A Rasch-Validation Study of a Novel Speaking Span Task

Bartolo Bazan
bazanlinkin2@gmail.com
Ryukoku University Heian Junior & Senior High School

Abstract

Working Memory refers to the capacity to temporarily retain a limited amount of information that is available for manipulation by higher-order cognitive processes. Several assessment instruments, such as the speaking span task, have been associated with the measurement of working memory span. However, despite the widespread use of the speaking span task, no study, to the best of my knowledge, has attempted to validate it using Rasch Measurement Theory. Rasch analysis can potentially shed light on the dimensionality of a complex construct such as working memory as well as examine whether a collection of items is working together to construct a coherent and reliable measure of a targeted population. This pilot study reports a Rasch analysis of a novel speaking span task, which was administered individually to 31 Japanese junior high school students and scored using a newly developed scoring system. Two separate analyses were conducted on the task: an analysis of the individual items using the Rasch dichotomous model and an analysis of the super items (sets) using the partial credit model. The results indicate that the task measures a coherent unidimensional latent variable and is thus a useful tool for measuring the construct. Moreover, Rasch analysis was shown to be suitable method for evaluating working memory tests.

Keywords: working memory, speaking span task, validation, Rasch measurement theory, Japanese

Working Memory (WM) can be defined as the capacity to temporarily store a limited amount of information that is available for manipulation by higher-order cognitive processes, such as language comprehension and production (Baddeley, 2012). Research in both first and second language (L2) acquisition has demonstrated that limitations in WM capacity may constrain the processes involved in language acquisition (Daneman & Green, 1986; Fortkamp, 1999; Gathercole & Baddeley, 1993; Guará-Tavares, 2008; Martin & Ellis, 2012). It has been acknowledged that individuals with higher WM capacity experience fewer difficulties than individuals with lower WM capacity in their attempts to learn an L2 successfully because of their increased aptitude to learn (Linck, Osthus, Koeth, & Bunting, 2014). Valid and reliable measurement of WM span is therefore essential in L2 research.

In the cognitive psychology literature, several methods of assessing WM capacity based on Baddeley's (2000) WM Model have been proposed, such as the speaking span task or the listening span task, and these have been adopted and adapted by L2 researchers. However, despite the widespread use of these instruments, no study, to the best of my knowledge, has attempted to validate them using Rasch model theory (Rasch, 1960). Rasch analysis can potentially shed light on the dimensionality of a complex construct, such as working memory, and whether an assessment instrument measures the hypothesized construct of WM span as intended (Bond & Fox, 2015). Moreover, Rasch analysis allows test developers to examine whether a collection of items is working together to construct a coherent and reliable measure of a targeted population.

This paper reports an analysis of a WM measurement instrument, namely a speaking span task, using Rasch model theory to tackle the validity issue of WM tests. Another purpose of this study is to obtain a baseline for the development of an improved second test. Validity is defined here as the inferences about a human ability that can be made from an observed performance on a task (Bond & Fox, 2015).

The Rasch Model

The Rasch model is a measurement model that uses test takers' responses on a test (correct and incorrect) to calculate ability level in terms of the measured construct relative to the difficulty of items on the test. The Rasch model converts raw scores into equal-interval scale data points, which allows for a more precise estimation of the target construct (Bond & Fox, 2015). Rasch measurement analyses offer a number of advantages over traditional techniques (Bond & Fox, 2015). First, Rasch analyses provide fit statistics to both evaluate the contribution of individual items to the measurement of the target construct and to investigate if test takers' performances match the model expectations. Second, Rasch analysis techniques provide variable maps, also known as Wright maps (Bond & Fox, 2015), which are plots that visually represent the person ability-item difficulty relationship on a single scale. The Wright maps are useful to evaluate the item difficulty hierarchy along the measured construct. Third, Rasch analyses provide reliability indices for both items and persons, which indicate the degree to which replicability of the item and person hierarchy along the variable is possible if the test were administered to a similar sample. Fourth, Rasch analyses produce person separation measures that estimate the number of statistically different ability groups into which a sample of test takers can be separated. Finally, the Rasch principal components analysis (PCA) serves as a way to identify deviations from the construct unidimensionality criterion underlying the Rasch model or, in other words, if the items measure a unidimensional construct. The PCA is complemented by the item fit graph, which provides a visual representation of how well the items adhere to the measurement of a single latent variable.

Baddeley's Working Memory Model

The concept of WM has received considerable attention in the cognitive sciences since the early 1970s and a large number of models have been proposed (Miyake & Shah, 1999). The most widely accepted model of WM is Baddeley and Hitch (1974) and Baddeley's (2000) multiple-component model. In this model, WM is composed of a limited capacity attentional control system known as the central executive and two subsidiary systems known as the phonological loop and the visuo-spatial sketchpad. The phonological loop involves the momentary storage of sounds and the rehearsal of aural information through inner speech and the decoding and storage of written language in phonological form. It has been shown to be a predictor of vocabulary learning, word recognition, and early reading skills (Gathercole & Baddeley, 1993). The visuo-spatial sketchpad performs the same functions of storage and rehearsal as the phonological loop but with visual images and spatial relations. The central executive, also known as executive control, regulates attentional resources and is the source of conscious processing, the creation of solutions to problems, and monitoring. A fourth subsidiary WM component, the episodic buffer, was later proposed by Baddeley (2000). The episodic buffer is responsible for linking WM with long-term memory (learned information). Another purpose of the episodic buffer is to chunk information in order to facilitate processing.

Executive Working Memory Measures

A number of tests, called span tasks, have been developed to measure the different components of Baddeley's (2000) WM model. There are two separate WM measurement paradigms (Wen, 2016). One, the simple memory span tasks paradigm, is comprised of storage-only versions of WM span tasks. Simple span tasks range from tasks involving the serial recall of digits or letters to the more recently proposed non-word recognition and non-word repetition span tasks (Gathercole, 1994). The construct that these tasks are intended to measure is the phonological loop (Baddeley, 2000) or phonological WM (PWM). The other is the dual-task paradigm which encompasses complex span tasks that impose concurrent demands on participants, such as the simultaneous storage and manipulation of information. These measures are designed to tax the executive functions of WM, or Executive WM (EWM), which is the construct represented in the task items of this pilot study. Current validity evidence for EWM tasks has thus far been limited to people with damaged frontal lobe brain regions (Miyake, Friedman, Emerson, Witzki, & Howerter, 2000). The present study, therefore, intends to extend the validity evidence available for EWM tasks. Among the different complex span tasks, one of the most commonly used is the speaking span task.

The original version of the speaking span task (Daneman & Green, 1986) required participants to silently read 70 seven-letter words arranged in five sets each of two, three, four, and five words. The words were presented consecutively on a computer screen for one second each with a 10 milliseconds gap between words. Participants continued to read the words until a blank screen was displayed, accompanied by a tone that signaled the end of the set. At this point, participants produced individual sentences for each of the words in the set, which were audio-recorded. Although there were no restrictions on the length of the sentences or the position of the stimulus word, the sentences were required to make sense both grammatically and semantically. A credit was given only when the sentence was grammatical and contained the target word in its original form of appearance. The participants' speaking span was determined by the total number of lexical items for which a grammatical utterance was produced, with a maximum possible score of 70. The participants' speaking span scores were compared to their scores on several measures of vocabulary production fluency showing positive correlations between the participants' EWM spans and their ability to produce synonyms from contextual cues. The authors concluded that the EWM span obtained with the speaking span task was a predictor of language (L1) fluency.

It may be argued, however, that it is unnecessary to employ a speaking span task with such a large number of items (70 items; 10 practice and 60 test items divided into five sets of two to five words) to obtain a valid measurement of EWM. In fact, extensive encounters with task items may degrade the purity of a complex span task (Miyake et al., 2000) because participants' performance with the relatively easier sets (i.e., the practice sets, the five sets of two words, and the five sets of three words) may prompt them to engage strategies to meet the requirements of the subsequent sets. In other words, the inclusion of so many items may induce the use of strategies, which would pollute the measurement of EWM. Moreover, including so many items may increase the likelihood of fatigue effects, which would also introduce random variance into the measurement. Furthermore, not only did the researchers make no attempt to control for either abstractness or frequency/familiarity of the lexical items, but they also included verbs inflected in past the tense (rather than plain verbs), which may have influenced the degree of difficulty of the task. This was illustrated in the two-word set example (Set 1: *quarter, battled*) Daneman and Green (p.11, 1980) provided.

The Speaking Span Task in L2 Research

Adaptions of Daneman and Green's (1986) EWM measure have been widely implemented in L2 research (e.g., Fortkamp, 1999, 2003; Guará-Tavares, 2008; Weissheimer & Mota, 2009; Wen, 2016). Fortkamp (1999) developed a version of the speaking span task based on Daneman's (1991) for use with a nonnative-English speaker population of Brazilian university students. The purpose of the research was to investigate whether the correlations between EWM capacity and L1 fluency found by Daneman (1991) were also true for L2 fluency. Participants' EWM span and L2 fluency were assessed by means of the speaking span task and a picture description/narration task, respectively. The speaking span task was administered individually on a computer and the stimuli words were in English (the participants' L2). This task was composed of 40 unrelated mono-syllabic English words. Participants were given approximately one second to read each word in the set until a blank screen was displayed accompanied by a tone that signaled the end of the set. Then, participants produced an English sentence with each word they could recall in the original order or presentation and in the original form of the word (i.e., noun). The speech-eliciting tasks were also carried out individually and were audio-recorded for an analysis of temporal variables and disfluency markers, such as rate of speech, amount of speech, filled pauses, and hesitation. The results of this study replicated those of Daneman's (1991) L1 investigation with high correlation coefficients.

Adopting Fortkamp's (1999) methodology, Guará-Tavares (2008) looked at the relationship between EWM and L2 performance under planned and unplanned conditions. L2 performance was operationalized in terms of the complexity, accuracy, and fluency paradigm. The participants' EWM span was measured in the same way as in Fortkamp's (1999) study. That is, the task was administered individually on a computer and the 50 Brazilian participants were required to read words and produce sentences in English. However, this speaking span task contained a larger number of items (20 practice items and 60 test items) in comparison to that of Fortkamp's (1999). The next phase of the experiment consisted of a picture-narration task in which participants in the planned group were given 10 minutes of planning time. Conversely, participants in the spontaneous group were requested to perform the task immediately after observing the pictures for 50 seconds. The scores of the speaking span task were correlated with those of the measures of L2 performance. The results indicated that EWM capacity was highly correlated with accuracy on the spontaneous condition and with fluency and complexity on the planned condition. Guará-Tavares concluded that individuals with higher EWM capabilities produce more fluent and complex speech under planned conditions.

The empirical studies reviewed in this section all share the same potential methodological problems as those identified in Daneman and Green's (1986) speaking span task, namely presenting a large number of items and not controlling for the nature of the vocabulary (i.e., word familiarity and concreteness). An additional issue has been the language in which the speaking span tasks were administered, the participants' L2, which is likely to be a confounding variable with EWM abilities (Linck et al., 2014). That is, the speaking span tasks listed above have probably indexed not only EWM, but also L2 proficiency and thus provide an impure measurement of EWM capacity (Miyake et al., 2000). In the present study, these potential confounds were accounted for when developing the novel speaking span presented herein.

Although adaptions of the original speaking span task have been extensively implemented in L2 research (Fortkamp, 1999, 2003; Guará-Tavares, 2008; Weissheimer & Mota, 2009; Wen, 2016), empirical studies on the validity of these measures are lacking. This raises doubts about whether the instruments do in fact measure a common latent variable (i.e., EWM). This pilot study addresses this issue by analyzing a novel speaking span task using Rasch theory. The results of this initial investigation will serve as the basis to develop a more refined speaking span test in future.

Research Questions

This pilot study attempts to answer the following research questions to provide validity evidence for the speaking span task as a measure of EWM. Validity is examined through Rasch analyses, which provide indicators of whether the instrument measures the intended variable (Bond & Fox, 2007).

1. Does the dataset show acceptable fit to the Rasch model?
2. Does the difficulty of the items increase as the sets become larger?
3. Is item reliability sufficient to suggest replicability of the item difficulty hierarchy if the test is given to a similar sample?
4. Is person reliability sufficient to suggest a similar spread of participants with higher and lower ability across similar samples?
5. Do the items separate participants into higher and lower ability in the latent construct?
6. Is the instrument sufficiently unidimensional?

Methodology

Participants

The participants for this study were a group of Japanese second-year junior high school (8th grade) students ($N = 31$), aged between 13 and 14 years old, at a private junior and senior high school in Western Japan. At this institution, students were streamed into high-, intermediate-, and low-level classes by academic level. The participants who took part in the speaking span task belonged to the intermediate class. Of these students, 18 were female and 13 were male. All participants' first language was Japanese.

Instrument and Administration

For the purpose of this study, I constructed a speaking span task that contained two novel features, thus differentiating it from previously proposed speaking span instruments. First, unlike the widely used standard form of the instrument, in which the stimuli are presented visually on a computer screen, I decided to present the target words in auditory form. The rationale behind this way of presenting the stimuli was to avoid mixing academic skills (i.e., reading) with communicative skills (i.e., speaking) as this may confound EWM with L1 reading proficiency. The audio was recorded by a female Japanese native speaker. The task was carried out in the participants' native language because tasks conducted in the L2 could measure not only EWM span but also L2 proficiency (Linck et al., 2014). The task was trialed with a group of 35 first-year high school students and functioned as expected. That is, the participants found the sets increasingly challenging and there were participants performing at different levels. Second, the task was shorter than its (L1) predecessors. The test consisted of 40 unrelated Japanese words that were randomly arranged into two sets of two, three, four, five, and six words (see Appendix A). Each word was followed by a one-second gap.

All words contained two or three mora and were nouns in Japanese (e.g., *eki, station*) though some may be verbalized by adding the suffix (*-suru*) (e.g., *ryokou, travel*). An attempt was made to control for the familiarity of items by including words that are well known to junior high school students, as confirmed by two Japanese speakers. Furthermore, memory research has shown that concrete words are easier to recall than abstract words (Gathercole & Baddeley, 1993), so I included 12 abstract words (see Appendix A) in order to increase the discriminatory power of the measure, that is, to increase the sensitivity of the measure to separate participants into various levels of EWM span.

Contrary to the standard speaking span task, this test did not include practice sets. The speaking span task did not seem as complex as to require practice trials in comparison with other EWM tasks, such as the Tower of Hanoi (Miyake et al., 2000) or the Wisconsin Card Sorting Test (Monchi, Petrides, Petre, Worsley, & Dougher, 2001). Furthermore, the test was not a computerized test and thus practice to understand the functions of the buttons was not necessary. Instead, the participants read the directions in Japanese, received a Japanese verbal explanation followed by some time to ask clarification questions, and saw an example performed by the researcher.

The participants performed the task individually in a quiet classroom. Each administration of the task took approximately 10 minutes and was audio-recorded. First, the participants were given the directions of the task in Japanese. They were instructed to remember the words in the sets and produce utterances containing the words in their order of appearance. Inflectional changes of the target words were accepted as correct. At the beginning of each set, the number of words that the set contained was made explicit. No restrictions were imposed on length or complexity of the utterances.

In order to score performance, a new scoring system was created. A point was given to each utterance produced correctly (i.e., made sense in Japanese) and in the order of appearance until failure to recall in order. That is, if on a set of five items, a participant produced sentences in the correct order from item one to three, failed to recall item four, but succeeded on item five, she would get three points on the set. This scoring system differs from the commonly used maximum set size (i.e., a credit is given for a set if all the items are recalled correctly) and total score performance procedures (i.e., a credit is given for each item recalled) (Wen, 2016) because it takes into account the ability to recall the positions of the items within the set. As participants have to hold in memory not only the items themselves, but also their positions within the test, this scoring procedure may provide a stricter estimate of the construct. The rationale is that after memory failure to recall items in order, participants are likely to engage in idiosyncratic strategies to recall the rest of the words in the list such as using the word's initial mora as a retrieval cue or attempting to recall the last word in the set before preceding words. Thus, this way of scoring the task prevents the last items in the sets, which are theorized to be the most difficult, from benefiting from recency effects (Kahana, Howard, Zaromb, & Wingfiled, 2002). For these reasons, the words that were recalled out of order were not scored.

This scoring procedure does, however, create problems of local dependency among the items in the sets because participants are not given a credit for items unless they have been successful with the preceding items. Table 1 confirmed that pairs of items within sets frequently showed correlated residuals. The assumption of local independence of items of the Rasch model was, therefore, violated. To partial out the effects of item local dependency, a partial credit analysis of the super-items (i.e., the sets treated as items) followed the dichotomous analysis of the individual items.

Table 1
Residual correlations used to identify dependent items

Correlation	Entry	Item	Entry	Item
1.00	3	I2.1, hari, needle	4	I2.2, soujiki, vacuum cleaner
.82	6	I3.2, yuubinkyoku, post office	7	I3.3, mesamachi, alarm clock
.71	22	I7.4, kiken, danger	37	I10.3, chikara, strength
.69	27	I8.4, kippu, ticket	32	I9.4, netsu, fever
.66	9	I4.2, randoseru, schoolbag	10	I4.3, keisatsu, police
.65	32	I9.4, netsu, fever	38	I10.4, eki, station
.65	38	I10.4, eki, station	39	I10.5, hige, moustache
.63	36	I10.2, piano, piano	37	I10.3, chikara, strength
.62	13	I5.3, yakusoku, promise	14	I5.4, ningen, people
.61	17	I6.3, hanashi, talk	18	I6.4, yubiwa, ring
.58	22	I7.4, kiken, danger	23	I7.5, takara, treasure

Rasch Analysis

Data were examined using Winsteps 4.3.1 Rasch software (Linacre, 2018). Two separate analyses were conducted on the measure: an analysis of the individual items using the Rasch dichotomous model and an analysis of the sets (super items or testlets) using the partial credit model. To explore Research Question 1, person and item fit statistics were examined. To explore Research Question 2, the Wright Map (Bond & Fox, 2015) was examined. Research Questions 3, 4, and 5 were investigated by looking at the item reliability, person reliability and separation indices, respectively. For Research Question 6, the item fit graph was inspected and a principal components analysis (PCA) of item residuals was conducted. All of the previously mentioned indicators reveal the degree to which the instrument is measuring a coherent unidimensional latent variable.

Results

Speaking Span Task Items

Person and item fit statistics

To examine whether the observed participants' performance matched the expectations of the model, person fit was examined. The infit mean square fit statistics, which are weighted non-standardized statistics, provide information about participants whose probability of getting an item correct is close to the difficulty of the items.

Less than 0.50 and above 1.50 mean-squares (MNSQ) is the rule of thumb to identify misfitting participants (Linacre, 2007). A visual inspection of the Winsteps person-statistics output table indicated that most participants behaved as expected by the model (see Table 2). Two participants (Persons c201 and c225, infit 1.87 and 1.60, respectively) were above the 1.50 cut-off value. However, 1.87 and 1.60 are not values that raise alarms as values between 1.50 and 2.00 do not degrade measurement (Linacre, 2007). There was, however, one participant (Person c226) with an extreme infit MNSQ value of 2.19. This participant also had an extreme outfit value of 9.90 (see Table 2), which suggests he or she may have been a low performer who used an idiosyncratic strategy (e.g., initial word mora recall or word chaining strategy) to correctly recall several items that were beyond his or her actual ability. Participants c226 and c225 were removed one at a time and the data was reanalyzed, but this did not improve the quality of the measure. For example, with participant c226 deleted, the person separation index and the variance explained by the measure decreased slightly (2.09 and 58.30, respectively). Although item and person reliability remained the same relative to the first analysis (.93 and .81, respectively), the item hierarchy showed departures from the hypothesized order of difficulty. That is, the last items of the sets (the theorized most difficult items) lay below the middle items on the Wright map. An additional drawback was that the measure of the participants did not improve. That is, a number of participants, who fitted the model on the first analysis, were found to misfit on the second. For these reasons, these two misffiting participants were retained and the first analysis was continued.

Table 2

Person statistics of the speaking span task (individual items)

Entry	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Person
24	2.12	0.84	2.19	2.90	9.90	4.40	c226
1	0.89	0.75	1.87	2.40	2.03	1.10	c201
23	-0.58	0.69	1.60	1.80	1.61	0.90	c225
9	1.19	0.65	1.39	1.30	0.80	0.20	c209
22	0.89	0.63	1.31	1.10	1.04	0.40	c224
5	-0.58	0.62	1.30	1.00	1.12	0.40	c205
29	2.12	0.62	1.21	0.70	0.84	0.30	c233
7	-1.18	0.60	1.19	0.70	0.83	0.20	c207
2	1.49	0.55	1.00	0.10	1.14	0.50	c202
8	0.01	0.57	1.11	0.50	0.65	-0.20	c208
11	-1.48	0.58	1.11	0.40	0.66	0.10	c211
4	-1.78	0.58	1.09	0.40	0.69	0.20	c204
17	-3.12	0.61	1.02	0.20	0.47	0.00	c217
16	0.30	0.55	1.01	0.10	0.66	-0.20	c216
12	0.01	0.55	1.00	0.10	0.95	0.20	c212
19	-1.18	0.55	0.95	-0.10	0.64	0.00	c219
3	1.19	0.55	0.90	-0.20	0.45	-0.20	c203
31	-1.18	0.55	0.77	-0.70	0.89	0.30	c235
26	0.59	0.55	0.88	-0.30	0.81	0.10	c222
20	-1.78	0.56	0.86	-0.40	0.53	0.00	c220
25	-0.29	0.55	0.86	-0.40	0.72	-0.10	c221
27	-0.29	0.55	0.86	-0.30	0.49	-0.50	c227
15	2.44	0.57	0.83	-0.50	0.41	-0.10	c215
18	0.89	0.55	0.70	-1.00	0.40	-0.40	c218
30	-0.29	0.55	0.67	-1.10	0.35	-0.80	c234
14	-1.78	0.56	0.65	-1.20	0.34	-0.20	c214
13	0.89	0.55	0.60	-1.50	0.30	-0.60	c213
6	-1.78	0.56	0.59	-1.50	0.28	-0.30	c206
10	-0.88	0.55	0.43	-2.30	0.23	-0.90	c210
21	1.49	0.55	0.43	-2.30	0.22	-0.50	c223
28	1.49	0.55	0.43	-2.30	0.22	-0.50	c228

Note. MNSQ = mean-squared; ZSTD = Standardized z-scores.

Item infit MNSQ statistics indicate the extent to which an item contributes to the measurement of the underlying construct (Bond & Fox, 2007). These indices reveal whether single, unidimensional construct is measured. Table 3 shows that the items contributed to measure the construct of interest. All the items are within the parameters (Infit MNSQ) 0.50 and 1.50 with the exception of one (item 7.3, *undou, exercise*), which is slightly above 1.60, but this value is not of concern. In contrast, item 3.2 (*yuubinkyoku, post office*), although within the infit MNSQ criteria, appears to have an extreme outfit MNSQ value (9.90). A visual examination of the table reveals that this item is the longest item on the test, which suggests that word length may have the cause.

Table 3
Item statistics of the speaking span task (individual items)

Entry	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Item
6	-4.21	1.15	1.20	0.50	9.90	3.90	I3.2, *yuubinkyoku*
7	-1.44	0.47	0.99	0.00	1.69	1.30	I3.3, *mezamashi*
21	1.03	0.56	1.60	2.50	1.57	1.30	I7.3, *undou*
16	-1.67	0.57	1.36	1.40	1.33	0.70	I6.2, *kusuri*
20	-1.22	0.50	1.17	0.80	1.35	0.90	I7.2, *tera*
29	-2.90	0.75	1.31	0.80	1.17	0.50	I9.1, *shiken*
30	0.10	0.48	1.24	1.20	1.27	0.90	I9.2, *omocha*
22	2.86	0.70	1.17	0.50	0.75	0.10	I7.4, *kiken*
19	-3.41	0.83	1.14	0.40	0.78	0.30	I7.1, *shigoto*
24	-3.41	0.83	1.14	0.40	0.60	0.10	I8.1, *kusa*
11	-4.21	1.10	1.11	0.40	0.70	0.20	I5.1, *kagami*
23	4.13	1.07	1.07	0.40	0.86	0.40	I7.5, *takara*
26	2.49	0.58	0.95	0.00	1.06	0.40	I8.3, *kizu*
17	0.46	0.44	1.04	0.30	0.97	0.00	I6.3, *hanashi*
12	-1.44	0.47	1.01	0.10	0.89	0.00	I5.2, *wasabi*
34	4.13	1.04	1.01	0.30	0.48	0.00	I9.6, *sekken*
31	2.18	0.54	0.92	-0.20	0.77	0.00	I9.3, *ryokou*
35	-2.19	0.54	0.92	-0.20	0.89	0.10	I10.1, *byouin*
37	1.91	0.51	0.91	-0.30	0.56	-0.50	I10.3, *chikara*
13	0.28	0.43	0.90	-0.40	0.89	-0.30	I5.3, *yakusoku*
25	-0.45	0.43	0.90	-0.40	0.84	-0.40	I8.2, *kutsushita*
32	4.13	1.03	0.90	0.20	0.27	-0.30	I9.4, *netsu*
39	4.13	1.03	0.90	0.20	0.27	-0.30	I10.5, *hige*
18	1.23	0.45	0.84	-0.70	0.62	-0.80	I6.4, *yubiwa*
10	-1.22	0.46	0.81	-0.80	0.63	-0.80	I4.3, *keisatsu*
38	3.35	0.76	0.80	-0.20	0.28	-0.30	I10.4, *eki*
36	1.23	0.45	0.75	-1.20	0.55	-1.00	I10.2, *piano*
14	0.84	0.44	0.73	-1.40	0.56	-1.20	I5.4, *ningen*
27	3.35	0.76	0.73	-0.30	0.25	-0.40	I8.4, *kippu*
3	-4.21	1.05	0.66	-0.20	0.12	-0.60	I2.1, *hari*
4	4.21	1.05	0.66	-0.20	0.12	-0.60	I2.2, *soujiki*
9	-1.67	0.49	0.66	-1.50	0.43	-1.10	I4.2, *randoseru*

Note. MNSQ = mean-squared; ZSTD = Standardized z-scores.

Wright map

Figure 1 shows a visual representation of participants' EWM capacity/item difficulty relations side by side on a common logit scale. Participants are located on the left along the scale based on their EWM span (the higher up the plot, the higher the participants' EWM score) and are represented by an 'X'. Items are placed on the right (the higher up the scale, the more difficult the item). The most difficult items were items 10.6, 9.5, and 8.5. This is not surprising because sets 10 and 9 were the most demanding (same level of difficulty) followed by set 8. As can be seen, the item difficulty hierarchy plotted on the Wright map aligns with the theoretical ordering of the items. In other words, the further the position of the item within the set, the more difficult the item is to answer. The items in each set are listed in descending order from the most difficult to answer (the last item in the set) to the easiest (the first item in the set). Set 9 does not follow this pattern because item 9.5 (*uwagi*, *coat*) and item 9.6 (*sekken*, *soap*) exchanged places in the difficulty hierarchy. That is, item 9.5 was above the difficulty level of the theorized most difficult item (item 9.6). This is probably because no student succeeded on either item (as an examination of Table 13 of the Winsteps output revealed), which meant that the items were not estimated. Consequently, Winsteps assumed that item 9.5 was as difficult as item 9.6. The easiest sets (sets 1 and 2), which contain two items each, were also not ordered in descending difficulty because they were well within the participants' ability. As expected, all participants completed set 1 and only one participant failed to complete set 2 (as Table 13 of the Winsteps output showed), which made the two items in each set lie next to each other at the same level of difficulty.

An examination of the map reveals that the speaking span task was difficult for the sample. There were nine items (items 10.6, 10.5, 10.4, 9.6, 9.5, 9.4, 8.5, and 8.4) that targeted no participant in the sample as they were above the EWM capacity of the most capable participant, which suggests that the sample needed participants with higher EWM spans. Six of those items belonged to the most difficult sets (set 10 and set 9), which was unsurprising. Of the top three most difficult items that participants could answer (item 10.3 *chikara,* item 9.3 *ryokou*, and item 8.3 *kizu,* which mean *strength*, *trip*, and *wound*), two were abstract words (*ryokou* and *chikara*) and one could be concrete or abstract depending on the context (*kizu*, *wound*). This suggests that abstract words can help differentiate persons with more of the construct from persons with less of the construct. The spread of the items was larger than the spread of the participants. However, the test was adequate to measure EWM capacity as it seems, at least visually, that it provided a sufficient level of discrimination between the participants in the sample.

```
MEASURE PERSON - MAP - Item
         <more>|<rare>
    5          +  I10.6, shiai       I8.5, ike          I9.5, uwagi
               |
               |
               |
               |
               |  I10.5, hige        I7.5, takara       I9.4, netsu
                  I9.6, sekken
    4          +
               |
               |
               |
               |  I10.4, eki         I8.4, kippu
               |
    3          +
               |  I7.4, kiken
              T|S
            X  |  I8.3, kizu
               |
           XX  |  I9.3, ryokou
    2          +
               |  I10.3, chikara
               |
          XXX  |
              S|
           XX  |  I10.2, piano       I6.4, yubiwa
    1          +  I7.3, undou
         XXXX  |  I5.4, ningen
            X  |
               |  I6.3, hanashi
            X  |  I5.3, yakusoku
                  I9.2, omocha
    0       XX M+M
               |
          XXX  |
               |  I8.2, kutsushita
           XX  |
            X  |
   -1          +
          XXX  |  I4.3, keisatsu    I7.2, tera
              S|
            X  |  I3.3, mezamashi   I5.2, wasabi
               |  I4.2, randoseru   I6.2, kusuri
         XXXX  |
   -2          +
               |  I10.1, byouin
               |
               |
               |S
              T|  I9.1, shiken
   -3          +
            X  |
               |  I7.1, shigoto     I8.1, kusa
               |
               |
               |
   -4          +
               |  I2.1, hari        I2.2, soujiki      I3.2, yuubinkyoku
                  I5.1, kagami
               |
               |
               |
               |
   -5          +  I1.1, kushi       I1.2, okane        I3.1, gomi
                  I4.1, shirokuma   I6.1, tenki
         <less>|<freq>
```

Figure 1. Wright map for the individual items of the speaking span task.

Person and item reliability and separation

The person reliability values indicate the level of replicability of person ordering if the participants were given a similar test of EWM. In contrast, the item reliability values indicate the level of replicability of item location along the scale if the speaking span test was given to a different sample. Person reliability or test reliability was estimated at .81 (see Table 4) and the reliability of the items was .93 (see Table 5). These values are above the cut-off guideline of .80 (Linacre, 2007), which indicates moderate person reliability and high item reliability, respectively. This means that participants or items with higher measures are likely to have higher measures than participants or items estimated with lower measures were the test given again. This also means that the speaking span task had an adequate difficulty range to discriminate between participants with different EWM spans. In other words, item difficulty sufficiently covered the range of person abilities. The Rasch person separation was calculated at 2.10 (see Table 4), suggesting the participants could be divided into two levels of EWM spans. In other words, the instrument had enough sensitivity to distinguish the participants with high EWM from those with low EWM spans. In contrast, the item separation was calculated at 3.61 (see Table 5), indicating that the measure divides the items into three levels of difficulty. These results provide evidence to support construct validity and reasonable confidence of replicability of the person and item ordering across similar samples. Tables 4 and 5 provide the infit MNSQ, which is expected to have a mean of 1.00 (Bond & Fox, 2015). The person fit and item fit statistics for the speaking span task were close to the ideal value of 1.00 (person fit infit MSNQ = 0.99 and outfit MNSQ = 0.99 and item fit infit MNSQ = 0.98 and outfit MNSQ = 1.04, respectively). As these values did not deviate substantially from the Rasch-modeled expectation of 1.00, the measurement can be said to contain little distortion or random noise (Linacre, 2018).

Table 4

Summary of the speaking span task analysis(persons)

	Total Score	Count	Measure	Real *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	21.00	40.00	-0.01	0.59	0.99	-0.10	0.99	0.10
P. *SD*	4.60	0.00	1.37	0.07	0.39	1.20	1.67	0.90
S. *SD*	4.60	0.00	1.40	0.07	0.40	1.30	1.70	0.90
Max	29.00	40.00	2.44	0.84	2.19	2.90	9.90	4.40
Min	11.00	40.00	-3.12	0.55	0.43	-2.30	0.22	-0.90
REAL *RSME*		0.59	TRUE *SD*	1.24	SEPARATION	2.10	PERSON RELIA.	.81
MODEL *RSME*		0.55	TRUE *SD*	1.26	SEPARATION	2.27	PERSON RELIA.	.84
SE OF PERSON MEAN = 0.25								
PERSON RAW SCORE-TO-MEASURE CORRELATION = 1.00								
CRONBACH ALPHA (*KR-20*) PERSON RAW SCORE “TEST” RELIABILITY = .82								
SEM = 1.94								

Note. P. SD = Population standard deviation; S. SD = Sample standard deviation; Max. = maximum value; Min = minimum value; RSME = square-root of the average error variance; SD = Standard deviation; RELIA. = reliability; SE = Standard error; SEM = standard error of the mean.

Table 5

Summary of the speaking span task analysis (individual items)

	Total Score	Count	Measure	Real *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	15.50	31.00	0.00	.69	0.98	0.10	1.04	0.10
P. *SD*	10.70	0.00	2.73	.25	0.22	0.10	1.64	0.90
S. *SD*	10.90	0.00	2.77	.25	0.22	0.80	1.67	0.90
Max	30.00	31.00	4.13	1.15	1.60	2.50	9.90	3.90
Min	1.00	31.00	-4.21	.43	0.66	-1.50	0.12	-1.20
REAL *RSME*		0.73	TRUE *SD*	2.63	SEPARATION	3.61	PERSON RELIA.	.93
MODEL *RSME*		0.70	TRUE *SD*	2.64	SEPARATION	3.74	PERSON RELIA.	.93
SE OF ITEM MEAN = 0.49								

Note. P. SD = Population standard deviation; S. SD = Sample standard deviation; Max = maximum value; Min = minimum value; RSME = square-root of the average error variance; SD = Standard deviation; RELIA. = reliability; SE = Standard error.

PCA of item residuals and item fit graph

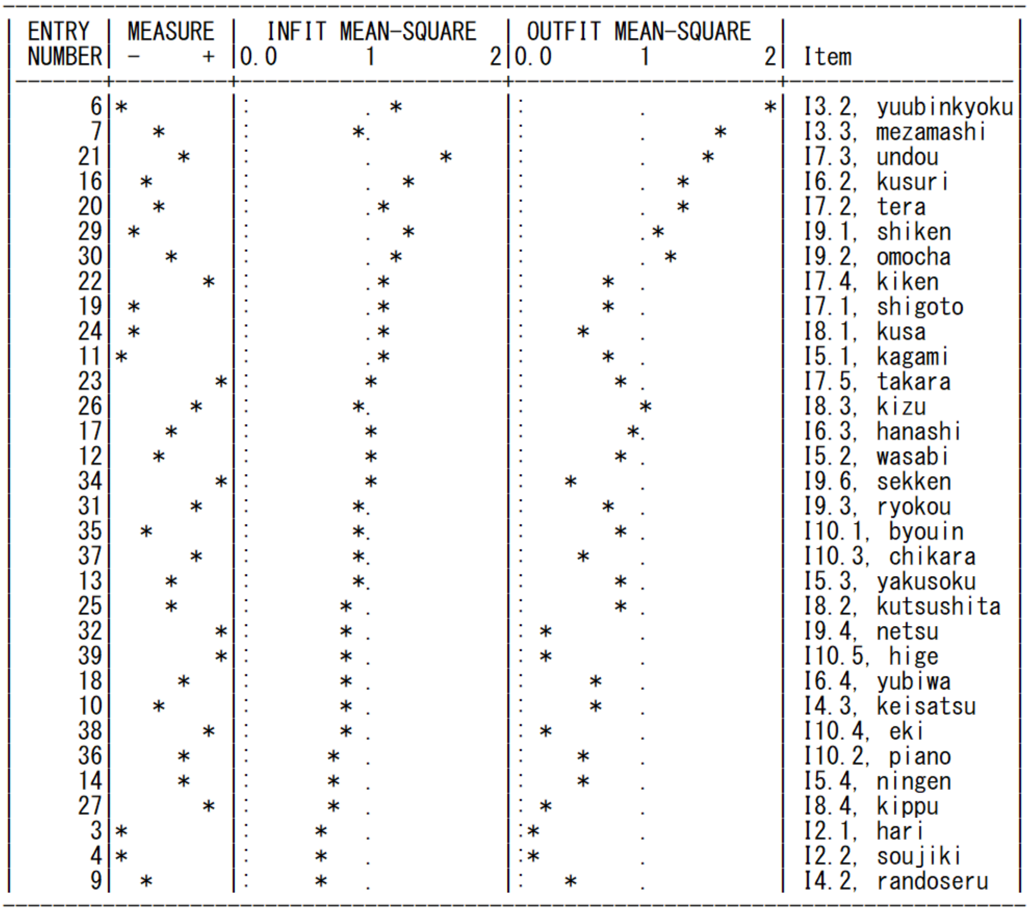
The unidimensionality of the construct was investigated through a PCA item residuals analysis. A unidimensional construct should adhere to two criteria: first, it should explain 20.00% or more of the variance in the data (Reckase, 1979). Second, unexplained variance in the first residual contrast should have an eigenvalue below 2.00 (Linacre, 2018) and represent less than 10.00% of the total variance (Linacre, 2007). The results of the analysis showed that the measure accounted for 59.40% of the total variance in the data with an eigenvalue of 46.82, and that the first principal contrast had an eigenvalue of 4.65 and accounted for 5.90% of the variance (see Table 6). The variance explained by the construct was above the cut-off value of 20.00%. The total variance explained by this dichotomous model was very similar (as shown below) to that of the partial credit analysis of the sets (super items), meaning that the items contributed to the measurement of a single latent variable independently of whether they were examined individually or as part of a set. The items' contribution to the measurement of a single construct is displayed in Figure 2. The linear ordering of the items shown by the infit mean-square pathway (Bond & Fox, 2015) suggests that the items relate to a single latent variable.

The eigenvalue of the first contrast (4.65) suggested that a second dimension may exist. Therefore, the content of the contrasted items was analyzed. Although word length appeared as a potential additional dimension (*netsu*, *kippu*, and *eki* vs. *yuubinkyoku*, *mezamashi*, and *yubiwa*, which mean *fever*, *ticket*, and *station*, and *post office*, *alarm clock*, and *ring*, respectively) (see Appendix A), the high eigenvalue seems to be the result of a test effect rather than a substantive construct. This is because the variance explained (5.90%) was not large enough (i.e., > 10.00%) to negatively affect the measurement of the main construct.

The other contrasts that had high eigenvalues (i.e., the 2nd to 5th contrasts) were examined but these explained little variance (see Table 6). Together, all five contrasts accounted for less variance (21.60%) than did the items (39.20%), which provides additional evidence to refute the presence of a second dimension.

Table 6
Speaking span task (individual items) standard residuals in eigen values

	Eigenvalue	Observed	Expected
Total Raw variance in observations	78.82	100.00%	100.00%
Raw variance explained by measures	46.82	59.40%	59.10%
Raw variance explained by persons	15.93	20.20%	20.10%
Raw variance explained by items	30.89	39.20%	39.00%
Raw unexplained variance (total)	32.00	40.60%	40.90%
Unexplained variance in 1st contrast	4.65	5.90%	
Unexplained variance in 2nd contrast	4.06	5.20%	
Unexplained variance in 3rd contrast	3.18	4.00%	
Unexplained variance in 4th contrast	2.60	3.30%	
Unexplained variance in 5th contrast	2.50	3.20%	



```
----------------------------------------------------------------------------------
| ENTRY  | MEASURE |  INFIT MEAN-SQUARE  | OUTFIT MEAN-SQUARE  |                 |
| NUMBER | -     + |0.0       1         2|0.0       1         2| Item            |
|--------+---------+---------------------+---------------------+-----------------|
|      6|*        |:         .  *       |:         .         *| I3.2, yuubinkyoku|
|      7|   *     |:        *.          |:         .     *    | I3.3, mezamashi  |
|     21|     *   |:         .     *    |:         .    *     | I7.3, undou      |
|     16|  *      |:         .  *       |:         .  *       | I6.2, kusuri     |
|     20|   *     |:         .*         |:         .  *       | I7.2, tera       |
|     29| *       |:         .  *       |:         .*         | I9.1, shiken     |
|     30|    *    |:         . *        |:         . *        | I9.2, omocha     |
|     22|       * |:         .*         |:      *  .          | I7.4, kiken      |
|     19| *       |:         .*         |:      *  .          | I7.1, shigoto    |
|     24| *       |:         .*         |:    *    .          | I8.1, kusa       |
|     11|*        |:         .*         |:      *  .          | I5.1, kagami     |
|     23|        *|:         *          |:       * .          | I7.5, takara     |
|     26|      *  |:        *.          |:         *          | I8.3, kizu       |
|     17|    *    |:         *          |:        *.          | I6.3, hanashi    |
|     12|   *     |:         *          |:       * .          | I5.2, wasabi     |
|     34|        *|:         *          |:   *     .          | I9.6, sekken     |
|     31|      *  |:        *.          |:      *  .          | I9.3, ryokou     |
|     35|  *      |:        *.          |:       * .          | I10.1, byouin    |
|     37|      *  |:        *.          |:    *    .          | I10.3, chikara   |
|     13|    *    |:        *.          |:       * .          | I5.3, yakusoku   |
|     25|    *    |:       * .          |:       * .          | I8.2, kutsushita |
|     32|        *|:       * .          |: *       .          | I9.4, netsu      |
|     39|        *|:       * .          |: *       .          | I10.5, hige      |
|     18|     *   |:       * .          |:     *   .          | I6.4, yubiwa     |
|     10|   *     |:       * .          |:     *   .          | I4.3, keisatsu   |
|     38|       * |:       * .          |: *       .          | I10.4, eki       |
|     36|     *   |:      *  .          |:    *    .          | I10.2, piano     |
|     14|     *   |:      *  .          |:    *    .          | I5.4, ningen     |
|     27|       * |:      *  .          |: *       .          | I8.4, kippu      |
|      3|*        |:     *   .          |:*        .          | I2.1, hari       |
|      4|*        |:     *   .          |:*        .          | I2.2, soujiki    |
|      9|  *      |:     *   .          |:   *     .          | I4.2, randoseru  |
----------------------------------------------------------------------------------
```

Figure 2. Item fit graph for the speaking span task (individual items).

Speaking Span Task Super Items

Following the item analyses, as the scoring system created local dependency among the items, I conducted an examination of the sets or super items (total scores on the set). The same research questions were investigated with regards to the sets.

There were 10 sets that were treated as 10 individual items. For this analysis, I used a partial credit model with codes ranging from zero (minimum score) to six (maximum possible score on the largest sets). Sets one and two were composed of two individual items so their score range was from zero to two points; sets three and four contained three items so their score range was zero to three points; and so on, up to sets 9 and 10, which had a score range of zero to six.

Person and item fit statistics

A person fit examination was conducted. An expected finding was that the same two participants (c225 and c226), who were found to misfit on the analysis of the individual items, were again identified as misfitting on the analysis of the super items with infit MNSQ values of 2.34 and 1.88, respectively (see Table 7). As occurred in the previous analysis, participant c226 had an extreme outfit value (7.54). Interestingly, these participants took the test on the same day and one after the other, which seems to indicate that performance may have been influenced by an external factor. However, another possible explanation for the participants' misfit may be the use of a small sample size (Boone & Noltemeyer, 2017). Nevertheless, overall, the participants' performances fit the expected model (see Table 7).

Table 7

Person statistics of the speaking span task (super items)

Entry	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Person
24	1.87	0.66	2.34	2.00	7.54	2.50	c226
23	0.25	0.61	1.88	1.50	1.73	0.90	c225
22	1.16	0.53	1.61	1.20	1.19	0.60	c224
29	1.87	0.53	1.54	1.00	0.85	0.40	c233
1	1.16	0.52	1.53	1.00	0.99	0.50	c201
9	1.34	0.51	1.50	1.00	1.18	0.60	c209
7	-0.16	0.51	1.23	0.60	1.16	0.50	c207
4	-0.61	0.53	1.19	0.50	1.16	0.50	c204
12	0.63	0.46	1.14	0.40	0.92	0.40	c212
11	-0.38	0.50	1.11	0.40	1.12	0.40	c211
8	0.63	0.44	1.05	0.30	1.07	0.50	c208
20	-0.61	0.49	1.04	0.30	0.93	0.10	c220
16	0.81	0.43	1.01	0.20	0.76	0.30	c216
19	-.16	0.46	1.00	0.20	0.93	0.20	c219
26	0.99	0.42	0.97	0.10	0.65	0.20	c222
25	0.44	0.44	0.89	0.00	0.77	0.20	c221
31	-0.16	0.46	0.88	-0.10	0.61	-0.40	c235
15	2.06	0.44	0.85	-0.10	0.52	0.10	c215
2	1.51	0.42	0.73	-0.40	0.70	0.20	c202
3	1.34	0.42	0.73	-0.40	0.51	0.10	c203
27	0.44	0.44	0.70	-0.50	0.55	-0.10	c227
5	0.25	0.44	0.69	-0.50	0.63	-0.10	c205
17	-1.61	0.51	0.68	-0.70	0.66	-0.60	c217
18	1.16	0.42	0.60	-0.70	0.46	0.00	c218
13	1.16	0.42	0.58	-0.80	0.42	-0.10	c213
14	-0.61	0.48	0.50	-1.00	0.48	-0.90	c214
30	0.44	0.44	0.48	-1.10	0.46	-0.20	c234
6	-0.61	0.48	0.44	-1.20	0.39	-1.20	c206
21	1.51	0.42	0.43	-1.20	0.25	-0.30	c223
28	1.51	0.42	0.43	-1.20	0.25	-0.30	c228
10	0.04	0.45	0.30	-1.70	0.29	-0.90	c210

Note. MNSQ = mean-squared; ZSTD = Standardized z-scores.

A Rasch item fit analysis followed the person fit analysis. Item fit statistics (see Table 8) showed that the sets had infit MNSQ values well within the criterion values (0.50 and 1.50), demonstrating the contribution of the sets to the measurement of a single construct. This was not true, however, for set 2, which was found to be overfitting (i.e., functioning better than expected by the Rasch model) with an infit MNSQ value of 0.51 and an outfit MNSQ value of 0.07. These values indicated that the set was redundant and did not contribute to the measurement of the construct, probably because the set (2 items) was too easy for the sample as all the participants completed it. This suggests that future implementations of this speaking span test should begin with set 3 (3 items). Nevertheless, as overfitting items (in this case super items) do not degrade measurement (Bond & Fox, 2015), set 2 was retained in the analysis.

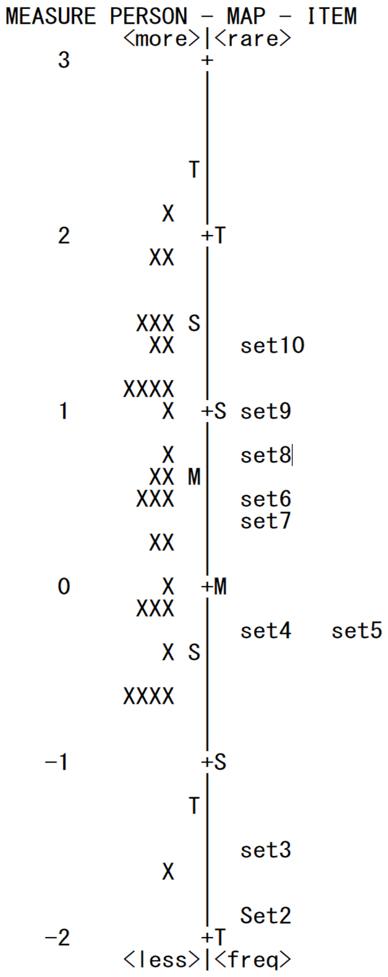
Table 8
Item statistics of the speaking span task (super items)

Entry	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Item
3	-1.49	0.39	1.06	0.30	2.25	2.40	set3
7	0.32	0.26	1.24	1.00	1.23	1.00	set7
6	0.52	0.22	1.15	0.70	1.13	0.60	set6
9	1.02	0.24	1.10	0.50	1.09	0.40	set9
5	-0.26	0.19	0.98	0.00	0.94	-0.10	set5
8	0.70	0.23	0.83	-0.60	0.88	-0.40	set8
10	1.33	0.19	0.81	-0.60	0.68	-1.00	set10
4	-0.30	0.26	0.73	-1.10	0.49	-0.90	set4
2	-1.83	0.56	0.51	-0.40	0.07	-0.80	set2

Note. MNSQ = mean-squared; ZSTD = Standardized z-scores.

Wright map

Figure 3 shows that the hierarchy of the items is close to the hypothesized level of difficulty of the sets. The most difficult sets, sets 10 and 9, are at the top and the easiest set, set 2, is at the bottom. It is important to note that set 1 was not reported by Winsteps, probably due to a ceiling effect as all 31 participants completed this set (as shown by an inspection of Table 13 of the Winsteps output). Sets 6 and 7 appeared to be flipped with respect to their theorized order. This may be due to some perceived relationship between words in set 7, which made recalling the words somewhat easier. However, the measurement error shown in Table 8 indicates that set 6 (*SE* = .22) and set 7 (*SE* = .26) are approximately 0.20 logits apart and thus, their difficulty levels cannot be statistically separated (Linacre, 2020). The map also shows that the sample found the speaking span sets relatively easy, with most participants falling within the range of 0.00 and 2.00 logits. Nevertheless, the considerable spread of the sample suggests that the instrument effectively separated the participants into levels of EWM ability.



```
MEASURE PERSON - MAP - ITEM
         <more>|<rare>
    3          +
               |
               |
               |
               |
              T|
               |
            X  |
    2          +T
           XX  |
               |
               |
          XXX S|
           XX  |  set10
               |
         XXXX  |
    1       X  +S set9
               |
            X  |  set8
           XX M|
          XXX  |  set6
               |  set7
           XX  |
               |
    0       X  +M
          XXX  |
               |  set4    set5
            X S|
               |
         XXXX  |
               |
               |
   -1          +S
               |
              T|
               |
               |  set3
            X  |
               |
               |  Set2
   -2          +T
         <less>|<freq>
```

Figure 3. Wright map for the speaking span task super items analysis.

Person and item reliability and separation

The Rasch person reliability value decreased to .71, which is natural since there were only 10 items (sets). The separation value also decreased to 1.57 and was influenced by the low number of sets (see Table 9). It is possible, however, that the true person separation and reliability values lie somewhere between those found in the analysis of the super items (separation = 1.57, person reliability = .71) and the higher values found in the dichotomous analysis of the individual items (separation = 2.10, person reliability = .81) as the dichotomous model likely overestimated separation and reliability due to local dependence of items within a set. In contrast to the person reliability and separation indices, item reliability barely decreased and was estimated at .91 (see Table 10), which provides evidence pointing to the replicability of the spread of items if the test were to be administered to a similar group. The Rasch item separation estimate was 3.21, indicating that the instrument separates items into three distinct levels as it was suggested by the analysis of the individual items. Tables 9 and 10 show that the infit MNSQ and outfit MNSQ for persons (infit MNSQ = 0.97 and outfit MSNQ = 0.97) and items (infit MNSQ = 0.94 and outfit MSNQ = 0.97), respectively were close to the expected average of 1.00. Consequently, the data seem to fit the Rasch model relatively well.

Table 9

Summary of the speaking span task (super items) analysis(persons)

	Total Score	Count	Measure	Real *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	21.00	10.00	0.57	0.47	0.97	0.00	0.97	0.10
P. *SD*	4.60	0.00	0.89	0.06	0.46	0.90	1.25	0.60
S. *SD*	4.60	0.00	0.90	0.06	0.47	0.90	1.27	0.70
Max	29.00	10.00	2.06	0.66	2.34	2.00	7.54	2.50
Min	11.00	10.00	-1.61	0.42	0.30	-1.70	0.25	-1.20
REAL *RSME*		0.48	TRUE *SD*	0.75	SEPARATION	1.57	PERSON RELIA.	.71
MODEL *RSME*		0.44	TRUE *SD*	0.77	SEPARATION	1.74	PERSON RELIA.	.75
SE OF PERSON MEAN = 0.16								
PERSON RAW SCORE-TO-MEASURE CORRELATION = 1.00								
CRONBACH ALPHA (*KR-20*) PERSON RAW SCORE “TEST” RELIABILITY = .69								
SEM = 2.54								

Note. P. SD = Population standard deviation; S. SD = Sample standard deviation; Max = maximum value; Min = minimum value; RSME = square-root of the average error variance; SD = Standard deviation; RELIA. = reliability; SE = Standard error; SEM = standard error of the mean.

Table 10

Summary of the speaking span task (super items) analysis(items)

	Total Score	Count	Measure	Real *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	65.30	31.00	0.00	0.28	0.94	0.00	0.97	0.10
P. *SD*	13.50	0.00	1.02	0.11	0.22	0.70	0.57	1.00
S. *SD*	14.40	0.00	1.09	0.12	0.23	0.70	0.60	1.10
Max	84.00	31.00	1.33	0.56	1.24	1.00	2.25	2.40
Min	44.00	31.00	-1.83	0.19	0.51	-1.10	0.07	-1.00
REAL *RSME*		0.30	TRUE *SD*	0.98	SEPARATION	3.21	ITEM RELIA.	.91
MODEL *RSME*		0.30	TRUE *SD*	0.98	SEPARATION	3.29	ITEM RELIA.	.92
SE OF ITEM MEAN = 0.36								

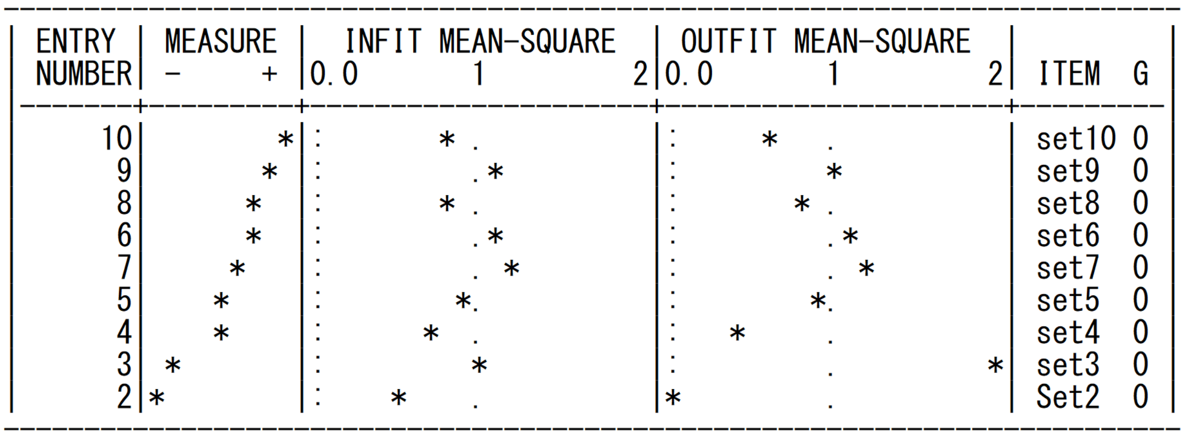
Note. P. SD = Population standard deviation; S. SD = Sample standard deviation; Max = maximum value; Min = minimum value; RSME = square-root of the average error variance; SD = Standard deviation; RELIA. = reliability; SE = Standard error.

PCA of item residuals

A Rasch PCA of item residuals was conducted for all 10 sets. Results showed that 56.50% (eigenvalue = 11.69) of the variance was accounted for by the construct (above the 20.00% criterion). The principal contrast explained 9.50% (eigenvalue = 1.97) of the variance, which suggests that unidimensionality held up. In addition, all the other four contrasts had eigenvalues below 2.00 and accounted for less than 10.00% of the unexplained variance (see Table 11), thus supporting the unidimensionality of the measure. Figure 4 illustrates the unidemsional construct graphically. As can be seen, all super items lie close to the ideal straight line indicative of unidimensionality (Bond & Fox, 2015).

Table 11
Speaking span task (super items) standard residuals in eigen values

	Eigenvalue	Observed	Expected
Total Raw variance in observations	20.69	100.00%	100.00%
Raw variance explained by measures	11.69	56.50%	55.40%
Raw variance explained by persons	5.60	27.10%	26.60%
Raw variance explained by items	6.08	29.40%	28.90%
Raw unexplained variance (total)	9.00	43.50%	44.50%
Unexplained variance in 1st contrast	1.97	9.50%	
Unexplained variance in 2nd contrast	1.68	8.10%	
Unexplained variance in 3rd contrast	1.54	7.50%	
Unexplained variance in 4th contrast	1.05	5.10%	
Unexplained variance in 5th contrast	0.95	4.60%	



```
-------------------------------------------------------------------------
| ENTRY | MEASURE |  INFIT MEAN-SQUARE  | OUTFIT MEAN-SQUARE  |         |
| NUMBER| -     + |0.0       1         2|0.0       1         2| ITEM  G |
|-------+---------+---------------------+---------------------+---------
|     10|        *|:       * .          |:     *   .          | set10 0 |
|      9|       * |:         .*         |:         *          | set9  0 |
|      8|      *  |:       * .          |:       * .          | set8  0 |
|      6|      *  |:         .*         |:         .*         | set6  0 |
|      7|     *   |:         . *        |:         . *        | set7  0 |
|      5|    *    |:        *.          |:        *.          | set5  0 |
|      4|    *    |:      *  .          |:   *     .          | set4  0 |
|      3| *       |:         *          |:         .         *| set3  0 |
|      2|*        |:    *    .          |*         .          | Set2  0 |
-------------------------------------------------------------------------
```

Figure 4. Item fit graph for the speaking span task (super items).

Discussion

One of the goals of this pilot study was to validate a novel speaking span task using Rasch model theory in order to facilitate the development of a more fine-grained task. The research questions concerned whether the results of the Rasch analyses provided validity evidence for the newly designed measure. Two separate Rasch analyses were conducted on the speaking span task: an analysis of the individual items using the Rasch dichotomous model and an analysis of the sets or super items using the Rasch partial credit model, with the latter being conducted to control for the influence of item dependency created by the scoring system on the dichotomous analysis.

The Wright maps of both analyses (Figures 1 and 3) show that the difficulty of the speaking span test seems to line up with the theoretical expectation that the further in a set an item is, the more difficult it should be and, in the same way, the longer a set is, the more difficult it should be. In addition, abstract words may help differentiate participants with higher EWM capacity from those with lower EWM capacity. Figure 1 revealed that the most difficult items were abstract words (i.e., items 10.3, 9.3, and 8.3, which mean *strength*, *trip*, and *wound*, respectively) and that these words are all middle items rather than last items, which are hypothesized to be the most difficult in the sets. This suggests that it was their abstract nature rather than their position in the set which caused them to be difficult to recall.

The results of the Rasch Analysis for the individual items indicated that all of the items were within the model fit parameters (0.50 and 1.50 infit MNSQ values), suggesting that the data fit the model's expectations and that the items adhered to the measurement of a unidimensional construct. However, one item was found to be largely outfitting, which was assumed to have occurred due to it being the longest word on the test. This points to the importance of stricter control over word length in future studies. Taken together, this finding and that of the concreteness of stimuli indicate the importance of lexical characteristics of the stimuli used in the task.

Additionally, the results of the partial credit analyses of the sets or super items revealed an overall pattern of item (set) fitness. An interesting finding was that the first two sets did not contribute to the measurement of EWM as demonstrated by the small infit and outfit values (set 2) and a ceiling effect (set 1). This suggests that future administrations of this test to similar samples should be further shortened by excluding sets of two items. This finding may also generalize to other speaking span tasks when administered to similar populations as in the present study.

The PCA of residuals revealed that the measure accounted for 59.40% of the variance on the dichotomous analyses and 56.50% on the partial credit, which provided further evidence of unidimensionality. The item fit graphs support the numerical evidence of the PCA by showing that the items and sets lie close to the ideal straight line of the unidimensional continuum (Bond & Fox, 2015). However, the principal contrast of the residual error analysis of the dichotomous model found an unwanted construct present in the data as the eigenvalue of the contrast was higher than the criterion of 2.00. A further examination of the content of the items confirmed that the secondary dimension was word length. Interestingly, the super items (sets) analysis demonstrated that when the items were put into sets, the second dimension disappeared. This means that the sets masked the impact of word length on the individual items. However, as noted above, word length should be explicitly controlled for in future versions of the instrument.

The item reliability coefficients provided evidence that similar results would be produced across similar samples. Despite the fact that local dependency among the items may have inflated the item reliability coefficient (.93) of the dichotomous model, this coefficient did not significantly differ from that of the partial credit model (.91) for which the influence of item dependency was partialled out. This means that the ordering of items is likely to be replicated if the speaking span test is given to other similar samples.

Regarding the person measures, the results of this pilot study, as they stand, do not provide convincing evidence to expect person replicability on future administrations of the test. Although the analysis of the individual items revealed moderate person reliability at .81, this figure decreased to .71 in the subsequent partial credit analysis of the super items. As there was local dependency among the items, the dichotomous model of the first analysis may have inflated person reliability, and thus a more realistic value may lie between the reliability coefficient obtained in the dichotomous analysis (.81) of the items and that of the partial credit analysis (.71). In all likelihood, the low reliability was a product of the homogeneity of the sample. This explanation is numerically supported by the low person separation index of the partial credit analysis, which indicates that the sample could not be separated into different levels of the construct. Thus, the speaking span test may prove to have higher reliability if administered to a more heterogeneous sample. Another possible explanation is that there were more items than participants in the dataset, which tends to result in low person reliability estimates (Bond & Fox, 2015).

All in all, these results showed that the instrument measured a unidimensional variable, providing preliminary validity evidence for the speaking span task. Therefore, it is reasonable to claim that the newly developed speaking span task, which links listening and speaking, provides effective measurement of the hypothesized construct. This suggests that carefully designed speaking span tasks that take into account word abstractness and frequency/familiarity, and which are shorter and lack practice sets, could be used as an alternative to longer and less practical tests. Daneman and Green's pioneer task consisted of 70 lexical items (almost twice the number of items contained in the present task) and thus involved an extensive degree of performance, which may risk the purity of the measure as the more one performs the task, the more likely they are to engage in strategic behavior, such as recalling by word association. The present task restricts the opportunities to adopt idiosyncratic strategies since it does not provide trial items until mastery of the task procedure and it contains a smaller number of items. These results also have implications for the way these complex span tasks are scored. The new scoring parameters, giving a credit to the string of items correctly retrieved in order or appearance until memory failure and discarding items retrieved after failure, appear to provide precise EMW spans.

This pilot study is not without limitations. First, the sample size is too small as to give the study enough statistical power to make claims about the generalizability of the findings. Therefore, caution must be taken when interpreting the results. Second, the instrument lacks practicality. It took about ten minutes to collect data from each participant, plus a similar amount of time to analyze and score the data of each participant. Future research should develop speaking span tasks that can be administered to whole groups of participants in one sitting. Follow-up studies should also investigate the influence of the lexical characteristics of stimuli, such as abstractness/concreteness, on the performance of speaking span tasks.

Conclusion

EWM measures such as the speaking span task continue to be widely employed in both cognitive psychology and L2 research. However, despite their popularity, further validation studies are needed.

This pilot study provides preliminary validity evidence that the novel speaking span task, which is shorter than its predecessors, allows for a measurement of EWM that can be more quickly obtained. Having a number of abstract words intermixed with content words of high familiarity seems to increase the power of the measure to differentiate between participants with more and less of the construct. Finally, based on the results, the new scoring system which involved giving credit to each item in the set recalled in order of appearance until memory failure and ignoring the items recalled afterwards, appears to be a precise measure of participants' speaking spans.

The purposes of the study were to obtain some preliminary validity evidence for the use of EWM tasks and to pilot a speaking span task in order to obtain some benchmarks for the development of an improved follow-up measure. The results of this study contribute preliminary evidence towards the establishment of the validity of EWM tasks. In addition, future versions of the speaking span task would benefit from the following modifications: employing words that are even in length, replacing words that are potentially perceived as being related, excluding sets of two items from the measurement, and being administered to a larger more heterogeneous sample. With such modifications, it is hoped that the test presented here can be used by researchers to further improve the measurement of working memory.

Acknowledgements

Many thanks to my advisor, Prof. James Sick, for his methodological guidance and feedback. Thanks also to my classmate and friend, Clint Denison, for his feedback on earlier drafts of this paper and for many interesting discussions regarding the topic of the study. I am also grateful to my friend, Andrew Wright, for his unconditional willingness to provide language assistance and proof-read the manuscript. Finally, I would like to thank the teachers and students who made this research possible.

References

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4,* 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D. (2012). Working memory: Theories, models and controversies. *Annual Review of Psychology, 63,* 1–30. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (ed.) *The Psychology of Learning and Motivation* (pp. 47–89). New York, NY: Academic Press.

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

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

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

Daneman, M. (1991). Working memory as a predictor of verbal fluency. *Journal of Psycholinguistic Research, 20,* 445–464. https://doi.org/10.1007/BF01067637

Daneman, M. & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language, 25,* 1–18. https://doi.org/10.1016/0749-596X(86)90018-5

Fortkamp, M. B. M. (1999). Working memory capacity and aspects of L2 speech production. *Communication and Cognition, 32,* 259–296.

Fortkamp, M. B. M. (2003). Working memory capacity and fluency, accuracy, complexity, and lexical density in L2 speech production. *Fragmentos, 24*, 69–104.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language.* Hove, England: Lawrence Erlbaum Associates Inc.

Guará-Tavares, M. G. (2008). Working memory capacity and L2 speech performance in planned and spontaneous conditions: a correlational analysis. *Trabalhos em Linguística Aplicada (UNICAMP), 52,* 09–29. https://doi.org/10.1590/S0103-18132013000100002

Kahana, M. J., Howard, M. W., Zaromb, F., & Wingfiled, A. (2002). Age dissociates recency and lag effects in free recall. *Journal of Experimental Psychology*, *28*(3), 530–540. https://doi.org/10.1037/0278-7393.28.3.530

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

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

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

Linacre, J. M. (2020). Standard errors: model and real. Retrieved from www.winsteps.com/winman/webpage.htm

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

Martin, K. I., & Ellis, N. C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition, 34*, 379–413. doi:10.1017/S0272263112000125

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

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

Monchi, O., Petrides, M. Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin card sorting revisited: Distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *The Journal of Neuroscience*, *21*(19), 7733–7741. https://doi.org/10.1523/JNEUROSCI.21-19-07733.2001

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

Weissheimer, J., & Mota, M. B. (2009) Individual Differences in Working Memory Capacity and the Development of L2 Speech Production. *Issues in Applied Linguistics*, *17*, 34–52.

Wen, Z. (2016). Phonological and executive working memory in L2 task-based speech planning and performance. *The Language Learning Journal*, *44*(4), 418–435. https://doi.org/10.1080/09571736.2016.227220

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

Appendix A

Speaking Span Task

Set 1			
Item Number	Japanese Word	Romanized Version	English Translation
I1.1	くし	kushi	comb (C)
I1.2	お金	okane	money (A)

Set 2			
Item Number	Japanese Word	Romanized Version	English Translation
I2.1	針	hari	needle (C)
I2.2	掃除機	soujiki	vacuum cleaner (C)

Set 3			
Item Number	Japanese Word	Romanized Version	English Translation
I3.1	ごみ	gomi	garbage (C)
I3.2	郵便局	yuubinkyoku	post office (C)
I3.3	目覚まし	mezamashi	alarm clock (C)

Set 4			
Item Number	Japanese Word	Romanized Version	English Translation
I4.1	白熊	shirokuma	polar bear (C)
I4.2	ランドセル	randoseru	backpack (C)
I4.3	警察	keisatsu	police (C)

Set 5			
Item Number	Japanese Word	Romanized Version	English Translation
I5.1	鏡	kagami	mirror (C)
I5.2	わさび	wasabi	wasabi (C)
I5.3	約束	yakusoku	promise (A)
I5.4	人間	ningen	people (C)

Set 6			
Item Number	Japanese Word	Romanized Version	English Translation
I6.1	天気	tenki	weather (A)
I6.2	薬	kusuri	medicine (C)
I6.3	話	hanashi	talk (A)
I6.4	指輪	yubiwa	ring (C)

Set 7			
Item Number	Japanese Word	Romanized Version	English Translation
I7.1	仕事	shigoto	job (A)
I7.2	寺	tera	shrine (C)
I7.3	運動	undou	exercise (A)

I7.4	危険	kiken	danger (A)
I7.5	宝	takara	treasure (C)

Set 8			
Item Number	Japanese Word	Romanized Version	English Translation
I8.1	草	kusa	weeds (C)
I8.2	靴下	kutsushita	socks (C)
I8.3	傷	kizu	wound (C)
I8.4	切符	kippu	ticket (C)
I8.5	池	ike	lake (C)

Set 9			
Item Number	Japanese Word	Romanized Version	English Translation
I9.1	試験	shiken	exam (A)
I9.2	おもちゃ	omocha	toy (C)
I9.3	旅行	ryokou	trip (A)
I9.4	熱	netsu	fever (A)
I9.5	上着	uwagi	jacket (C)
I9.6	石鹸	sekken	soap (C)

Set 10			
Item Number	Japanese Word	Romanized Version	English Translation
I10.1	病院	byouin	hospital (C)
I10.2	ピアノ	piano	piano (C)
I10.3	力	chikara	strength (A)
I10.4	駅	eki	station (C)
I10.5	ひげ	hige	moustache (C)
I10.6	試合	shiai	match (A)

Note. C = concrete word; A = abstract word.