The Construction and Validation of a New Listening Span Task

Bartolo Bazan
bazanlinkin2@gmail.com
Department of English, Ryukoku University Heian Junior & Senior High School
https://doi.org/10.37546/JALTSIG.TEVAL25.1-4

Abstract

The listening span task is a measure of working memory that requires participants to process sets of increasing numbers of utterances and store the last word of each utterance for recall at the end of each set. Measures to date have contained an exceedingly demanding processing component, possibly leading to insufficient resources to meet the word recall requirement, which may have affected the sensitivity of the measure to distinguish different levels of working memory. Further, tasks thus far have asked participants to verify the content utterances based on knowledge, which may have confounded the measurement of working memory capacity with world knowledge. An additional weakness is that they lack sound psychometric construct validity evidence, which clouds what these tools actually measure. This pilot study presents a listening span task that accounts for preceding methodological shortcomings, which was administered to 31 Japanese junior high school students. The participants listened to ten sets (two sets of equal length of two, three, four, five and six utterances) of short casual utterances, judged whether they made sense in Japanese, and recalled the last word of each utterance in the set. Performance was assessed through a scoring procedure new to listening span tasks in which credit is given for the words recalled in order of appearance until memory failure. The data was analyzed through the Rasch model, which produces evidence for different aspects of validity and indicates if the items in a test measure a unidimensional construct. The results provided validity evidence for the use of the new listening span task and revealed that the instrument measured a single unidimensional construct.

Keywords: working memory, listening span task, validation, Rasch model, Japanese

Working Memory (WM) refers to a mental workspace where information, retrieved from long-term memory, is held and simultaneously manipulated (Baddeley et al., 2002). WM has been closely associated with performance on a wide range of cognitive skills such as first and second language use (Gathercole & Baddeley, 1993; Linck et al., 2014).

A number of complex span tasks, which are tasks that tap both the processing and storage functions of WM, have traditionally been employed to measure WM capacity. An example of such tasks is the listening span task, which is the focus of this study, and in which an individual is asked to verify the grammaticality of each of a series of utterances at the same time as retaining the sequence of the final words of preceding utterances (Daneman & Carpenter, 1980). However, despite the general acceptance of these tasks as measures of WM capacity, few efforts have been made to revise the tasks and/or to collect validity evidence for their use. This is problematic for three reasons. First, the theoretical premise underlying complex WM tasks implies that individuals with efficient processing capacity will also have larger storage capacity (Daneman & Merickle, 1996). Thus, WM tasks, including the listening span task, have been designed to measure WM storage with the interference of perhaps excessively demanding processing components. If test-takers need to maximally employ their available resources to meet the task processing requirement, they may have insufficient capacity to temporarily store target items. This has resulted in a narrow spread of item-recall scores, which suggests that the tasks may not be sensitive enough to differentiate people with different WM levels. Second, the tasks have not been developed to account for confounding methodological factors such as potential knowledge biases in the utterance verification component of listening span tasks. That is, tasks that involve judging the plausibility of utterances based on knowledge may measure both WM span and general knowledge and may therefore provide inaccurate WM measurement. Third, the tasks' construct validities have not been well supported as the available validity evidence has been limited to people with frontal lobe lesions (Miyake et al., 2000) as well as the fact that WM instruments predict performance on a wide range of tasks, such as following directions, note-taking, reading, and writing (Conway et al., 2005). It is thus unclear what WM tasks actually assess.

Based upon these three issues, it is fair to state that there is a paucity of both rigorous methodological revisions of WM tasks as well as psychometrically sound validity evidence, such as that provided by the Rasch model (Rasch, 1960). The goal of this pilot study is to (a) develop a complex span task, namely a listening span task, which addresses the flaws of its predecessors, as well as to (b) collect validity evidence for its use through Rasch analyses. Rasch analysis provides detailed information about different aspects of validity and can reveal whether a set of items is functioning to measure a single underlying construct, such as WM capacity. In this paper, validity refers to the strength of evidence in support of inferences about a human trait that can be made from an observed performance on a task. Validity is assessed by analyzing the person and item fit and the person and item reliability and separation (Bond & Fox, 2015). The results of this pilot study will be

used as a baseline to develop a computerized listening span task that can be administered to multiple participants simultaneously rather than individually as in the current task procedure.

The Listening Span Task

The listening span task is a complex span task that was originally developed by Daneman and Carpenter (1980) to gauge WM capacity, which is the capacity to store information with the interference of processing demands. The task was constructed with 60 utterances of between nine and 16 words in length which had been taken from quiz books and whose content covered a variety of knowledge domains such as literature, biological sciences, and geography. To illustrate, one of the utterances participants heard was *You can trace the languages English and German back to the same roots* (p. 458). The participants were required to listen to 15 sets of utterances (three each composed of two, three, four, five, and six utterances, respectively) and hold in memory the final word of each utterance after having judged the truthfulness of the statement. Half of the utterances were true, while the other half were false. Participants had to decide whether the presented utterance was true or false and were given a second and a half to attempt to store the last word so that they did not have time to rehearse the words in their minds. If participants did not verify the utterance within the time given, they were pushed to answer quickly or, if they did not know the answer, they were presented with the next utterance. The true-or-false component was added as a distractor and was not included in the score. Upon completion of the set, participants heard a beep signaling that they could start to recall the final words in the sets in order of appearance. The participants' WM span was defined as the utterance-level at which they were correct on two of the three sets. If the participants recalled all the words of only one set of the three, they were awarded half a credit. The test was finished when the participants could not recall any of the words in all three sets at a particular level. For example, if a participant was correct on two sets at the two-, three-, and four-utterance-levels but was incorrect on all three sets at the fifth level, they were given a span of 4.00. If the participants performed correctly on one of the three five-level utterance sets, they were given a score of 4.50. However, due to the fact that the task was excessively demanding on the participants, as Daneman and Carpenter themselves acknowledged, a credit was given for any set at which all of the words were recalled, regardless of their order of appearance.

Drawing on Daneman and Carpenter's (1980) test specifications, Osaka et al. (2003) designed a Japanese version that required participants to hear three sets of four utterances, judge their semantic plausibility, and store in memory the first word of each utterance in the set for oral recall at the end of the set. The utterances were six seconds long and were presented at one second intervals. The rationale for using the initial word in the utterances was that utterances tended to finish with a verb in Japanese. The high performing participants obtained a mean word recall score of 96.90 whereas the low-performing participants' mean score was 89.10. However, because an explanation of the scoring procedures was not reported, it is unclear how these average scores were computed. In addition, example utterances were lacking.

Another Japanese version of the task constructed by Komori (2016) contained 70 utterances between 35 and 46 mora[1] long (M = 41.77) that were divided into five sets of two, three, four, and five utterances. Similar to Daneman and Carpenter's (1980) listening span task, Komori's task required participants to listen to the increasingly longer sets and judge whether the utterances were true or false based on general knowledge. Half of the utterances were true and the other half were false. Simultaneously, as in Osaka et al.'s (2003) task, the participants had to remember the initial word (a noun) of each utterance and recall them at the end of each set. Although participants were allowed to recall the target words in any order, they were not allowed to start the recall with the last target word in the set. An example set of two utterances is as follows: (1) *migi tede chokiwo tsukuri hidari tede paawo tsukuruto orarete iru teno yubiwa nihonto naru* [When you make scissors with the right hand and paper with the left hand while playing rock-paper-scissors, the number of folded fingers is two] and (2) *denwawa onseiwo shingou henkashite hanareta aiteni tsutaeru monode, keitai gatamo aru* [Telephones are devices that encode voices as signals to communicate with a distant person, and include mobile phones] (Komori, 2016, p. 4).

The scoring system utilized by Komori (2016) was similar to that used by Osaka et al. (2003), and WM capacity was calculated as the maximum set size at which the participants could recall all the words in three of the five sets. An additional half credit was given if the participants were successful on two of the following difficulty sets. For example, if participants recalled all of the words in three sets of two utterances and two sets of three utterances (the following difficulty level), they were awarded 2.50 points. An inspection of the descriptive statistics table, however, revealed that the task was difficult for the sample. A subgroup of participants, classified as high spans, obtained a mean word recall score of 0.96 (SD = 0.08), 0.96 (SD = 0.06), 0.91 (SD = 0.07), 0.84 (SD = 0.07) for the sets of two, three, four, and five utterances, respectively. In other words, they recalled on average less than one word per difficulty level.

In fact, the three listening span tasks reviewed above all seemed to be highly demanding due to the length of the utterances found therein, which not only affected the amount of information that needed to be held in memory for processing, but also increased the duration of the retention interval, possibly resulting in the decay of the words temporarily stored in memory (Towse, et al., 2000). Thus, if the listening span task requires the participants to allocate an excessively large amount of resources to the processing component, it is likely that they will be left with insufficient storage capacity to meet the item-recall component requirement effectively. There is some evidence to support this hypothesis. In a study that compared individuals with and without aphasia under different WM conditions, Ivanova and Hallowell (2014) found that long utterances negatively impacted word-recall by the non-aphasic participants as opposed to the aphasic participants.

Furthermore, a highly demanding test would yield a narrow spread of scores, as occurred in Komori's (2016) study, which in Rasch measurement would translate into low item and person reliability. According to Bond and Fox (2016), this is because the test-takers with lower WM capacity would not have items that targeted their WM level, whereas the hardest sets of items would not allow test-takers with sufficient WM capacity to provide information about their functioning. Consequently, the person ability separation index would be low, which is an indication that the task may be insufficiently sensitive to distinguish people with different WM levels.

An additional methodological limitation is that the true/false verification component of previous listening span tasks was based on general knowledge, thus confounding WM performance with world knowledge. That is, knowledgeable test-takers may have scored higher and less knowledgeable test-takers lower than would be expected in a WM measure without this knowledge bias, which suggests that previous listening span tasks have provided an imprecise measurement of WM capacity. Insofar as the true/false component entails a judgement based on knowledge, the listening span task would measure both WM capacity and knowledge rather than true WM capacity. Thus, while the original authors did attempt to control for knowledge by selecting content that would be likely known to all potential test-takers, it is still possible that it could impact performance.

Lastly, along with these task-requirement limitations, the scoring methods utilized by the authors of the preceding studies allowed the participants to recall the target words in any order. This may have impacted the hypothesized hierarchy of difficulty of the items (because the further in the set that the words appeared, the more difficult the item should be to recall). For example, as participants could begin by recalling the hypothesized most difficult items (the last items in the sets), the difficulty level of those final items would fall below the difficulty level of the preceding items in the set, which are theorized to be easier. The present study has been designed to address these weaknesses.

Research Questions

The goal of the current study was to address the weaknesses of previously published listening span tasks by developing a new task and collecting validity evidence for its use through Rasch analysis. With this in mind, the present study was guided by the following research questions:

1. Do the items within the new listening span task (NLST) sets gradually increase in difficulty as hypothesized (i.e., the further their position within the set, the more difficult they should be)?
2. Does the NLST data fit the Rasch model?
3. Is the NLST item reliability sufficient to suggest replicability of the item difficulty hierarchy if the listening span task is administered to a similar sample?
4. Is the NLST person reliability sufficient to suggest a similar spread of participants with higher and lower levels of the construct (WM capacity) if the same participants were administered a similar item sample?
5. Does the NLST separate the assessed participants into different levels of the construct (higher and lower WM capacity)?
6. Is the NLST unidimensional?

Each Research Question addresses different aspects of construct validity (Bond & Fox, 2015). Together, the findings for each of these questions provide evidence towards validity claims for the listening span task.

Methodology

Participants

This study took place at a private junior and senior high school in Western Japan. At this institution, students were streamed into classes by academic level, comprising high-, intermediate-, and low-level classes. The 31 participants who performed the listening span task came from the second grade of junior high school and were aged 13 and 14 years old. The sample was composed of nine volunteer students from the low-level class, 10 from the intermediate class, and 12 from the high-level class. There were 20 female students and 11 male students, who were more or less equally distributed among the three levels. All participants were native speakers of Japanese.

Instruments and Administration

A shortened listening span task, which consisted of 40 unrelated casual utterances about daily-life situations (see Appendix A), was developed for the purposes of this study. The test differed from previous versions in several ways. First, utterance length was shorter with a range of between three and five words. This modification was made to keep the processing component of the task from exceeding WM capacities. Second, the task contained fewer items (20 less than the original version and 30 less than Komori's [2016] test) and there was no practice session prior to the test. The rationale for these changes was that the longer the task, the more likely it would be that participants would engage in idiosyncratic strategies to complete it (Miyake et al., 2000), which may confound measurement. Further, longer tasks have the disadvantage that participants may become tired, which would negatively impact their performance. Additionally, practice trials were not included because, unlike other span tasks such as the Tower of Hanoi or the Wisconsin sorting test (Miyake et al., 2000), listening span measures are relatively simple. Also, the task was not computerized because this step may have required additional trials for the participants to become familiar with the functions of the keys.

Third, not only did the NLST test items differ from those in previous WM tests in terms of length and number of items. They also differed in terms of the content verification component and target words for recall. Instead of a content verification component based on knowledge, this study used a grammaticality judgement test that required participants to verify if the utterances made sense in Japanese, which may have accounted for the knowledge bias of previous versions. Half of the utterances were grammatical and the other half ungrammatical (incorrect word order) and they were randomly arranged into two sets of two, three, four, five, and six utterances. Fourth, in contrast to its Japanese predecessors, the utterance-final words served as the to-be remembered items. This change had the benefit of reducing the level of memory decay as the duration of target-word retention was minimized by its final position in the utterances, thus reducing the likelihood of reliance on verbal rehearsal strategies to recall the words. Although the reason for previous listening span measures to use the initial words as target words was that Japanese utterances tend to end in verbs, which may make recall easier, casual Japanese utterances can also end in adjectives. Furthermore, ungrammatical utterances do not need to end in verbs. In this task, the target words were 12 adjectives, 11 nouns, 11 verbs, three quantifiers, and three adverbs, most of which were two or three mora long (see Appendix A). Further, an attempt was made to control for the complexity and frequency of the words, by including only words that the researcher deemed easy and highly frequent. Two students at the same institution who were unrelated to the study confirmed that all words were morphologically simple and likely to be known to participants. The utterances were audio-recorded by a female Japanese native speaker. The task was preceded by a demonstration of how to perform it.

The shortened listening span task was administered by the author of the present study one-on-one in a quiet room and was conducted entirely in Japanese. Before the test began, the participants received written and oral instructions that asked them to judge the grammaticality of each utterance and recall the final word in each utterance in the set in the correct order at the end of the set. It was explained that if the utterance's final word had a particle attached to its end (i.e., *kireida*), participants did not have to recall the particle. Similarly, if the utterance ended in a verb inflected in the past tense, participants could recall it in its plain form. After the instructions, the participants had the opportunity to ask clarification questions.

The audio stimuli were presented by the author using Windows Media Player on a laptop computer and the test began with the two sets of two utterances, gradually moving up to the two sets of six utterances. Immediately after each utterance, the audio was paused and the participants judged if the utterance made sense in Japanese. At the end of each set, participants recalled the to-be-remembered words in their order of presentation. Each participant's performance was audio-recorded for scoring. The scoring procedure was adopted from Bazan (2020) and consisted of giving a credit for each word recalled in

a string in the correct order of appearance until memory failure to recall in order. For example, if on a set of five utterances, participants correctly recalled the last word of the first and second utterances, failed to recall the last word of the third utterance (i.e., *sanbanmewo oboete nai* [I don't remember the third word]), but succeeded in recalling the last word of the fourth and fifth utterances, they were given two points (one for utterance 1 and another for utterance 2) and the rest of their responses did not count.

The utterance verification component was used as a distractor to make sure participants processed the utterances, and thus was not scored (participants were not made aware of this information). This scoring system had the added advantage of preventing participants from benefiting from recency effects as participants had to recall the words in the exact order of appearance rather than in free recall. This design maintained the hypothesized order of difficulty of the words (i.e., the further in the set, the more difficult they should be to recall). Contrary to previous scoring procedures (Friedman & Miyake, 2005), where the test was terminated when participants failed to perform perfectly on a particular set, participants in this study were administered all of the sets, regardless of how many words they could recall, which gave participants equal opportunities.

Rasch Analysis

The data were entered into a spreadsheet, which was imported to Winsteps 4.3.1 (Linacre, 2018) for analysis using the Rasch dichotomous model (i.e., items scored as right or wrong). Research Question 1 was addressed by an examination of the Wright map. Research Question 2 was answered by looking at the item and person fit indices. Research Questions 3, 4, and 5 were explored using the person and item reliability and the separation indices, respectively. Research Question 6 involved a principal components analysis (PCA) of item residuals and an inspection of the item fit graph. These are dimensionality indicators about whether the test assesses a single construct.

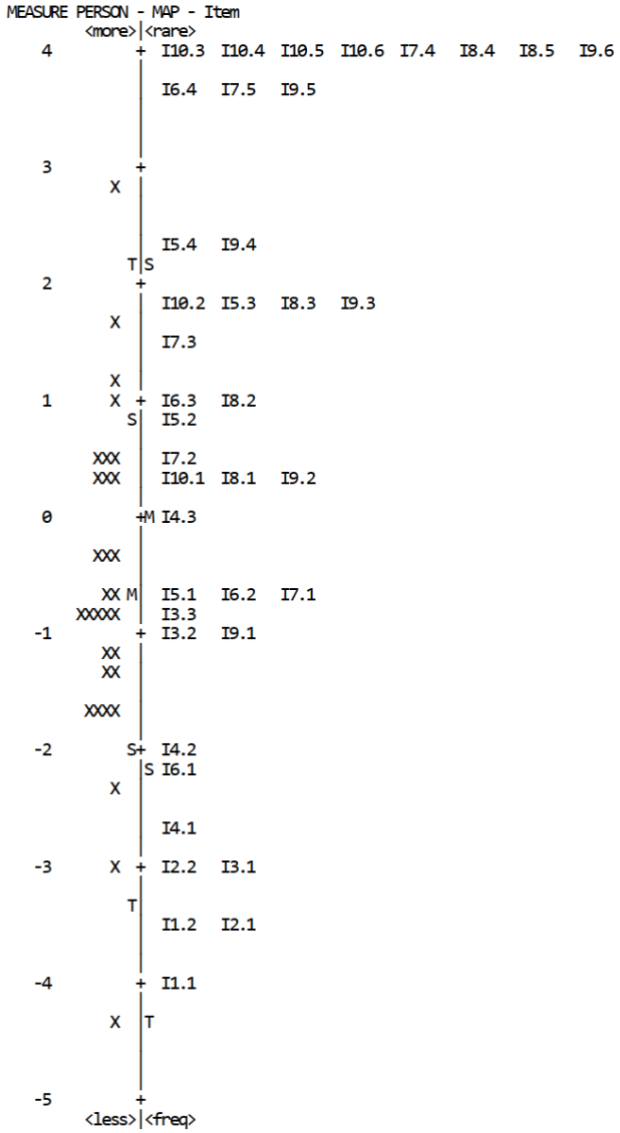
Results

Item-Person Map or Wright Map

Figure 1 shows the item-person relationships plotted on the map of the listening span task. The line in the center of the map represents the distances between points on the logit scale, which is an interval scale measurement, and which is shown numerically on the left. In other words, the distances between the data points along the line are thought to represent equal amounts of WM capacity. The participants, who are each represented by an 'X', are located on the left side in ascending order of hypothesized WM capacity. That is, the higher up the map, the higher the participants' scores on the listening span task. Similarly, the items are spread on the right side in ascending order of difficulty. In other words, the higher up the map, the more difficult the item. The plot shows that the individual items within each set are ordered in accordance with the theoretical expectation that the later the item appears in the set, the more difficult it should be. For example, the last item of set 3 (item 3.3) is higher than the second (item 3.2), which is higher than the first (item 3.1). As shown, a number of items in sets 7, 8, and 10 (items 7.4, 7.5, 8.4, and 8.5 and items 10.3, 10.4, 10.5, 10.6, respectively) do not align with the theorized difficulty hierarchy. For example, items 10.3, 10.4, 10.5, and 10.6 are shown to be equally difficult. This is explained by the fact that the fit statistics of those items were not estimated by Winsteps as no participant was successful on them. An alternative explanation is that there were not enough participants at this level to discriminate the difficulty hierarchy of the more difficult items. The distribution of participants is heavier in the lower half of the figure (below the 0.00 logit measure to the left of the persons, representing the mean item difficulty), which suggests that a greater number of higher WM-span participants were needed. The participants were, however, well spread out over approximately seven logits along the WM logit scale.

Figure 1

Wright map for the listening span task individual items analysis



Note. "X" represents each individual participant's performance, "I" are the items, which are followed by the set number and the item number, the logit scale is all the way on the left, under Measure. The line down the middle separates items and persons, locating these facets on a common frame of reference in keeping with the Rasch model.

Person and Item Fit Statistics

The Rasch fit statistics are quality-control indicators that are useful to evaluate the degree to which the data meet the model's expectations. Rasch provides two aspects of fit, namely infit mean-square (MNSQ), which is a weighted unstandardized form of fit, and outfit MNSQ, which is a non-weighted standardized fit statistic that is sensitive to outliers (Linacre, 2002; Bond & Fox, 2015). As the outfit statistic is affected by outliers (unexpected performances of participants who manage to succeed on items above their abilities), infit MNSQ tends to be the statistic that guides the assessment of fit (Bond & Fox, 2015). In this study too, decisions about fit will be made based on infit MNSQ, but outfit values will also be examined to investigate unexpected performances of persons and items. Based on Linacre's (2002; 2007) guidelines, fitting persons and items were defined as those with infit MNSQ values of between 0.50 and 1.50 with perfect fit being indicated by a value of 1.00.

The person infit MNSQ indices for the participants in the listening span (see Table 1), revealed that all participants but one (participant c301, infit MNSQ = 1.81) had infit MNSQ values within the acceptable parameters of 0.50 and 1.50, indicating

that the sample behaved as expected by the model. The person infit MNSQ values, excluding person c301, ranged from 0.52 (person c102) to 1.44 (person c112), which shows that the participants' performance had acceptable fit.

The high infit MNSQ statistic for participant c301 (infit MNSQ = 1.81) is accompanied by a large outfit MNSQ value of 2.65, which suggests that this participant's performance was unexpected by the model. This was also true for participants c112, c136, and c229 who had large outfit MNSQ indices of 4.03, 3.83, and 2.89, respectively. An examination of their individual data revealed that participants c112, c136, and c301 were low performers (logit WM measures of -0.86, 0.53, and 0.66, respectively) who, perhaps through the use of an idiosyncratic strategy such as initial word mora recall or word chaining, managed to succeed on items above their WM level such as items 6.3 or 10.2 (difficulty measures of 1.00 and 1.86, respectively). In contrast, participant c229 was a capable participant (logit WM measure of 1.21) who unexpectedly failed on some easy items such as items 4.1, 4.2, and 4.3 (difficulty measures of -2.71, -1.96, and 0.00, respectively). The source, however, of these participants' poor outfit values seems to be the small size of the sample (N = 31) because a few unexpected responses can make the participants misfit in small samples (Boone & Noltemeyer, 2017).

Table 1

Person statistics for the listening span task

Person	Measure	*SE*	Infit MNSQ	Outfit MNSQ
c112	-0.86	0.6	1.44	4.03
c136	0.53	0.57	1.44	3.83
c229	1.21	0.55	1.29	2.89
c301	-0.62	0.66	1.81	2.65
c303	0.31	0.51	1.13	1.43
c125	0.98	0.57	1.42	1.13
c201	-1.12	0.58	1.32	1.06
c306	-1.66	0.61	1.31	0.93
c134	-0.86	0.57	1.29	1.06
c313	0.31	0.48	0.83	1.28
c225	-1.66	0.6	1.26	1.26
c307	-0.86	0.53	1.12	1.18
c114	0.53	0.5	1.12	0.85
c113	1.69	0.52	1.07	0.75
c321	-0.86	0.5	0.97	0.9
c102	0.31	0.48	0.93	0.69
c120	-0.86	0.5	0.89	0.67
c212	-2.25	0.56	0.85	0.53
c124	-1.38	0.52	0.85	0.49
c324	-4.34	0.8	0.79	0.21
c224	-1.38	0.52	0.76	0.5
c222	-0.62	0.49	0.75	0.51
c234	0.53	0.48	0.69	0.47
c103	-0.38	0.49	0.67	0.43
c108	-1.66	0.53	0.62	0.35
c314	-1.66	0.53	0.6	0.33
c210	-0.38	0.49	0.59	0.41
c203	-1.12	0.51	0.59	0.36
c209	2.84	0.58	0.58	0.28
c302	-0.38	0.49	0.53	0.34
c102	-2.94	0.61	0.52	0.21

Note. SE = standard error; MNSQ = mean-squared.

The item infit MNSQ values (see Table 2) showed a similar pattern of relatively well-behaved data. All values were within the cut-off parameters of 0.50 and 1.50 (i1.1 had the highest value, infit MNSQ = 1.42, and i6.1 had the lowest, infit MNSQ = 0.69), meaning that the items in the listening span task were successful in measuring the intended construct, thus providing evidence for construct validity (Bond & Fox, 2015).

The high outfit of items i1.1 (outfit MNSQ = 3.27), i4.1 (outfit MNSQ = 2.76), and i.6.4 (outfit MNSQ = 2.91) is explained by the fact that these items elicited unexpected performance by a few participants as an inspection of the Winsteps tables of item and persons responses revealed. For example, i1.1 had the lowest difficulty measure (-4.00 logits) and was supposed to be within all participants' WM capacities, yet two participants (c212 and c136) were unexpectedly unsuccessful on the item, which caused the high outfit (outfit MNSQ = 3.27).

Table 2

Item statistics for the listening span task

Item	Measure	*SE*	Infit MNSQ	Outfit MNSQ
i1.1	-4	0.96	1.42	3.27
i1.2	-3.45	0.68	1.01	1.08
i2.1	-3.45	0.68	0.93	0.72
i2.2	-3.04	0.6	0.97	0.67
i3.1	-3.04	0.69	1.34	1.12
i3.2	-1.03	0.41	0.85	0.75
i3.3	-0.86	0.41	0.97	0.94
i4.1	-2.71	0.61	1.25	2.76
i4.2	-1.96	0.48	1.09	1.59
i4.3	0	0.42	0.88	0.77
i5.1	-0.69	0.41	0.87	0.81
i5.2	0.77	0.47	0.86	0.75
i5.3	1.86	0.6	0.81	0.42
i5.4	2.26	0.67	0.76	0.37
i6.1	-2.19	0.48	0.69	0.49
i6.2	-0.69	0.47	1.31	1.25
i6.3	1	0.54	1.21	1.06
i6.4	3.63	1.2	1.27	2.91
i7.1	-0.69	0.41	0.99	0.91
i7.2	0.56	0.45	0.98	0.86
i7.3	1.53	0.55	0.82	0.68
i7.4	4.94	1.85	***	***
i7.5	3.63	1.18	1.22	0.93
i8.1	0.37	0.44	0.76	0.74
i8.2	1	0.49	0.96	0.94
i8.3	1.86	0.66	1.24	0.66
i8.4	4.94	1.85	***	***
i8.5	4.94	1.85	***	***
i9.1	-1.03	0.41	0.93	0.79
i9.2	0.37	0.44	0.96	0.95
i9.3	1.86	0.6	0.76	0.61
i9.4	2.26	0.71	1.14	0.87
i9.5	3.63	1.07	1.01	0.26
i9.6	4.94	1.85	***	***
i10.1	0.37	0.46	1.09	1.02
i10.2	1.86	0.6	1.01	1.07
i10.3	4.94	1.85	***	***
i10.4	4.94	1.85	***	***
i10.5	4.94	1.85	***	***
i10.6	4.94	1.85	***	***

Note. SE = standard error; MNSQ = mean-squared; *** = maximum measure.

Person and Item Reliability and Separation

The person and item reliability indicate the degree to which replicability of the person and item hierarchy is possible if the listening span test is given to a sample with similar characteristics. The higher the reliability value, the more confidence that can be placed in obtaining a similar ordering of persons and items across samples. The data revealed a person reliability estimate of .84 (the maximum possible value is 1.00), which is above the cut-off value of .80 (Linacre, 2007) and suggests that the probability of obtaining a similar spread of participants' WM capacities in similar samples is high. In addition, the person separation was estimated at 2.28 (see Table 3), which indicates that the sample was separable into three different levels of WM capacity (high, average, and low) as separation indices above 2.00 distinguish three distinct levels of the variable investigated (Duncan et al., 2003).

Table 3

Summary of the listening span task analysis (persons)

	Total Score	Count	Measure	Real *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	13.4	40	-0.6	0.55	0.97	-0.1	1.03	0.1
P. SD	5.2	0	1.37	0.07	0.33	1.2	0.98	1.1
S. SD	5.3	0	1.39	0.07	0.34	1.2	0.99	1.1
Max	27	40	2.84	0.8	1.81	2.6	4.03	2.9
Min	2	40	-4.34	0.48	0.52	-2.1	0.21	-1.4
REAL *RSME*	0.55	TRUE *SD*	1.25	SEPARATION	2.28	PERSON RELI.	0.84	
MODEL *RSME*	0.52	TRUE *SD*	1.27	SEPARATION	2.43	PERSON RELI.	0.86	
SE OF PERSON MEAN = 0.25								
PERSON RAW SCORE-TO-MEASURE CORRELATION = .99								
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .84								
SEM = 2.09								

Likewise, item reliability was .91 and item separation was calculated at 3.21 (see Table 4), meaning that the item difficulty hierarchy and spacing of items is highly replicable and that the listening span task separates items into four difficulty groups (Duncan et al., 2003). It is worth noting here that the total number of items provided in Table 4 (i.e., $N = 31$) is lower than the total number of items of the complete test (i.e., $N = 40$) because extreme scores are excluded. In any case, these results provide supporting evidence for construct validity (Bond & Fox, 2015).

Table 4

Summary of the listening span task analysis (items)

	Total Score	Count	Measure	Real *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	13	31	0	0.6	1.01	0	1.03	0.1
P. SD	9.3	0	2.15	0.21	0.19	0.7	0.68	0.6
S. SD	9.4	0	2.18	0.22	0.19	0.7	0.69	0.6
Max	29	31	3.63	1.2	1.42	1.9	3.27	1.8
Min	1	31	-4	0.41	0.69	-1.2	0.26	-0.8
REAL *RSME*	0.64	TRUE *SD*	2.05	SEPARATION	3.21	ITEM RELI.	0.91	
MODEL *RSME*	0.61	TRUE *SD*	2.06	SEPARATION	3.4	ITEM RELI.	0.92	
SE OF ITEM MEAN = .39								

PCA of Item Residuals and Item Fit Graph

A PCA analysis of the item residuals was conducted in order to examine the unidimensonality of the construct. There are two PCA requirements for unidimensionality. First, a unidimensional construct should account for 20.00% of the variance

or more (Reckase, 1979). Second, the principal contrast should produce an eigenvalue below 2.00 (Linacre, 2018) and explain less than 10.00% of the variance (Linacre, 2007). As shown in Table 5, the WM construct accounted for 51.30% (eigenvalue = 33.67) of the total variance, indicating that the instrument measured a single construct (the criterion value was at least 20.00%). Additionally, the principal contrast accounted for 6.50% of the unexplained variance, satisfying the unidimensionality criterion (< 10.00%). However, despite the fact that the first contrast explained less than 10.00% of the variance, its high eigenvalue (4.30) suggested the possibility of a second dimension.

Table 5

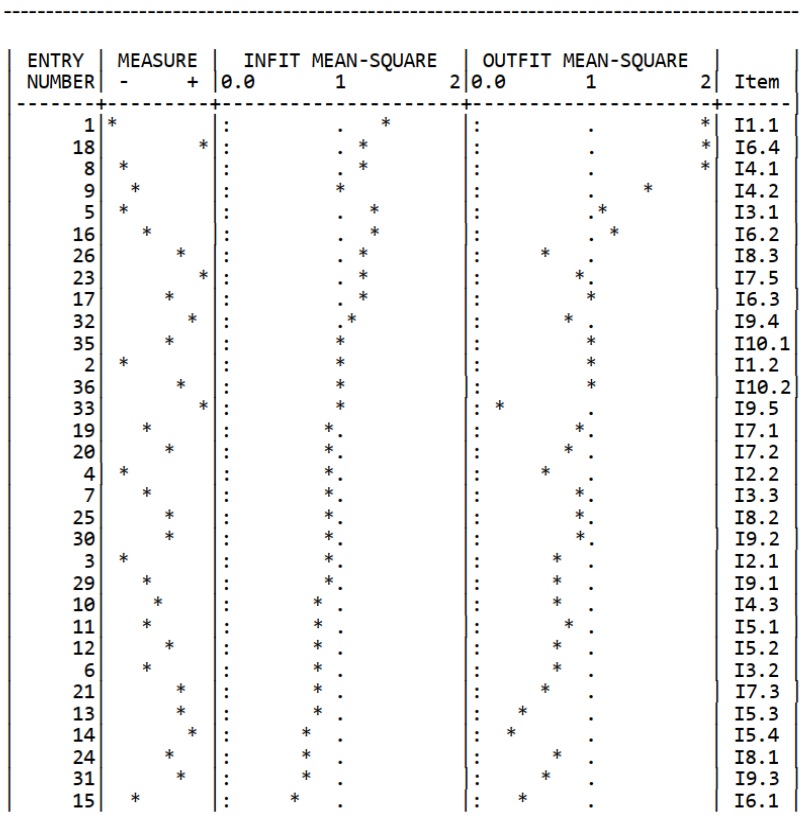
Listening span task standard residuals in eigen values

	Eigenvalue	Observed	Expected
Total Raw variance in observations	65.67	100.00%	100.00%
Raw variance explained by measures	33.67	51.30%	55.50%
Raw variance explained by persons	10.64	16.20%	16.00%
Raw variance explained by items	23.02	35.10%	34.60%
Raw unexplained variance (total)	32	48.70%	49.50%
Unexplained variance in 1st contrast	4.3	6.50%	
Unexplained variance in 2nd contrast	3.77	5.70%	

In any case, the item fit graph (see Figure 2) ruled out the possibility of an underlying second dimension. The infit MNSQ section of the figure depicts a unidimensional path. The straight vertical dotted line in the middle of the path represents the hypothesized unidimensional construct of WM capacity. As seen, the items of the listening span task, represented by asterisks, appear to be aligned along the ideal straight line. In addition, no item is outside the delimiting lines of the path, which would be of concern for the unidimensionality of the measure. Therefore, it seems reasonable to conclude that the listening span task is potentially unidimensional. However, a larger sample is required to make a clear conclusion.

Figure 2

Item fit graph for the listening span task



Discussion

The purpose of this study was to address the methodological issues of previous instruments by designing a listening span task and collecting sound psychometric validity evidence for its use through the Rasch model. It was argued that previous tests contained an exceedingly demanding processing component. The utterances in Osaka et al.'s (2003) instrument were six seconds long and those of Komori's (2016) ranged between 35 and 46 mora. In the latter study, the average recall rate was similar across sets (less than one word), which suggests that a set of two utterances was as difficult as a set of five. Additionally, the utterance verification of previous tasks such as found in Daneman and Carpenter's (1980) pioneering task, may have confounded WM measurement with world knowledge. The present listening span task addressed these shortcomings by controlling for utterance length and by having a grammaticality judgement test as the utterance verification component. Two additional new features were that the task lacked practice sets because the more practice, the more likely participants are to engage in strategies to complete the task (Miyake et al., 2000), and that it contained fewer items than its predecessors, which helps increase the practicality of its administration. Performance was scored adopting a scoring procedure that accounted for order of appearance to prevent participants from free recall, which is likely to involve strategic behavior.

Research Question 1 examined whether the items within the sets gradually increased in difficulty as expected based on theory (Daneman & Carpenter, 1980). An examination of the Wright map revealed that, overall, the difficulty of the items matched the theoretical expectations as the items were ordered along the map in ascending order of difficulty from initial to final set items. This means that, for example, recalling the target word of the third utterance in a set of three was more difficult than recalling the target word in the first utterance. This hierarchy of item difficulty is in contrast to that of Osaka et al.'s (2003) and Komori (2016). In these studies, the average recall rate was less than one item per set, which suggests that most items had a similar level of difficulty. The items in those studies may have been overly difficult, which may have impacted the precision of the measurement. These contrasting item difficulties can be explained by the impact of utterance length of the processing component of the task on the word storage component. The longer utterances used by Osaka et al. (2003) and Komori (2016) are likely to have produced greater interference and longer retention duration, potentially causing the to-be-remembered words to fade more easily. This explanation is in line with Cowan's (1999) embedded processes model of WM, which posits that WM is limited not only by the capacity to hold information but also, by the time the information can be held.

Research Question 2 asked whether the data fit the Rasch model. The majority of the items displayed good fit to the Rasch model, as none of the items had infit figures outside the established range (0.50 and 1.50). Three items (i1.1, i4.1, and i6.4), showed poor outfit (3.27, 2.76, and 2.91), but this was probably due to the unexpected performance of some participants, perhaps caused by a lack of concentration or nervousness at the beginning of the test. A most likely explanation for these high outfit values is, however, that no practice items were given to acclimate the participants to the task prior to its performance. This could have induced the participants to fail those items due to a lack of familiarity with the task procedures rather than due to a lack of ability.

Likewise, the participants performed close to the expectations of the model. One participant (c301) was identified as having large infit (1.81) and outfit (2.65) indices and three others (c112, c136, and c229) had large outfit indices (4.03, 3.83, and 2.39), which was explained by their off-target performance on several items probably due to the lack of practice trials. An alternative explanation is that the participants may have used idiosyncratic strategies such as initial mora recall to succeed on items that were above their level of ability.

All in all, the fit of the data to the Rasch model suggests that the construction of the task was, in general terms, effective. First, the grammaticality judgement task is likely to have served as a tool to make sure that the participants fully processed each utterance and that they did not simply focused on retaining the target words while ignoring the utterances (Turner & Engle, 1989). Second, although there are advantages to selecting the target words based on a corpus, such as a stricter control for word frequency and the elimination of a possible confound (i.e., the recall errors may be due to difficult word recognition rather than WM capacity), the intuitive approach used in the current study produced data that largely conform to the predictions of the Rasch model. Third, the results of the current pilot study lend support to the use of the scoring system corroborating the findings obtained by Bazan (2020). It is important to note, however, that this scoring system has the disadvantage that it requires participants to attempt all the trials (i.e., from Set 1 to Set 6), which may cause frustration once the task advances beyond the ability of the participants. In addition, this scoring system ignores one of the sources of the data, that of the processing component, as the grammatical verification of the utterances is not scored.

Research Questions 3 and 4 regarded the item and person reliability indices, respectively. These data revealed an item reliability coefficient of .91, which suggests that the spread of items along the WM continuum would likely be replicated if the NLST were given to another sample of similar characteristics. Similarly, the person reliability coefficient (.84) suggested high likelihood of reproducibility of the person hierarchy (Linacre, 2007). In other words, the participants would likely be placed at a similar level of ability if they were given a similar listening span task. These findings are consistent with the high reliability that complex span tasks have been shown to have based on split-half correlations or test-retest methods (Conway et al., 2005; Waters & Kaplan, 2004).

Research Question 5 investigated whether the listening span task separated the participants into different levels of ability. The participant performance was shown to be separable (separation = 2.28) into three levels of WM capacity (high, average, and low). This separation suggests that the top and bottom 23% of the sample had high and low ability, respectively, whereas the remaining 54% had average ability (Linacre, 2013).

According to Conway et al. (2005), the most common way of separating participants in the complex span task paradigm is quartile splits, in which the top and bottom quartiles of a distribution of WM scores are categorized as high and low span, respectively. This is the process of separation that both Komori (2016) and Osaka et al. (2003) used to split their respective samples for follow-up analyses. However, this process is problematic because it forces the separation groups to be equal in size, thereby treating participants, who may have different ability levels, as if they had the same ability level (Conway et al., 2005). For example, a group categorization based on quartile splits may give two groups of 30 participants each, but there may be different spreads of abilities within each group. The Rasch separation index is an alternative that is likely to yield a more precise separation and consequently more precise follow-up analyses. This is because the Rasch separation shows how many statistically differentiable ability levels exist in the population (Linacre, 2013), whereas quartiles might overestimate or underestimate the number of levels.

Research Question 6 explored the dimensionality of the measure. This dichotomous model explained 51.30% of the total variance (above 20.00%), which is one of the criteria of unidimensionality (Reckase, 1979). However, the results of the Rasch PCA contrast showed a concerning eigenvalue of 4.30, indicating the possible existence of a secondary dimension. Therefore, I examined the content of the items with the standardized residual loadings for the items in the Winsteps output to see if they hinted at a pattern (see Table 6). The items that were indicative of a possible subdimension seemed to share a common theme which could be called *familiarity* or *relevance to the participants' life*, as they were clearly broken into a cluster of utterances that had to do with the participants' everyday lives and a cluster of utterances that did not. For example, i3.2 *kouende tomodachito asobu* [I play with my friends in the park] vs. i9.4 *otousanwa inuga sukida* [My father likes dogs] or i6.3 *ekiga chikakattara benrida* [It is convenient to have the station close-by] vs i2.*1 kodomotachiga okashiwo kau* [children buy snacks]. This meant that the task could be confounded by the degree to which participants found the utterances related or not to their lives. There were, however, items that did not support this interpretation such as i6.1 *nihonwa supeinyori semai* [Japan is smaller than Spain] vs. i7.3 *utaimasu tomodachiga jouzuni* [sings my friend well], which is an ungrammatical utterance.

The possibility of this second dimension, however, was dismissed by the linear alignment of the items in the item fit graph. In general terms, these dimensionality results seem to provide support for unitary models of WM as opposed to models that consist of multiple separable subsystems (Miyake & Shah, 1999). Importantly, however, the sample size of this study is not large enough as to make a robust claim about the unitary versus the non-unitary nature of WM. Nevertheless, these results provide validity evidence for the measure and suggest new ways of developing and scoring listening span tasks.

Limitations

This study presents a number of limitations that should be addressed in future investigations. First, the sample size was too small to give sufficient statistical power to the results and therefore corroborating evidence from larger samples is necessary. Second, the grammatically incorrect utterances of the grammaticality judgement may have inadvertently altered the nature of the task because storing and recalling words as part of natural utterances may be fundamentally different from doing so with ungrammatical utterances (i.e., the former reflects natural processing and thus may benefit from correctly ordered word sequences). In other words, the reading comprehension system utilizes the predictability of upcoming words based on collocations and context so using jumbled utterances may lead to a processing deficit. This can be addressed by replacing the grammaticality judgement test with an affirmative-negative judgement. In other words, the task would only be composed of natural affirmative and negative utterances.

Table 6

Standardized residual loadings for the first contrast

Loading	Item	Loading	Item
0.71	I3.3, *tsukaimashou mizuwo taisetsuni* [water let's use wisely]*	-0.67	I9.3, *nomanai sakewo amari* [much alcohol I don't drink]*
0.65	I3.2, *kouende tomodachito asobu* [I play with my friends in the park]	-0.66	I9.4, *otousanwa inuga sukida* [my father likes dogs]
0.58	I3.1, *ikenai isshouni bokuwa* [can't go with you I]*	-0.51	I10.2, *furu ashitawa amewo* [it will tomorrow rain]*
0.49	I6.3, *ekiga chikakattara benrida* [It is convenient to have the station close-by]	-0.44	I2.1, *kodomotachiga okashiwo kau* [children buy snacks]
0.49	I6.4, *dekiru konohendewa hanamiwa* [We can in this area do *hanami* (cherry-blossom viewing)]*	-0.43	I9.2, *ryokouwa denshade iku* [I am going to travel by train]
0.41	I1.2, *amakunai wasabi zenzen* [wasabi at all isn't hot]*	-0.36	I2.2, *umeboshiya nattoga kiraida* [I dislike *umeboshi* (salted plums) and *natto* (fermented beans)]
0.39	I4.3, *kaerimasu seitowa aruite* [go home the students on foot]*	-0.33	I7.1, *honwo yomisugiruto mega tsukareru* [When I read too much, my eyes get tired]
0.2	I1.1, *sono eigawa kowai* [the movie is scary]	-0.29	I9.1, *oniichanwa yakyuuwo yameru* [my brother is going to quit baseball]
0.18	I5.4, *eigono shikenwa kantanda* [the English exam is easy]	-0.28	I7.3, *utaimasu tomodachiga jouzuni* [sings my friend well]*
0.17	I6.1, *nihonwa supeinyori semai* [Japan is smaller than Spain]	-0.24	I7.2, *nerutokini denkiwo kesu* [I turn off the lights when I go to bed]
0.13	I7.5, *jibunno mochitai misega* [my own shop I want to run]*	-0.23	I10.1, *komaru tsukattara okanewo* [money I will be troubled if I spend]*
0.08	I6.2, *aitia hitowo atarashi* [a person new I want to meet]*	-0.2	I8.1, *maketa shiaiwa kinouno* [the game yesterday we lost]*
0.05	I5.1, *kotoga aru shinpaina* [there is worries me something]*	-0.11	I8.3, *oishii totemo gohanwa* [very good the food is]*
0.03	I4.2, *taberu bokuwa ringowo* [eat I apples]*	-0.11	I9.5, *koutsujikoga mainichi aru* [there are traffic accidents every day]
0.03	I5.3, *arukinikui kono kutsuwa totemo* [it's very hard on these shoes to walk]*	-0.03	I8.2, *heiyaga totemo kireida* [the room is very clean]
0	I5.2, *hashiruto ashiga itai* [it hurts when I run]	-0.02	I4.1, *kenkounotameni undousuru* [I exercise to stay healthy]

Note. *Translation written in incorrect English word order to reflect the ungrammatical Japanese sentences.

Third, the scoring system may have created interdependence among the items, inflating the reliability coefficients. To address this issue, future investigations should include a polytomous analysis of the superitems (sets treated as items) following the analysis of the individual items. In addition, future scoring systems should account for the processing component of the task, and perhaps, the word recall interval (i.e., the time between the recall of one word and the next). Fourth, this study did not account for the abstract or concrete nature of the target words. In future investigations, the number of abstract words should be controlled for as they tend to be more difficult to recall. Similarly, the target words should be selected from frequency lists because if words with overly different frequencies are mixed together in the same task, it is difficult to know if it is WM or word recognition what is causing the recall errors. Finally, future research should also examine the impact of utterance complexity on performance.

Conclusion

Despite the popularity of listening span tasks since Daneman and Carpenter's (1980) pioneering research, no study has attempted to revise and/or collect evidence to support their construct validity. This study provides initial psychometric validity evidence for a new listening span task that was constructed to address the shortcomings of length of utterance and knowledge bias identified in previous tests. The Rasch model appeared to be a suitable approach to investigate the functioning and validity of listening span tasks and, perhaps, WM measures in general. It is this author's hope that researchers employ this listening span task to further improve the assessment of WM capacity.

Notes

[1] A mora is defined as a minimal unit of sound of metrical time in the Japanese phonological system.

Acknowledgements

Many thanks to my advisor, Prof. James Sick, for his methodological guidance. I am also grateful to the reviewers and editor for their constructive feedback. Any remaining errors are my own. I would also like to express my gratitude to Andrew Wright, Clint Denison, and Sachiko Aoki for their assistance, and to all the teachers and students who made this research possible.

References

Baddeley, A. D., Kopelman, M. D., & Wilson, A. B. (2002). *The handbook of memory disorders* (2nd. Ed). Wiley.

Bazan, B. (2020). A Rasch-validation study of a novel speaking span task. *Shiken, 24*(1), 1–21. Retrieved from http://teval.jalt.org/node/95

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd. ed.). Erlbaum.

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners, *Cogent Education*, *4*(1),1–13. https://doi.org/10.1080/2331186X.2017.1416898

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd. ed.). Erlbaum.

Conway, A. R. A., Kane, M. J., Buntig M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–102). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422–433. https://doi.org/10.3758/BF03214546

Duncan, P. W., Bode, R. K., Lai, S. M., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation, 84*, 950–963. https://doi.org/10.1016/S0003-9993(03)00035-2

Friedman, N., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*(4), 581–590. https://doi.org/10.3758/bf03192728

Gathercole, S. E., & Baddeley, A. D. (1993). *Essays in cognitive psychology. Working memory and language.* Erlbaum.

Ivanova, M. V., & Hallowell, B. (2014). A new modified listening span task to enhance validity of working memory assessment for people with and without aphasia. *Journal of Communication Disorders*, *52*, 78–98. https://doi.org/10.1016/j.jcomdis.2014.06.001

Komori, M. (2016). Effects of working memory capacity on metacognitive monitoring: A study of groups differences using a listening span test. *Frontiers in Psychology*, *7*, 1–9. https://doi.org.10.3389.2016.00285

Linacre, J. M. (2002). What do infit, outfit, mean-square, and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2007). *A user's guide to WINSTEPS: Rasch-model computer program.* MESA.

Linacre, J. M. (2013). Reliability, separation, and strata: Percentage of sample in each level. *Rasch Measurement Transactions*, *26*(4), 1399. https://www.rasch.org/rmt/rmt264g.htm

Linacre, J. M. (2018). Dimensionality: Contrasts and variances. www.winsteps.com/winman/webpage.htm

Linacre, J. M. (2018). Winsteps (Version 4.3.1) [Computer Software]. Winsteps.com.

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,* 861–883. https://doi.org/10.3758/s13423-013-0565-2

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contribution to complex frontal lobe tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100. https://doi.org/10.1006/cogp.1999.0734

Miyake, A. and Shah, P. (Eds) (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.

Osaka, M., Osaka, N., Kondo, H., Morishita, M., Fukuyama, H., Aso, T., & Shibasaki, H. (2003). The neural basis of individual differences in working memory capacity: An fMRI study. *Neuroimage*, *18*, 789–797. https://doi.org/10.1016/S1053-8119(02)00032-0

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* University of Chicago.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207–230. https://doi.org/10.2307/1164671

Towse, J. N., Hitch, G. J., & Hutton, U. (2000). On the interpretation of working memory span in adults. *Memory and Cognition*, *28*, 341–348. https://doi.org/10.3758/BF03198549

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*, 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Waters G. S., & Caplan, D. (2003). Verbal working memory and on-line syntactic processing: Evidence from self-paced listening. *Quarterly Journal of Experimental Psychology*, *57*, 129–163. https://doi.org/10.1080/02724980343000170

Appendix A

Listening Span Task

Set	Item	Grammaticality		Sentence
Set 1	I1.1	✓	Japanese	その映画は怖い
			Romanized Version	*sono eigawa kowai*
			English Translation	the movie is scary
	I1.2	×	Japanese	甘くないワサビは全然
			Romanized Version	*amakunai wasabi zenzen*
			English Translation	wasabi at all isn't hot*
Set	Item	Grammaticality		Sentence
Set 2	I2.1	✓	Japanese	子供たちがお菓子を買う
			Romanized Version	*kodomotachiga okashiwo kau*
			English Translation	children buy snacks
	I2.2	✓	Japanese	梅干しや納豆が嫌いだ
			Romanized Version	*umeboshiya nattoga kiraida*
			English Translation	I dislike *umeboshi* (salted plums) and *natto* (fermented beans)
Set	Item	Grammaticality		Sentence
Set 3	I3.1	×	Japanese	行けない一緒に僕は
			Romanized Version	*ikenai isshouni bokuwa*
			English Translation	can't go with you I*
	I3.2	✓	Japanese	公園で友達と遊ぶ
			Romanized Version	*kouende tomodachito asobu*
			English Translation	I play with my friends in the park
	I3.3	×	Japanese	使いましょう水は大切に
			Romanized Version	*tsukaimashou mizuwo taisetsuni*
			English Translation	water let's use wisely*
Set	Item	Grammaticality		Sentence
Set 4	I4.1	✓	Japanese	健康のために運動する
			Romanized Version	*kenkounotameni undousuru*
			English Translation	I exercise to stay healthy
	I4.2	✓	Japanese	食べる僕はリンゴを
			Romanized Version	*taberu bokuwa ringowo*
			English Translation	eat I apples
	I4.3	×	Japanese	帰ります生徒はあるいて
			Romanized Version	*kaerimasu seitowa aruite*
			English Translation	go home the students on foot*
Set	Item	Grammaticality		Sentence
Set 5	I5.1	×	Japanese	ことがある心配な
			Romanized Version	*kotoga aru shinpaina*
			English Translation	there is worries me something*
	I5.2	✓	Japanese	走ると足が痛いだ
			Romanized Version	*hashiruto ashiga itai*
			English Translation	it hurts when I run
	I5.3	×	Japanese	歩きにくいこの靴はとても
			Romanized Version	*arukinikui konokutsuwa totemo*
			English Translation	it's very hard on these shoes to walk*
	I5.4	✓	Japanese	英語の試験は簡単だ
			Romanized Version	*eigono shikenwa kantanda*
			English Translation	The English exam is easy
Set	Item	Grammaticality		Sentence
	I6.1	✓	Japanese	日本はスペインより狭い
			Romanized Version	*nihonwa supeinyori semai*
			English Translation	Japan is smaller than Spain
			Japanese	会いたい人を新しい

Set 6	I6.2	×	Romanized Version	*aitai hitowo atarashii*
			English Translation	a person new I want to meet*
			Japanese	駅が近かったら、便利だ
	I6.3	✓	Romanized Version	*ekiga chikakattara, benrida*
			English Translation	It is convenient to have the station close-by
			Japanese	出来るこの辺では花見は
	I6.4	×	Romanized Version	*dekiru konohendewa hanamiwa*
			English Translation	We can in this area do *hanami* (cherry-blossom viewing)*

Set	Item	Grammaticality		Sentence
			Japanese	本を読みすぎると目が疲れる
	I7.1	✓	Romanized Version	*honwo yomisugiruto mega tsukareru*
			English Translation	When I read too much, my eyes get tired
			Japanese	寝る時に電気を消す
	I7.2	✓	Romanized Version	*nerutokini denkiwo kesu*
			English Translation	I turn off the lights when I go to bed*
Set 7			Japanese	歌います友達は上手に
	I7.3	×	Romanized Version	*utaimasu tomodachiwa jouzuni*
			English Translation	sings my friend well
			Japanese	僕はお金が欲しいだ
	I7.4	✓	Romanized Version	*bokuwa okanega hoshii*
			English Translation	I want money
			Japanese	自分の持ちたい店が
	I7.5	✓	Romanized Version	*jibunno mochitai misega*
			English Translation	my own shop I want to run

Set	Item	Grammaticality		Sentence
			Japanese	負けたしあいは昨日の
	I8.1	×	Romanized Version	*maketa shiai kinouno*
Set			English Translation	the game yesterday we lost*
8			Japanese	部屋がとてもきれいだ
	I8.2	✓	Romanized Version	*heiyaga totemo kireida*
			English Translation	the room is very clean
			Japanese	美味しいとてもご飯は
	I8.3	×	Romanized Version	*oishii totemo gohanwa*
			English Translation	very good the food is*
			Japanese	この飲みにくい薬は
	I8.4	×	Romanized Version	*kononominikui kusuriwa*
			English Translation	hard to swallow the medicine is*
			Japanese	間に合う次の電車に
	I8.5	×	Romanized Version	*maniau tsugino denshani*
			English Translation	in time we are for the next train*

Set	Item	Grammaticality		Sentence
			Japanese	お兄ちゃんは野球を辞める
	I9.1	✓	Romanized Version	*oniichanwa yakyuuwo yameru*
			English Translation	my brother is going to quit baseball
			Japanese	旅行は電車で行く
	I9.2	✓	Romanized Version	*ryokouwa denshade iku*
			English Translation	I am going to travel by train
			Japanese	飲まない酒をあまり
	I9.3	×	Romanized Version	*nomanai sakewo amari*
			English Translation	much alcohol I don't drink
Set			Japanese	お父さんは犬が好きだ
9	I9.4	✓	Romanized Version	*otousanwa inuga sukida*
			English Translation	my father likes dogs
			Japanese	交通事故が毎日ある
	I9.5	×	Romanized Version	*koutsujikoga mainichi aru*
			English Translation	there are traffic accidents every day*
			Japanese	授業はもう始まった

	I9.6	✓	Romanized Version	*jugyouwa mou hajimatta*
			English Translation	the class has already started
Set	Item	Grammaticality		Sentence
Set 10	I10.1	×	Japanese	困る使ったらお金を
			Romanized Version	*komaru tsukattara okanewo*
			English Translation	money I will be troubled if I spend*
	I10.2	×	Japanese	降る明日は雨が
			Romanized Version	*furu ashitawa amega*
			English Translation	it will tomorrow rain*
	I10.3	×	Japanese	この問題は難しいです
			Romanized Version	*konomondaiwa muzukashii desu*
			English Translation	this question is difficult
	I10.4	✓	Japanese	山にヤギがいた
			Romanized Version	*yamani yagiga ita*
			English Translation	there were goats in the mountains
	I10.5	×	Japanese	僕は帰ると思う
			Romanized Version	*bokuwa kaeruto omou*
			English Translation	I think I will go home*
	I10.6	×	Japanese	くれた送って友達が
			Romanized Version	*kureta okutte tomodachiga*
			English Translation	home took me my friend*

Note. *Translation written in incorrect English word order to reflect the ungrammatical Japanese sentences.