SHIKEN

Volume 24 • Number 1 • June 2020

Contents

1. A Rasch-Validation Study of a Novel Speaking Span Task
Bartolo Bazan

22. Shiken: Past and future
David Allen

37. Voices in the field: An interview with Yuko Goto Butler
David Allen

Testing and Evaluation SIG Newsletter

ISSN 1881-5537

Shiken

Volume 24 No. 1
June 2020

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.

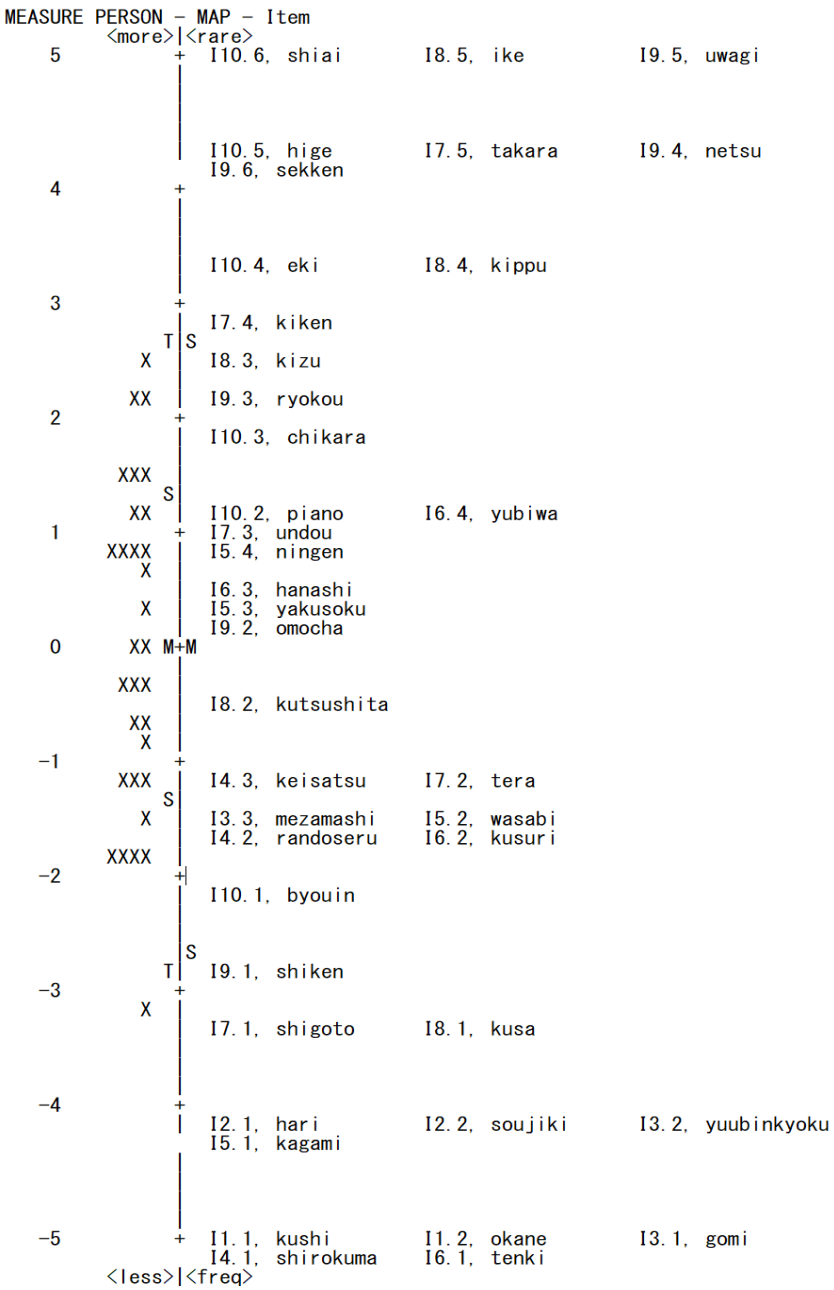

Figure 1. Wright map for the individual items of the speaking span task.

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%	

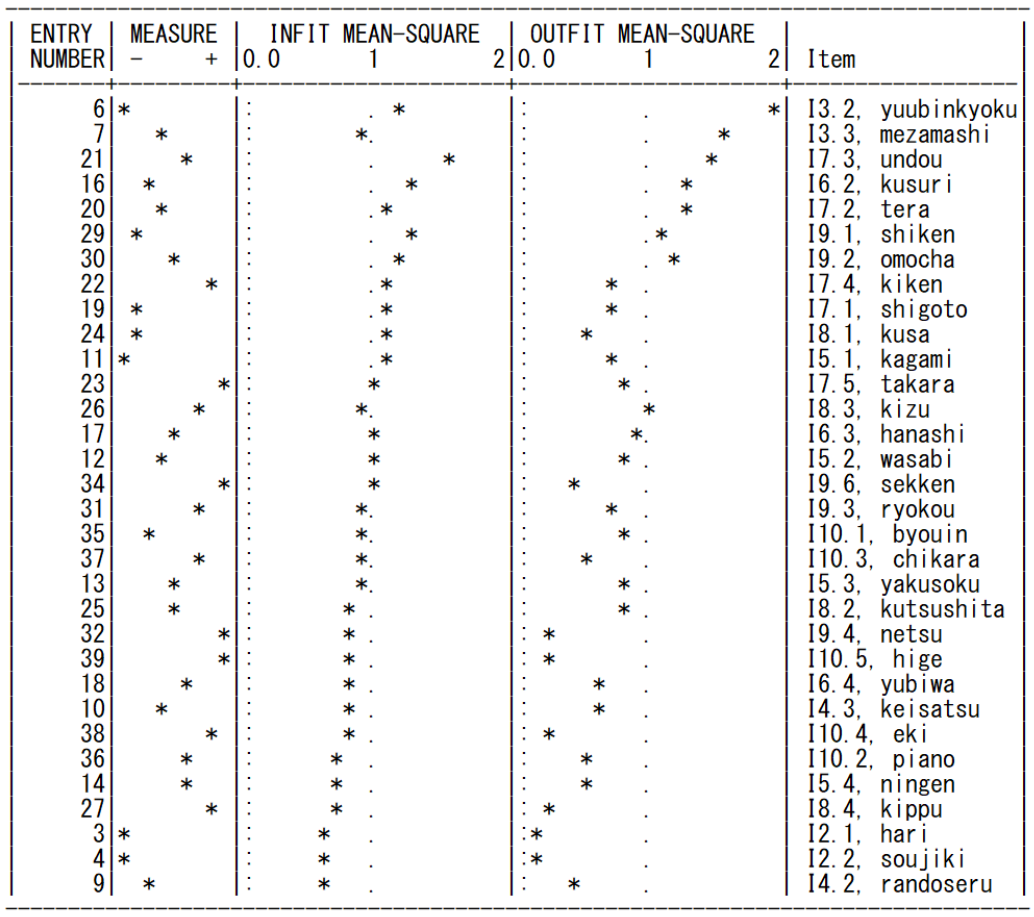

Figure 2. Item fit graph for the speaking span task (individual items).

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.

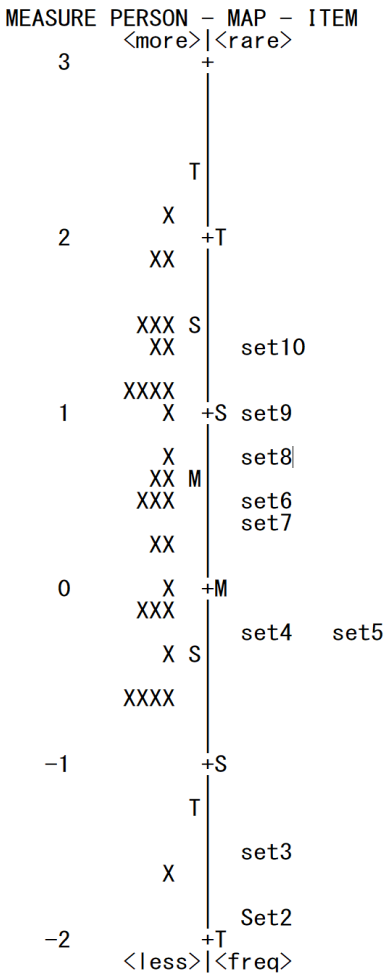

Figure 3. Wright map for the speaking span task super items analysis.

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%	

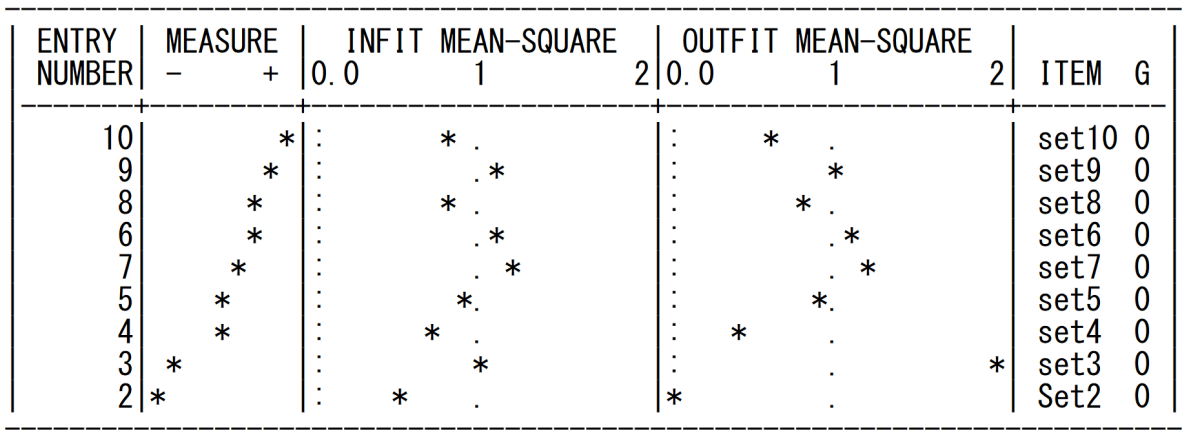

Figure 4. Item fit graph for the speaking span task (super items).

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.

Shiken: Past and future

David Allen
allen.david[at]ocha.ac.jp

Ochanomizu University

Abstract

This article presents a history of *Shiken* since it was first published in 1997 until 2019, followed by suggestions for areas of future research in assessment to which the publication may be well suited to contribute. In the historical overview, data is presented about the following: the origins, titles, editors, and distribution; the article types; the contents of research articles and the design and methodologies they have employed. Regarding research article content, four prominent themes were identified: mass market tests, entrance exams, statistics, and validity/reliability. Regarding design and methods, research articles have tended to focus on English language tests with university students in Japan, while utilizing test and/or instrument data and quantitative methods of analysis. Recommendations for future research areas include investigations into the validity of test interpretations and uses of four-skills, vocabulary and other tests used in Japan, and language assessment literacy. Recommendations for future research design and methods include focusing more on a range of test stakeholders; various contexts, such as pre-tertiary education; and the use of qualitative and mixed methods.

Keywords: *Shiken*, TEVAL, newsletter history, journal history, four-skills tests, entrance exams

In 1996, a group of teachers and academics founded a new Special Interest Group (SIG), Testing & Evaluation (TEVAL), at the Japan Association of Language Teachers (JALT) and began a newsletter that was first published in 1997. This newsletter, *SHIKEN: JALT Testing & Evaluation SIG Newsletter*, is now in its 24th year of publication. Having reached the grand age of almost a quarter of a century, it seems fitting to present its history, to chart the ground covered so that we may look forward to new horizons of yet unexplored territory.

I was inspired to take on this task by reading a recent issue of *Assessing Writing*, in which the editor from 2002 to 2017, Liz Hamp-Lyons, and the current editor David Slomp, comment independently on the 'ideas, questions and concerns' explored in the articles featured in the journal between 2000 and the present day (Hamp-Lyons, 2019; Slomp, 2019). In the same volume, Zheng and Yu (2019) overviewed the trends in the journal and what they tell us about the field of writing assessment. These articles served as a timely reminder to me that journals are historical documents that chart the changes in academic thinking, research interests and methods. Simultaneously, as the incoming editor I needed to better understand the character of *Shiken*, particularly what has been done in the journal since its inception, in terms of the kinds of articles published and the topics covered. Thus inspired, I dug deep into the *Shiken* archives and documented my findings; the sum of which is presented in the following article.

This article is organized into two sections covering the past and the future, respectively. In the first section, I overview the contents, and highlight a number of themes that emerged from a comprehensive survey of all published issues. In terms of methodology, I adopted a straightforward approach to cataloging the digital versions of all issues and then analyzing them, using spreadsheet software. Through reading and rereading the issues, I became aware of various trends and themes, which then became subject to a more focused analysis. In this way, the texts provided the raw data, from which frequencies were tallied (e.g., the number of interviews) and categories and sub-categories were created (e.g., a category was formed for articles dealing with entrance exams, and sub-categories were formed for those that deal with this topic in a discursive manner and those that collected empirical data). I have strived to be accurate in analyzing and presenting the data, and have tried not to overindulge my own inherent biases (i.e., to my own research areas and beliefs about language teaching, learning and assessment). I have also invited and received multiple reviews from experts, some of whom have been involved in the SIG and this publication since its early days. As a result, I believe the outcome presented here is a fair and accurate representation of the publication's history.

In the second section, I look to the future of *Shiken* and in what ways the publication can contribute to language testing and evaluation in the Japan. Based on the foregoing history, and considering current trends in the field, I present a number of areas that are important for future research in the Japanese context. It is therefore hoped that

this research will be of practical value in not only encouraging researchers to conduct much needed assessment research but also to submit it to *Shiken*, thereby contributing to the continued success of the publication.

The Past: *Shiken* between 1997 and 2019

General publication information

Shiken began in 1997 at the initiative of Leo Yoffe, Jeffrey Hubbell and JD Brown, and has continued up to this first issue of 2020, which will be the 24th year of publication. The publication was originally (and formally at least still is) entitled *SHIKEN*: *JALT Testing & Evaluation SIG Newsletter*. In 2012 it was temporarily renamed *Shiken Research Bulletin* but since 2014 the abbreviated title *Shiken* has been commonly used. These name changes coincided with a number of changes to the editorship, which has developed as follows: Paul Jaquith (1997 to 1998), Tim Newfields (1999 to 2011), Aaron Olaf Batty (2012), Jeffrey Stewart (2013), Trevor Holster and J. W. Lake (2014), and Trevor Holster (2015 to 2019). Although the journal was originally distributed only in printed form, all back issues were eventually made available online in HTML and PDF format by Tim Newfields. Initially, TEVAL policy was to mail printed versions to SIG members and to make back issues publicly available online one year after publication. In 2012, the TEVAL officers decided to drop printed distribution altogether and make *Shiken* an open-access, online journal, with all articles immediately available online to TEVAL members and non-members alike.

Regarding output, 49 issues in 23 volumes were published, with two issues every year for 16 years, except for two when there was only one issue (1998, 2014), and five that saw three issues per volume (2000-2003, 2009). A variety of article formats have featured with varying frequency, which is taken up in the next section.

Overview of the contents

According to the titles of the published articles, between 1997 and 2019, there were 75 *Articles*; four *Opinion Pieces*; 27 *Book Reviews*; 27 *Interviews*; 49 *Statistics Corner* articles by JD Brown; eight *Rasch Measurement in Language Education* articles; 11 *Assessment Literacy Quizzes*; one *Test Review*, though test reviews have usually been published as *Articles*; and three *Software Corner* articles. In addition, there have been a number of other sections for communicating news and events. Finally, other than during a short period during 2012 and 2015, issues have typically not featured an editorial or foreword.

From the overview of article types, a number of observations can be made. Firstly, while there have been some regular contents, particularly *Articles* and *Statistics Corner*, other article types have appeared during particular periods of time, such as *Book Reviews* and *Interviews*, which largely coincided with Newfields' editorship, and the two mini-series (i.e., the *Rasch Measurement* articles and *Assessment Literacy Quiz*). Other article types have been notably infrequent (i.e., *Opinion Pieces* and *Software Corner*). In the following, a selection of article types is described in more detail.

A special mention is required for the *Statistics Corner* articles, which have been a regular feature of *Shiken* and which for many have become synonymous with the publication. JD Brown, a founding member of the TEVAL SIG in 1997, contributed one *Statistics Corner* article every issue until his retirement in 2019. Each article responds to a question about statistics in a form that is accessible to readers with minimal experience of statistics and quantitative methods. These articles were later compiled into a book, *Statistics Corner* (2016), which is provided free-of-charge to every new member of the SIG; in other words, even after retirement, JD continues to support the next generation of language teachers and assessment researchers in Japan.

Similarly, a special note is needed for both the *Rasch Measurement in Language Education* series and the *Assessment Literacy Quizzes* contributed by Jim Sick and Tim Newfields, respectively. A series of eight articles on the topic of Rasch analysis was contributed between 2008 and 2013 by the SIG's current coordinator, Jim Sick. These articles provide an accessible introduction to Rasch measurement from an applied perspective. Similarly, Newfields' interest in developing the assessment literacy of teachers and researchers resulted in his

series of quizzes between 2006 and 2011. Questions were raised on topics ranging from quantitative issues in assessment to test administration and then answered in detail with suggestions for further reading. It is noteworthy that these two series, together with the *Statistics Corner* series, indicate that a primary concern of *Shiken* has been to instruct its readers, many of whom are newcomers to the field, on methods for quantitative analysis of test-related data.

Finally, while *Research Articles* and *Book Reviews* tend to make up the bread and butter of most academic periodicals, the *Interview* series illustrates a somewhat novel aspect of *Shiken*, and also other JALT publications, such as in *The Language Teacher*. Between 2001 and 2011 interviews with language assessment specialists working in Japan and elsewhere were featured. These were mainly conducted by Newfields, and anecdotal evidence suggests that they are highly praised by the *Shiken* readership. Perhaps the reason for their popularity lies in their ability to reveal the person behind the research: only in interviews can the reader get a sense of the academic as a person who got degrees, took jobs, did research and made a career in language assessment. For readers who are ambivalent about the attraction of language assessment as a field in which to develop a career, the interviews are undoubtedly one of the most stimulating and enlightening of all the article formats. To date, 24 illustrious individuals have been interviewed, creating an incomplete who's who of contemporary language assessment research, many of them with strong ties to Japan. For the purpose of back-cataloging, but also as a reference for future interviews, here is the comprehensive, chronological list of interviewees: Leo Yoffe, Randy Thrasher, Dan Douglas, Gholamreza Hajipournezhad, Liz Hamp-Lyons, Michihiro Hirai, JD Brown (interviewed twice), Lyle Bachman, Kenji Ohtomo, Robert Gardner, Kazuhiko Saito, Yoshinori Watanabe, Michael Todd Fouts, Barry O'Sullivan, Trevor Bond, Glenn Fulcher, Carsten Roever, David Beglar, Jessica Wu, George Engelhard, Spiros Papageorgiou, Shozo Kuwata (in two languages), Alaistar van Moere (in two parts), and Meg Malone.

Articles: Content

In this section, I provide an overview of the content of *Shiken* articles. The observations presented here were made from a content analysis of all issues of the publication from 1997 to 2019. Overall, a wide range of topics are covered in the journal, though often only once or twice; these latter include vocabulary issues in assessment (Beglar, 2000; MacDonald, 2019; Trace & Janssen, 2014), test-taking strategies (Paton, Howarth & Cameron, 2018; Yoshida, 2006), ongoing assessment (Carbery, 1999; Croker, 1999), needs analysis (Kikuchi, 2005), cognitive diagnostic assessment (Aryadoust, 2011a, 2011b), rating scales (Venema, 2002) and rubric design (Duarte, 2016; Marshall, 2014). In contrast, the following four topics were addressed in multiple studies: 1) mass market tests, 2) entrance exams, 3) statistics, and 4) reliability and validity issues.

Mass market tests

A variety of mass market tests have featured in research in *Shiken* over the years, in the form of a test review, as the subject of research or as a method of assessing test-taker performance. The Test of English for International Communication (TOEIC, including TOEIC Bridge and TOEIC LPI) has been featured more than any other test, that is, eight times as the focus of the research, most recently in Paton et al. (2018), and twice as a comparison test (Hirai, 2002; Kanzaki, 2015). However, it is noteworthy that a number of these articles have been highly critical of the test itself (Chapman, 2003; Chapman & Newfields, 2008) or the organizations that administer it (McCrostie, 2010). McCrostie's (2010) article, *The TOEIC in Japan: A scandal made in heaven*, details in journalistic style the history of the Institute for International Business Communication (IIBC), which oversees test administration in Japan. In his *TOEIC®: Tried but undertested*, Chapman (2003) lamented the lack of research into the test; and Chapman and Newfields' (2008) sardonically titled *The 'New' TOEIC®* comments: 'what's remarkable about the new [2006] version of this test is how much is unaltered' (p. 33).

Only one article has focused on the development of the Test of English as a Foreign Language (TOEFL) (McNamara, 2001). However, more recently two articles have appeared on the TOEFL ITP (Institutional Testing Program), a quasi-official version of the TOEFL based on the Listening, Reading, and Structure sections of the

older, paper-based TOEFL. One article focused on the interpretation of ITP scores in pretest-posttest designs (Koizumi et al., 2015) and another on teacher perceptions of the test (Collins & Miller, 2018). We predict, and hope, that given the increasingly widespread use of TOEFL ITP in Japanese higher education institutions, often for multifarious purposes (e.g., placement, progress monitoring, exit testing), research into its use, or misuse, will appear in future.

Other mass market tests have received limited attention. The International English Language Testing System (IELTS) Academic exam was featured in Aryadoust (2011b) and Boddy (2001). The Business Language Testing Service (BULATS) exam appeared once (Hirai, 2002), while the pre-two level of the EIKEN Test was reviewed by Plumb and Watanabe (2016). The Cambridge Young Learners of English Test was described by MacGregor (2001), and the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) was reviewed by Yoffe (1997).

Entrance exams

Entrance exams in Japan have been researched and discussed specifically in eight articles (i.e., roughly one in ten). Six of these articles provide an overview and/or discussion of the issues (Brown, 2000; Guest, 2009; Murphey, 2000/2003, 2009; Sage, 2007) and three conduct some form of empirical research into the exams (Akiyama, 2003; MacDonald, 2019; Mitchell, 2017). Regarding the discursive articles, JD Brown (2000) discussed strategies for creating positive washback from entrance exams, while in the same year an opinion piece by Murphey (2000) presented *Excerpts from an open letter submitted to the Japanese government concerning education and university entrance exams*, an article which was republished in 2003. A decade later, Murphey (2009) discussed the need to revise the system of exams for university entrance, while Guest (2009) discussed validity and reliability issues of Japanese university entrance exams. Guest's article was introduced as 'controversial' in the foreword and responses were invited, though none materialized; moreover, it was originally intended to be a two-part article, though no follow up article appeared. Finally, Sage's (2007) article critiques MEXT's 2003 Action Plan and the Ministry of Education's approach to assessing communicative competence.

Turning to the three empirical studies, Akiyama (2003) investigated the assessment of students' English speaking ability in junior high schools, the results of which are used for selection in high schools. Mitchell (2017) investigated senior high school teachers' and assistant language teachers' (ALTs') perceptions of language education pressures. Finally, MacDonald (2019) investigated the vocabulary level of the reading passages featured in the national Center Test between 2014 and 2019. In sum, it is clear that the entrance exams are a hot topic in Japan and that washback (i.e., the effect of a test on teaching and/or learning) is central to the issue. However, the majority of articles published in *Shiken* have tended to focus less on collecting primary data through empirical research and more on presenting arguments and opinions, often based on secondary sources of evidence.

Statistical and measurement issues

Eleven articles have taken up statistical issues as the primary focus of study. Six of the articles tackle these in a discussion or tutorial format, such as Smiley's (2015) account of the learning curve he encountered when studying Rasch analysis; Holster and Lake's (2015) account of using Rasch for analyzing classroom test data; and Molloy and Newfields' (2004, 2005a, 2005b) three-part article dealing with statistics in SPSS. These articles further highlight the instructional role of *Shiken*. Other studies tackle measurement issues in the format of empirical research. These include Stewart and Gibson's (2010) use of item response theory to equate pre and post-tests in the classroom; Koizumi et al.'s (2015) investigation into the regression to the mean effect as evident in TOEFL ITP scores; and Stubbe and Stewart's (2012) investigation of scoring formulas for Yes/No tests.

Reliability and validity

Perhaps the largest single, yet admittedly broad, category of articles is those that deal with the validity and reliability issues of tests. Following contemporary inclusive views on the nature of test validity (e.g., Messick, 1989; Kane, 2013; Weir, 2005), studies may be counted in this category if they either present evidence for, or

discuss the nature of, any of the following aspects: construct validity (i.e., theorized dimensions of the construct being assessed), content validity (i.e., test content and format), scoring validity (i.e., reliability), criterion-related validity (i.e., concurrent and predictive validity) or consequential validity (i.e., impact and washback). At least 55 of the 79 articles appear to fall into this category, including the majority of the articles that deal with mass market tests, entrance exams, and statistical and measurement issues mentioned previously. Moreover, many articles discuss multiple aspects of validity.

It may suffice here to provide a few representative examples of validity-related studies: Jia and Zhang (2007) investigated the construct validity of an English language test for PhD applicants; Stewart, Gibson and Fryer (2012) investigated the reliability of the TOEIC Bridge test; Kanzaki (2015) investigated the validity of the Minimal English Test (see Maki, 2018) by conducting item analysis and correlational analysis with TOEIC scores; Hirai's (2002) study dealt with criterion-based validity by comparing performance scores on TOEIC and BULATS tests; and Collins and Miller (2018) investigated teachers' perceptions of TOEFL ITP, which pointed towards possible washback on teaching from the test. While these studies all dealt with a specific test, which is the most common format for articles in this category, others have sought to discuss validity-related phenomenon more generally. For instance, in opinion pieces, Roberts (2000) and Newfields (2002) both discussed the notion of face validity; Cubilo (2014) discussed the applicability of Kane's (2006) argument-based framework to classroom and program contexts; and Pan (2008) critically reviewed five washback studies.

Articles: Research design and methodologies

In this section I summarize the features of research articles in *Shiken* from the perspective of research design and methodologies. Specifically, I focus on the language being researched, macro- and micro-contexts, participants, data collection methods and methods/approaches used in analyzing data.

Every article that has focused on assessment of a particular language, has focused on English. There have been no articles about assessing other languages, such as Japanese as a second language, or French, German or Chinese as foreign languages. This is despite the fact that *Shiken*, similar to JALT and the TEVAL SIG, is not specifically focused on the English language.

In terms of the macro-context in which research was conducted, the following regions and countries have featured. Unsurprisingly, most studies were conducted in Japan (n = 47). However, there were two studies conducted in Asia: one including participants from various Asian countries (Aryadoust, 2011b) and one with a focus on innovative testing in Asia (Murphey, 2009); two studies in Australia: one discussed entrance exams in Australian universities (Gruba & Hill, 1997) and another conducted a study with participants in an Australian copper mine (Marshall, 2014); one study focused on a doctoral entrance exam in China (Jia & Zhang, 2007); one three-part study analyzed complexity ratings in Iran (Hajipournezhad, 2001, 2002a; 2002b); one collected test data in Colombia (Trace & Janssen, 2014); and two described becoming a lecturer of testing in the US (Gorsuch, 2000a, 2000b). The remainder of articles have been largely context-independent (i.e., focused on theoretical or statistical issues).

In terms of specific micro-contexts, the vast majority of studies were conducted in college/university contexts (n=42), while two were conducted in junior high schools (Akiyama, 2003; Duarte, 2016), two in high schools (Koizumi & Yano, 2019; Mitchell, 2017), and two in private companies (Hirai, 2002; Marshall, 2014). These data highlight that most research occurs at the tertiary level despite the fact that assessment is equally if not more important during pre-tertiary education. As with all academic research, this is most likely due to the relative accessibility of university participants to researchers working in tertiary institutions.

Focusing on participants, the vast majority of empirical studies recruited students, of whom the majority were at college/university (n=29), while the minority were either junior high school students (Akiyama, 2003; Duarte, 2016) or adult learners (Hirai, 2002; Marshall, 2014). One study focused on PhD applicants (Jia & Zhang, 2007), while four studies investigated teachers, two of which involved university teachers (Collins & Miller, 2018; Kikuchi, 2005) while the others involved junior high school teachers (Akiyama, 2003, as survey respondents and

test interlocutors) and high school teachers and ALTs (Akiyama, 2003, as raters; Mitchell, 2017). Hajipournezhad (2002b) also reported teachers' ratings, but provided little information about the educational context.

In terms of data collection methods, of the 36 articles utilizing primary data, test performance data and/or instrument data (e.g., self-ratings on the CEFR-J in Runnels, 2013) made up the majority of data sources for analyses (n=30). Survey data featured in six articles (twice without additional data: Kikuchi, 2005; Collins & Miller, 2018; and four times with additional data sources: Akiyama, 2003; Koizumi & Yano, 2019; Mitchell, 2017; Harrison & Vanbaelen, 2013). Semi-structured interviews conducted by email were used once (Mitchell, 2017). There have been no articles in *Shiken* that have conducted and presented data from oral interviews or focus groups.

Of the 36 articles that collected primary data, 33 used quantitative analytic methods, one used qualitative analysis (Marshall, 2014), and two used mixed methods (Koizumi & Yano, 2019; Mitchell, 2017). Of all the quantitative analytic approaches, none is more synonymous with testing and assessment than item-response theory and particularly the Rasch approach. Since 1997, Rasch analysis has been used in 15 articles, beginning with Akiyama (2003) and most recently by Patterson (2019). From the data, over the 27 years of publication, the frequency of articles employing Rasch analytic techniques has increased, with all except Akiyama (2003) appearing within the last eleven years.

Only one article in *Shiken* has involved the collection and analysis of solely qualitative data: Marshall (2014) used student presentation video data to inform his development of a rating rubric for formatively assessing students' speaking performance. Additionally, two studies utilized mixed-methods. In his survey of high school teachers, ALTs' and first-year university students' perceptions of the pressures of English education at high school, Mitchell (2017) collected and analyzed quantitative survey data while following up on responses from the high school teachers in semi-structured email correspondence. In Koizumi and Yano (2019), the authors investigated an assessment of English oral presentations at a senior high school. The authors analyzed score data through quantitative (many-facet Rasch) analyses and student perceptions of the test through a survey which included Likert scale and open-ended items. Together these two studies employed different combinations of participants (i.e., high school teachers, ALTs and students), data formats (i.e., survey and email data; or test and survey data) and analyses (i.e., quantitative and qualitative, though primarily the former).

Final notes

Although space does not permit a thorough definition and analysis of the term 'quality' and how it pertains to the articles published in *Shiken*, it is certainly fair to say there has been a change in the kinds of articles published. Earlier editions had room for comic pieces, satires, and parodies, such as *Research parody: The Templin 1/2k* by Stephen Templin and Audie O'Lingual (2001). There was also the journalistic reporting of McCrostie (2010), who published a related article in the Japan Times (August 2009) entitled, *TOEIC: Where does the money go?* In contrast, more recent issues have typically only featured academic research, which has, as noted earlier, tended to be increasingly focused on quantitative analyses of test data.

The future: *Shiken* in 2020 and beyond

Based on the hitherto documented history, content and orientation of *Shiken* from 1997 to 2019, it is now possible to point to the future. In this section, I will make suggestions for the future scope and orientation of *Shiken*.

Overview

Shiken has always aimed to be a publication in which both expert and novice researchers may publish their work, and which focuses primarily, though not exclusively, on the Japan context. This aim entails a number of characteristics that distinguish *Shiken* from major journals in language testing. Particularly, *Shiken* accepts pilot studies and exploratory studies that seek to provide a basis for future research, in addition to completed research projects. Moreover, *Shiken* welcomes articles that are highly context specific and therefore not necessarily

generalizable outside of Japan, or indeed even across other micro-contexts within Japan. While such articles may often be less appropriate for international journals, they constitute an important source of evidence on test use in Japan. These locally-focused articles can inform practitioners and test developers in Japan and contribute directly to the debates that are occurring in this context. However, because *Shiken* is freely available online and articles can be promoted using academic-networking sites such as researchgate.net and academi.edu, the research published therein can also attain global reach and significance.

The primary article format in *Shiken* is the peer-reviewed research article, for which some recommendations are detailed below. In addition, *Shiken* continues to consider interviews with assessment researchers and practitioners, as well as mini-series. (Readers wishing to submit interviews or a mini-series of articles are encouraged to contact the editor prior to submission).

Research areas for future studies

In this section I will highlight a number of areas for research that should be addressed in future issues of *Shiken*. These research areas are not new; in fact, they represent largely a continuation of the main themes identified previously, that is, mass market tests, entrance exams and validity studies. Perhaps most crucially, the present situation of English language education and assessment in Japan demands research into a particular combination of these subjects; that is, the use of four skills English tests for college/university admissions and the validity of the interpretations and uses of these test scores in the Japanese context. In addition, other areas include research into vocabulary tests and assessment literacy. Importantly, these suggestions are necessarily selective and are, ultimately, only suggestions. Research is required in many areas of language assessment and evaluation and *Shiken* will continue to publish a wide range of topics in accordance with the journal guidelines.

Validation of four-skills mass market tests for use as entrance exams

The main area for future research concerns the use of a selection of four skills tests for university admissions as proposed by the Ministry of Education, Culture, Sports, Science and Technology (MEXT, 2016). These include the Cambridge Assessment exams (e.g., B1 Preliminary), EIKEN exams, GTEC exams, IELTS Academic, TEAP (including TEAP CBT), and TOEFL IBT (see http://4skills.jp/index.html for up-to-date information). The purpose of this innovation is to stimulate positive washback on the learning and teaching of productive skills in pre-tertiary English education and to improve the articulation between pre-tertiary and tertiary English education. In other words, assessing all four skills equally on high-stakes entrance exams, rather than primarily reading as has been the case with the National Center Test and the various university entrance exams (i.e., *nijishiken*; Brown & Yamashita, 1996; Kikuchi, 2006), is intended to generate positive impact on English education.

Naturally, the scale of potential impact of the proposed reform has resulted in considerable discussion among both experts and non-experts alike. A noteworthy example of expert commentary is the open letter sent by the Japan Language Testing Association (JLTA, 2017) to MEXT, which praises the general aim of the innovation (i.e., positive impact) but also highlights significant shortcomings that need to be addressed if the intended impact is to be realized. For instance, much of the criticism of the proposal has focused on the significant issues of accessibility of test centers for students living in remote areas and the financial cost of taking the tests. Recently, due to harsh public criticism of the proposal, MEXT has delayed the initial implementation stage (MEXT, 2019).

The issue of using these mass market tests for Japanese university entrance purposes in the Japanese context is essentially one of validity. In other words, are the proposed interpretations and uses of test scores valid in the context of Japanese university admissions, and are the expected consequences of introducing these tests beneficial and acceptable? Although MEXT has recommended the tests, it has yet to be demonstrated that they are suitable for the purpose of Japanese university entrance. For instance, while the interpretation and use of test scores from 'gold standard' (Weir, 2020) international tests, such as IELTS, TOEFL and the Cambridge Assessment exams, may be valid in the context for which they were designed, such an argument does not automatically transfer to other contexts of use; in other words, the one-size-fits-all argument is not supportable (O'Sullivan, 2020; Weir, 2020). Likewise, even for locally developed tests, such as EIKEN, and for 'glocal' tests, such as TEAP, which

have been developed locally with international expertise (Weir, 2020), evidence is required that they are fit for purpose.

To determine if a specific test is appropriate for use in a specific university admission process in Japan, evidence must be gathered that either supports or rejects the use of the test scores for that purpose in that context. Such evidence is crucial because, firstly, the recommended tests differ greatly in terms of, *inter alia*, their intended purpose, the specific skills and knowledge required, and the degree of cultural appropriateness of their content; and secondly, the contexts of use will also differ in terms of the test taker characteristics (e.g., average proficiency level) and the language needs of the university program to which they will enter. Therefore, a single broad validity argument cannot simply be made for all tests in all tertiary admissions contexts; the arguments must be test and context specific.

Although the testing agencies may be assumed to be at least partly responsible for conducting validity research, the language education and assessment community in Japan should also shoulder some of this burden. Following O'Sullivan's (2020) view on localization, local stakeholders must be involved in the process of validation if the interpretations and uses of tests are to be justified in a specific local context. Local stakeholders are most accessible to local researchers and thus there are countless opportunities for contributions by language assessment researchers in Japan. Key areas for research include:

- **Domain/Needs analyses**: It is necessary to identify the language knowledge, skills, and abilities valued in the language use situations and the relevant tasks to elicit them (Im et al., 2019). Relevant questions to ask here may include: What are the English needs in universities in Japan? How relevant and appropriate are the tasks used in the various tests to these contexts? Research has begun in this area by investigating university teachers' and students' perspectives of English language needs (e.g., Sawaki, 2016, and Tahara, 2018, respectively). However, more research is required to gain a fuller understanding of needs in various micro-contexts in Japan.
- **Curriculum alignment**: It is widely understood that assessment must be viewed as one element in a learning system, which also involves the curriculum and the delivery of that curriculum. This system should be guided by a unified approach to language learning, teaching and assessment. Considering the Japanese national educational system, the function of the to-be-replaced National Center Test was to provide an indication of achievement at the end of the national course of study. The new four-skills tests must thus also provide such an indication, which requires them to be sufficiently aligned in terms of content and level with this curriculum. Consequently, research is needed to demonstrate the extent of this alignment. The only external investigation to my knowledge is Shiratori (2018), who considered the validity of using Cambridge B1 Preliminary in admissions at Hokusei Gakuen University Junior College.
- **Uses of test scores:** Research should also investigate test developers' intended and test users' actual meanings and uses of test scores, by juxtaposing them (Im et al., 2019). In this case, questions may include: How are tests being used in various contexts in Japan? How do stakeholders interpret the scores? How are the scores being used to make decisions, and do these uses conflict with the developers' intended uses? These questions may be pursued both prior to and following implementation of the tests in specific contexts.
- **Consequences**: The primary purpose of the introduction of the four skills tests is to generate positive impact on English education in Japan. Clearly, then, research must thus be conducted into the existence, intensity, direction (i.e., positive or negative) and nature of impact. Although this research will primarily be conducted following test implementation, it can also be undertaken based on previous studies and an analysis of a context and its stakeholders in order to predict impact. As a rule, it should be assumed that both intended and unintended consequences of test use will be predicted and observed; impact research is thus essential because it may provide opportunities to intervene and promote intended positive consequences (e.g., though teacher training and assessment literacy initiatives).

Approaches to test validation

In addressing these issues, researchers will need to adopt an approach to investigating validity. The argument-based approach based on the work of Toulmin (1958) is perhaps the most widely cited approach to test validation. Variations of this approach can be seen in the work of Bachman and Palmer (2010), Chapelle, Enright and Jamieson (2008), Kane (2006, 2013), and Kunnan (2018). Bachman and Damböck (2017) provide an accessible introduction that is aimed at classroom teachers.

An alternative framework is the socio-cognitive approach, which was initially developed in the U.K. (Chalhoub-Deville & O'Sullivan, 2020; O'Sullivan, 2011, 2020; Weir, 2005) and underlies many well-established tests, including IELTS, the Cambridge Assessment exams, and TEAP, all of which are recommended four skills tests for the Japanese university admissions context. The appealing features of the socio-cognitive model include its focus on the test taker and the context, which have proved crucial in the development and validation of tests in various contexts, such as TEAP in Japan (see Dunlea et al., 2020) and the General English Proficiency Test in Taiwan (see Wu et al., 2020), as well as Aptis (see O'Sullivan, 2012) which is the first 'localizable' test in that it can be modified according to specific needs of the context.

Validation of additional test uses in Japan

In addition to the aforementioned mass-market tests for entrance purposes, some of the above research questions can be applied to other test use contexts; for example, to other tests specifically developed for university entrance, such as the newly developed British Council - Tokyo University of Foreign Studies (TUFS) Speaking Test for Japanese Universities (BCT-S; see https://www.britishcouncil.jp/exam/bct-s/about); or to tests, such as TOEFL ITP and Pearson Versant (see https://www.pearson.com/english/versant.html), which are being used for placement, progress monitoring, and/or exit testing purposes.

Another type of test that requires validation research is the vocabulary knowledge test. In recent years, a large number of vocabulary tests have been developed, including the New Vocabulary Levels Test (McLean & Kramer, 2015), described in *Shiken*. Such vocabulary tests are typically designed in accordance with one of an ever-increasing number of frequency-based wordlists. It has been argued, however, that the development and validation of many vocabulary tests has been lacking rigor and systematicity (Schmitt, Nation & Kremmel, 2019). Moreover, the approach to developing vocabulary tests that are based (exclusively) on word frequency may also be questioned. Because many teachers and researchers in Japan utilize vocabulary tests for various purposes, it is likely that this is an area to which *Shiken* can contribute.

Assessment literacy

Another key area for research is understanding and promoting the assessment literacy of teachers and other stakeholders. An interesting example of this is Berry, Sheehan and Munro (2019), who investigated the assessment literacy of teachers in Europe. Through interviews with teachers, the authors illustrated how teachers treated assessment and testing as different concepts; the former was often characterized as part of good practice in teaching, while the latter referred to formal testing, with which the teachers lacked confidence and deferred to exam boards. It would be interesting, in fact, it is essential, to investigate how teachers in Japan perceive exams (i.e., school exams, university entrance exams, and four-skills tests) in relation to the national course of study and their own teaching practices. *Shiken* has a strong history of promoting assessment literacy among its readership and thus research of this kind would be warmly received.

Research design and methodology

Turning to the nuts and bolts of assessment research, three key aspects of research design and methodology can be highlighted.

Firstly, future assessment research is needed in a greater diversity of micro-contexts, particularly at pre-tertiary levels of formal education (i.e., elementary school, junior and senior high school), but also in the private sector,

particularly within the so-called shadow education system, for example, in cram schools (*juku*), prep schools (*yobikou*) and conversation schools (*eikaiwa*). Small-scale studies will likely be highly context-dependent as they are situated in specific schools or other micro-contexts, where idiosyncratic factors of participants will play a crucial role. While this is acceptable, follow-up studies in related contexts will be necessary for comparison. Larger-scale studies will require principled sampling from a number of related contexts and will likely have much greater generalizability.

Secondly, future research needs to involve a broader range of stakeholders. Regarding the most important stakeholder, the test-taker, not only university students but also test takers at various ages, especially young learners, should be the focus of future research. Moreover, although *Shiken* articles have tended, like most language assessment research, to focus on test takers and their scores, other stakeholders are also implicated in test use and therefore their perceptions and behaviors should be the subject of future research. These stakeholders include parents and guardians, employers, teachers, school principals, school administrators, school boards, examination boards, test administrators, education boards, policy makers, test developers and lawyers (Chalhoub-Deville & O'Sullivan, 2020).

Thirdly, future research should use various data collection methods and analytic approaches. The present review has shown that researchers have rarely utilized data collection methods that result in data suitable for qualitative analysis (i.e., interviews or open-ended survey items). For many areas of assessment research, such as exploring the nature of test use and consequence in society, qualitative data is fundamental. Similarly, in addition to quantitative analyses, research is needed that utilizes qualitative and mixed methods approaches.

Conclusions

This review has highlighted the scope and trends of research in *Shiken* between 1997 and 2019. It has also indicated a number of potentially fruitful avenues for future assessment research, though it should be emphasized again that these are simply suggestions rather than delimiters for research.

As one of the few specialized publications that is dedicated to assessment research in Japan, *Shiken* has played an important role in disseminating assessment research over the last almost quarter of a century. In 2020, the need for research into testing and evaluation in Japan is as imperative as it has ever been. In this context, we hope that *Shiken* will continue to serve its purpose as a vehicle for research into test validity; we hope it will continue to contribute to the important debates in language testing; and we hope through research that it will help to promote positive consequences of test use for the millions of test stakeholders in Japan.

Acknowledgements

I am grateful to the numerous reviewers who provided feedback on this paper.

References

Akiyama, T. (2003). Assessing speaking in Japanese junior high schools: Issues for the senior high school entrance examinations. *Shiken: JALT Testing and Evaluation SIG Newsletter, 7*(2), 2-11.

Aryadoust, V. (2011a). Cognitive diagnostic assessment as an alternative measurement model. *Shiken: JALT Testing and Evaluation SIG Newsletter, 15*(1), 2-6.

Aryadoust, V. (2011b). Application of the fusion model to while-listening performance tests. *Shiken: JALT Testing and Evaluation SIG Newsletter, 15*(2), 2-9.

Bachman, L. F., & Palmer, A. S. (2010). Language assessment in practice. Oxford: Oxford University Press.

Bachman, L. F., & Damböck, B. (2017). *Language assessment for classroom teachers.* Oxford: Oxford University Press.

Beglar, D. (2000). Estimating vocabulary size. *Shiken: JALT Testing and Evaluation SIG Newsletter, 4*(1), 2-5.

Berry, V., Sheehan, S., & Munro, S. (2019). What does language assessment mean to teachers? *ELT Journal*, *73*(2), 113-123.

Boddy, N. (2001). The revision of the IELTS speaking test. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *5*(2), 2-5.

Brown, J. D. (2000). University entrance examinations: Strategies for creating positive washback on English language teaching in Japan. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *3*(2), 2-7.

Brown, J. D. (2016). *Statistics corner: Questions and answers about language testing statistics.* Tokyo: JALT Testing and Evaluation Special Interest Group.

Brown, J. D., & Yamashita, S. O. (1995). English language entrance examinations at Japanese universities: What do we know about them? *JALT Journal, 17*(1), 7–30.

Carbery, S. (1999). Practicalities of ongoing assessment. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *3*(1), 2-9.

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical Development and Integrated Arguments*. British Council Monograph Series. London & Sheffield: British Council & Equinox Publishing.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language.* New York: Routledge.

Chapman, M. (2003). TOEIC®: Tried but undertested. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *7*(3), 2-7.

Chapman, M., & Newfields, T. (2008). The 'New' TOEIC®. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *12*(2), *3*2-37.

Collins, J. B., & Miller, N. H. (2018). The TOEFL (ITP): A survey of teacher perceptions. *Shiken: JALT Testing and Evaluation SIG Newsletter 22*(2), 1-13.

Croker, R. (1999). Fundamentals of ongoing assessment. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *3*(1), 10-16.

Cubilo, J. (2014). Argument-based validity in classroom and program contexts: Applications and considerations. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *18*(1), 18-24.

Duarte, A. (2016). An alternative to the traditional interview test: The observed pair interview. *Shiken: JALT Testing and Evaluation SIG Newsletter 20*(2), 44-49.

Dunlea, J., Fouts, T., Joyce, D., & Nakamura, K. (2020). EIKEN and TEAP: How two test systems in Japan have responded to different local needs in the same context, in L.W. Su, C. Weir, & J. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. 131-161). New York: Routledge.

Green, A. (2019). Restoring perspectives on the IELTS test. *ELT Journal, 73(2),* 207-215.

Gorsuch, G. J. (2000a). On becoming a testing teacher: Preliminary notes (Part 1). *Shiken*: *JALT Testing and Evaluation SIG Newsletter*, *3*(2), 8-17.

Gorsuch, G. J. (2000b). On becoming a testing teacher: Preliminary notes (Part 2). *Shiken*: *JALT Testing and Evaluation SIG Newsletter*, *4*(1), 9-21.

Green, A. (2007). *IELTS Washback in Context: Preparation for academic writing in higher education (Studies in Language Testing 25)*. Cambridge: Cambridge University Press.

Gruba, P., & Hill, K. (1997). An overview of Australian university entry requirements for international students. *Shiken: JALT Testing and Evaluation SIG Newsletter, 1*(2), 14-17.

Guest, M. (2008). Some new proposals and responses in ascertaining the reliability and validity of Japanese university entrance exams (part 1). *Shiken: JALT Testing and Evaluation SIG Newsletter, 12*(1), 7-13.

Hajipournezhad, G. (2001). Reading complexity judgments - Episode 1. *Shiken: JALT Testing and Evaluation SIG Newsletter, 5*(3), 2-6.

Hajipournezhad, G. (2002a). Reading complexity judgments - Episode 2. *Shiken: JALT Testing and Evaluation SIG Newsletter, 6*(1), 8-14.

Hajipournezhad, G. (2002b). Reading complexity judgments - Episode 3. *Shiken: JALT Testing and Evaluation SIG Newsletter, 6*(2), 6-9.

Hamp-Lyons, L. (2019). Reflecting on the past, embracing the future. *Assessing Writing*, 42, 100423.

Harrison, J. J., & Vanbaelen, R. (2013). Brown's approach to language curricula applied to English communication courses. *Shiken: JALT Testing and Evaluation SIG Newsletter 17*(2), 2-12.

Hirai, M. (2002). Correlations between active skill and passive skill test scores. *Shiken: JALT Testing and Evaluation SIG Newsletter, 6*(3), 2-8.

Holster, T., & Lake, J. W. (2015). From raw scores to Rasch in the classroom. *Shiken: JALT Testing and Evaluation SIG Newsletter, 19*(1), 32-41.

Im, G-H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia, 9*(14).

Jia, Y., & Zhang, W. (2007). Evaluating the construct validity of an EFL test for PhD candidates: A quantitative analysis of two versions. *Shiken: JALT Testing and Evaluation SIG Newsletter, 11*(1), 2-16.

Japan Language Testing Association (JLTA). (2017). Proposal for handling English testing within the 'Prospective University Entrance Scholastic Abilities Evaluation Test [provisional name]'. Retrieved February 14, 2020, from http://jlta2016.sakura.ne.jp/wp-content/uploads/2017/04/JLTA_proposal2017E.pdf

Kane, M. (2006) Validation. In R. Brennan (Ed.), *Educational Measurement*, 4th ed. (pp. 17-64), Westport, CT: American Council on Education and Praeger.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kanzaki, M. (2015). Minimal English Test: Item analysis and comparison with TOEIC scores. *Shiken: JALT Testing and Evaluation SIG Newsletter, 19*(2), 12-23.

Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities After a decade. *JALT Journal, 27*(1), 77–96.

Koizumi, R., In'nami, Y., Azuma, J., Asano, K., Agawa, T., & Eberl, D. (2015). Assessing L2 proficiency growth: Considering regression to the mean and the standard error of difference. *Shiken: JALT Testing and Evaluation SIG Newsletter, 19*(1), 3-15.

Koizumi, R., & Yano, K. (2019). Assessing students' English presentation skills using a textbook- based task and rubric at a Japanese senior high school. *Shiken: JALT Testing and Evaluation SIG Newsletter 23*(1), 1-33.

Kunnan, A. J. (2018). *Evaluating Language Assessments*. New York: Routledge.

MacDonald, E. (2019). An analysis of vocabulary level in reading passages of the National Center Test. *Shiken: JALT Testing and Evaluation SIG Newsletter, 32*(2), 19-27.

MacGregor, L. (2001). Testing young learners with CYLE: The new kids on the block. *Shiken: JALT Testing and Evaluation SIG Newsletter, 5*(1), *4-7.*

Maki, H. (2018). *The Minimal English Test* kenkyuu (saisho eigo tesuto): Tokyo: Kaitakusha.

Marshall, P.A. (2014). Diagnosing students' proficiency on a spoken performance assessment. *Shiken: JALT Testing and Evaluation SIG Newsletter 18*(1), 10-17.

McCrostie, J. (August, 2009). TOEIC: where does all the money go? *Japan Times*. Available at: https://www.japantimes.co.jp/community/2009/08/18/issues/toeic-where-does-the-money-go/ Last accessed 26/01/2020.

McCrostie, J. (2010). The TOEIC® in Japan: A scandal made in heaven. *Shiken: JALT Testing and Evaluation SIG Newsletter, 14*(1), 2-10.

McLean, S., & Kramer, B. (2015). The creation of a New Vocabulary Levels Test. *Shiken: JALT Testing and Evaluation SIG Newsletter, 19*(2), 1-11.

McNamara, T. (2001). The challenge of speaking: Research on the testing of speaking for the new TOEFL. *Shiken: JALT Testing and Evaluation SIG Newsletter, 5*(1), 2-3.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education & Macmillan.

MEXT (2016). *Koudai setsuzoku shisutemu kaikakukaigi: Saishu houkoku* [The final announcement of reports on discussions regarding the improvement of the upper secondary school-university articulation]. Retrieved February 14, 2020, from https://www.mext.go.jp/b_menu/shingi/chousa/shougai/033/toushin/1369233.htm

MEXT (2019). *Reiwa sannendo daigakunyugakusha sentaku ni kakaru daigaku nyushi eigo seiseki teikyo shisutemu unei taikou no haishi ni tsuite (tsuchi)* [Regarding the abolition of the operating rules for the system for providing English scores for university entrance examinations related to the selection of university enrollees in 2021]. Retrieved February 14, 2020, from https://www.mext.go.jp/a_menu/koutou/koudai/detail/1397731.htm

Mitchell, C. (2017). Language education pressures in Japanese high schools. *Shiken: JALT Testing and Evaluation SIG Newsletter, 21*(1), 1-11.

Molloy, H. P. L., & Newfields, T. (2004). Some preliminary thoughts on statistics and background information on SPSS (Part 1). *Shiken: JALT Testing and Evaluation SIG Newsletter, 8*(2), 2-5.

Molloy, H. P. L., & Newfields, T. (2005a). Some preliminary thoughts on statistics and background information on SPSS (Part 2). *Shiken: JALT Testing and Evaluation SIG Newsletter, 9*(1), 2-5.

Molloy, H. P. L., & Newfields, T. (2005b). Some preliminary thoughts on statistics and background information on SPSS (Part 3). *Shiken: JALT Testing and Evaluation SIG Newsletter, 9*(1), 2-7.

Molloy, H. P. L. (2009). Testing the test: Using Rasch person scores. *Shiken: JALT Testing and Evaluation SIG Newsletter, 13*(3), 6-12.

Murphey, T. (2000). Excerpts from an open letter to the Japanese government concerning education and university entrance exams. *Shiken: JALT Testing and Evaluation SIG Newsletter, 4*(1), 5-8.

Murphey, T. (2003). Excerpts from an open letter to the Japanese government concerning education and university entrance exams. *Shiken: JALT Testing and Evaluation SIG Newsletter, 4*(1), 5-7.

Murphey, T. (2009). Innovative school-based oral testing in Asia. *Shiken: JALT Testing and Evaluation SIG Newsletter 13*(1), 14-20.

Newfields, T. (2002). Challenging the notion of face validity. *Shiken: JALT Testing and Evaluation SIG Newsletter 6*(3), 14.

O'Sullivan, B. (2011). Language Testing, in J. Simpson (Ed.) *The Routledge Handbook of Applied Linguistics* (pp. 259-273). New York: Routledge.

O'Sullivan, B. (2012). *Aptis test development approach (ATR-1)*. Retrieved from British Council website: https://www. britishcouncil.org/sites/default/files/aptis-test-dev-approach-report.pdf

O'Sullivan, B. (2020). Localization [Foreword], in L. W. Su, C. Weir, & J. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. xiii-xxviii). New York: Routledge.

Pan, Y-C. (2008). A critical review of five language washback studies from 1995-2007: Methodological considerations. *Shiken: JALT Testing and Evaluation SIG Newsletter, 12*(2), 2-16.

Paton, S. M., Howarth, M.W., & Cameron, A. (2018). Test-taking strategy instruction for Part 3 of the TOEIC Bridge. *Shiken: JALT Testing and Evaluation SIG Newsletter, 22*(1), 1-6.

Plumb, C., & Watanabe, D. (2016). A critique of the Grade 2 EIKEN test reading section: Analysis and suggestions. *Shiken: JALT Testing and Evaluation SIG Newsletter, 20*(1), 12-17.

Roberts, D. M. (2000). Face Validity: Is there a place for this in measurement? *Shiken: JALT Testing and Evaluation SIG Newsletter, 4*(2), 6-7.

Sage, K. (2007). MEXT's 2003 action plan: Does it encourage performance assessment? *Shiken: JALT Testing and Evaluation SIG Newsletter, 11*(2), 2-5.

Sawaki, Y. (2017). University faculty members' perspectives on English language demands in content courses and a reform of university entrance examinations in Japan: A needs analysis. *Language Testing in Asia, 7*(13).

Schmitt, N., Nation, P., & Kremmel, B. (2019). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching,* 1-12. https://doi.org/10.1017/S0261444819000326

Shiratori, K. (2019). Supporting English education reform in Japan: The role of B1 Preliminary. *Cambridge Assessment English - Research Notes: 73.* Cambridge: UCLES.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*(4), 373-391.

Slomp, D. H. (2019). Complexity, consequence, and frames: A quarter century of research in Assessing Writing. *Assessing Writing*, 42, 100424.

Smiley, J. (2015). Classical test theory or Rasch: A personal account from a novice user. *Shiken: JALT Testing and Evaluation SIG Newsletter, 19*(1), 16-31.

Stewart, J., & Gibson, A. (2010). Equating classroom pre- and post-tests under item response theory. *Shiken: JALT Testing and Evaluation SIG Newsletter, 14*(2), 11-18.

Stewart, J., Gibson, A., & Fryer, L. (2012). Examining the reliability of a TOEIC Bridge practice test under 1- and 3-parameter item response models. *Shiken: JALT Testing and Evaluation SIG Newsletter, 16*(2), 8-14.

Stubbe, R., & Stewart, J. (2012). Optimizing scoring formulas for yes/no vocabulary tests with linear models. *Shiken: JALT Testing and Evaluation SIG Newsletter, 16*(2), 2-7.

Tahara, T. (2018). Japanese university students' perspectives on English language needs in secondary school and university education. *Bulletin of the Graduate School of Education of Waseda University, 26*(1), 153-169.

Templin, S. A., & O'Lingual, A. (1998). Research parody: The Templin 1/2k. *Shiken: JALT Testing and Evaluation SIG Newsletter, 5*(1), 8-9.

Toulmin, S. (1958). The uses of argument. Cambridge: Cambridge University Press

Trace, J. W., & Janssen, G. (2014). Corpus-informed test development: Making it about more than word frequency. *Shiken: JALT Testing and Evaluation SIG Newsletter, 18*(1), 3-9.

Venema, J. (2002). Developing classroom specific rating scales: Clarifying teacher assessment of oral communicative competence. *Shiken: JALT Testing and Evaluation SIG Newsletter, 6*(1), 2-6.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13*(3), 318-333.

Weir, C. J. (2005). *Language test validation: An evidence-based approach*. Oxford: Palgrave.

Weir, C. J. (2020). Global, local or "glocal": Alternative pathways in English language test provision, in L. W. Su, C. Weir, & J. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. 193-225). New York: Routledge.

Wu, R. Y-F. (2020). The General English Proficiency Test in Taiwan: Past, present and future, in L. W. Su, C. Weir, & J. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. 9-41). New York: Routledge.

Yoffe, L. (1997). An overview of the ACTFL proficiency interview: A test of speaking ability? *Shiken: JALT Testing and Evaluation SIG Newsletter, 1*(2), 2-13.

Yoshida, K. (2006). Theoretical frameworks of testing in SLA: Processing perspectives and strategies in testing situations. *Shiken: JALT Testing and Evaluation SIG Newsletter 10*(1), 1-6.

Zheng, Y., & Yu, S. (2019). What has been assessed in writing and how? Empirical evidence from *Assessing Writing* (2000–2018). *Assessing Writing, 42,* 100421.

Voices in the field: An interview with Yuko Goto Butler

By David Allen
Ochanomizu University

Bio

Yuko Goto Butler is Professor of Educational Linguistics at the Graduate School of Education at the University of Pennsylvania. She is also the director of the Teaching English to Speakers of Other Languages (TESOL) program at Penn. She received her Ph.D. in Educational Psychology from Stanford University. Her research interests are primarily focused on the improvement of second/foreign language education among young learners in the U.S. and Asia as well as assessment methods for them in response to the diverse needs of an increasingly globalizing world.

Keywords: Self-assessment, formative assessment, young learners

Interview

Can you describe the different areas of work you've done in language assessment?

I'm looking at assessment for young learners (YLs). The definition of YLs is a little bit tricky, but I tend to work with children from two to 12 years old (up to the end of primary school) who are learning an additional language or languages in an instructional setting. I'm not good at handling teenagers so I'm going to stop there! When it comes to assessment, I'm primarily interested in formative assessment. I'm looking at self-assessment, task-based assessment and speaking assessment. I'm interested in second language learners of English as well as learners of other languages. Most recently I started working with YLs of Japanese as a second language.

Can you tell us how you got into the field of language assessment?

I've always been interested in students' language development; my master's thesis was on first language and bilingual acquisition among children. Since I did my graduate work in the U.S., I was working with bilingual children in the U.S. originally. After I got a job at Penn, I realized that I was the only Asian or Asian American faculty member in the School of Education. At that time, the school started having more and more Asian students and people began sending them to me! I realized that even though I'm originally from Japan, I didn't know much about Asian education. So, I became interested in looking at language learning and teaching in Asian contexts. Coincidentally at that time, many Asian countries started introducing English at the primary school level. Since I'd been working with bilingual children and second language YLs in the US, it was a natural transition for me to start looking at Asian primary school students. Soon I realized that assessment is the least researched area in primary school language education: Teachers are looking for answers in what kind of assessment they should use – there are lots of questions but few answers. That's how I began to look at assessment for YLs.

You've researched self-assessment, mainly for use with young learners of English (e.g., Butler & Lee, 2010; Butler, 2018) – what place do you think self-assessment has in language learning situations in general and specifically in Japan?

I believe that self-assessment (SA) has potential as a type of formative assessment. SA is aligned very well with the notion of learner-centered approaches to language learning. SA should promote self-regulated learning. In principle, SA should help learners develop greater autonomy and self-regulatory abilities, which in turn should help them learn better in the long run. SA is also considered less stressful for learners. This is particularly important for YLs because we don't want to give them unnecessary stress especially at the beginning of their English language learning. People talk about SA being suitable in so-called 'exam-oriented cultures' because in those cultures learners tend to have high anxiety. Since SA is considered less stressful, it's well suited for learners who have high anxiety for English language learning. Finally, SA is less constrained by time and a large class size. Considering that Japanese classes still tend to have many students in general, SA has some practical merits as well.

But SA is not the ultimate solution for everything – we have to be very careful when we implement it. There are a series of considerations that are required for successful implementation of SA. First, especially when you

administer it to YLs, items need to be constructed very carefully in order to avoid confusion. One of my studies (Butler, 2018) indicated that children's interpretations of the meanings of some of the commonly used words in SA – such as 'understand' or 'comprehend' – vary greatly. In addition, teachers often ask 'do you like English classes?', but these kinds of items do not really give you meaningful information because children tend to please adults. Moreover, how people respond to SA items is not really well understood, even among adults, not to mention among children. There are many questions unanswered regarding SA.

Teachers often ask how reliable SA is and whether children can be responsible for assessing their own abilities. There are some developmental and experiential effects – their age and the amount of experience with SA are very influential factors. So, if you just show up and say 'here's a SA, just fill this in'; children would not be able to handle it very well. It's very important to contextualize SA. It is better to ask about children's performance in a particular task immediately after they complete the task. Feedback is also very important. But currently as a profession, we do not know what kind of feedback is the most useful for children after SA.

Some researchers invite learners to create SA items. That's not something that Japanese schools have explored much – it's a really new idea for most Japanese teachers. But when I started asking children about this possibility in my study, they seemed to be very excited about it. It turned out that the children already had a very good idea about what an assessment should look like and what kinds of assessment they wanted. Children's insights can be incorporated as meaningful input when teachers construct assessment items. Inviting students as part of the assessment development process is something that teachers can explore in the future. That would help children to be more responsible for their own learning as well.

Perhaps at this point some readers may be wondering about the reliability of SA. Is it fair to assume that SA should only be used as low stakes formative assessment, not medium/high stakes summative assessment?

It can be used for both purposes. For adult learners, we know that the reliability is reasonably high in general so that my university, for example, uses SA for placement purposes for foreign language classes. The correlation with other measures of proficiency is high enough, so SA can be used reliably.

When it comes to YLs, the reliability of SA can vary depending on the children's age and their experience with SA. If a SA is not reliable for YLs, it is usually not because YLs do not have the ability to self-assess their performance, but because we, as researchers or teachers, do not administer SA in such a way that is appropriate for their developmental level. As long as we can administer SA appropriately, I think it can work well with children even for summative purposes. But, in my view, the real benefit of SA is to use it as a pedagogical tool rather than for a measurement purpose. That way of using SA is probably more powerful for learners than using it for summative purposes.

What can young learners tell us about tests?

My colleagues and I have been working on a project looking at YLs' language assessment literacy (Butler, Peng, & Lee, under review). Of course, there's substantial individual differences in to what extent they know about language assessment. But we found that the Chinese students with middle class backgrounds that we worked with knew so much about assessment based on their experiences. They could tell us much about what kinds of topics, formats, and procedures that they felt comfortable with. One of the children expressed her frustration with the current assessment practice at school saying that 'we want to have humanized assessment'! These children wanted to have more meaning-focused and communicative-based assessment. They wanted the assessment content to be practical and fun, often in a story-based format, and cognitively challenging. They also wanted the assessment format to be more accessible to children. Some of the children could articulate construct irrelevant factors in the tests that they experienced.

It was very surprising for us to find out how much they knew about assessment based on their experiences; they actually drew on their experiences of assessments from other subjects such as Chinese and science. This finding got me thinking that we – language assessment professionals – do not have much conversation with assessment professionals in other subject areas, but that might be very fruitful. We also found that the children had actually taken many tests outside of the classroom, including international proficiency tests such as Cambridge and

TOEFL tests for YLs. These tests tend to be based on more recent communicative-based approaches. The children who had been exposed to such tests can be very critical of some of the traditional assessment practices at school. It's very important for teachers to listen to YLs; they know so much more than what teachers expect.

The children also told us that the kinds of topics that we often dealt with in assessment did not seem to be very exciting for them. We must better understand what kinds of topics are most suitable for reading and listening assessment tasks for YLs. The children are looking for something exciting, meaningful, or new to them so that they can learn something new by taking a test.

What do you think is in store for the future of assessing young learners?

Technology will play a greater role in assessment for sure. YLs are growing up with digital technology from very early stages in their lives, and they increasingly use language through technology. Technology is very important as a means to assess language performance and also as a target of language use. We don't yet have a very clear idea of whether their cognitive processes have changed due to digital technology – we still need to have basic research on that – but I wouldn't be surprised if YLs' cognitive processing and preferred learning strategies are somewhat different from previous generations. Because the target language use is changing in the era of digital technology, construct definitions need to be changed as well.

Game-based assessment is also a possibility. At this point, I don't have a very clear idea about what game-based assessment should look like, but I think the elements of gamification can be incorporated in assessment. I conducted a game study with primary school students here in Japan (Butler, 2017). One of the fascinating findings was that the children incorporated unexpected elements in their game plans. In a car racing game, for example, your car suddenly breaks down towards the end of the race – this is an unexpected element. We, adults, tend to think that there should be a linear relationship between learning and scores; the more you learn, the higher score you get. That's the ground rule. But the children do not necessarily seem to subscribe to the rule when it comes to gamification. By having unexpected elements in games, students who learned most don't necessarily get the highest score. You wonder why their game plans included these unexpected elements. The children had a good reason for it. It's because such elements make games more exciting and more motivating; and so even someone who is not so good at English can still win the game. Isn't that fascinating? It sounds a little bit radical, but when it comes to formative assessment, it may be possible to incorporate some ideas from gamification which can excite and motivate children. YLs are looking for something exciting, even in assessment.

Do you have any advice for researchers in Japan who wish to conduct studies with young learners?

Conducting any research concerning children in Japan is very challenging because schools are generally not very open. But we cannot blame schools – this is partly our fault as researchers because we tend to conduct research for the sake of research. In the child developmental field, research *with* children as opposed to research *on* children, has been advocated and is very popular. Instead of just treating children as an object or subject of research, we should try to give them more agency. Some researchers even talk about having a child as a co-researcher, although admittedly it is not always easy and it's very complicated both theoretically and logistically. But the general spirit of respecting children's agency is an important concept. Researchers working with children need to keep in mind that children are not merely the subject of research. Researchers need to listen to their voices more seriously.

Related to that, I think we need to have more pedagogically oriented research for YLs, simply because we don't have much research supporting teachers' pedagogical concerns. Rod Ellis and Natsuko Shintani (Ellis & Shintani, 2013) have talked about the teachers having a very difficult time finding pedagogical solutions in SLA research. This was because researchers tend to follow their theoretical interests when conducting research and only include a paragraph of practical implications in the very last section of their paper – as if it is a peripheral issue for them. Researchers need to be more active in inviting teachers when they plan and execute research. That, in the long run, will probably help to facilitate more cooperation from the schools.

In essence, there are two types of research that need to be promoted more: More child-centered research and more pedagogically-oriented research that is accessible to teachers.

Thank you, Professor Butler!

Selected Publications

Butler, Y. G. (2007). Factors associated with the notion that native speakers are the ideal language teachers: An examination of elementary school teachers in Japan. *JALT Journal, 29*(1), 7-40.

Butler, Y. G. (2015). Task-based assessment for young learners: Old meets new cultures, in M. Thomas & R. Hayo (Eds.) *Contemporary task-based language teaching in Asia* (pp. 348–365). London: Bloomsbury.

Butler, Y. G. (2016). Self-assessment of and for young learners' foreign language learning, in M. Nikolov (Ed.) *Assessing young learners of English: Global and local perspectives* (pp 291–31). New York: Springer.

Butler, Y. G. (2017). Motivational elements of digital instructional games: A study of young L2 learners' game designs. *Language Teaching Research*, *21*(6), 735-750.

Butler, Y. G. (2018). The role of context in young learners' processes for responding to self- assessment items. *Modern Language Journal, 102*(1), 242–261.

Butler, Y. G. (2019). Assessment of young English learners in instructional settings, in X. Gao (Ed.), *Second Handbook of English Language Teaching* (pp. 477-496). Switzerland: Springer.

Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessment among Korean elementary school students studying English. *Modern Language Journal, 90*(4), 506–518.

Butler, Y. G., & Lee, J. (2010). The effect of self-assessment among young learners. *Language Testing 17*(1), 1–27.

Butler, Y.G., Peng, X., & Lee, J. (under review). "'I want humanized assessment!': Young learners' voices for a learner-centered approach to language assessment literacy.

Butler, Y. G., & Zeng, W. (2014). Young foreign language learners' interactions during task-based paired assessment. *Language Assessment Quarterly, 11*, 45–75.

Ellis, R., & Shintani, N. (2013). *Exploring language pedagogy through second language acquisition research*. London: Routledge.

Call for Papers

Shiken is seeking submissions for future issues on an ongoing basis. After checking the guidelines for publication below, please send your submission to the editor at tevalpublications@gmail.com.

Overview

Shiken aims to publish articles concerning language assessment issues relevant to assessment professionals, researchers, classroom practitioners and language program administrators. *Shiken* is dedicated to publishing articles on assessment in various educational contexts in and around Japan.

Article formats include research papers, replication studies, and review articles, all of which should typically not exceed 7000 words; as well as informed opinion pieces, technical advice articles, and interviews, all of which should typically not exceed 3000 words. Please contact the editor if you wish to submit an article that differs from these formats.

Novice researchers are encouraged to submit, but should aim for papers that focus on a single main issue. Please review recent issues of *Shiken* to understand the level of detail, depth of discussion and provision of empirical evidence that is required for acceptance.

Format

To facilitate double-blind peer review, the name(s) of the author(s) should not be mentioned in the manuscript. All citations and references to the author or co-author's work should read (Author, Year) e.g. "(Author, 2017)."

Articles should be formatted according to the guidelines of the *Publication Manual of the American Psychological Association (APA), 7th Edition*. Submissions should be formatted in Microsoft Word (.docx format) using 12-point Times New Roman font. The page size should be set to A4, with a 2.5 cm margin. The body text should be block justified, and single-spaced. Paragraphs should be separated by a blank line space. Each new section of the paper should have a section heading. Please review recent issues of *Shiken* for examples of section headings.

Research articles must be preceded by an abstract that succinctly summarizes the article content in less than 200 words. The reference section should begin on a new page immediately following the body text. Authors are responsible for comprehensive and accurate referencing. Tables and figures should be numbered and titled. Separate sections for tables and figures should follow the references. Any appendices should be numbered and should follow the tables and figures.

Evaluation

All papers are double-blind peer-reviewed by at least two expert reviewers. Initial evaluation is usually completed within four weeks. The whole process from submission to publication is normally completed within six months. Submissions should be sent to the editor at tevalpublications@gmail.com.