et al, 2002) as a general factor of working memory, which indicates persistent positive correlations among diverse WM tasks, signifying a cohesive underlying construct. Additionally, evidence suggests that WM capacity represents individual variances in the executive aspects of working memory, especially concerning executive attention and cognitive control (e.g.: Engle & Kane, 2004; Kane & Engle, 2002; etc.).

Conversely, domain-specific WM is adept at processing and retaining information within particular domains. For example, in the domain of verbal information, this type of WM is essential for language-related tasks. Similarly, it is crucial for mental navigation or spatial tasks in the spatial information domain, and for tasks requiring visual imagery in the visual information domain. The proficiency of domain-specific WM systems is evident in their capacity to efficiently process information pertinent to their respective areas, thereby offering a targeted approach to memory handling.

The Process Overlap Theory (Kovacs & Conway, 2016) provides an important framework for understanding how domain-general executive processes (e.g., attention and cognitive control) interact with domain-specific processes (e.g., linguistic ability) in tasks such as the LST. According to this hypothesis, WM measurement tasks engage both domain-general executive functions, which are required to maintain and manage information, and domain-specific processes, which are required for interpreting and processing language.

Importantly, the distinction between domain-general and domain-specific WM has profound implications for the selection of assessment methodologies. Kovacs, et al. (2019) demonstrate that ability differentiation occurs in complex span tasks but not in simple span tasks (see Table 1 in the next section for the details of these tasks). This supports the theoretical distinction between simple and complex span tasks (Conway & Kovacs, 2013; Engle et al., 1999). Thus, assessing WM capabilities necessitates careful consideration of the cognitive domain relevant to the task. Measurement tasks and instruments are often carefully designed to focus on WM within particular domains. Therefore, the choice of an appropriate assessment tool, one that aligns with the cognitive function under scrutiny, is vital. This alignment ensures that the WM assessment is not only contextually relevant but also accurately mirrors the intricacies of the cognitive processes in question. This approach is supported by Peng and Swanson (2022), who highlight that individuals demonstrate more efficient processing capabilities in domains where they have substantial knowledge, compared to domains with which they are less familiar.

Table 1. WM measurements (according to Wen, 2016)

Domain	Memory span	Measurement tasks	Component
Domain general	Simple span	Digit Span Task	Phonological WM
	Complex span	Operation Span Task	Executive WM
Domain specific (language)	Simple span	Word/Nonword repetition task	Phonological WM
		Sentence Recall Task	Phonological WM/Episodic buffer
	Complex span	Reading/Listening/Speaking span task	Executive WM/Episodic buffer task

Working memory and L2 proficiency

The interaction between WM and long-term memory raises a question about what exactly WM tasks are measuring. In the context of L2 learning, effective comprehension and cognitive processing require not only retrieving L2 knowledge from LTM but also integrating this knowledge seamlessly with information currently held in working memory. Ericsson and Kintsch's (1995) concept of retrieval structures highlights the importance of L2 knowledge stored in LTM, emphasizing how learners use this stored knowledge to facilitate language tasks. The interaction between L2 knowledge in LTM and domain-specific WM processes is crucial for tasks that require learners to process and manipulate L2 stimuli. Verbal WM tasks can therefore be viewed as measures of skill efficiency in maintaining and ordering verbal information rather than measures of separate memory capacity, as suggested by Schwering and MacDonald (2020).

Learners with higher proficiency are likely to have more developed retrieval structures, allowing them to access and integrate their L2 knowledge more efficiently. Baddeley's episodic buffer within the WM model plays a critical role by temporarily storing and incorporating elements from LTM, allowing for the simultaneous activation of L2 knowledge and WM processes, thereby facilitating comprehension. Randall (2007) suggests that learners with lower proficiency rely more heavily on WM to manage fundamental language tasks, such as phoneme recognition, syntax, and semantics. As proficiency increases, these processes become more automatic, freeing up WM resources to focus on higher-order tasks like interpreting meaning and understanding complex syntax. This shift reflects greater processing efficiency, enabling learners to better manage and chunk information.

We chose listening as our focus because listeners must process incoming language continuously, which may impose larger demands on the efficiency of working memory. Since listeners are unable to pause and review material during the comprehension process, the cognitive load and resource utilization become more demanding compared to reading, where information can be revisited.

Working memory measurements

In the assessment of working memory, two distinct types of tasks are commonly utilized: simple span tasks and complex span tasks. Each of these tasks targets different aspects of working memory, revealing various facets of its functionality. The common methods for measuring WM and the components that these tasks assess are shown in Table 1.

Simple span tasks are represented by the digit span test introduced by Jacobs in 1887. These tasks are primarily designed to measure the phonological short-term store and articulatory rehearsal capabilities of working memory. In such tasks, participants are typically presented with a sequence of items, like digits or words, and are asked to recall them in the order presented. These tasks assess the basic ability of an individual's WM to maintain phonological information through rehearsal processes, focusing on the capacity to store and reproduce information without significant transformation or manipulation.

On the other hand, complex span tasks, like the RST and the LST, require simultaneous processing and storage of information, placing a higher cognitive demand on the individual. Participants engage in processing activities, such as reading or listening to sentences, while also being required to remember specific elements from what they have processed, like the final word in each sentence. The dual demands of complex span tasks engage participants in a series of cognitive processes, namely updating, switching, and inhibition, each playing a critical role in the task's successful execution. During the task, as new sentences are presented, participants are required to process this novel information while concurrently updating their memory with details they need to remember. Additionally, the complex span tasks also demand proficient switching between various cognitive operations. Participants are expected to comprehend the meaning of each new sentence while also remembering specific words from these sentences. Another essential aspect of these tasks is the ability to suppress irrelevant information. Participants must concentrate on remembering only the target words, actively inhibiting the recall of non-target words contained within the sentences. This selective attention and inhibition are essential to prevent the WM from being overloaded with unnecessary information. This careful coordination of cognitive processes ensures the successful completion of complex span tasks.

In summary, while simple span tasks focus on the basic storage capacity of working memory, complex span tasks like the RST and LST provide a more comprehensive assessment. They evaluate not only the storage capacity but also the processing efficiency, including the ability to update, switch, and inhibit information. These tasks highlight the dynamic and multifaceted nature of working memory, especially in contexts demanding high cognitive functioning, such as language comprehension and learning.

Issues of varying scoring methods of WM measurements

Different scoring schemes, such as the SRT, the LST, and the RST, are employed for WM assessments, with each having specific implications and considerations. The original scoring strategy for the RST and LST, which were created by Daneman &

Carpenter (1980), was to determine the span or the longest string of phrases during which the participant could accurately recall the target word in each sentence. However, subsequent research indicated that alternative scoring methods might offer greater reliability. For instance, Friedman & Miyake (2005) argued that scoring based on word recall, rather than span, could provide a more reliable indicator of WM capacity. This method counts the total number of correct words recalled across all sentence sets, regardless of whether the entire set was correctly recalled.

Caplan & Waters (2005) proposed a more comprehensive grading method that includes not only recall but also errors and reaction time, creating a composite score. This approach recognizes that WM capacity involves not just the ability to recall information but also the efficiency and accuracy with which this information is processed and retrieved. Thus, including errors and reaction times in the scoring provides a more thorough understanding of WM performance, particularly in tasks that require both the processing and recall of information, such as the RST and LST.

The SRT, often used in language studies, traditionally employs an all-or-nothing scoring approach. In this method, a participant's recall is considered correct only if an entire sentence is recalled accurately. However, this approach can lead to a floor effect, especially in second-language contexts. A floor effect occurs when the task is too challenging for most participants, resulting in a majority of scores clustering at the lower end of the scoring range. This is a significant issue in L2 settings, where participants may struggle more with language-based tasks, thus limiting the effectiveness of the all-or-nothing scoring method in accurately assessing their WM capacity.

Therefore, the choice of scoring method in WM assessments is crucial and can significantly impact the interpretation of results. While original methods focused on span, subsequent research suggests that incorporating additional measures like word recall, errors, and reaction times can provide a more comprehensive assessment of working memory. Furthermore, the suitability of scoring methods may vary depending on the context, such as in L2 applications, highlighting the need for adaptable and nuanced approaches in WM research.

Issues of reliability and validity of WM measurements

Reliability and validity are especially important concerns when it comes to WM assessments. The dynamics of WM and its implications in different cognitive processes can be better understood using measurement tasks. However, ensuring that these assessments reliably and validly measure what they are intended to is crucial for the integrity of research findings. Both reliability and validity are foundational to the utility and applicability of WM assessments in cognitive research and practical applications.

Many studies have focused on the reliability of WM assessments, particularly when first-language and second-language use are involved. Waters & Caplan (2003) conducted an extensive review of existing studies to summarize the reliability of various WM assessments in L1 contexts. Their work provides valuable insights into how consistently these tasks measure WM capacity. However, there has not been as much research done on how reliable these activities are in L2 environments. This gap in research is highlighted by Shin's (2020) meta-analysis, which looked at thirty-

seven papers and discovered that just nine of them documented the reliability of the Reading Span Task (RST). Importantly, these studies included assessments in both L1 and L2 contexts. This finding indicates a lack of specific data on the reliability of the RST in L2 contexts. Understanding the effectiveness and application of these WM measures across varied linguistic backgrounds is greatly constrained by the lack of specific reliability data for L2 learners, given the extra cognitive challenges associated with processing a second language.

The validity of WM assessments, especially the RST, has also come under scrutiny. Turley-Ames and Whitfield (2003) raised concerns about the validity of the RST as a measure of WM capacity. They found that when participants were taught specific strategies, their performance on the RST improved. The validity of the RST as a single indicator of WM is called into doubt by this study, which suggests that it may be influenced beyond WM capacity, such as test-taking strategies or learning effects. Such results suggest that the RST might be tapping into a broader set of cognitive skills than initially intended. To address these validity concerns, Friedman and Miyake (2004) emphasized the need for employing well-established measures of criterion validity when evaluating WM tasks. They asserted that for a WM measure like the RST to be considered valid, it should exhibit a strong correlation with other recognized measures of cognitive processes it intends to assess. This methodology is crucial to guarantee that WM tasks appropriately assess the cognitive constructs they claim to measure and decrease the impact of external factors.

In summary, while tasks like the RST are invaluable tools in cognitive research, their reliability, especially in L2 contexts, and their validity as true measures of WM capacity require further investigation. The insights from researchers such as Waters and Caplan (1996), Shin (2020), Turley-Ames and Whitfield (2003), and Friedman and Miyake (2004) point toward the necessity for continuous evaluation and improvement of WM assessment methods. Ensuring the reliability and validity of these tools is necessary for advancing our understanding of WM and its role in cognitive functions.

Purpose of the present study

The purpose of the present study is to critically evaluate the reliability and validity of various WM measurements, with a focus on auditory processing tasks. This investigation is driven by existing concerns regarding the reliability and validity of widely used tasks in WM research.

Research questions

The present study addresses the following two research questions:

1. Which of the WM measurements is the most reliable in terms of test-retest reliability internal consistency, and Rasch person and item reliability?
2. Which of the WM measurements has the highest criterion-related validity with the listening comprehension ability in L2 English?

Methods

Participants

The participants were recruited on a university campus in Japan through flyers. A total of 39 undergraduate and postgraduate students who had been learning English as a foreign language since primary school responded. Of these 39 participants, 19 were Japanese and 20 were Chinese, with an average age of 23.54. One participant did not partake in the retest, and two completed the second test after 2 months, a significantly longer interval compared to other participants who retested after a month. Their listening skills were assessed using the listening subsection of the TOEIC Listening Test, with scores showing enough variation for analysis (median score of 49, ranging from 24 to 59). The scores indicate that they possess a certain level of English proficiency, with an average score of 47.15 out of 60 ($sd = 8.29$). Although we recruited participants with different L1 backgrounds, separate t-tests revealed no significant differences in their scores on any task. Since our goal is to verify the method of measuring WM efficiency, which could contribute to English listening ability, we did not take into account the participants' knowledge of a third language or their L1 dialects as contributing factors.

Procedures

Participants were administered the TOEIC Listening Test and 5 WM measurement tasks in the first test, and re-took the tasks again after about one month. It took approximately 90 to 120 minutes in the first test and 60 minutes in the retest. There was a break during each task to avoid an overwhelming burden. All the tasks were created by HSP 3.6, Psychopy v3.2.4, and Psychopy v2023.1.2.

TOEIC Listening Test

The TOEIC Listening Test was chosen as the benchmark for English proficiency due to the specific tasks in Parts 2 and 3. In Part 2, participants listened to a sentence followed by three responses to choose from. Since the sentence and three answer choices were not written, task required participants to temporarily store them and process the meaning. In Part 3, participants processed and remembered conversation content, selecting answers to questions about it. These tasks inherently involve using WM to process and retain language information, aligning well with the requirements of domain-specific WM measurement. This enhances the TOEIC Listening Test's suitability as a criterion for evaluating English ability.

Digit span task (DST)

The Digit Span Task¹ (DST) is a cognitive test commonly used to measure WM capacity. Unlike the other four domain-specific WM measurements, the DST was administered to assess participants' domain-general capacity of WM for comparison and control.

The DST consisted of digit sequences of varying lengths presented in participants' L1, which is more appropriate for measuring domain-general WM. Specifically, the task encompassed sequences ranging from 4 to 10 digits, with three

sequences administered at each level of complexity. Participants were told to pay close attention as the program auditorily presented them with various digit sequences. Following each sequence's presentation, participants were charged with recalling and repeating the sequence in the exact order in which it was delivered.

To assess performance on the DST, we adopted Baddeley et al.'s (2009) scoring method. Each recalled digit was considered correct if it appeared in the appropriate position relative to an adjacent recalled digit. The absolute serial position was taken into account only for the first and last words in a sequence. These words were scored as correct if they were produced in their respective positions within the recall sequence.

Listening span task (LST)

In this study, we employed the LST as a measurement of domain-specific WM that is widely used in the literature. The decision to use the LST rather than the more common RST is our focus for assessing WM efficiency in dealing with auditory stimuli and evaluating the validity of the measurement in relation to listening ability. This task is chosen for its capability to engage participants in simultaneously processing and remembering linguistic information, thereby making it particularly suitable for assessing WM with a focus on the domain-specific aspects of language processing. This task is more cognitively demanding than simple DST, as it necessitates participants to both store and manipulate information while comprehending sentence structure. Since Daneman and Carpenter's (1980) LST was originally created for native language (L1) users, including difficult words for second language learners, sentences from Ushiro and Sakuma's (2000) RST and LST were utilized with their permission.

The task was created using PsychoPy 3.2.4 and was administered individually on a computer with participants wearing headphones. The LST has multiple levels, ranging from 3 to 6 sentences, with each level including three sets. In this task, participants were given a series of sentences to process. After each sentence, they were instructed to press the "left" (No) or "right" (Yes) button to indicate whether or not the sentence made sense. Following that, participants were asked to memorize a single word. After finishing each set, participants were asked to recall every word they had been instructed to memorize. Crucially, regardless of the word sequence, participants were given credit for every word they correctly remembered from every set. Participants were given a break after completing every single level.

In the LST, we recorded various measures, including participants' reaction times for their judgments, the error of their judgments, and the total number of words correctly recalled. The overall evaluation of the LST in our study thus encompasses both the processing and storage components. Following Waters and Caplan's (1996) study, we divided the data into two sections: the processing section (RT + judgment error) and the storage section (word recall accuracy) along with the overall evaluation of LST (processing + storage). This approach aligns with Waters and Caplan's recommendation to use a composite Z score that incorporates word recall, judgment error, and judgment RT for scoring tasks like the RST and LST.

Sentence recall task (SRT)

SRT demonstrates an advantage attributed to the automatic engagement of long-term language structure, leading to a more effective binding of information within the episodic buffer, as emphasized by Baddeley (2000) and Baddeley, Allen, and Hitch (2010). The task's significance in capturing language processing-specific episodic buffer functions is further supported by this observation. Additionally, the utilization of the SRT is supported by studies such as Alloway et al. (2004) and Baddeley et al. (2009), who, in their L1 research, have highlighted its appropriateness for assessing methods suitable for measuring the domain-specific (L2) episodic buffer function. Even though it is acknowledged, there is still variation in how it is administered and scored, indicating that it needs to be improved and standardized.

In this task, thirty statements ranging in length from 4 to 20 words were chosen from junior high school textbooks. To ensure that the meaning of the sentences was general and familiar to participants, these sentences were carefully chosen to eliminate any proper nouns. Both the written SRT (WrittenSRT) and spoken SRT (SpokenSRT) involved auditory presentation of sentences, but they differed in the mode of response required from the participants. Participants were instructed to recall the sentences immediately following the presentation of each sentence.

For the SRT, we used two scoring methods: one that counted the overall number of words properly recalled (Total), and another that determined the maximum continuous count of words a participant could reconstruct (Max).

Sentence recognition task (SRecogT)

While the SRT has been utilized as a measurement of WM in previous studies (Alloway et al., 2004; Baddeley et al., 2009; Pham & Archibald, 2023), it has the potential deficit of being difficult to score. The recalled sentences must be transcribed to identify the accuracy of recall, making automatic scoring impossible. To address this issue, Moustapha (2022) devised a Sentence Recognition Task in her unpublished Master's thesis, where participants heard a sentence followed by visually presented words and phrases that they had to judge whether they appeared in the sentence. Studies (Gathercole et al., 2001; Hulme et al., 1997; Jefferies et al., 2006; Macken et al., 2014) have demonstrated that the use of recognition memory paradigms can simplify the item retrieval process. This simplification often results in reduced or missing lexicality effects, where the expected differences in memory performance based on the lexical properties of words are either less pronounced than usual or entirely absent.

In this task, participants were presented with a total of 19 sentences. Following the listening of each sentence, a sequence of words or phrases was shown one at a time. Participants were then asked to judge whether each word or phrase appeared in the sentence by clicking the "Y" (Yes) or "N" (No) button. For analysis, reaction times as well as the accuracy were recorded.

Analysis method

The analysis was conducted using R ver. 4.2.0. This study's data analysis includes both reliability and validity assessments. For reliability testing, Cronbach's Alpha Coefficients were calculated to assess the internal consistency of test items, and Rasch person and item reliability as well as the separation index were calculated to determine how well each measurement assesses individual's ability and item difficulty. Rasch person reliability assesses the consistency of individual scores across different items on a test or assessment. Specifically, it indicates the extent to which a person's performance is stable and replicable, accounting for the difficulty of the items and the ability of the individual. A higher reliability value suggests that the measurement is dependable, indicating that the person's scores reflect their true abilities rather than being influenced by random error or item characteristics. Item reliability, on the other hand, assesses how accurately the difficulty levels of the test items are estimated. It compares the variance in item difficulty with measurement error. A high item reliability (typically above 0.9) suggests that the items are well-targeted and cover a wide range of difficulty levels, demonstrating the robustness of the test's design. Furthermore, test-retest reliability was calculated to ascertain the stability of each measurement.

The study also sought to establish criterion-related validity by examining the relationship between the WM indices used and the TOEIC Listening Ability. The extent to which WM indices explain variations in listening ability was also determined with the help of a Generalized Linear Mixed Effects Model (GLMM). As stated earlier, a primary objective of this study was to identify the specific type of WM measurement that exhibits the strongest correlation with English listening proficiency. Through investigating these aspects, the study aimed at strengthening our comprehension of the function of WM in English listening proficiency and provide significant perspectives for subsequent investigations in this domain.

All the continuous variables were converted to standardized scores to facilitate comparison. The reaction time data was log-transformed before standardization to normalize the distribution and to improve the linearity of the data.

Results

Descriptive statistics

Table 2 presents the descriptive statistical analysis for each performance metric across different measurements. All indices will be reported separately in the following session.

Cronbach's alpha and Rasch reliability

Table 3 presents the results of Cronbach's alpha, which shows that the WrittenSRT Max, WrittenSRT Total, and SpokenSRT Total have high internal consistency. This result implies that the items on these scales evaluate the same construct consistently, demonstrating high reliability in assessing sentence recall ability. Similarly, the DST, LST recall accuracy, SpokenSRT Max, and SRecogT all show a high level of internal consistency. However, the LST Judgment shows that its items have a moderate level

Table 2. Descriptive statistics in working memory measurement tasks

		First test				Second test			
		min	max	M	sd	min	max	M	sd
DST	digits	95	147	127.59	14.55	99	147	132.27	12.24
LST	Judgment error	1.00	25.00	10.36	5.77	2.00	32.00	8.92	6.71
	RT(s)	6.12	21.05	8.65	2.41	6.49	16.52	8.14	1.81
	Recall words	22.00	49.00	35.54	7.03	16.00	48.00	36.05	8.54
SpokenSRT	Max	1.60	7.27	4.25	1.45	0.00	9.33	5.00	1.86
	Total	2.60	10.67	6.64	2.05	0.00	11.33	7.08	2.20
WrittenSRT	Max	1.33	9.13	5.01	2.00	1.47	8.27	4.94	1.52
	Total	2.47	11.00	7.32	2.20	2.13	11.27	7.90	2.02
SRecogT	Judgment accuracy	100.00	154.00	128.28	12.86	100.00	150.00	126.51	12.39
	RT(s)	.80	1.80	1.31	.21	.86	1.76	1.24	0.21

Note: DST: Digit Span Task. LST: Listening Span Task. SRT: Sentence Recall Task. SRecogT: Sentence Recognition Task. Full score of word recall for LST is 54. RT: response time
 Max: Maximum length of continuous words recalled correctly. Total: Total words recalled correctly.
 The full score of accuracy for SRecogT is 162.

of internal consistency. While not as high as the preceding scales, this level of consistency may still be considered acceptable depending on the unique context and research aims. It shows that the judgment portion of the LST may have slightly more variability in its measurements when compared to other scales.

The results of the Rasch Model analysis for various tasks, as shown in Table 4, indicate varying levels of reliability and separation indexes for both persons and items.

Both scoring methods of WrittenSRT tasks exhibit high person reliability (Max: .96, Total: .97) and strong separation indices (Max: 4.98, Total: 5.44), indicating excellent individual performance consistency and ability to differentiate between various abilities. Item reliability for these two tasks is also high (Max: .92, Total: .91), suggesting effective measurement of item difficulty. Similarly, the SpokenSRT tasks show high person reliability (Max: .94, Total: .96) and good separation (Max: 3.90, Total: 5.19), reflecting strong consistency and ability to distinguish between different levels of SpokenSRT task performance. The item reliability and separation are similarly robust. For the LST Judgment error, the person separation (1.70) and item separation (1.42) are relatively low, indicating that this measure has limited capability to differentiate between individual abilities and the difficulty levels of test items. The moderate person reliability (.74) and lower item reliability (.67) further support this interpretation, indicating a challenge in consistently assessing both individual abilities and item difficulties. Similarly, while the LST Recall Accuracy test shows slightly better indices, with person separation at 1.96 and item separation at 2.49, the person separation index is also not enough to discriminate the individual's ability. The SRecogT Judgment demonstrates effectiveness in measuring individual abilities, as evidenced by its high person separation (2.59) and person reliability (.87) result. Despite the satisfactory item

Table 3. Cronbach's alpha

	Cronbach's alpha value
Toeic	.89
DST	.85
LST Judgment error	.78
LST Recall accuracy	.82
WrittenSRT Max	.93
WrittenSRT Total	.94
SpokenSRT Max	.88
SpokenSRT Total	.92
SRecogT Judgment	.87

Table 4. Result of rasch model analysis

	Person		Item	
	Separation index	Reliability	Separation index	Reliability
Digit Span Task	.35	.11	2.04	.81
LST Judgment error	1.70	.74	1.42	.67
LST Recall accuracy	1.96	.79	2.49	.86
WrittenSRT Max	4.98	.96	3.48	.92
WrittenSRT Total	5.44	.97	3.18	.91
SpokenSRT Max	3.90	.94	3.18	.91
SpokenSRT Total	5.19	.96	3.43	.92
SRecogT Judgment	.88	.44	1.46	.68

dependability (.76), the item separation score of 1.93 indicates that it may not be able to separate the item difficulty into more than two levels.

While all the domain-specific tasks exhibited high Rasch person and item reliability, the DST revealed surprisingly low person reliability. Upon closer examination of the person measures, it was observed that some individuals could recall 10 digits perfectly but still missed recalling one of the 4 digits. The person separation index was also low, indicating that the variance was too small to discriminate among individuals' abilities to recall digits. However, item reliability and separation index were satisfactory, suggesting that larger digits are more difficult to recall than the smaller digits.

Table 5. Statistic result of test-retest reliability

	correlation
DST	.71
LST RT	.92
LST Judgment error	.80
LST Recall accuracy	.76
LST processing (Judgment error + RT)	.87
LST composite (Judgment error + Recall + RT)	.90
Written Max	.89
Written Total	.93
Spoken Max	.88
Spoken Total	.91
SRecogT Judgment	.88
SRecogT RT	.78

Test-retest reliability

Achieving test-retest stability would be important when assessing the reliability of a test, which means the test results should be stable when people take the test over time. The reliability of each measurement in this study is demonstrated through the correlations presented in Table 5.

The Digit Span Test exhibited the test-retest correlation of .71, which is considered to be moderately correlated. The strong correlation of the LST Judgment error and LST Recall accuracy, which are .80 and .76, respectively, indicates that there is consistency in these measures. The processing of the Listening Span Test, which combined both judgment and reaction time, also showed a significant correlation of .87 between the two test administrations, highlighting the stability of this measure. The strong correlation of .92 for LST reaction times strengthens the measure's reliability, suggesting response time stability. The reliability of the combined assessment was reinforced by the strong correlation of .90 which was shown in the overall LST composite score. Notably, strong test-retest reliability was also demonstrated by the evaluations in both spoken and written modalities, as shown by the high correlations for measures of WrittenSRT (with a max of .89 and a total of .93) and SpokenSRT (with a max of .88 and a total of .91). Additionally, the SRecogT produced a constant .78 correlation in the reaction time, and the test also produced a dependable .89 correlation in the judgment of SRecogT. These results corroborate the assessments' usefulness in assessing WM by confirming their stability and reliability across time. This study's investigation of the connections between WM and second language acquisition outcomes across several task modalities is made possible by the measures' significant test-retest reliability, which increases the credibility of the results.

Table 6. Relationship between the TOEIC Listening Test and WM measurement tasks

	correlation
DST	.22 n.s
LST RT	.02 n.s
LST Judgment error	-.71***
LST Recall accuracy	.58***
Written Max	.75***
Written Total	.84***
Spoken Max	.70***
Spoken Total	.71***
SRecogT Judgment	.76***
SRecogT RT	-.18 n.s

Note: *** $p < .01$.

Correlation between TOEIC Listening Test and WM indices

From Table 6, we can observe the relationship between the TOEIC Listening Test and the WM measurements. It appears that aside from the DST and the reaction time of LST and SRecogT, all the administered tasks had a strong association with TOEIC Listening Test scores. The WrittenSRT has emerged as the most highly connected with TOEIC performance among these measurements.

GLMM

To assess the extent to which each WM measurement can explain the variance in TOEIC listening scores, we conducted analyses using Generalized Linear Mixed Effects Models. All the continuous variables were converted to standardized scores to facilitate comparison. The reaction time data was log-transformed before standardization to normalize the distribution of the data. Table 7 presents the models that have been analyzed. The models encompass the fixed effect of each WM measurement, along with random intercepts for participants and TOEIC test items. Additionally, participant random slopes are included for each WM measurement, based on the hypothesis that individual variations in the relationships between WM measurement and TOEIC listening scores may exist and contribute to the overall model's explanatory power.

Table 8 shows the results for different scales and constructs, which gave us important insights into the efficacy of our statistical models and the variables affecting the phenomena we were studying. In our study, the primary goal was to determine which measurements provide the most reliable relationship between WM efficiency in L2 and L2 proficiency. To achieve this, we opted to run separate models for each measurement, rather than combining all the tasks into a single model. This

Table 7. Analysis codes for Generalized Linear Mixed Effects Model

	alpha
	TOEICaccuracy~
DST	(1 + DST.z ID) + (1 ToEICItem) + DST.z
LST RT	(1+LSTrtLog.z ID)+(1 ToEICItem)+LSTrtLog.z
LST Judgment error	(1 + LSTerror.z ID) + (1 ToEICItem) + LSTerror.z
LST Recall accuracy	(1 + LSTrecall.z ID) + (1 ToEICItem) + LSTrecall.z
LST processing	(1 + LSTprocessing ID) + (1 ToEICItem) + LSTprocessing
LST composite	(1 + LSTcomposite ID) + (1 ToEICItem) + LSTcomposite
Written Max	(1 + WrittenSRTmax.z ID) + (1 ToEICItem) + WrittenSRTmax.z
Written Total	(1 + WrittenSRTtotal.z ID) + (1 ToEICItem) + WrittenSRTtotal.z
Spoken Max	(1 + SpokenSRTmax.z ID) + (1 ToEICItem) + SpokenSRTmax.z
Spoken Total	(1 + SpokenSRTtotal.z ID) + (1 ToEICItem) + SpokenSRTtotal.z
SRecogT Judgment	(1 + SRecogT.z ID) + (1 ToEICItem) + SRecogT.z
SRecogT RT	(1+SRecogTrtLog.z ID)+(1 ToEICItem)+SRecogTrtLog.z

Table 8. Statistic results of the Generalized Linear Mixed Effects Model

Fixed Effects	Estimate	SE	t	p	Conditional R ²	Marginal R ²
Intercept						
DST	.24	.17	1.42	.16 n.s	.70	.02
LST Judgment error	-.78	.10	-7.594	3.11e-14***	.70	.18
LST Recall accuracy	.61	.15	4.068	4.75e-05***	.69	.11
LST processing	-.44	.13	-3.49	.000488***	.68	.12
LST composite	.37	.07	.07	1.35e-07***	.69	.16
Written Max	.87	.14	6.04	1.51e-09***	.67	.25
Written Total	.97	.09	10.51	<2e-16***	.71	.27
Spoken Max	.82	.16	5.22	1.82e-07***	.69	.21
Spoken Total	.79	.15	5.30	1.17e-07***	.69	.19
SRecogT Judgment	.80	.133	6.04	1.59e-09***	.67	.21

approach avoids potential confounding effects that could arise from including all tasks in a single model, allowing a clearer understanding of how each task independently relates to WM and proficiency. After running these separate models, we compiled the results into a single table for easier interpretation and comparison of the measurements' effects.

The DST, while showing a positive trend (Estimate = 0.24), was not statistically significant ($p = 0.16$), suggesting that, on its own, DST may not be a reliable predictor of WM efficiency in the L2 context. However, several other tasks show significant results. LST Judgment error had a strong negative effect (Estimate = -0.78 , $p < .05$), suggesting that individuals with fewer errors in judgment tasks demonstrate better WM performance in L2. LST Recall accuracy was positively associated with WM efficiency (Estimate = 0.61, $p < .05$), implying that better recall accuracy leads to higher efficiency. The LST composite score, which combines different aspects of the task, further confirms a positive relationship with WM (Estimate = 0.37, $p < .05$). In addition to LST measures, written tasks were strong predictors. Written Max and Written Total both had highly significant positive effects (Estimate = 0.87, $p < .05$ and Estimate = 0.97, $p < .05$, respectively), indicating that higher written task performance strongly correlates with better WM in L2. Similarly, Spoken Max and Spoken Total were also significant predictors (Estimate = 0.82, $p < .05$ and Estimate = 0.79, $p < .05$) showing that proficiency in spoken tasks is linked to improved WM efficiency. Lastly, SRecogT Judgment exhibited a significant positive effect (Estimate = 0.80, $p < .05$), reinforcing the importance of accurate judgment in supporting working memory.

The conditional R-squared values, consistently around .70 for various scales, indicate that statistical models, which include both fixed and random effects, are effective at explaining a significant portion of the variability. In other words, these models account well for the complex and individual variations. Notably, WrittenSRT Max stood out for having an explanatory power of .71. These results underlined how thorough our approach was in recognizing the complexity of the data, and they served as a strong foundation for our interpretations and implications that followed.

This finding emphasizes the significance of the WrittenSRT as a promising predictor of participants' TOEIC listening proficiency. It also emphasizes the task's potential importance in assessing and improving English listening abilities in the context of language learning and assessment. A detailed investigation into the specific components and processes underlying this strong association might provide insightful findings for research.

Discussion

Reliability of the working memory measurements

One prominent finding is the consistently high test-retest reliability observed across various assessments and modalities. In the DST, a moderately correlated performance suggests reasonable stability across repeated administrations. High correlations in the LST Judgment error and word recall indicate commendable stability, reinforcing their reliability. The LST processing, combining judgment error and reaction time, reveals a significant correlation between repeated administrations, suggesting a stable measure for assessing working memory. The overall LST composite score maintains a strong correlation, affirming a comprehensive and consistent evaluation of working memory. The SRT, both spoken and written tasks, demonstrate strong test-retest reliability. The SRecogT also maintains stability, evident in reaction time and overall test correlations.

By using Cronbach's alpha, the tasks' internal consistency is further emphasized, demonstrating their reliability in measuring the same underlying construct. The consistently high internal consistency observed in tasks related to sentence recall and recognition strengthens the reliability of these measurements. The WrittenSRT, both in its maximum and total scores, demonstrates exceptionally high internal consistency with Cronbach's alpha values. Similarly, the SpokenSRT, in both maximum and total scores, exhibits strong internal consistency. The DST and SRecogT also show high internal consistency, with alpha values indicating reliable measurements of WM and phrase recall skills. The LST, while exhibiting a moderate level of internal consistency, raises interesting considerations. Although not as high as some other scales, the acceptable consistency level suggests that the LST maintains reliability. Depending on the research goals and context, this level of consistency may still be deemed acceptable.

The Rasch Model analysis reveals a comprehensive picture of the reliability and validity of various WM measurements. High reliability and separation indices in tasks like WrittenSRT Max and WrittenSRT Total indicate robustness in measuring WM capacity. Moderate reliability in LST suggests that it is somewhat reliable but less discriminative. The SRecogT also demonstrates good reliability and ability to differentiate item difficulty. Overall, the result of Rasch analysis suggests that most WM measurements used are reliable and valid tools for assessing WM in second language acquisition (SLA) contexts, with SRT showing stronger discriminative power than others.

Validity of the working memory measurements

The substantial correlation observed between WM measurements and TOEIC Listening Test scores underscores the close connection between domain-specific WM and language proficiency. Notably, the WrittenSRT emerges as the measurement with the tightest correlation, indicating its potential to predict participants' listening proficiency.

However, the non-significant relationship between the DST and the TOEIC Listening Test prompts exploration of potential factors influencing these findings. The unique characteristics of domain-specific WM in English language proficiency may not be fully captured by domain-general WM tests, according to our findings. This suggests that the domain-general WM capacity is irrelevant to the comprehension of second language with the participants in the present study. However, it is possible that the DST used in this study may not have adequately measured domain-general working memory. Further research incorporating longer digit sequences or a more demanding backward digit span requiring manipulation could yield different results. These revelations highlight the necessity of taking subtle factors into account when analyzing the connection between linguistic competency and working memory.

The LST investigation in the study emphasizes the trade-off between ecological validity and cognitive involvement. The LST is deemed ecologically invalid because it requires participants to remember a sentence-final word or a word irrelevant to the sentence being processed, which is not reflective of natural language use.

In real-world communication, listeners typically focus on the meaning of the entire sentence or conversation rather than isolating sentence-final words. Furthermore, Schwering and MacDonald (2020) point out that LST is based on the theoretical assumption that memory for words is separable from memory for order. However, they argue that linguistic representations in long-term memory consist of statistical regularities of language, which inherently encode the serial ordering of words, challenging this assumption. Despite this, the controlled and ecologically invalid setup of the LST allows it to engage key cognitive processes, such as inhibiting competing words, updating words for recall, and switching between sentence processing and word recall. Despite concerns over ecological validity, the LST showed a strong correlation with English language proficiency, indicating that it remains a valuable tool for tapping into the executive WM processes crucial for language comprehension and retention. This correlation highlights the task's utility in capturing essential aspects of language processing, providing important insights into the cognitive foundations of language skills.

The SRT simulates naturalistic language comprehension and recall, providing a valuable means to examine how WM operates during real-world language processing tasks. Both the spoken and written versions demonstrated a significant relation with the TOEIC task, albeit with the written version exhibiting a higher correlation. This might be attributed to the multifaceted nature of the spoken task, requiring individuals to hold auditory information in working memory, formulate verbal responses, and produce speech. The heightened cognitive load involved could subjectively increase the task's difficulty. Notably, this perception contrasts with the expectation that the spoken task might be easier due to the absence of the need to spell out words. Additionally, the perceived burden of speaking in front of the experimenter may impact cognitive performance. Moreover, while speaking requires immediate recall of a sentence, participants can read and rewrite the sentence in a written mode during recall. This gives the written mode an advantage in further utilizing the episodic buffer to reconstruct the sentence. However, speaking involves processing various phonological features, such as tone and pronunciation, posing unique challenges compared to writing. This finding draws attention to the complexity of WM engagement in different language modalities.

The SRecogT evaluates participants' ability to remember, process, and comprehend sentence-level information, providing a multifaceted perspective on language-related working memory. Given its potential as a reliable measure of domain-specific WM efficiency, it showed the second-highest correlation with TOEIC scores. This significant link could be explained by the task's increased ecological validity, and the usage of displayed words may help to lessen processing burden. Thus, it becomes evident that the SRecogT is a useful instrument for evaluating language memory in a way that is relevant to the context.

Conclusion

This study places substantial emphasis on the parameters of WM measurement within the realm of second language acquisition. Our study adds to a fundamental knowledge of how WM measures affect language acquisition research by examining

the nuances of these evaluations. The results offer significant insights into the subtleties of WM measurement, contributing to the field of second language acquisition in its provision of refined and contextually relevant WM measurement methods. By bridging the gap in the effectiveness and reliability of these assessment approaches, this research helps evaluate the subtle interactions between WM and language learning. It offers valuable insights into the functioning of WM during second language acquisition, paving the way for more effective training methods and learning strategies.

Understanding, problem-solving, adjusting to new information, and learning a new language are all significantly impacted by working memory, which is an essential cognitive skill. Considering variables like language proficiency, cognitive demands, and language exchanges, adapting the research and ideas from first language acquisition to the unique setting of second language learning can be challenging. This paper draws attention to the shortcomings of common WM tasks, which frequently have insufficient validation and verification, making it difficult to use them to assess how well second language WM functions.

To sum up, this study extends our understanding of the reliability and validity of WM measurements in the context of second language acquisition. It highlights how important it is to use specific WM tasks that account for the intricacies of language acquisition. By examining the validity and reliability of different assessment tools in the context of language learning, we offer guidance to researchers and practitioners for a more accurate understanding and evaluation of the role of WM in language acquisition.

Limitations and future research directions

However, it is crucial to acknowledge the limitations of this study. The backgrounds and language abilities of the participants may have an impact on how broadly applicable the findings of the research can be. To fully understand the underlying mechanisms and factors of the profound connection between language proficiency and working memory, further research is necessary. Additionally, while the emphasis on listening skills aligns with the TOEIC test and the study's objectives, it restricts the generalizability of the findings to other language modalities, such as speaking, reading, and writing. For instance, the cognitive processes involved in listening, like auditory processing and retention, might differ significantly from those in reading, where visual processing and comprehension play a more prominent role. Examining the impact of task modality on perceived difficulty and cognitive performance could provide valuable insights into the complex interactions within language-related working memory.

Practically speaking, this research holds substantial significance. Educators and researchers can employ these validated WM measurement methods to develop tailored instructional materials targeting specific components of working memory. It is worth noting that recent research trends are increasingly focused on developing effective WM training methods to enhance language learning outcomes (e.g.: Karousou & Nerantzaki, 2022; Marashi & Sadinezhad, 2022; Santacruz & Ortega, 2018; Peng & Swanson, 2022; etc.). However, existing WM tasks used for training do

not effectively enhance language skills. Schwering and MacDonald (2020) and Peng and Swanson (2022) suggest that training focused on domain-general attentional control and overall capacity is unlikely to be effective. Instead, they argue that WM is deeply interconnected with language LTM, and training is only effective when it directly targets improvements in language skills.

Existing domain-specific tasks, such as RST and LST, are also unlikely to serve as effective training tasks for improving language skills. Moreover, their impracticality for classroom use further limits their applicability. In this context, SRT shows potential as both a measurement and training tool. Teachers may use this task to encourage students to vocally reproduce longer sentences, leveraging the interaction between WM and LTM.

Having reliable and valid measures of WM is critical, as they serve as the foundation for evaluating the effectiveness of these training programs. Our study contributes to this by validating WM measurement methods, which are essential for researchers aiming to improve language learning interventions through targeted WM training. It can also be applied to research studies that employ WM as a measure to assess the effectiveness of different instruction strategies or L2 learning methodologies.

Furthermore, the WrittenSRT is considered a reliable indicator of WM efficiency, which influences the listening comprehension, emphasizing its importance in language education and assessment. Additionally, the slightly higher correlation for the written version suggests that researchers may consider exclusively using the written task in future studies testing working memory. This approach simplifies data processing and reduces the burden of transcribing audio data.

In summary, this study provides a reliable framework for assessing WM in the context of second language acquisition. The findings recognize the relevance of assessment techniques and emphasize the critical role that WM plays in language competency.

Note

¹ While Chinese numbers are all one-syllable long, Japanese numbers can include both one- and two-syllable forms. To account for this difference, we examined whether participants' first language influenced their performance in the DST. A t-test comparing the DST scores of Japanese and Chinese participants showed no significant difference ($t = -0.985$, $p = 0.33$). These results indicate that syllabic variations between the two languages do not significantly affect participants' DST performance, as numbers appear to be processed as single units of information rather than by their syllable count.

Reference

- Alloway, T. P., Gathercole, S. E., Willis, C., & Adams, A. M. (2004). A structural analysis of working memory and related cognitive skills in young children. *Journal of Experimental Child Psychology*, *87*(2), 85–106.
- Baddeley, A. (2000). The episodic buffer: a new component of working memory?. *Trends in Cognitive Sciences*, *4*(11), 417–423.
- Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189–208.
- Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2009). Working memory and binding in sentence recall. *Journal of Memory and Language*, *61*(3), 438–456.

- Baddeley, A., Allen, R. J., & Hitch, G. (2010). Investigating the episodic buffer. *Psychologica Belgica*, *50*(3), 223–243.
- Caplan, D., & Waters, G. (2005). The relationship between age, processing speed, working memory capacity, and language comprehension. *Memory*, *13*(3–4), 403–413.
- Cowan, N. (1999). An embedded-processes model of working memory. *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, *20*(506), 1013–1019.
- Conway, A. R., Cowan, N., Bunting, M. F., Theriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–183.
- Conway, A. R., & Kovacs, K. (2013). Individual differences in intelligence and working memory: A review of latent variable models. *Psychology of Learning and Motivation*, *58*, 233–270.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466.
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309.
- Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. *Psychology of learning and motivation*, *44*, 145–200.
- Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211.
- Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*(4), 581–590.
- Friedman, N. P., & Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory and Language*, *51*(1), 136–158.
- Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *The Quarterly Journal of Experimental Psychology Section A*, *54*(1), 1–30.
- Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: evidence for a reintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1217.
- Ibarra Santacruz, D., & Martinez Ortega, D. (2018). Can Working Memory Strategies Enhance English Vocabulary Learning?. *How*, *25*(2), 29–47.
- Jacobs, J. (1887). Experiments on “prehension”. *Mind*, *12*(45), 75–79.
- Jefferies, E., Frankish, C. R., & Ralph, M. A. L. (2006). Lexical and semantic binding in verbal short-term memory. *Journal of Memory and Language*, *54*(1), 81–98.
- Jefferies, E., Ralph, M. A. L., & Baddeley, A. D. (2004). Automatic and controlled processing in sentence recall: The role of long-term and working memory. *Journal of Memory and Language*, *51*(4), 623–643.
- Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic bulletin & review*, *9*(4), 637–671.
- Karousou, A., & Nerantzaki, T. (2022). Phonological memory training and its effect on second language vocabulary development. *Second Language Research*, *38*(1), 31–54.
- Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*(3), 151–177.
- Kovacs, K., Molenaar, D., & Conway, A. R. (2019). The domain specificity of working memory is a matter of ability. *Journal of Memory and Language*, *109*, 104048.
- Macken, B., Taylor, J. C., & Jones, D. M. (2014). Language and short-term memory: The role of perceptual-motor affordance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1257.
- Marashi, S. M., & Sadinezhad, M. A. (2022). The effect of working memory training on vocabulary recall and retention of Iranian EFL learners: the case of dual N-Back task. *Journal of English Teaching*, *8*(1), 36–48.
- Peng, P., & Swanson, H. L. (2022). The domain-specific approach of working memory training. *Developmental Review*, *65*, 101035.
- Pham, T., & Archibald, L. M. (2023). The role of working memory loads on immediate and long-term sentence recall. *Memory*, *31*(1), 61–76.
- Randall, M. (2007). Memory, psychology and second language learning.

- Schwering, S. C., & MacDonald, M. C.** (2020). Verbal working memory as emergent from language comprehension and production. *Frontiers in Human Neuroscience*, *14*, 68.
- Shin, J.** (2020). A meta-analysis of the relationship between working memory and second language reading comprehension: Does task type matter?. *Applied Psycholinguistics*, *41*(4), 873–900.
- Turley-Ames, K. J., & Whitfield, M. M.** (2003). Strategy training and working memory task performance. *Journal of memory and language*, *49*(4), 446–468.
- Ushiro, Y., & Sakuma, Y.** (2000). Modifying reading and listening span tests for group testing. *JLTA Journal*, *3*, 67–82.
- Waters, G. S., & Caplan, D.** (1996). The capacity theory of sentence comprehension: critique of Just and Carpenter (1992).
- Waters, G. S., & Caplan, D.** (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 550–564.
- Wen, Z.** (2016). *Working memory and second language learning: Towards an integrated approach*. Multilingual matters.
- Wen, Z., Mota, M. B., & McNeill, A.** (Eds.). (2015). *Working memory in second language acquisition and processing* (Vol. 87). Multilingual Matters.
- Moustapha, M.I.** (2022). Can speech recognition in Japanese L2 learners of English be improved with hearing-in-noise practice? [Unpublished master dissertation]. Nagoya University.