SHIKEN

A Journal of Language Testing and Evaluation in Japan

Volume 28 • Number 1 • November 2024

https://doi.org/10.37546/JALTSIG.TEVAL28.1

Contents

Testing and Evaluation SIG

ISSN 1881-5537

Shiken: A Journal of Language Testing and Evaluation in Japan

Volume 28 No. 1
November 2024

https://doi.org/10.37546/JALTSIG.TEVAL28.1

Article
Shiken: A Journal of Language Testing and Evaluation in Japan
2024, Vol. 28(1) 1–57
DOI: 10.37546/JALTSIG.TEVAL28.1
ISSN: 1881-5537

Investigating the assessability of speaking proficiency in a group discussion context

Paul Garside
garsidepaul@hotmail.com
Meiji University

Abstract

The main purpose of this exploratory study was to attempt to measure the construct of speaking proficiency in a group discussion context. Although peer-discussion activities are commonly used in ESL/EFL classrooms, little is known about how to adapt this format for testing purposes and whether it can be done so reliably. In this study, an analytic rubric was used to assess the proficiency of Japanese university students during group discussions. Rasch (MFRM) analysis was then conducted to investigate the extent to which the students, raters, and category items (i.e., subcategories of the rubric) fit the model. Results showed that although the raters differed in terms of severity, they maintained internal consistency, therefore allowing MFRM to control for this disparity. Following this procedure, students could be separated into approximately three levels of proficiency. Furthermore, all category items fit the model sufficiently well to conclude that a single construct was being measured. These findings support the idea that group oral testing can be conducted reliably as an aspect of L2 speaking assessment.

Keywords: group speaking assessment, Rasch analysis, facets, MFRM

Whether for high-stakes examinations or in-class testing, the performance-based assessment of L2 speaking has become increasingly common over recent decades. This performance is usually evaluated in accordance with a scoring rubric—sometimes referred to as a rating scale—which can either be holistic or analytic. In the former, a single global score is assigned; in the latter, the construct is subdivided into several related categories with a separate score assigned for each one (Green, 2014). The main advantage of analytic rubrics is that they offer a more reliable assessment of proficiency, as they provide specific information about a learner's strengths and weaknesses regarding the construct of interest (Hamp-Lyons, 2016). When designing such a scale for assessment purposes, rating categories should be chosen that reflect the theoretical conception of the construct (Spaan, 2006). In the case of speaking assessment, speech elicitation tasks that allow candidates to fulfill the stated criteria are then selected. For example, if the rating scale mentions the ability to give and support opinions—as in the current study—the assessment task(s) should be presented in such a way that candidates are clearly required to do so.

The use of rubrics or rating scales for assessment inevitably involves an element of subjectivity, as raters bring different perspectives and levels of expertise that can lead to different scoring outcomes (Pill & Smart, 2020). For example, in an experimental study Duijm et al. (2018) found that linguistically-trained expert raters focused more on accuracy of output, whereas untrained raters focused more

on fluency. Such differences in rater behavior introduce confounds that can threaten the reliability of a test if they are left unaccounted for. They can, however, be mitigated post-assessment via many-facet Rasch measurement (MFRM), which is a statistical technique that identifies the effect of variables (or facets) such as rater severity and item difficulty, and adjusts scores accordingly (Ockey, 2022). MFRM was used in the current study to analyze the consistency of four expert raters when assessing learners in a Japanese EFL context.

As well as rater judgments, the other facets modeled were test-taker performance and the functioning of the assessment instrument, which consisted of scaled items for the following five categories: fluency, accuracy, strategy use, active listening, and content. Analysis of the students' scores was intended to reveal differences in performance, which could then be used for grading purposes. Analysis of the category items was intended to reveal whether they form a unidimensional construct; that is, whether they tap into the same measurement domain and can therefore be measured by using the same task. As misfitting categories do not belong to the same underlying construct, rating-based assessors of group speaking proficiency can use such information when considering which items to include or remove from their own scoring rubrics.

Literature Review

This section begins by defining speaking proficiency in both psycholinguistic and interactional terms. The former element focuses on the internal mechanisms of the individual, whereas the latter highlights the reciprocal nature of speaking in context, reflecting how conceptions of the construct have expanded over time. A brief history of L2 speaking assessment is then outlined, with its development traced from interview tests to pair and group activities in which candidates interact with each other instead of the examiner. Finally, the role of MFRM in rater-based language assessment is addressed.

Defining Speaking Proficiency

Testing aspects of language use entails defining the underlying constructs to be measured (Spaan, 2006). In the case of speaking proficiency, it requires understanding the nature of L2 speech production. Following pioneering work from Skehan (1998), research on speech production has commonly been divided into the three psycholinguistic components of complexity, accuracy, and fluency (CAF). First, complexity relates to the range of lexis, morphology, and syntax used by a speaker. Next, accuracy is gauged by comparison with target language norms of correctness (Pallotti, 2020). Finally, fluency refers to the speed and smoothness with which a speech sample is produced. Fluency can be evaluated either objectively, using measures of speech rate, repair, and pausing phenomena, or subjectively, with raters asked to give their impression of a speaker's performance

(Segalowitz, 2010). The CAF framework has come to play an important role in language testing and assessment, with combinations of linguistic measures frequently used as criteria in rating scales (Kuiken & Vedder, 2020). By addressing multiple distinct aspects of language use, the multifaceted nature of speaking proficiency can be better reflected in measurement.

The CAF framework focuses on the formal linguistic characteristics of speech production. It does not, however, address the issue of communicative adequacy, defined as "the degree to which a learner's performance is more or less successful in achieving the task's goals efficiently" (Pallotti, 2009, p. 596). As achieving one's goals is fundamental to any speech act, this aspect should not be overlooked when operationalizing L2 speech (Tavakoli & Wright, 2020), or when evaluating learner output (Pallotti, 2009). In short, fluent speech that is irrelevant to the task or difficult to comprehend, even if accurate and complex, could not be described as communicatively adequate.

A further criticism of psycholinguistic approaches is that they are concerned with speech produced in isolation rather than talk as a shared social activity (Luoma, 2004). Therefore, to establish a theoretical basis for assessing speaking activities based on real-world interaction, it is necessary to examine models that incorporate an interactive dimension. Perhaps the most influential of such models has been Canale and Swain's (1980) communicative competence, which includes strategic and sociolinguistic elements, in addition to a grammatical component. Strategic competence refers to the ability to overcome communication breakdowns, whereas sociolinguistic competence pertains to the pragmatic and sociocultural norms of language use in context. The model was later expanded to encompass discourse competence, which relates to the coherence and cohesion of extended stretches of speech across various genres.

It is clear, therefore, that speaking proficiency is highly contingent on the context of the interaction and the behavior of other participants (Young, 2011). Accordingly, the term interactional competence has become widely used to emphasize the dynamic, co-constructed nature of talk in local, practice-specific contexts (He & Young, 1998). The construct of speaking proficiency has thus been expanded to include such inherently social aspects as turn and topic management, active listening, and non-verbal behavior, in addition to breakdown repair (Galaczi & Taylor, 2018), highlighting the complexity of L2 interaction. However, acknowledging the intertwined role of speakers and interlocutors has to be recognized both pedagogically and for assessment purposes.

Assessing Speaking Proficiency

Practically, the main issues to be addressed when assessing speaking are whether to have candidates talk together or with an examiner, and whether to use a holistic or analytic rubric. The classic speaking test format is the oral proficiency interview (OPI), in which an examiner poses questions to individual candidates for the purpose of eliciting samples of speech sufficient to judge their speaking ability (Nakatsuhara et al., 2020). Originally devised with a holistic rating scale, it was revised to incorporate five distinct components of proficiency—accent, comprehension, fluency, grammar, and vocabulary—representing an important step towards the reliable assessment of a multifaceted speaking construct (Fulcher, 2003). Nevertheless, the OPI format has been criticized for producing interaction that is asymmetrically initiated and controlled by the examiner, with the role of the candidate simply to answer each question in turn (Van Lier, 1989). According to this view, the traditional OPI does not resemble realistic communication, in which participants take joint responsibility for shaping and maintaining conversations. Moreover, as Roever and Ikeda (2021) have pointed out, if interactional abilities are not elicited or assessed in a speaking test, inferences regarding the ability to participate in real-world interaction are undermined, thus raising issues of test authenticity.

Some testing organizations, such as Cambridge Assessment English, have responded to such criticisms by introducing a paired testing element (Vidaković & Galaczi, 2013). In this format, candidates have to interact with each other for at least part of the exam and are required to exchange opinions during a task in order to reach a decision. As a result, paired speaking tests elicit a wider variety of talk than interview tests, as participants are required to initiate and manage turns during interaction (Swain, 2001). An additional benefit is the positive washback that occurs when assessment conditions are reflected in curriculum goals and classroom activities that simulate the test (Harsch & Malone, 2020). Paired speaking assessment therefore creates a virtuous cycle, as it resembles language use in the real world more closely than traditional testing formats.

Extending this principle further, learners can also be assessed during group discussion tasks without any interaction with the examiner. This learner-centered, multi-party format heightens the need for participants to manage and direct their own interaction, thus allowing more aspects of interactional competence to be elicited and measured (Galaczi & Taylor, 2020). Furthermore, from a pedagogical perspective it promotes optimal washback as students need to learn to collaborate without the intervention of an instructor in order to prepare for the test (Linn, 1993). In practical terms, group oral tests are also more cost effective and time efficient, as several candidates can be tested simultaneously.

However, group oral testing has not received a great deal of attention in the literature and claims about its reliability have been mixed. Shohamy et al. (1986) found that group oral test results had the lowest correlation with results of other speaking tasks—consisting of an OPI, a role play, and a reporting task—implying that a different construct was being measured in the group context. In contrast, Fulcher (1996) found that scores on a group oral task did generalize to two oral interview tasks undertaken by the same examinees. He concluded that all three tasks were operating on a unidimensional scale, and that large task effects are more likely to be an artifact of the rating scale than underlying properties of the test item. Furthermore, Bonk and Ockey (2003) achieved rater and scale reliability in group discussion tests by including a large number of observations (see below for a more detailed account). While acknowledging the potentially wide variety of unexamined variables inherent in this format (e.g., social status, personality factors, and proficiency level) the authors concluded that, given the prevalence of peer discussion in language classrooms, some form of examinee-controlled discourse has become essential when conducting oral assessment. To sum up, although group oral testing introduces additional noise that could affect test performance, it also has major benefits in terms of efficiency, washback, and applicability to real-world contexts. Moreover, if this kind of testing can be conducted reliably, as some studies have indicated, it reinforces the idea that group oral testing should be included as an aspect of L2 assessment.

Many-Facet Rasch Measurement

One way to increase the reliability of rater-based assessment is to use MFRM. The inherent subjectivity of human judgments means that that test takers' scores are likely to be affected by differences in rater severity; that is, how strict individual raters are when assigning scores (Pill & Smart, 2020). MFRM estimates the magnitude of this effect and automatically accounts for it when scoring student performance (Ockey, 2022). Furthermore, inconsistent raters can be identified and provided with formative feedback.

In a study based on the rating of writing samples, Weigle (1998) used MFRM to investigate rating patterns before and after training was provided. Although some differences in severity persisted after training, fewer extreme scores were produced and internal consistency improved among both experienced and inexperienced raters. The author concluded that rater training promotes intra-rater reliability (i.e., internal consistency), which can then be controlled for by MFRM as long as differences between raters are systematic. Moreover, this process can be used even if raters have different conceptions of the construct being tested.

Bonk and Ockey's (2003) study, mentioned above, used MFRM to examine two iterations of a large-scale group oral test in a Japanese university. The facets modeled were: examinee, question prompt, rater, and the five category items used in the rating scale (pronunciation, fluency, grammar, vocabulary/content, and communicative skills/strategies). Apart from examinee ability, rater severity had the largest effect on test scores, prompting the authors to conclude that failing to control for this variable would be irresponsible in high-stakes testing, especially in cases when only one judge assigns a rating to each candidate. Furthermore, all category items fit the model sufficiently well such that unidimensionality remained strong across both data sets. This combination of interactional and linguistic variables was, therefore, considered to form one underlying construct.

Gaps and Research Questions

Two gaps in the literature are addressed in this exploratory, cross-sectional study. The first concerns the reliability of group oral testing which, despite the prevalence of peer discussions in EFL contexts, has been under researched as a testing format. The second gap relates to the nature of speaking proficiency. As speaking tests have become more diversified, the construct of L2 interaction has expanded to include interactional competence, and therefore variables associated with interlocutors as well as speakers (Galaczi & Taylor, 2018). As scoring rubrics reflect this development, it is important to investigate whether the various category items form part of the same underlying construct.

The research questions (RQs) are stated as follows:

1. To what extent can speaking proficiency be assessed reliably by raters in a group discussion context?
2. To what extent do the facets modeled fit the conception of speaking proficiency in a group discussion context as a unidimensional construct?

Methods

This section describes the participants and methods of data collection. Next, the theoretical justification for the categories included in the scoring rubric is outlined. Finally, the concept of fit in MFRM analysis, and how it pertains to the current study, is explained.

Participants

16 first-year university students (nine male and seven female) from a competitive, co-educational university in Tokyo participated in the study. All students were enrolled in my weekly speaking classes for the semester during which it took place. Informed consent was obtained from each participant. They were all familiar with the group discussion format as such activities were conducted

regularly in class. All participants were non-English majors and were selected at random from four separate classes, representing three different linguistic proficiency levels. One class was a high beginner level, two were low intermediate, and one was intermediate, with participants having been assigned to these classes on the basis of a standardized placement test (TOEIC Listening and Reading). However, as the placement test contained no oral component the classes were of relatively mixed speaking abilities. All four raters were experienced native-speaker teachers of English in Japanese universities, familiar with the group discussion format. Pseudonyms have been applied except in the case of Paul (the researcher).

Recorded Discussions

Groups of four members from intact classes were video recorded during lessons over one week. As each class consisted of either seven or eight members, the remaining members engaged in a parallel discussion activity at the other end of the classroom. Each discussion lasted 16 minutes; the instructor did not intervene once the discussion had begun, so that participants were given the fullest possible opportunity to display their interactional skills. Written prompts, used as the basis for the discussion, were provided and read by the participants. Students had already discussed questions related to the topic in pairs, but no specific preparation time was provided before the group discussion began.

For all groups, the question prompts were:

1. What is important to be happy?
2. Do you think people in Japan are happy?

The recorded discussions were then viewed and rated by four native speakers (two from the U.K. and two from the U.S.) who all have extensive experience of teaching Japanese university speaking classes. The raters were made aware of the context and purpose of the study, and opportunities were provided to discuss and ask questions about the rating scale. Each rater watched two of the four videos, so each group was evaluated by two different raters. The rating plan was designed to ensure sufficient overlap between raters and therefore avoid disjointed subsets (see Table 1), which is necessary to maintain the validity of MFRM.

Table 1

Rating Plan

Rater	Groups
Paul	1 & 2
Neil	1 & 3
Aiden	2 & 4
Calvin	3 & 4

For scoring purposes, the raters were provided with a rubric containing level descriptors (Appendix A) and a mark sheet (Appendix B). These ratings were then used to conduct MFRM analysis using FACETS version 4.1.4 (Linacre, 2024).

The Scoring Rubric

Measuring speaking proficiency in a communicative context, such as a group discussion, needs to account for psycholinguistic research in SLA, as linguistic knowledge and cognitive processing skills have been found to contribute significantly to communicative ability (De Jong et al., 2012). It should also account for the role of interactional competence (e.g., turn-taking and interlocutor variables) in effective L2 interaction (Galaczi & Taylor, 2018). This perspective informed the attempt to categorize and describe the elements of speaking proficiency in a group discussion context shown in the rubric (Appendix A). Although the lack of empirical evidence and theoretical consensus regarding the development of language acquisition presents a major challenge when devising such a scale (Kuiken & Vedder, 2020), the following five category items were chosen to reflect the multifaceted nature of the construct: fluency, accuracy, strategy use, active listening, and content (see below for the theoretical justification). Each category was then subdivided into five levels, with related descriptors, for the purposes of standardization and consistency of assessment (Weigle, 2002). These categories are broadly similar to the ones used by Bonk and Ockey (2003), but with active listening replacing pronunciation. Given that participants are by definition likely to spend more time listening than speaking during a group discussion, it is necessary to ascertain whether unidimensionality is maintained upon the inclusion of this category.

Fluency

Beginning with the CAF model, fluency is included in the rubric because speed of output is essential to maintaining the flow of interaction. Patience is demanded of listeners if speech becomes excessively halting and fragmented, with pauses that appear mid-clause more strongly associated (i.e., negatively correlated) with human ratings of fluency than those that appear at clause-end boundaries (Suzuki & Kormos, 2020). Pausing phenomena and speech rate have consistently been found to correlate with subjective ratings of fluency (Pallotti, 2020); therefore, both of these elements were included in the descriptors for that category. Filled pauses can, however, serve important communicative functions, such as signaling an intention to hold the floor (Segalowitz, 2010), hence the descriptors at higher levels refer to hesitation at appropriate points as well as to speaking at natural speed.

Accuracy

Accuracy of grammatical and lexical forms—another element of the CAF model—is also included as it indicates proximity to target language norms. In a communicative context, however, the amount and frequency of mistakes is of less importance than whether they hinder comprehensibility, and hence communicative effectiveness (Pallotti, 2020). This consideration is therefore reflected in the descriptors used for that category, with performance at higher levels marked by either very few mistakes or mistakes that rarely impede communication. Accurate use of both lexical and grammatical structures demonstrates the linguistic knowledge necessary to deal with a variety of topics and situations, thus justifying their inclusion in this category's descriptors. However, complexity—the third element of the CAF framework—was not included in this scale as it is valued more highly in formal contexts, such as academic writing, than in communicative interaction. Moreover, the overuse of complex structures can impede communication, especially if they do not match the interlocutor's level of comprehension (Pallotti, 2020).

Strategy Use

In addition to the above psycholinguistic items, operationalizing speaking proficiency in a group discussion context requires the inclusion of interactional features (Galaczi & Taylor, 2020). The first of these is strategy use, which encompasses breakdown repair and turn taking. Repair is a common feature of spontaneous speech (Riggenbach, 1998), therefore how learners deal with miscommunications and breakdowns often determines their communicative success. Moreover, skillful participants can pre-empt potential breakdowns by checking whether their contributions have been comprehended during or after their turn. High performance in this category also involves effective turn-taking management, as taking and ceding the floor—as well as encouraging others to contribute—facilitates the smooth and efficient functioning of interaction (Wong & Waring, 2021).

Active Listening

The other interactional category included is active listening, reflecting the fact that interlocutors are integral to interaction (Galaczi & Taylor, 2020). The contingent nature of spoken communication means that participation is not limited to producing and managing one's own output; rather, the ability to respond appropriately also forms part of the construct of speaking proficiency in this context. High performance in this category involves asking open-ended questions, indicating agreement or disagreement, and using reactions to demonstrate interest and empathy. Indeed, it is hard to imagine a successful group discussion taking place without a steady stream of such listener-based contributions. Although non-verbal behavior, such as eye contact and facial expression, is also an

important element of interaction (Galaczi & Taylor, 2018), it was not included in the rubric to avoid placing an unrealistic burden on the raters.

Content

The ability to generate content is fundamental to communicative success and the efficient achievement of a task's goals (Pallotti, 2009). Speaking cannot exist in any meaningful sense without content, and effective participation in a group discussion requires contributions that are related to the topic. It also entails supporting opinions (e.g., with reasons or examples), while using appropriate phrases or discourse markers to manage the flow of interaction. For example, phrases like *In my opinion* or even just *I think* show that the speaker can differentiate opinion from fact, which can help to avoid ambiguity. Performance at higher levels therefore entails using such features appropriately. It additionally involves the ability to initiate discourse—also reflected in the descriptors for this category—as initiating is a necessary precursor to generating content.

Item Fit in MFRM

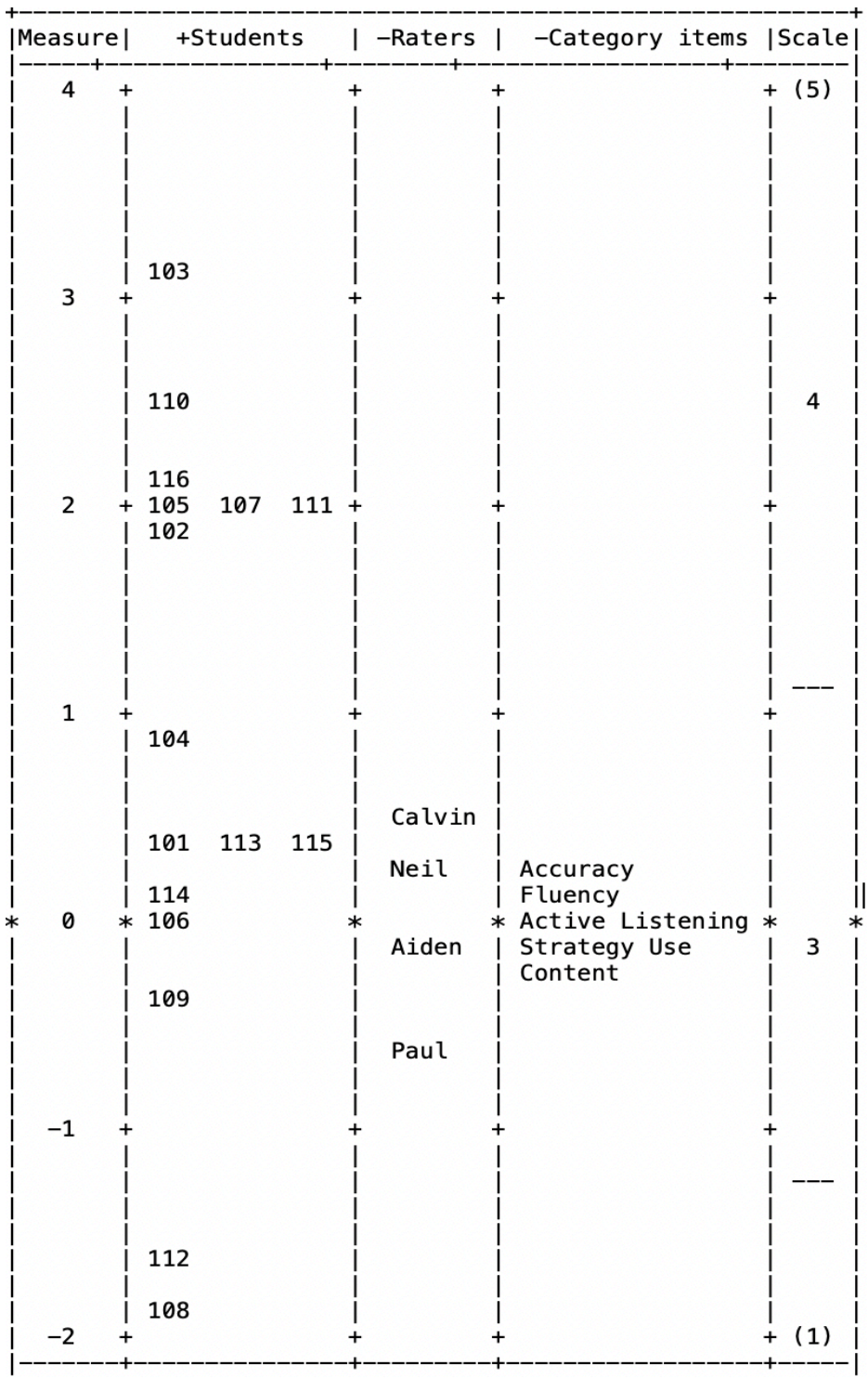
Item fit is an important assumption of Rasch modeling. Items that fit the model should have infit and outfit mean square (MNSQ) values close to the expected 1.0, although Linacre (2002) has argued that a range of .5 to 1.5 is productive for measurement purposes, and therefore acceptable in lower stakes or exploratory contexts such as the current study. Rater misfit threatens test reliability—which relates to RQ1—as it indicates atypical or random patterns of behavior, which have a major impact on all other facet measure estimates (Bonk & Ockey, 2003). Moreover, unlike with rater severity, Rasch modeling cannot control for raters who do not maintain internal consistency. Fit statistics are also relevant to unidimensionality—which relates to RQ2—as category items that tap into the same measurement domain, and therefore form part of the same underlying construct, should have values within the expected range.

Results and Discussion

Figure 1 depicts all three facets modeled in the analysis. Wright maps use a logit—short for log odds—scale, which produces standardized interval measurements (as seen in the left-hand column) based on statistical probability. They provide a graphic illustration of the amount of variance within each facet, and the common scale allows for comparison with the other facets. Upon visual inspection it is clear that the greatest amount of variance is found among the students, followed by the raters, and finally the category items. Each facet is examined in detail below.

Figure 1

All-facet Wright Map for the MFRM Analysis



Note. N = 16. Measure values are in Rasch logits.

For speaking in a group discussion context to be measured reliably—relating to RQ1—students have to be differentiated by the degree of the construct they are able to demonstrate during the task. In this case, student ability varied from a minimum of -2.02 logits to a maximum of 3.28 logits, representing a wide spread of abilities among the 16 participants. The separation value was 2.78, suggesting that the participants can be approximately divided into three proficiency levels based on this activity. Furthermore, a Rasch reliability statistic of .89 indicates that these figures are highly reproducible. The logits presented in Figure 1 are based on *fair average* scores, automatically generated to control for differing levels of severity among raters who do not assess all the same students. The fair average scores differ slightly from the observed (i.e., unadjusted) scores, as shown in Table 2, although this adjustment is essential to avoid test reliability from being undermined by differences in rater severity.

Table 2

Rating Scores Based on Recorded Group Discussions

Rank	Student number	Proficiency level	Rater 1 raw total	Rater 2 raw total	Observed average	Fair average
1	103	1	22	21	4.30	4.22
2	110	3	21	18	3.90	4.04
3	116	2	19	19	3.80	3.88
4=	105	2	19	21	4.00	3.85
4=	107	2	21	19	4.00	3.85
4=	111	3	19	18	3.70	3.85
7	102	1	23	16	3.90	3.82
8	104	1	19	17	3.50	3.41
9	101	1	19	15	3.30	3.21
10=	113	2	14	17	3.10	3.19
10=	115	2	16	15	3.10	3.19
12	114	2	17	13	3.00	3.09
13	106	2	16	16	3.20	3.03
14	109	3	14	13	2.70	2.88
15	112	3	12	10	2.20	2.38
16	108	2	12	12	2.40	2.23

Note. Raw scores represent the total of all 5 categories (maximum = 25). Observed and Fair averages represent the average of all categories across both raters (maximum = 5).

In terms of fit, three students fell outside of the acceptable range. However, Bonk and Ockey (2003) argued that person misfit is unlikely to be a major problem in this kind of data set, as the nature of the task precludes misfit based on lucky guessing or examinee inattention. Rather, it is more likely to reflect the fact that some participants have a marked disparity between their strong and weak points. Accordingly, no unusual behavior that could have contributed to person misfit was

observed, as all participants remained on task throughout the recorded discussions.

Investigation of individual cases is further revealing. For example, student 116, who had the highest infit MNSQ of 2.06, was awarded 5 points for active listening and strategy use but only 2 points for accuracy by one rater. This was the only data point to be flagged as unexpected in the analysis, although it could simply reflect the fact that this student has a lower level of accuracy in comparison with other elements of their speaking proficiency. Therefore, as unexpected disparities between participants' strengths and weaknesses do not necessarily indicate misfit, these data were retained in the model and were not judged to threaten the reliability of the test.

The four discussion groups were formed from three different linguistic proficiency bands, although scores on this task did not correspond closely with those levels. For example, student 103—who received the highest score—was from the highest proficiency band (Level 1), but student 110, who ranked next highest, was from the lowest band (Level 3). In addition, student 108—who ranked the lowest overall—was from the middle band (Level 2). In general, the students were distributed relatively evenly, regardless of their linguistic proficiency (see Table 2), implying that the construct of speaking proficiency in a group discussion context is distinct from general linguistic proficiency. This finding calls into question the validity of using standardized tests without a speaking component—such as TOEIC Listening and Reading—to stream students into different levels of speaking classes. Speaking—especially in a group context—requires interactional skills that could be more related to issues of personality than formal linguistic proficiency. For example, Nakatsuhara (2013) found that extraverts performed better than introverts on an open-ended group speaking test, suggesting that freer spoken interaction, with its potential for heightened stress, favors extraverted personality types. In pedagogical terms, making students aware of the importance of active listening, and teaching strategies to deal with communication problems, could improve their ability to interact regardless of their linguistic knowledge.

Raters

Internal consistency among raters is another prerequisite for reliable measurement, enabling differences in severity to be controlled for. Table 3 shows a relatively wide disparity in terms of severity, with Calvin, at .6 logits, the most severe, whereas Paul, at -.72, was the most lenient. The observed and fair averages verify this divergence, as does the separation value of 2.03, which could be partly explained by the lack of a formal calibration or norming session. The fixed chi-square value of 15.2 was significant at $p = <.001$, confirming the differences in severity. Looking at individuals, the raw scores in Table 2 show that

student 102 received a potentially alarming difference of 7 points between the two raters. Nevertheless, raters awarded the same score to the same student in 45-55% of cases (see Table 3), which is above Wolfe and Smith's (2007) recommended criterion of 40%. Moreover, the raters demonstrated sufficient consistency in their scoring, with fit statistics ranging from .62 to 1.4, allowing the fair averages produced by FACETS to control for disparities in rater severity, thus maintaining test reliability. This technique can also be adopted for relatively low-stakes or classroom assessment if, for example, individual teachers grade each other's tests, either in real time or via video recordings, thus providing the multiple measures required for MFRM analysis.

Table 3

Rater Severity and Model Fit

Rater	Measure (logits)	Observed average	Fair average	Exact agree (%)	Infit MNSQ	Outfit MNSQ
Calvin	.63	3.08	3.16	50	.62	.62
Neil	.27	3.38	3.30	55	.74	.74
Aiden	-.15	3.35	3.46	45	1.40	1.41
Paul	-.75	3.72	3.68	50	1.18	1.16

Note. Observed and Fair averages represent the average score awarded across all students and categories (maximum = 5).

Category Items

Regarding unidimensionality—which relates to RQ2—the five rating categories all demonstrated acceptable fit (see Table 4). This finding suggests that all items belonged to a general construct of speaking proficiency, corroborating Bonk and Ockey's (2003) finding, although the categories used were not exactly the same. Active listening produced the 'noisiest' score (infit MNSQ = 1.28), which perhaps reflects the fact that it is the item least directly related to speaking proficiency. The level descriptors refer to asking questions, using reactions, and indicating agreement or disagreement, all of which—as the category title implies—depend on a degree of listening ability. Furthermore, active listening is arguably the category most related to personality factors. For instance, a learner can be called on by others to offer an opinion (i.e., content) and to clarify a comment (i.e., strategy use), but deciding whether to ask a question or react to a contribution depends on the initiative of the individual. As a result, less proactive or more introverted participants are perhaps likely to score lower in this category.

The category items displayed considerably less variability than the student and rater facets and did not prove difficult for the majority of the students (see Figure 1). Table 4 shows that the full range of difficulty was just over half of one logit, from a maximum of .27 (Accuracy) to a minimum of -.32 (Content), suggesting that all categories were of approximately equal difficulty. The fact that accuracy had

slightly lower scores than other categories could be interpreted as evidence of learners paying less attention to that aspect, given the communicative context of the activity. However, no firm conclusions can be drawn in this regard as Rasch separation and reliability statistics of 0 confirm that these items could not be divided into distinct levels of difficulty.

Table 4

Model Fit of Rubric Category Items

Category item	Measure	Model *SE*	Infit MNSQ	Outfit MNSQ
Accuracy	.27	.29	.98	1.01
Fluency	.18	.29	.89	.87
Active listening	.02	.29	1.28	1.28
Strategy use	-.15	.29	.97	.95
Content	-.32	.29	.78	.80

Note. All statistics are based on Rasch logits.

Conclusion

The results of this exploratory study suggest that speaking proficiency in a group discussion context can be measured reliably using ratings based on an analytic rubric, supported by MFRM analysis. It also holds up as a unidimensional construct, even though a variety of theories and models—including the CAF framework and interactional competence—were drawn on when devising the rubric, reflecting the complex nature of L2 spoken interaction. A large amount of variance was observed among the student participants, with results suggesting approximately three distinct levels of performance, despite the low sample size. This degree of separation indicates that participants could be reliably separated by ability, which is necessary for the kind of classroom assessment upon which this study is based. However, proficiency displayed in a group discussion is only one aspect of speaking proficiency as important differences exist with other speaking contexts, such as a role play or even an OPI. It is therefore essential to adapt rubrics and rating scales used for assessment to the specific demands of each task.

There are many advantages to group oral testing, despite the large number of variables it presents (e.g., personality, status, gender, and age of co-participants), and the potential for inconsistent rating. From a practical point of view, it is more efficient and less time consuming than conducting oral interviews, especially among larger classes. Moreover, it simulates the kind of autonomous behavior that learners need to replicate beyond the classroom, where learners are required to take responsibility for managing their own interactions. Testing these behaviors not only allows inferences to be drawn about the kinds of real-world skills that learners require, it also promotes positive washback and encourages these skills to

be taught and practiced in language classrooms. If further studies can confirm the reliability of group oral testing, such findings could have many practical and pedagogical benefits.

Declaration of competing interests:
P. Garside has declared no competing interests.

References

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model* (3rd ed.). Routledge.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral task. *Language Testing, 20*(1), 89–110. https://doi:10.1191/0265532203lt245oa

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47. https://doi.org/10.1093/applin/l.1.1

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5–34. https://doi.org/10.1017/S0272263111000489

Duijm, K., Schoonen, R., & Hulstijn, J. H. (2018). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing 35*(4), 501–527. https://doi.org/10.1177/0265532217712553

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing 13*(1), 23–51. https://doi.org/10.1177/026553229601300103

Fulcher, G. (2003). *Testing second language speaking*. Routledge.

Galaczi, E. D., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Galaczi, E. D., & Taylor, L. (2020). Measuring interactional competence In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 338–348). Routledge.

Green, A. (2014). *Exploring language assessment and testing*. Routledge.

Hamp-Lyons, L. (2016). Farewell to holistic scoring. Part two: Why build a house with only one brick? *Assessing Writing, 29*, A1–A5. https://doi.org/10.1016/j.asw.2016.06.006

Harsch, C., & Malone, M. E. (2020). Language proficiency frameworks and scales. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 33–44). Routledge.

He, A., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young, & A. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1–24). Benjamins.

Kuiken, F., & Vedder, I. (2020). Scoring approaches: Scales/rubrics. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 125–134). Routledge.

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

Linacre, J. M. (2024). FACETS Rasch measurement computer program (Version 3.87.0). [Computer software]. Winsteps.com

Linn, R. L. (1993). Educational assessment: expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1–16. https://doi.org/10.2307/1164248

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Nakatsuhara, F. (2013). *The co-construction of conversation in group oral tests*. Peter Lang.
Nakatsuhara, F., Inoue, C., & Khabbazbashi, N. (2020). Measuring L2 speaking. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 285–294). Routledge.
Ockey, G. J. (2022). Item response theory and many-facet Rasch measurement. In G. Fulcher, & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 462–476). Routledge.
Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045
Pallotti, G. (2020). Measuring complexity, accuracy, and fluency (CAF). In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 201–210). Routledge.
Pill, J., & Smart, C. (2020). Raters: Behavior and training. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 135–144). Routledge.
Riggenbach, H. (1998). Evaluating learner interaction skills: Conversation at the micro level. In R. Young, & A. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 53–67). Benjamins.
Roever, C. & Ikeda, N. (2021). What scores from monologic speaking tests can(not) tell us about interactional competence. *Language Testing, 39*(1), 7–29. https://doi.org/10.1177/02655322211003332
Segalowitz, N. (2010). *Cognitive bases of second language fluency.* Routledge.
Shohamy, E., Reves, E., & Bejarno, Y. (1986). Introducing a new comprehensive test of oral proficiency. *ELT Journal, 40*(3), 212–220. https://doi.org/10.1093/elt/40.3.212
Skehan, P. (1998). *A cognitive approach to language learning.* Oxford University Press.
Spaan, M. (2006). Test and item specifications development. *Language Assessment Quarterly 3*(1), 71–79. https://doi.org/10.1207/s15434311laq0301_5
Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition, 42*(1), 143–167. https://doi.org/10.1017/S0272263119000421
Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing 18*(3), 275–302. https://doi.org/10.1177/026553220101800302
Tavakoli, P., & Wright, C. (2020). *Second language speech fluency: From research to practice.* Cambridge University Press.
Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*(3), 489–508. https://doi.org/10.2307/3586922
Vidaković, I. & Galaczi, E. D. (2013). The measurement of speaking ability 1913-2012. In C. J. Weir, I. Vidaković, & E. D. Galaczi (Eds.), *Measured constructs: A history of Cambridge English language examinations 1913-2012* (pp. 257–346). Cambridge University Press.
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287. https://doi.org/10.1177/026553229801500205
Weigle, S. C (2002). *Assessing writing*. Cambridge University Press.
Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II–Validation activities. In E. V. Smith & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 243–290). JAM Press.
Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 426–443). Routledge.

Appendix A

Scoring Rubric with Level Descriptors

	Fluency	Accuracy	Strategy Use	Active Listening	Content
Five	Speaks at natural speed; only occasional hesitation at appropriate points; speech is easy to follow.	Vocabulary and grammatical structures used accurately; very few mistakes evident.	Uses strategies to effectively deal with real or potential communication breakdowns; confidently manages turn-taking.	Demonstrates active listening by asking open-ended questions, using natural reactions, and indicating (dis)agreement.	Gives and supports opinions effectively; uses appropriate discourse markers; can confidently initiate interaction.
Four	Speaks slightly below natural speed; occasional hesitation mid-sentence; speech generally easy to follow.	Vocabulary and grammatical structures sufficiently accurate to deal with all topics; mistakes rarely impede communication.	Attempts strategies to deal with real or potential communication breakdowns; sensitive to turn-taking.	Demonstrates active listening by asking questions, using natural reactions, and indicating (dis)agreement.	Gives and supports opinions generally effectively; usually uses appropriate discourse markers; can initiate interaction.
Three	Speaks slowly; noticeable hesitation at various points; sometimes demands patience from listeners.	Vocabulary and grammatical structures sufficiently accurate to deal with basic topics; mistakes occasionally impede communication.	Limited attempts to deal with communication breakdowns; turn-taking may be formulaic.	Demonstrates active listening by asking simple questions, using reactions, and indicating (dis)agreement.	Able to give and support opinions; sometimes uses appropriate discourse markers; can respond when prompted.
Two	Speaks very slowly; frequent hesitation at various points; frequently demands patience from listeners.	Very limited accuracy of vocabulary and grammatical structures; frequent mistakes.	Struggles to deal with communication breakdowns; turn-taking may be awkward and hesitant.	Demonstrates active listening by using reactions and / or indicating (dis)agreement.	Able to give simple opinions; may lack discourse markers; may struggle to respond when prompted.

Usability of a speaking assessment portal for Japanese teachers of English

Rie Koizumi, Makoto Fukazawa, and Chihiro Inoue
koizumi.rie.ge@u.tsukuba.ac.jp
fukazawa@edu.u-ryukyu.ac.jp
chihiro.inoue@beds.ac.uk
University of Tsukuba
University of the Ryukyus
University of Bedfordshire

Abstract

Against a backdrop of insufficient training for pre-service and in-service teachers, as well as limited access to materials and resources related to speaking assessment (SA), this study reports on the development and usability of an SA Portal, drawing upon the perceptions of teachers who used the website. The Portal is intended for Japanese senior high school teachers of English as a way to equip them with a wider range of relevant resources. It includes tips for conducting speaking tests; SA examples and explanations, including tasks, rubrics, and videos; and useful websites and resource. There were two phases in this preliminary usability study, and teacher perceptions were collected in each phase. We found that teachers received the content of the Portal positively. Teachers also provided numerous points for improvement from micro and macro levels. Most of these suggestions have been implemented in the Portal, while the remaining ones will be considered in the future. The practical implications of the Portal itself and the use of feedback from its users are also outlined. Specifically, soliciting input from users with diverse backgrounds, employing various open-ended questions, and allowing sufficient time for multiple revisions can lead to valuable feedback that contributes to effective improvements.

Keywords: second language speaking assessment, online resources, teacher training, rubric, teacher perception

Speaking assessment (SA) in classrooms is an indispensable element of language education (e.g., Poehner & Inbar-Lourie, 2020). It can be used for formative and summative purposes, helping both teachers and students understand the students' learning status, strengths, weaknesses, and other features. In this context of classroom-based speaking assessment, teachers act as the primary test developers, administrators, raters, and providers of feedback. The Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) strongly encourages second language (L2) English teachers in Japan to use performance-based SA in classrooms to assess knowledge and skills; thinking, judgment, and expression; and a proactive attitude towards learning (i.e., linguistic accuracy, content appropriateness, and willingness to communicate; National Institute for Educational Policy Research, 2023). However, the frequency and quality of SA need to be improved (Koizumi, 2022a, 2022b; Tando, 2023; see Kaneko, 2019, for teachers' voices for this issue). To enhance the quantity and quality of SA in Japan, two issues must be addressed with urgency (e.g., Koizumi, 2022b). First, there are insufficient opportunities for both pre-service and in-service Japanese teachers to receive proper training. Second, there is a lack of materials and resources related to SA, especially those freely available online in Japanese. As a result, teachers often lack opportunities to learn how to select appropriate SA formats and rubrics from various options and how to use them consistently and formatively (e.g., Koizumi, 2022b). Increasing the availability of resources for teachers would help enhance teachers' L2 assessment literacy by incorporating these resources into teacher training programs or help build consensus among teachers within and across their respective schools.

To address these issues, we developed a Speaking Assessment (SA) Portal (the Portal, hereafter), available online for Japanese teachers of English, particularly in senior high schools. It is expected that teachers will use the Portal for teacher training, independent self-study, and teacher meetings at local and regional

levels, and that they can enhance their L2 assessment literacy to implement SA effectively in the classroom. While reporting on the development of the Portal, we also present a preliminary usability study based on teachers' perceptions of the Portal, using two datasets.

Literature Review

Multiple online resources are available for L2 teachers and those interested in learning about language assessment and SA. For example, *Language Testing Resources Website* (Fulcher, 2024) has been prominent in disseminating essential knowledge, including useful videos, explanations, and discussion topics on language assessment in English. Assessment & Evaluation Language Resource Center, Georgetown University (2024) provides a summary of resources for teachers to learn about language assessment in English. British Council (2024b) provides a practical glossary and videos on language assessment in English. They also provide helpful videos in Japanese with a focus on the assessment of four skills (British Council, 2024a). British Council (n.d.) also hosts useful teacher training kits in English. The Japan Language Testing Association (n.d.) also hosts various functional workshop videos and online tutorials, primarily in Japanese. Another existing resource is Tools to Enhance Assessment Literacy (TEAL, 2024c), which focuses on broad aspects of language assessment, particularly in the context of teaching additional languages (e.g., Vietnamese and Tagalog) in Australia.

Many English testing resources are open access (e.g., *Language Testing Resources Website*; Fulcher, 2024), which help teachers acquire the fundamental knowledge of language assessment. Among these, TEAL (2024c) is considered the most beneficial in comprehensively providing not only guidelines on how to implement SA but also numerous concrete examples (TEAL, 2024a). These include 21 SA tasks (e.g., "Role play: Giving advice to a friend"); and for each task, a rubric; three to seven learners' videos; and a commentary for each video.

However, the usability of TEAL's (2024a) task and other examples are limited for Japanese teachers, primarily because MEXT recommends using a specific rubric format in classroom SA, which differs from that in TEAL (MEXT, 2022; National Institute for Educational Policy Research, 2023). Moreover, speech samples in the videos are not always similar to those typically produced by Japanese learners. Building on the need for SA resources that specifically cater to Japanese teacher audiences, we created the online Portal and conducted a usability study by obtaining feedback from teachers to refine the quality of the Portal.

The research questions (RQs) are as follows:

1. After using the Portal, how do Japanese teachers of English perceive its usefulness?
2. What do these teachers indicate as areas for improvement?

Method

Speaking Assessment (SA) Portal

Using TEAL as a model, we developed the Portal (https://sites.google.com/view/speaking-assessment/) including the principles and practices of developing and conducting SA (i.e., tasks, rubrics, speech samples, explanations of how they are scored), and resources for further learning. Figure 1 shows the top page of the Portal, which is freely accessible to anyone. As shown in Table 1, the Portal includes sections such as "Tips for conducting speaking tests," "SA examples and explanations," "Useful websites and resources," and "Frequently Asked Questions" on test development, administration, scoring, feedback, and other SA matters.

Figure 1
Top Page of the Speaking Assessment Portal

Table 1
Structure of Speaking Assessment (SA) Portal (as of September 2024)

Section	Content
1. Tips for conducting speaking tests [J]	How to develop speaking tests How to administer speaking tests How to score elicited spoken performance Types of feedback to provide
2. SA examples and explanations [J]	14 tasks (including both monologues and dialogues): role plays with a teacher, teacher-led interviews, oral interaction in pairs, and short speeches with questions and answers among paired students Each task includes a task description, a rubric, a worksheet, and six to 10 videos of learner speech samples per task. 120 videos in total[a]. Each video is edited to blur parts that could reveal personal information. Each video is accompanied by scores based on the rubric, a transcription, and a rationale for the given scores.
3. Useful websites and resources (to direct users to resources for further learning) [J]	Videos, scoring criteria, and sample scores from various speaking tests and assessment practices, with each video classified by proficiency levels (e.g., Graded Examinations in Spoken English [GESE] and Integrated Skills in English by Trinity College London)

	Resources for developing one's own speaking tests (e.g., task examples in various tests, analytical tools, Interactional Competence checklist [full and brief versions of Nakatsuhara et al., 2018, translated into Japanese], and materials created by municipal boards of education[b]) Resources for updating L2 assessment literacy related to SA (e.g., free online courses and resources such as Instructional Topics in Educational Measurement Series, and Language Assessment in the Classroom]) Resources for learning about automated scoring[b]
4. Frequently Asked Questions (FAQ)[b] [J]	Answering questions regarding test development, administration, scoring, giving feedback, and other SA matters
5. Research meetings[b] [J]	Language Learning Assessment Research Meetings
6. Digest of SA in Japanb [E]	Videos demonstrating and explaining optimal scoring practices for SA
7. Project members [J]	Introduction of the members who contributed to the development of the Portal

Note. [ ] = Language used; J = Japanese; E = English. [a] *In Phase 1, the Portal had five tasks and approximately five videos per task.* [b] *Uploaded after Phase 1.*

The Portal mainly differs from TEAL in that it specifically focuses on SA tasks and rubrics that adhere to MEXT guidelines, and speech sample videos with first language (L1) Japanese speakers learning English as an L2. The Portal is tailored to the Japanese context of learning English as a foreign language. It utilizes the learners' L1 and it addresses narrower ranges of English proficiency levels and learner profiles. Tasks vary from role plays with a teacher, teacher-led interviews, oral interaction in pairs, and short speeches with questions and answers among the paired students. Role-play tasks with teachers were originally developed as part of the CEFR-J project and linked to CEFR-J levels (see Tono & Negishi, 2020, for task development; see Koizumi, 2022a; Tono, 2022, for actual tasks).

Usability Study

The project to improve the website consisted of two phases, each involving Japanese teachers of English. In Phase 1 (April 2020 to March 2022), the Portal was planned and created by the authors, and tested by teachers through online questionnaires. The teachers' feedback was used for substantial revisions. We also presented the Portal and its development at a conference where we received additional feedback from the audience. In Phase 2 (April 2022 to November 2023), we further revised the Portal and sought feedback from another group of teachers. All instructions to the participants and feedback from them were provided in Japanese. All direct citations were translated from Japanese to English by the first author.

Participants and Procedures in Phase 1

We recruited six Japanese teachers of English with more than 10 years of teaching experience to participate in this study (Teachers A to F in Table 2). We intended to diversify the study participants to obtain feedback from various perspectives. An honorarium was provided, except for one participant who declined to receive it.

Table 2
Summary Statistics of Persons

Phase	Participant	Background (Approximate time spent in Phase 1)
1	Teacher A	Taught English in a senior high school (2 hours)
	Teacher B	Taught English in a senior high school (1.5 hours)
	Teacher C	Worked for a prefectural education in-service training center, responsible for training teachers; previously taught English in a senior high school (3 hours)
	Teacher D	Taught English language teaching at a university education department; previously taught English in a junior high school (3 hours)
	Teacher E	Retired from a university, specialized in language assessment; previously taught English in a senior high school (2 hours)
	Teacher F	Worked for a private company after teaching English in a junior high school (5.5 hours)
2	Teacher G	Taught English in an elementary school
	Teacher H	Taught English in a junior high school
	Teacher I	Taught English in a senior high school

The six teachers in Phase 1 answered Questionnaires 1 and 2, which included closed- and open-ended questions (see Appendices A to D for questions in Japanese and English). First, they watched a video with instructions regarding what they were going to do. They were informed that the main target users were senior high school English teachers, although the Portal may also provide useful information to English teachers at other types of schools. They were requested to answer as if they were teachers who administered speaking tests to students and to examine the usefulness, appropriateness, and ease of content to improve the quantity and quality of the Portal. Second, they were asked to spend 30 minutes browsing through the overall Portal and wrote their opinions and suggestions for improvement in Questionnaire 1. Third, they were requested to spend approximately 1.5 hours reading "SA examples and explanations" and answering the questions in Questionnaire 2. The participants spent approximately 1.5 to 5.5 hours reading the Portal and answering all the questions.

Participants and Procedures in Phase 2

We solicited additional feedback from three teachers (Teachers G to I in Table 2). Although we intended to create the Portal to primarily cater to senior high school teachers, we also included elementary and junior high school teachers to explore the potential usefulness and challenges of expanding our focus. We asked the teachers to provide an overall impression of the Portal focusing on its useful aspects and those that need to be improved. They presented their perspectives in a PowerPoint file and discussed their opinions at an online research meeting. An honorarium was provided afterward.

Analysis in Phases 1 and 2

Responses to the closed questions in Phase 1 of Questionnaires 1 and 2 were tallied. Verbal feedback in the open-ended format in Phases 1 and 2 was analyzed thematically.

Results and Discussion

Overall Perceptions of the Portal (RQ1)

We generally received positive responses from teachers in both Phases 1 and 2. Therefore, we report the results from both phases together in this section. As shown in Table 3, most teachers in Phase 1 found some of the content interesting and appropriate. "SA examples and explanations" was considered the most interesting by five teachers, followed by "Tips for conducting speaking tests" and "Useful websites and resources," each selected by four teachers. Similarly, most teachers found "SA examples and explanations" and "Useful websites and resources" the most appropriate (five teachers), followed by "Tips for conducting speaking tests" (three teachers).

Table 3
Number of Teachers Who Found the Portal Content Interesting and Appropriate in Phase 1

	Tips for conducting speaking tests	SA examples and explanations	Useful websites and resources	Project members
Interesting[a]	4	5	4	1
Appropriate[b]	3	5	5	2

Note. n = 6.[a] Based on Questionnaire 1, Item 2.[b] Positive comments were tallied based on responses in Questionnaire 1, Items 3 to 5 and Questionnaire 2, Item 3. See Appendices A to D for actual items.

Here is an example comment from a teacher, regarding SA examples and explanations:

> The videos provided cover various task formats to a certain degree. After understanding the key points about SA through the five tasks, teachers should be able to adapt the format to other tasks with different topics and situations. All the scoring procedures—specifically, how teachers use the rubric to score—are easy to understand with the provided transcriptions and rationales for the scores. The edited conditions of the videos were useful. Although parts of the videos were blurred, the atmosphere during the interaction was easily understandable. (Teacher D)

Teachers C and E mentioned that "Useful websites and resources" help teachers understand how to conduct interviews by providing level-specific videos and scoring rubrics from speaking tests in other countries. Teacher C also noted that this type of online resource is much more useful and effective than paper-based booklets.

In Phase 2, three teachers provided positive feedback on the topics and content of the Portal as found below:

> Essential points, such as how to conduct speaking tests, are summarized on the Portal. Many sample videos help teachers visualize the tests. Teachers can also learn how to score the tests by reading the transcriptions. (Teacher H)

> By watching video explanations of scores in each video, teachers can learn how to set a scene and situation in a speaking test and how teachers can respond and ask questions according to students' proficiency levels and their reactions. By reading the guidelines on the Portal, it was easy to see how to administer and score speaking tests. For example, normally it is difficult for teachers to find time to explain SA tests and conduct them during class time. However, the explanations on the Portal were useful for creating speaking tests with a balanced focus on validity, reliability, and practicality. Even for current teachers, the Portal can supplement insufficient in-service teacher training. (Teacher G)

> The explanations on development, administration, scoring, and giving feedback, as well as example rubrics and CEFR information, could be helpful when teachers attempt to relate teaching with assessment in the classroom and create tasks in their own contexts. The Portal helps teachers grasp the gist of SA before reading Koizumi (2022a) in detail. (Teacher I)

Areas for Improvement That We Addressed and Will Address (RQ2)

While teachers had a positive view of the overall design and content of the Portal, they also offered suggestions for further modifications during Phases 1 and 2. Below, we summarize these suggestions according to categories, rather than by phases, using direct quotations from the teachers. We also explain how we addressed these suggestions under "Solution." Areas for improvement that we have not yet addressed are summarized under "Future plan."

Overall design (from Phase 1)

- Making explanations easy to read
    - I understand that writers try to use simple language, but descriptions are sometimes difficult for high school teachers to understand. (Teacher D)
    - There are many words on the page. There should be a blank line between (1) and (2). (Teacher F)
        - Solution: We decreased technical terms, simplified the language, and added blank lines between the items. We also provided a subsection "Further reading" for those interested.
- Ensuring consistent use of terms:
    - Terms are not always consistent, such as *raters* and *scorers* (i.e., *hyokasha*, *saitensya*). (Teacher F)
        - Solution: We revised the site to use the term *saitensya* and other terms consistently.
- Using a clear and consistent layout:
    - The layout must be consistent to ensure a unified atmosphere. (Teachers D and F)
    - Font sizes should be larger and consistent across sections. (Teachers C and D)
    - Color and fonts should be used to show the highlights and make reading easy. (Teacher D)
    - The layout in 'SA examples and explanations' with bar buttons to show URLs is much clearer than 'Tips for conducting speaking tests.' The addition of illustrations is helpful. (Teacher D)
    - Regarding 'SA examples and explanations,' having a bar button in blue for Example 1 would enhance handling ease. Having red letters corresponding to the criteria, rather than simple black letters, makes it easier to understand. (Teacher D)
    - Illustrations in 'Useful websites and resources' are too large (Teacher F)
        - Solution: We modified the layout and used consistent colors and fonts across sections and items.
- Avoiding mechanical errors:

- While I understand that the Portal is under construction, there are many noticeable typos. (Teacher A)
    - Solution: We reviewed and corrected all content.

- Providing access to videos:
    - YouTube videos cannot be viewed in teacher rooms in certain prefectures. Although some teachers may be able to watch them on tablets, many schools only have tablets for students. Some teachers found it difficult to watch the YouTube videos. (Teacher B)
        - Solution: We added an explanation to the FAQ section that videos can be obtained by contacting us.
    - I could not view the videos in French, in the 'Useful websites and resources.' section (Teacher C)
        - Solution: We corrected the URL.

Suggestions for Tips for conducting speaking tests in the Portal (from Phase 1)

- Providing a brief introduction of speaking tests:
    - A section is needed to briefly explain what performance tests look like. Including various test formats and patterns, such as student-teacher and student-student conversations, would help young teachers transition to other more detailed pages. (Teacher D)
        - Solution: We provided this information.

- Explaining how to modify tasks and rubrics:
    - (Scores based on a rubric are provided for each video performance.) One of the biggest concerns is that teachers at lower-level schools may consider speaking tests unmeaningful when they see a video where all scores are cs [out of a, b, and c, with c being the lowest score]. They may think that all of their students would receive similar scores. Providing an explanation of how to modify the rubric and guidelines to conduct speaking tests according to their context would be helpful. (Teacher C)
        - Solution: We added an explanation to the FAQ section of the Portal.

- Explaining how to conduct speaking tests:
    - It is ideal to have case scenarios for conducting speaking tests for approximately 40 students per class, with five classes in one school per year. Conducting speaking tests to measure interaction (dialogues) and presentation (monologues) efficiently and fairly would be helpful. (Teacher C)[1]
    - Considerations for teachers to conduct a role play with a student would be helpful in understanding basics (e.g., how to take a neutral stance to avoid any effects of teachers on student performance; how to create a supportive atmosphere in which students can speak well). (Teacher F)
        - Solution: We added explanations to the FAQ section.

- Explaining how to score speaking and reach final scores:

    - It is easy to judge the number of sentences to score, but it is difficult to judge which is better: two short sentences without conjunctions or a long sentence with conjunctions. The explanation would be helpful in this regard. (Teacher D)
        - Solution: We added an explanation to the FAQ section.
    - An explanation and example video performances on how to finalize scores when they differ across teachers would be helpful. (Teacher D)
        - Solution: We added an explanation in the subsection of Scoring speaking tests.
        - Future plan: We will consider including examples in the future.
    - I would like to know how and where teachers diverge in scoring, even after discussing the rubric beforehand. (Teacher A)
        - Future plan: We will address this in the future.
- Explaining matters related to MEXT evaluation guidelines:
    - Having only three levels for all evaluations was unreasonable. Therefore, a fine-grained evaluation would be more appropriate. (Teacher E)
    - Evaluating the willingness to communicate is difficult. Students usually try to speak during a test, so they will eventually receive Score b [out of a, b, and c, with c being the lowest score]. (Teacher E)
        - Solution: Because the Portal follows MEXT's evaluation guidelines, we explained it as is. We also noted this on the top page of 'SA examples and explanations' and added it to the FAQ section.
- Explaining how to create videos or recordings:
    - An explanation would be helpful for technical matters important in developing and administering speaking tests, such as how to videotape and record performance. (Teacher E)
        - Solution: We added an explanation to the FAQ section.
- Providing concrete examples:
    - Examples of feedback explanations are needed, such as samples of feedback on sheets and a video on giving oral feedback, and examples and explanations of score reports, which would help teachers understand the image of this activity. (Teacher F)
        - Solution: We added the explanation to the section.
- Providing resources for further learning:
    - The Portal says that feedback should include not only the current speaking ability, but also how to improve it. Any website that helps increase speaking ability and is accessible to students during self-study would be helpful. (Teacher F)
        - Solution: We added an explanation to the FAQ section.
- Providing downloadable materials:

- I would like to have rubrics and worksheets (also feedback sheets and reflection worksheets) downloadable in Excel and Word, which I can modify according to my context. This will save time in developing them myself. (Teacher A)
- It is important to score while watching the videos. Providing a scoring worksheet would help teachers individually and, in a group, allow them to write scores and rationales. (Teacher C)
    - Solution: We uploaded the files to the Portal.

Suggestions for SA examples and explanations in the Portal (from Phase 1)

- Providing visual aids:
    - Along with verbal explanations, a flowchart of explanations and a video explaining SA procedures would be helpful in catering to teachers' individual needs and preferences. (Teacher D)
        - Future plan: This should be addressed in future revision.
- Providing additional examples:
    - Having examples from both analytic and holistic rubrics would be helpful. (Teacher E)
        - Solution: We included analytic examples in the format provided by MEXT as part of the test specification examples. We will consider including holistic rubrics in the future, but analytic rubrics would generally fit the teaching context in Japan.
- Adding an interactive mode:
    - It may be useful to have a section in which teachers can input their scores and check whether their scores are correct as part of the practice. This gamification may enhance teachers' interest. (Teacher C)
        - Solution: We considered this option but decided not to include it because such a function might imply that there are absolutely correct answers in scoring performance, which is not our intention. Since scoring rubrics should be tailored to students and various classroom contexts, our focus is on presenting the principles, possible options, and examples of SA practice.
- Changing the order of items:
    - Reading a rubric before watching a video is intuitively easy to understand. I do this during self-training and discuss the criteria with my colleagues. (Teacher A)
        - Solution: We changed the order to make the website more user-friendly for teacher training.
- Modifying the length of task explanation in the video:
    - Having a long time to read the task description is unnecessary. (Teacher E)
        - Solution: Originally, each slide was shown for 14 seconds, but we shortened it to 7 seconds in the videos.
- Improving video quality:

- Some videos were difficult to hear due to the recording quality and the students' voice volumes. (Teacher F)
    - Solution: We added the explanation to the FAQ section.[2]

Suggestions for SA examples and explanations in the Portal (from Phase 2)

- Providing more fine-grained task specifications:
    - Task descriptions are broadly written to make tasks generalizable across contexts, although some tasks have specific conditions. According to the Course of Study or the MEXT curriculum guidelines, setting a clear purpose (why you need to do this), scene (in what scene do you talk to), and situation (to whom you are talking) in which students need to communicate in English is important. It may be necessary to emphasize the need for teachers to set concrete and detailed purposes, scenes, and situations while considering class activities and observing students' reactions, NOT using the same task and the rubric from the Portal. (Teacher I)
        - Solution: We added an explanation to SA examples and explanations.[3]
- Providing a wider range of tasks, rubrics, topics, and examples:
    - The current Portal has more interactive tasks, which is nice. However, more monologic tasks would be helpful, as more teachers conduct monologic speaking tests. Furthermore, junior and senior high school students are expected to work on both daily and social topics. The Portal has more daily or familiar topics, and more examples of social topics, such as environmental issues, racial discrimination, and technology, which appear in textbooks, would increase its value. (Teacher I)
    - While I understand that the Portal is mainly intended for senior high school teachers, the SA format and rubric examples are beyond the level of elementary school students. If easier examples are provided, this will be more helpful. Furthermore, more tasks would allow teachers to understand task variation, such as using class interaction as part of assessment and evaluating recordings submitted by students. (Teacher G)
        - Future plan: We will consider including such tasks, rubrics, and examples in the future to cater to various needs.
- Providing additional examples and clarifications for the rubric:
    - Regarding 'Willingness to communicate' in the rubric, it is difficult to understand what behaviors and utterances are measured in evaluating students 'trying to communicate to the partner(s),' although this may depend on each school's context. Regarding "Content appropriateness" in the rubric, questions arise as to (a) whether utterances need to be sentences, not fragments, and (b) which is evaluated more highly: detailed utterances with grammatical errors OR brief utterances with correct grammar. More detailed examples of the rubric would also be helpful. (Teacher H)
        - Future plan: We will include such examples and clear explanations, although each teacher or school needs to plan practices themselves eventually.
- Explaining how to select representative videos:
    - It is great that the Portal contains 120 videos. However, it is difficult to watch all of them. It might be helpful to show a selection of a few tasks first or to display only the first video,

with the second one appearing after watching the first. This could make it easier for busy teachers to navigate. (Teacher G)

- Solution: We added an explanation to the FAQ section on how the videos are categorized and how a single video can be selected for viewing.
- Future plan: We will further consider creating a suggested entry point.

Suggestions for Useful websites and resources in the Portal (from Phase 1)

- Explaining technical terms:
  - I hope to read more explanations in 'Useful websites and resources' on, for example, what GESE is, and what it does. (Teacher D)
    - Solution: We included more information in a concise manner for teachers.
- Providing visual aids:
  - A concise table of the CEFR levels at the top of the page would be helpful. Teachers would like to examine the relationship between such levels and high school students' first- to third-year levels. (Teacher D)
    - Solution: We added a table along with Eiken grade information.
- Providing additional materials:
  - Practical, leading-edge examples from across Japan would be helpful. (Teacher D)
    - Solution: We included a summary of good practices and useful resources from the websites of municipal boards of education. We also added information on cutting-edge research such as automated scoring.

Suggestions for publicizing the Portal (from Phase 1)

- Taking strategic measures to publicize the Portal:
  - Taking strategic measures to publicize the Portal would attract more visitors. An example is asking municipal boards of education and educational centers across Japan to promote the Portal and actively use its contents in teaching training. Only a limited number of teachers read the monthly *English Teachers' Magazine* (by Taishukan Publishing). Annual training for first-, fifth-, and tenth-year teachers would be particularly beneficial. (Teacher C)
    - Future plan: Contacting university teachers who teach in programs that offer teaching certification courses may also be helpful. Therefore, these measures should be considered in the future. Some teachers already found the Portal and contacted us or reported using it, so we should also check for missing groups to be contacted.

As seen above, the teachers' comments focused on both micro and macro levels. The micro-level feedback included suggestions on visual design and the use of simple language to make the resources more user-friendly and enhance readers' understanding. The macro-level feedback involved recommendations for adding more value to the Portal, such as providing explanations on unexpanded topics and increasing awareness among the intended readers. Thanks to the productive feedback from teachers in Phases 1 and 2, we identified additional areas for improvement, detailed as follows:

- Adding visual aids:
  - We can include a flowchart that helps teachers select an appropriate task example and decide on their test specifications by choosing the ability to be measured, the test format, and/or test requirements.
  - We can hide scores and explanations when teachers watch the videos, revealing them by clicking on a bar. This would allow teachers to focus on watching the video and scoring them by themselves.
- Adding various examples:
  - We can include various rubrics and score examples based on a single video (e.g., providing cases of strict and lenient rubrics and scoring decisions).
  - We can include videos showing how students develop their speaking abilities over time, allowing teachers to intuitively understand the students' longitudinal progress (see Tamura, 2022, for such videos).
  - We can include content to help teachers to understand how students' spoken utterances differ depending on test formats during the same period.
- Adding a test task bank:
  - We can include a bank of test tasks and a rubric (i.e., test task bank), which is a concept similar to the Task-Based Language Teaching [TBLT] Language Learning Task Bank (Indiana University, 2024). However, a test task bank would differ from the TBLT Task Bank by providing information on test difficulty and other measurement details (see Koizumi, 2022b).
- Adding an interactive platform:
  - We can include an interactive platform for communication between the Portal developers and teachers, as well as teachers. One idea is to create a page for benchmarking criteria, asking questions, and sharing experiences and information, accessible only to registered teachers. This would allow for open discussions among teachers in a closed forum, similar to TEAL's (2024b) discussion forum.

Conclusion

We developed a Speaking Assessment (SA) Portal to address the needs of Japanese senior high school English teachers for online SA resources that can be used for teacher training and self-study. The Portal includes various videos, each consisting of a task, a rubric, rubric-based scores, a transcription, an explanation of the scores, and a worksheet in the "SA examples and explanations" section. Other sections include "Tips for conducting speaking tests" and "Useful websites and resources."

We then examined the usability of this Portal by using feedback from teachers who accessed the website to assess its quality and identify areas for improvement. The first research question explored how Japanese teachers of English perceive the Portal's usefulness after using it. The responses in Phase 1 indicated that most teachers viewed the Portal positively. In particular, "SA examples and explanations" was considered the most interesting and appropriate by the intended users.

The second research question investigated what areas of the Portal these teachers identified for improvement. Numerous suggestions were made, ranging from adding more information, materials,

examples, and visual aids, to including more diverse types of task formats and rubrics. We have addressed most of these suggestions from the participating teachers and plan to further develop the content and functionality of this Portal website.

There are limitations in the current preliminary usability study. First, we gathered feedback from a relatively small number of teachers. Involving a more diverse group of participants could provide a wider range of perspectives useful for revision. Second, we did not employ extensive questionnaires or interviews to collect data on teacher perceptions. A more comprehensive approach, using a mixed-methods research design to gather teachers' perceptions from various viewpoints, would provide more detailed insights and help improve the Portal (see Shen et al., 2015, for a method example).

Regarding the practical implications derived from the current study, the Portal can be useful for teacher training and independent study, as indicated by teachers' perceptions. Moreover, asking intended resource users to provide feedback through various open-ended questions is critical. To address teachers' interests and concerns, content developers should involve users from diverse backgrounds (e.g., current and former teachers, teacher trainers, teachers with and without knowledge of language assessment and speaking assessment, as was done in the current study). To effectively utilize the feedback from users, content developers need to plan ahead and allocate sufficient time to receive comments and revising resources at multiple stages of development. These efforts would facilitate communication between content providers and users, ultimately benefiting the dissemination of content—in this case, SA principles and practices.

Acknowledgement

This study was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Scientific Research (C), Grant Number 20K00894. We would like to thank Yuichiro Yokouchi and Masaru Yamamoto for their valuable contribution to the project, and Masashi Negishi, Yukio Tono, Emiko Kaneko, and other CEFR-J team members for letting us use CEFR-J tasks. We also appreciate the assistance and advice from Yumi Koyamada, Yoko Okamoto, Yo In'nami, as well as the students in the videos, and the teachers who provided their opinions in response to our usability survey.

Notes

1. The following is the description from the FAQ section of the Portal website.

- How can we assess speaking interaction (dialogues) and presentation (monologues) efficiently and fairly when there are approximately 40 students per class, with five classes in one school per year?
    - It is important to create a yearlong plan and decide when to conduct speaking tests in relation to teaching. Tests should be planned to focus on validity while also considering reliability and practicality. When considering practicality, the following questions arise: (a) How many lessons can be used for a class with 40 students? (b) Can tests be scored outside of class time? and (c) How many minutes per student can be allocated for conducting speaking tests?

      Based on answers to these questions, it is possible to select one out of four patterns, as described in "Administration of speaking tests" ((v) in-class administration and in-class scoring, (x) in-class administration and out-of-class scoring, (y) out-of-class administration and scoring on the spot, and (z) out-of-class administration and scoring after the test; see Koizumi, 2022b, p. 154). Using the selected option, we can concretely decide on a test format (e.g., teacher-led interviews, pair conversations, group discussions to measure oral interaction) and a rubric. We will also decide

whether to focus on presentation or interaction, as well as what specific abilities we would like to measure while considering teaching objectives and what activities are conducted in lessons.

For example, when we can only use (a) one 50-minute lesson (b) with tests scored within the lesson, we can use only 40 minutes for a speaking test because 10 minutes are needed for explanation. Then, we can use (c) one minute per student (maximum of 40 seconds for speaking time). Test formats that align with this requirement include speech to measure monologues, and group discussions to measure interactions. In Matsuo's practice (Matsuo, 2019), all 40 group discussions are tested and scored within a 50-minute lesson.

Regarding the number of speaking tests to be conducted in a year, some schools have one per term with three speaking tests in total in a year. Others have speaking tests around the time of the term tests, so they have five speaking tests in total per year.

Koizumi (2022a) shows speaking test samples conducted in the second year at a school and discusses how teachers maintain validity and reliability.

2. The following is the description from in the FAQ section of the Portal website.

- Could you please improve the situation where sample speeches are difficult to hear in some videos due to recording conditions and students' voice volumes?
    - Some videos were difficult to hear because actual test videos were used on the Portal. However, in real-test scoring, videos that do not have ideal conditions are still scored, so this difficulty may reflect real-life situations. Although we have recordings made with voice recorders and could overlay those sound files onto the videos, we chose not to do so and kept the original video sounds. This is because sounds that are undetectable by the ears may be picked up by voice recorders, which can differ substantially from what is heard during speaking tests.

3. Teacher: I provided an example of the club activity task of the SA Portal website.

- Current description:
    - Role play task: Talking about club activities and hobbies
    - Setting:
        - Teacher: An international student in the same class. The student wants to join a club activity, so they ask questions.
        - Student: A student who wants to make friends with the international student. Answers questions and asks questions.

In addition, more can be added to the situation related to an international student, as follows:

- Additional setting (see italics for the addition):
    - Teacher: An international student in the same class. The student wants to join a club activity, so they ask questions. They want to experience Japanese culture (or a sport specific to Japan) in a club activity. They do not have much money and would like to join a club without the need to buy tools for the club.

References

Assessment and Evaluation Language Resource Center (2024). *Resources*. https://aelrc.georgetown.edu/resources/

British Council. (2024a). *Doga de miru eigo yongino hyoka no pointo* [Tips for assessing English four skills through videos]. https://www.britishcouncil.org/exam/english/aptis/research/assessment-literacy

British Council. (2024b). *How language assessment works.* https://www.britishcouncil.org/exam/english/aptis/research/assessment-literacy

British Council. (n.d). *Language assessment resources*. https://www.teachingenglish.org.uk/professional-development/teachers/assessing-learning/articles/language-assessment-resources

Fulcher, G. (2024, September 11). *Language testing resources website*. https://languagetesting.info/

Indiana University (2024). *TBLT language learning task bank*. https://tblt.indiana.edu/tasks/index.html

Japan Language Testing Association. (n.d.). *Workshop video/web tutorial*. https://jlta2016.sakura.ne.jp/?page_id=808

Kaneko, J. (2019). *Chuko renkei wo fumaeta eigo jugyo niokeru activity to performance test kaihatsu nikansuru chosakenkyu* [Survey research into activities and performance test development for English lessons based on collaborations between junior and senior high schools]. 2018 report of a survey research project on teacher pre-service training. https://www.gakushubunka.jp/scholarship/kenkyugaiyou_kaneko_h30.pdf

Koizumi, R. (Ed.). (2022a). *Jitsurei de wakaru eigo speaking test sakusei gaido* [A practical guide for developing English speaking tests]. Taishukan Publishing.

Koizumi, R. (2022b). L2 speaking assessment in secondary school classrooms in Japan. *Language Assessment Quarterly, 19*(2), 142–161. https://doi.org/10.1080/15434303.2021.2023542

Matsuo, M. (2019). *Jirei hokoku: Tesuto ga totatsumokuhyo to shido niataeru eikyo: Semina repoto* [Case study: Effects of tests on course goals and teaching: Seminar report]. British Council. https://www.britishcouncil.jp/programmes/english-education/japan/report/assessment2018-seminar/case1

MEXT (Ministry of Education, Culture, Sports, Science and Technology). (2022). *Shido to hyoka no ittaika nimuketa koko gaikokugoka niokeru pafomansu tesuto sanko shiryo (shidosya yo shiryo)* [Reference guides of performance tests in teaching foreign languages at upper senior high schools: Toward the integration of teaching and assessment: Materials for instructors]. https://www.mext.go.jp/content/20220705-mxt_kyoiku01-1000021347-1.pdf

Nakatsuhara, F., May, L., Lam, D., & Galaczi, E. (2018). *Learning oriented feedback and interactional competence* (*Research Notes*, Vol. 70). Cambridge Assessment English. https://www.cambridge-exams.ch/research-notes-issue-70-learning-oriented-feedback-and-interactional-competence

National Institute for Educational Policy Research. (2023). *Shido shiryo jirei shu* [Handbook of teaching materials and case examples]. https://www.nier.go.jp/kaihatsu/shidousiryou.html

Poehner, M. E., & Inbar-Lourie, O. (Eds.). (2020). *Toward a reconceptualization of second language classroom assessment: Praxis and researcher-teacher partnership*. Springer.

Shen, H., Yuan, Y., & Ewing. R. (2015). English learning websites and digital resources from the perspective of Chinese university EFL practitioners. *ReCALL, 27*(2), 156–176. https://doi.org/10.1017/S0958344014000263

Tamura, T. (2022). *Korede wakaru, dekiru! Shogakko gaikokugo pafomansu tesuto: Hanasukoto yaritori no hyoka* [Understand and master it! Performance tests for a foreign language English in elementary schools: Assessing speaking Interaction; DVD].

Tando, H. (2023). The investigation of issues of speaking performance evaluation in elementary schools in Aomori Prefecture. *TELES (Tohoku English Language Education Society) Journal*, *43*, 84–96. https://doi.org/10.57539/telesjournal.43.0_84

TEAL (Tools to Enhance Assessment Literacy). (2024a). *Common oral assessment tool*. https://teal.global2.vic.edu.au/assessment-tools/common-oral-assessment-tasks/

TEAL. (2024b). *Discussion forum*. https://teal.global2.vic.edu.au/discussion-forum/

TEAL. (2024c). *Tools to Enhance Assessment Literacy for teachers of English as an additional language*. https://teal.global2.vic.edu.au/

Tono, Y. (2022). *CEFR-J CAN-DO tesuto: Sampuru bajon* [CEFR-J CAN-DO test: Sample version] (Version 1.0). Retrieved March 17, 2024, from https://www.cefr-j.org/download.html#cefrj_testasks

Tono, Y., & Negishi, M. (Eds.). (2020). *Kyozai tesuto sakusei notameno CEFR-J risosubukku* [The CEFR-J resource book: Reference level descriptions and test development]. Taishukan Publishing.

Appendix A

Questionnaire 1 in Japanese

Note. This was answered after reviewing the overall site.

1. お名前をお願いします。
2. Website の中で、最初にパッと見てみて、面白い、読んでみたいと思ったものを選んでください（複数回答可）。

   A. スピーキングテストのコツ
   B. スピーキングテストの実例と解説
   C. 役立つ Website・参考資料
   D. プロジェクトメンバー紹介

3. 「スピーキングテストのコツ」について感想をお願いします。（例：だいたい内容は知っていた。〜についてさらに知りたい。レイアウトは〜だ）
4. （「スピーキングテストの実例と解説」については、後で詳細に見ていただきますので、飛ばします。）「役立つ Website・参考資料」について感想をお願いします。（例 : だいたい内容は知っていた。〜についてさらに知りたい。レイアウトは〜だ）
5. 「プロジェクトメンバー紹介」について感想をお願いします。（例：だいたい内容は知っていた。〜についてさらに知りたい。レイアウトは〜だ）
6. 本 Website を、高校の先生方などに使っていただくために、何かあったらよいと思う内容や機能はありますか？あれば書いてください。

7. 他に何か感想かご意見があればよろしくお願いいたします。

Appendix B

Questionnaire 1 in English

Note. This was answered after reviewing the overall site. It was translated into English by the first author.

1. Please write your name.
2. Please select all the items that you found interesting or would like to read. (Multiple answers were allowed.)

   A. Tips for conducting speaking tests

   B. SA examples and explanations

   C. Useful websites and resources

   D. Project members
3. Please write your impression about "A. Tips for conducting speaking tests" (e.g., "I knew almost all the content"; "I want to know more about …"; "The layout is ….")
4. (Please skip "B. SA examples and explanations." You will be asked to read it later.) Please write your impression about "C. Useful websites and resources." (e.g., "I knew almost all the content"; "I want to know more about …"; "The layout is ….")
5. Please write your impression about "D. Project members." (e.g., "I knew almost all the content"; "I want to know more about …"; "The layout is ….")
6. Please write any content or functions, if any, that this Portal should have to facilitate the use from senior high school teachers and others.
7. Please write any other opinions.

Appendix C

Questionnaire 2 in Japanese

Note. This was answered after reviewing "B. SA examples and explanations."

1. お名前をお願いします。
2. 「Speaking test の実例と解説」の中でご覧になったタスクを選んでください（複数回答可）。

   A. 教員と Role play 映画　タスク 1：映画に誘う (CEFR-J A1.3)

   B. 教員と Role play 道案内 タスク 2：道案内をする (CEFR-J A2.1)

   C. 教員と Role play 学校　タスク 3：学校を紹介する (CEFR-J B1.1)

   D. 教員と Role play 教育　タスク 3：子どもの教育の改善を提案する (CEFR-J B2.1)

   E. 話すこと（発表・スピーチ）・話すこと（やり取り・ペアでの質疑応答）

3. 「Speaking test の実例と解説」について全体的な感想をお願いします。（例：だいたい内容は知っていた。～についてさらに知りたい。レイアウトは～だ）
4. 「Speaking test の実例と解説」の中の細かな点について気になった点等書いてください。
5. 「Speaking test の実例と解説」を、高校の先生方などに使っていただく（例：自己研究、校内研修、スピーキングテスト実施前の打ち合わせ）ために、何かあったらよいと思う内容や機能はありますか？あれば書いてください。
6. 他に何か感想かご意見があればよろしくお願いいたします。
7. 今回の Website 確認にかけてくださった時間はどのくらいですか？（例：～分、～時間）

（謝礼に関する質問は、省略）

Appendix D

Questionnaire 2 in English

Note. This was answered after reviewing “B. SA examples and explanations.” It was translated into English by the first author.

1. Please write your name.
2. Please select all the tasks that you saw in the “B. SA examples and explanations.” (Multiple answers were allowed.)

   Task 1: Inviting your friend to see a movie (CEFR-J A1.3)

   Task 2: Showing the way (CEFR-J A2.1)

   Task 3: Introducing your school (CEFR-J B1.1)

   Task 4: Suggesting a way to improve child education (CEFR-J B2.1)

   Task 5: Making a speech and asking questions and answer them in a pair
3. Please write your overall impression about “B. SA examples and explanations.” (e.g., “I knew almost all the content”; “I want to know more about …”; “The layout is ….”)
4. Please write any points in detail to improve the site in “B. SA examples and explanations.”
5. Please write any content or functions, if any, that this Portal should have to facilitate the use from senior high school teachers and others (e.g., self-study, within-school training, meeting before the administration of a speaking test).
6. Please write any other opinions.
7. How long did you spend reading the Portal and writing your opinions? (e.g., … minutes, … hours)

(Other questions related to honorariums were omitted here.)

A comparison of text analysis tools: Levels of agreement and disagreement

Mart Christine Johnston
martchristinevito2017@gmail.com
Takushoku University

Abstract

This paper explores agreement levels among text analysis tools and factors influencing text difficulty using 75 Grade 2 EIKEN texts. EIKEN is a Japanese proficiency test for high school graduates. Text analysis included word count, average sentence length, CEFR, CEFR-J levels (determined by Text Inspector and CVLA, respectively), Flesch Reading Ease Score, Flesch-Kincaid Grade Level, Lexile score, and total coverage (from AntWordProfiler) using the New General Service List (NGSL). Average sentence lengths had stronger correlations with other indices than word counts. Flesch-Kincaid Grade Level and Lexile reading levels correlated moderately due to reliance on sentence length. According to both tools, the texts are generally deemed appropriate for grades 8-9 in the US education system. Considering that text levels for second language classrooms are typically several levels lower than those in the US education system and recognizing that EIKEN Grade 2 is intended for high school graduates in Japan, it can be inferred that the texts align with expectations for high school graduates in Japan, equivalent to grades 8-9 in the US education system. CEFR and CEFR-J levels, although their text level assignments were similar, had only moderate correlations, reflecting metric differences between Text Inspector and CVLA. AntWordProfiler's total coverage showed weak correlations, focusing solely on word frequency. The results from this study show clear discrepancies in text difficulty depending on the type of measure used and call for varied approaches to text analysis.

Keywords: text analysis, EIKEN, text difficulty, text measures, Text Inspector, CVLA, AntWordProfiler, NGSL

The aim of reading in one's first (L1) and second (L2) languages is similar: to find the meaning of the text (Nuttall, 1996). However, languages differ in many ways, such as orthographies, phonologies, and morphologies, which can affect how L2 readers process L2 texts (Grabe and Yamashita, 2022). Some ways to address this L2 reading issue are creating graded readers (Waring, n.d.) and developing basal reading programs (Ocampo, 1997). Although young children who are native speakers of English, for example, may benefit from graded readers as they begin to learn how to decode, L2 readers may struggle more with comprehension than L1 readers. As such, graded readers are designed to control some variables of text difficulty, such as vocabulary items, sentence structures, and even text length, to avoid overestimation of the types of texts and difficulty levels L2 readers can process.

Literature Review

While a wide variety of metrics for text difficulty are available, it is not clear if they all classify texts in the same way. Hermosa (2002) argues that different formulas are expected to obtain different scores when other variables are considered. Student perceptions were used by Holster et al. (2017) and Arai (2022) to prove that correlation of indices to difficulty varies significantly.

To test such claims, it would be of benefit to run the same texts through multiple schemes used to classify to see if the ratings they give are consistent. Online text analysis tools have made it easier to determine readability and assess difficulty in vocabulary use and sentence structures. This study seeks to evaluate the level of agreement among a number of online tools available and identify factors influencing text difficulty using EIKEN Grade 2 texts as its sample.

Measures of Text Difficulty

This literature review starts with an overview of EIKEN and its relationship with the CEFR, or Common European Framework of Reference for Languages. It is followed by the introduction of CEFR and CEFR-

J levels based on metrics provided by online analysis tools, namely Text Inspector (n.d.) and CEFR-based Vocabulary Level Analyzer ver. 2.0 or CVLA (Uchida, n.d.). Other readability formulas to be the discussed are Flesch Reading Ease (FRE), Flesch Kincaid Grade Level (FKGL), Lexile, and token coverage.

EIKEN and CEFR

Dunlea and Matsudaira (2009) determined student performance on Pre-1 and Grade 1 EIKEN tests with the abilities described at each level in the CEFR. Their results indicated that students who passed the Grade 1 test exhibited strong performance described at the CEFR C1 level, while those who passed the Grade Pre-1 test were at the CEFR B2 level. It is then assumed that EIKEN Grade 2 passers correspond to B1 or one level below. However, this assumption warrants further research for confirmation. The present study, however, does not examine test scores or test-takers' abilities, nor does it refer to the Council of Europe's Manual. Instead, it will use Text Inspector to analyze EIKEN Grade 2 passages. Developed by a professor of Applied Linguistics, Stephen Bax, this online tool was chosen for this research as it relies on the English Vocabulary Profile or EVP (English Profile, n.d.), which categorizes words according to difficulty (Text Inspector, n.d.).

Text Inspector's Scorecard (CEFR)

Text Inspector (n.d.) is an online analysis tool that can assign CEFR levels. As there are debates on CEFR reliability (Runnels, 2014; Hong et al., 2020), Text Inspector does not completely rely on the Framework. The CEFR level that Text Inspector assigns is based on over 200 metrics such as readability, lexical diversity, and lexical sophistication. To determine lexical sophistication, corpora such as the British National Corpus and the Corpus of Contemporary American English, and word reference tools such as the Academic Word List and the English Vocabulary Profile (EVP), were used. The EVP, for example, assigns CEFR levels from A1 to C2 to single and multi-word items in a text. However, Text Inspector utilizes only EVP's word list at the word level. CEFR A1 is subdivided into two sub-levels, and there are two additional levels, D1 and D2, for texts of a higher academic level. The estimated CEFR levels provided by Text Inspector can also be compared with the estimated CEFR-J levels given by CVLA, a text analysis tool tailored to English education in Japan.

CVLA (CEFR-J)

CEFR-based Vocabulary Level Analyzer ver.2, or CVLA, created by Prof. Uchida Satoru, is a free online tool that assigns CEFR-J levels to texts (Uchida, n.d.). This was designed specifically for Japanese students, as they tend to fall within different sections of the lower levels of CEFR. It relies on Tono's (2022) CEFR-J wordlist, developed from a corpus that contains items from primary and secondary school textbooks from China, Korea, and Taiwan (MEXT, 2012; Tono, n.d.). While the wordlist covers levels A1 to B2, CVLA extracts its C-level words from the EVP, which is also used as a reference by Text Inspector (n.d.). Factors including readability index, verbs per sentence, average word difficulty, and the ratio of B-level content words to A-level content words contribute to the text's CEFR-J level assignment. The text's average scores using these metrics are totaled and converted into one of 12 levels, from Pre-A1 to C2. Despite focusing on single-word items, CVLA's other indices compensate for this limitation. Therefore, this study will employ CVLA to analyze EIKEN Grade 2 texts.

Flesch Reading Ease and Flesch Kincaid Grade Level

Flesch Reading Ease (FRE) and Flesch Kincaid Grade Level (FKGL) metrics help distinguish between the comprehension ease of different texts (Wallace, 1992, p. 77). FRE and FKGL rely on word and

sentence lengths, remaining traditional yet widely used tools (Flesch, n.d.). FKGL, calculated by (.39 x words/sentences) + (11.8 x syllables/words), correlates with US class grades, indicating higher scores for more complex texts (Kincaid et al., 1975). FRE, 206.845 - (1.015 x words/sentences) - (84.6 x syllables/words), rates from extremely difficult to very easy (Flesch, 1948). Because of the similar variables present in their respective formulas, both tools show an almost perfect linear correlation (Štajner et al., 2012).

FRE scores can be categorized from extremely difficult to very easy, as shown in Table 1. The rating scale ranges from 0 to 100, where 0 indicates text that is nearly impossible to read, while 100 signifies very easy readability (Flesch, 1948, p. 230). The corresponding grade level is also provided for each readability score. As mentioned earlier, FRE scores are dependent on the sentence, word, and syllable counts.

Table 1
Flesch Reading Scores Translated to School Grades

Readability Scores	Description	Grade Levels (US system)
90-100	Very Easy	5th grade
80-90	Easy	6th grade
70-80	Fairly Easy	7th grade
60-70	Plain English	8th and 9th grade
50-60	Fairly Difficult	10th to 12th grade (high school)
30-50	Difficult	college
0-30	Very Difficult	college graduate

Regarding FKGL, scores are compared to the grade levels in the US education system. Text Inspector (n.d.) states that an FKGL score below 12 signifies easy readability for the general public, while a score below 8 suggests very easy text comprehension. Grade-level adjustments are therefore suggested for second/foreign language learners (Linguapress, n.d.). For example, some studies show that Japanese high school reading materials are equivalent to US grades 5-9 (Browne, 1998; Chujo & Hasegawa, 2004; Kukita & Fukuda, 2015), while Sugiura et al. (2020) found that high school textbook units can cover larger grade levels from 4th to 11th. EIKEN Grade 2 tests target high school graduates, indicating that texts should align similarly with FKGL levels below Grades 11-12 in the US system. Because of these comparisons, FKGL and FRE are appropriate for EIKEN Grade 2 analysis, whose reliability can also be cross-checked with Lexile scores.

Lexile Reading

EIKEN used MetaMetrics' online Lexile® Analyzer (2016) to measure the complexity of the test forms (reading sections) administered in 2013 and 2014 for all levels. The tool employs a Lexile framework based on word frequency and sentence length. According to MetaMetrics (2024), a score of 200L or below indicates a beginner level, while a score of 1200 or above suggests advanced proficiency. EIKEN Grade 2 tests were between 1000L and 1020L, indicating higher difficulty (MetaMetrics, 2016). Similar to FKGL and FRE, the Lexile Framework is designed primarily in the context of the US education system (MetaMetrics, 2022). For instance, Grade 8 US students typically encounter books with 1010L scores, rising to 1185L by year-end and 1300L in college. Given that the EIKEN Grade 2 texts are designed for high school graduates in Japan, this study will interpret Lexile grade levels similar to the approach in analyzing FKGL scores. It is anticipated that the grade levels corresponding to Lexile scores will not surpass Grade 11 or 12, equivalent to Grades 8-9 in the American school system.

Total coverage

Total coverage is dependent on word frequency. Word frequency affects readers' word recognition speed and text comprehension (Haberlandt & Graesser, 1985). Text comprehension becomes more difficult with more low-frequency or unknown words (Nation, 2013). Word frequency lists such as the NGSL (Browne et al., 2013) can be beneficial in material development. These lists are developed from corpora, such as the COCA (Corpus of Contemporary American English), which can rank words based on frequency.

Laufer (1989, p. 321) found that learners need to know at least 95% of the text's word tokens for comprehension. Hirsh and Nation (1992) later stated that a reader needs around 98% token coverage to read for pleasure, which Hu and Nation (2000, p. 419) reaffirmed in a later study. This study determines whether 95% of the vocabulary found in EIKEN Grade 2 texts is covered by the NGSL.

Correlations between readability indices

As mentioned earlier, this study aims to determine the correlations between the variables and formulas used in text analysis. As previous studies only focus on finding relationships among traditional readability formulas such as FRE and FKGL, this study will investigate the correlations among the following text analysis variables and formulas: word count, average sentence length, CEFR and CEFR-J levels (determined by Text Inspector and CVLA, respectively), Flesch Reading Ease Score, Flesch-Kincaid Grade Level, Lexile score, and total coverage (determined by AntWordProfiler and using the NGSL). It is hypothesized that indices sharing similar text measures will exhibit stronger correlation scores. Thus, the research question is: Will there be agreements among text analysis tools in terms of the results?

Method

Fifteen EIKEN Grade 2 reading tests collected from the EIKEN website (2023) were used in this study. Each reading test contains six parts labelled 1, 2A, 2B, 3A, 3B, and 3C. The actual labels in the test were not used in this study. Section 2A will be named Section A; 2B will be named Section B, and so forth. As mentioned earlier, the first section, with 20 short texts, is excluded from the analysis.

The gaps in the first and second gap-fill passages were completed with the omitted phrases or words. The answer choices not selected were removed. The comprehension questions for the fourth and fifth texts were not analyzed either. For every C text (an email), other parts of the email message, such as the subject and sender’s email address, were removed, and only the body text was analyzed. This was done to avoid a false count of the total number of sentences. The title of every text was also removed from the analysis. This study refers to the passages by their year, number, and letter, such as 2018-1A (section A of the first test administered in 2018).

Text Inspector was used to determine the following metrics: word counts and average sentence lengths, Flesch Reading Ease (FRE) scores, Flesch Kincaid Grade levels, and scorecards or CEFR levels (based on the metrics set by this online analysis tool). To analyze texts, one can either copy and paste them into the provided box or upload a .txt document. Then a mode is chosen for analysis (reading, writing, and listening). For this study, I selected the reading mode for text analysis.

Regarding the CEFR-J level of the texts, I used CVLA: CEFR-based Vocabulary Level Analyzer (ver. 2.0). Users can simply copy and paste the text for analysis. In contrast to Text Inspector, it has only two modes for text analysis: reading and listening. To avoid confusion, the average scores for calculating four textual features (readability index, verbs per sentence, average word difficulty, and the ratio of B-level content words to A-level content words) for each CEFR-J level displayed in the results were derived from previous studies (Uchida and Negishi, 2018). The "Input" section presents the scores obtained from the text being analyzed.

Lexile® Analyzer (2024) was employed to obtain the Lexile grades of the texts. In addition to the range scores and grade levels, it recommends books aligning with the estimated scores. Due to limited space, I recorded only the equivalent grade levels in the results section and moved the estimated Lexile range of the texts to *Appendix A.*

To generate vocabulary profiles for the texts, I used a free software tool called AntWordProfiler 2.0.1 (Anthony, 2022). NGSL (1000, 2000, 2800) and AWL (960) (Browne et al., 2013) were imported into the software as reference lists. The program calculates the percentage of the text covered by each list. Texts can either be copied and pasted on the given input screen or encoded as .txt format.

During the analysis, the difficulty levels of the five text sections of each test were compared. Texts with the lowest scores were generally considered the easiest (except for Flesch Reading Ease). In addition, Pearson's correlation scores were determined using JASP software to evaluate the level of agreement among all text measures. Non-numerical data, e.g., CEFR-J levels, were changed to numerical values for analysis. Later, I used these data for further discussion.

Results

This section will demonstrate the results using the text analysis tools introduced in previous sections. Table 2 presents Pearson's correlation coefficients between various text analysis measures. Abbreviations include WC (word count), ASL (average sentence length), FKGL (Flesch Kincaid Grade Level), LE (Lexile Grade Level), FRE (Flesch Reading Ease), FRG (FRE equivalent grade level), TC (total coverage), CEFR (CEFR levels by Text Inspector), and CEFRJ (CEFR-J levels by CVLA).

Please refer to *Appendix B* as it presents the 75 texts alongside their word counts and average sentence lengths, as determined by Text Inspector. The fourth and fifth columns display the grade levels assigned to the texts by Text Inspector (FKGL) and Lexile Analyzer, respectively. The FRE scores, also analyzed by Text Inspector, are provided along with their corresponding grade-level equivalents in the sixth and seventh columns. The eighth column indicates the total coverages resulting from analyzing the texts using AntWordProfiler. Finally, the ninth and last columns show the CEFR-J and CEFR levels determined by CVLA and Text Inspector, respectively.

Table 2
Pearson's correlation scores between the different text analysis indices

	WC	ASL	FKGL	LE	FRE	FRG	TC	CEFR
ASL	.29*	-						
FKGL	.49***	.69***	-					
LE	.38***	.68***	.72***	-				
FRE	-.49***	-.43***	-.95***	-.61***	-			
FRG	.34**	.35**	.86***	.5***	-.92***	-		
TC	-.26*	.16	-.05	-.17	.14	-.06***	-	
CEFR	.44***	.36**	.65***	.45***	-.66***	.60***	-.13***	-
CEFRJ	.33**	.38***	.64***	.43***	-.64***	.57***	-.03***	.64***

**p < .05, ** p <.01, *** p < .001*

Discussion

To recall, the research question is: Will there be agreements among text analysis tools in terms of the results? First, this section will discuss the results displayed in *Appendix B.* Variables such as word count and sentence length are examined for their impact on text difficulty. The study compares text analysis measures that share similar variables: Lexile and FKGL; CEFR levels assigned by Text Inspector and CEFR-J levels by CVLA. FRE and total coverage are also compared because Strauss et al. (2007) found a strong correlation between word length and word frequency, and both tools assess the readability of texts. The section concludes with interpretations of the correlation scores (Table 2) among text analysis indices.

Ranges of word counts and sentence lengths

The word counts (column 2 of *Appendix B)* for texts A and B range between 255 to 276, while the email texts or C texts have the lowest word count, ranging from 201 to 240. The longest texts are D and E, with word counts ranging from 334 to 373. However, average sentence lengths do not have similar patterns to determine which texts receive the highest or lowest scores (column 3). The range of average sentence lengths of the texts are 13.00 and 22.99. Longer sentences can make a text more difficult, as they may contain dependent clauses, creating compound, complex, or complex-compound sentences. However, a longer sentence could also provide explanations or extra information to help the reader. Nevertheless, the longest sentences with scores close to 22.99 were predominantly from texts A, B, D, and E. Only a few C texts (18.0-19.0) fall into this range. This suggests that average sentence length is not consistently reliable in determining text difficulty. Moreover, based on the results, word count and average sentence length do not consistently align.

Ranges of Lexile Grade Levels and Flesch Kincaid Grade Levels

Determining the FKGL (column 4 of *Appendix B*) involves factors such as word length, sentence length, and syllable count of a text. Twenty-six texts scored between 8.00 and 8.99, suggesting suitability for average students in American Grades 8 to 9. The average FKGL of all 75 texts is 8.63, closely correlating with the FRE scores (columns 6 and 7). Individual text FKGLs ranged from 5.41 to 11.82, covering up to six academic year levels. However, none exceed the 11.82-grade level assignment, with texts 2019-1E, 2019-3A, and 2022-3E identified as the most challenging and designated for American Grade 11 students. The texts span a broader range of reading levels, from Grades 3 to 12 using the Lexile Analyzer. Lexile scores, determined by sentence length and word frequency (MetaMetrics, 2022), show that longer sentences and fewer familiar words correlate with higher difficulty levels. Among them, 46 texts are considered readable for Grades 5 to 12 in the US, with Lexile scores ranging from 1010 to 1200L (Please refer to *Appendix A*), while 27 are suitable for Grades 3 to 7, ranging from 810 to 1000L (column 5). Notably, 2019-3A and 2019-3B stand out with Lexile scores of 1210 to 1400, intended for Grades 10 to 12 students in the US system. According to MetaMetrics (2024), US high school textbooks typically have around 1100L, while postsecondary texts usually have 1300L. The EIKEN Grade 2 texts' average Lexile scores fall between 943.33 and 1133.33L, generally suitable for students in Grades 5 to 10 in the US. These students are expected to read books with an average score of 943.33L at the beginning of the year, increasing to 1133.33L by the end.

Comparison reveals that both FKGL and Lexile Analyzer generally assess EIKEN texts within Grades 11 and 12 in the US system, with only a few reaching those levels. Both measures concur that texts 2019-2C, 2019-3C, and 2022-3B are the easiest, with FKGLs ranging from 5.41 to 5.78 and Lexile levels averaging Grades 4-5. Texts 2019-3A and 2019-3B pose the greatest difficulty according to the Lexile Analyzer, placing them in Grades 10-12, which aligns with FKGLs of 11.82 and 10.09 respectively. These results

suggest that both indices utilize sentence length as a factor in scoring, potentially resulting in similar grade level assignments for texts.

Ranges of the Flesch Reading Ease scores and total coverage percentages

As previously discussed, FRE calculation involves three aspects: sentence count, word/token count, and syllable count. Out of 75 texts, 44 scored between 60 and 60.99, suitable for Grades 8-9 in the US system, akin to the Flesch Reading average of 8.63. Column 6 of *Appendix B* shows FRE score ranges across sections, with equivalent grade levels in Column 7. C texts, typically emails, are the easiest, with scores between 62.17 and 83.47, fitting Grade 6-9 readers. The word count may have played a significant role in this outcome, as each email text contains between 201 and 240 words. On the other hand, longer texts (A, B, D, E) suit students from 7th grade to college. Sections A and E may be more challenging. Only 2019-1E and 2019-3A were rated college-level. However, none of the texts were rated as readable for college graduates (0-30 FRE scores), suggesting that the texts might have been simplified or controlled to remain below a certain difficulty level.

The majority of the text reading scores align with 8th and 9th grades in the US education system, supporting EIKEN's assertion that Grade 2 texts are intended for high school graduates. This assumption stems from the belief that English books created in non-English speaking countries may be comparatively easier than those produced for native English speakers. Therefore, disparities in English materials between the US and Japan may span three to four grade levels. This is evident in the Flesch-Kincaid Grade and Lexile grade levels.

While Flesch Reading Ease scores are based on word, syllable, and sentence counts, total coverages are based on word frequency. Integrated with the NGSL, AntWordProfiler forms the vocabulary profiles of texts (Column 8). Only 27 out of 75 texts achieved 95% total coverage, suggesting many contain unfamiliar words, especially in science and healthcare topics. The results further show that 8 out of 15 text C or email texts are more readable due to higher high-frequency word usage. There were five texts in each of sections A, B, and E, and four texts in section D, all of which achieved 95% or higher in total coverage. Therefore, 41 texts across these categories do not exhibit readability solely based on total coverage. Additionally, comparing readability based on grade levels and word frequency may lead to the assumption that texts intended for Grade 8-9 students (US education system) generally have coverages below 95% and include more low-frequency or unfamiliar words, an aspect teachers should consider.

Ranges of the Text Inspector's Scorecard (CEFR) and CVLA- CEFR-J levels

Columns 9 and 10 of *Appendix B* show CEFR and CEFR-J levels determined by Text Inspector and CVLA. Both tools consider factors that include readability, word count, and sentence length, referring to corpora and word reference lists such as the BNC and EVP to categorize words into different difficulty levels. CVLA utilizes its own corpora for A and B levels but references EVP for C vocabulary.

According to both Text Inspector and CVLA, C texts are generally less challenging, not surpassing B2 and B2.1. CVLA assigns subcategories A2.1 and A2.2 to align with Text Inspector's A2+. Most texts labeled A2.1 and A2.2 by CVLA were also rated easiest by Text Inspector. Both analysis tools also agreed that 2019-2C was the easiest text. If CVLA A2.2 level will be considered, both tools also agreed that 2022-1C and 2022-3B can be categorized as very easy. For A and B texts, one per section received A levels from both tools. Specifically, 2022-3A got A2+ from both, while 2022-3B was A2+ by Text Inspector and A2.2 by CVLA.

The majority of the A, B, D, and E texts fell into CEFR/CEFR-J levels B or C. Notably, four texts—2018-3A, 2019-3A, 2019-3B, and 2022-3E—were labeled very difficult by both Text Inspector (B2+) and

CVLA (C1), suggesting alignment between B2+ and C1. CVLA, designed for English education in Japan, classified several texts at a C1 level, suggesting they may be overly challenging for Japanese learners. Therefore, CVLA deems what Text Inspector finds moderately difficult as very difficult. Despite this, no texts were categorized as C2, typically associated with news articles and academic papers. This trend mirrors Text Inspector's findings, which similarly did not identify any EIKEN Grade 2 texts at levels C1, C2, or D. Most texts were rated B level with Text Inspector assigning B2 to 22 texts, B1+ to 21 texts, and B1 to 17 texts, while CVLA rated 27 texts as B2.1, 10 as B2.2, 9 as B1.2, and 8 as B1.1. Factors such as word count likely influence difficulty, as longer texts (A, B, D, and E) were assigned higher CEFR or CEFR-J levels. Educators can note the presence of challenging vocabulary items in these texts. Wordlists such as the CEFR-J wordlist (Tono, 2022) serve as valuable references for identifying vocabulary items belonging to the B levels.

Correlations among Indices

The correlation between FKGL and word count is moderate (0.488), indicating that texts with more words tend to have higher FKGL scores. Conversely, there is a moderate negative correlation of -0.485 between word count and FRE, indicating that longer texts generally have lower FRE scores, suggesting increased difficulty. The Lexile scores and FKGL demonstrate strong to very strong negative correlations with FRE (-0.608 and -0.949 respectively), indicating that higher grade levels correspond to lower FRE scores, suggesting more challenging readability. Additionally, stronger positive relationships exist between FKGL and Lexile (0.722), FKGL and ASL (0.694), and Lexile and ASL (0.677), surpassing other indices in correlation strength. These scores shed light on why FKGL, Lexile, and FRE assigned similar grade levels to texts. However, the ASL's correlation with FRE is only moderate at -0.432. Despite this, ASL correlates better with other text measures than word counts, contrary to the earlier assumption that word count is the primary predictor of text difficulty.

Total coverage (TC) displays weak correlations with other text measures. Despite the assumed strong correlation between word length and frequency (Strauss et al., 2007), total coverage lacks even a moderate correlation with FRE (0.136) or Lexile scores (-0.169). This weak correlation likely stems from AntWordProfiler's sole focus on word frequency, neglecting sentence structure. Consequently, the previous assumption that texts for US Grade 8-9 students typically have coverages below 95% and include more low-frequency items remains inconclusive.

The analysis of CEFR levels by Text Inspector reveals strong correlations with other text measures. For instance, a strong linear relationship (0.654) exists between CEFR and FKGL, likely influenced by Text Inspector's comprehensive assessment of texts using 200 metrics, including readability. Additionally, CEFR shows a strong correlation with CEFR-J (0.639), possibly because both tools utilize wordlists to categorize vocabulary items. It is worth noting that CEFR-J (CVLA) and CEFR (Text Inspector) exhibit almost identical correlation scores with other measures. To recall, CVLA employs factors such as readability index, verbs per sentence, average word difficulty, and the ratio of B-level to A-level content words for text analysis. However, the correlation score between CEFR and CEFR-J is only at 0.639. This is unexpected, as one might anticipate a very strong correlation similar to that of FRE and FKGL (-0.949). The correlation between Text Inspector and CVLA could be attributed to the differing factors used in their text analyses. Conversely, FRE and FKGL, despite employing distinct formulas, share common measurement factors such as word, syllable, and sentence counts, which could explain their very strong correlation.

Limitations and Future Research

The outcomes of this study were informed by a purely quantitative approach to text analysis. While this quantitative analysis is valuable, teachers should also incorporate qualitative analysis to strengthen their judgment in selecting materials for proficiency test preparation. The approach in this research also had a number of limitations, for example, excluding comprehension questions, answer choices, and specific text elements may have influenced the study outcomes as assessed by text analysis tools. Not understanding the questions would of course present difficulty issues for test takers. Furthermore, this research focused on a small subset of EIKEN tests. Furthermore, it focused on EIKEN tests at a specific level. Analyzing texts from multiple EIKEN levels (e.g., EIKEN 1, pre-1, pre-2, 3, 4, and 5) alongside EIKEN 2 would offer a broader perspective on how text analysis tools perform across varying proficiency levels.

Another possibility for future research would be a separate and focused study on multi-word items present in texts that can affect text difficulty. This would help to avoid the limitations present when many of the analysis tools focus mainly on word count and variety. Additionally, a subsequent study where the findings could be validated by testing students using the same subject texts used. Careful analysis of student performance would provide important insight into the text difficulty.

Conclusion

The analysis conducted in this study provides valuable insights into the readability and difficulty levels of Grade 2 EIKEN texts intended for Japanese high school graduates, addressing the research question of whether there are agreements among various text analysis tools regarding the assessment of text difficulty. Word counts varied significantly among sections, with the email passages emerging as the easiest due to their lower word counts, while the longer texts presented the most challenging reading sections. While average sentence length does not follow such clear patterns, examining correlation scores reveals that ASL has a stronger linear relationship with other indices such as Flesch Kincaid and Lexile grades. Notably, there were results from obtaining FKGL and Lexile reading levels that were similar, most probably because both tools depend on sentence length. It is important to note that only a few texts reached grades 11 and above (US system). Most texts align with grades 8-9 in the US education system, as determined by both tools. Given that text levels for second language classrooms are typically several levels lower than those in the US education system and considering that EIKEN Grade 2 is designed for high school graduates in Japan, it can be inferred that the texts meet the expected proficiency levels for high school graduates in Japan, equivalent to grades 8-9 in the US education system. Texts used for language learning purposes are generally expected to be easier to read than those used in native language education. Regarding FKGL and FRE, they are considered to have the strongest correlation at -0.949, likely due to the similarities of textual factors being measured: syllable, sentence, and word counts, despite using different formulas. Similarly, the CEFR and CEFR-J levels of the texts (assigned by Text Inspector and CVLA respectively) were generally comparable, with email passages deemed the easiest and longer texts considered difficult, though the correlation is only 0.639. This discrepancy may be attributed to the differing factors utilized in their text analyses. Despite this, both tools exhibited almost identical correlation scores with other measures. Looking at the majority of the results of the texts B2 (Text Inspector) and B2.1 (CVLA) scores, teachers may consider finding texts with such levels when preparing students. Using these tools, the research also found out that there were vocabulary items that the Text Inspector found moderately difficult, whereas CVLA found them very difficult for Japanese learners. Regarding the analysis of total coverages, it was revealed that total coverage has weak correlations with other text measures. While total coverage focuses solely on word frequency, it does not consider other important factors such as sentence structure, leading to its limited utility in assessing text difficulty. Overall, the findings of this study highlight the need for educators and curriculum developers to consider

multiple factors when evaluating text difficulty. Hermosa (2002) suggests educators consider student variables, such as their responses using techniques like the cloze procedure, and encourages assessments and opinions provided by teachers regarding the learning materials. By integrating various text analysis tools and methodologies, educators can make more informed decisions when selecting texts for language learning purposes, ultimately enhancing the effectiveness of language instruction.

Acknowledgements

I would like to thank Professors Jonathan Rees of the University of Birmingham and Jeffrey Stewart of Tokyo University of Science/Takushoku University, as well as all the reviewers of the JALT Testing and Evaluation SIG, for their guidance and advice when writing this paper.

Declaration of competing interests:

M. Johnston has declared no competing interests.

References

Anthony, L. (2022). *AntWordProfiler* (Version 2.0.1*)* [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Arai, Y. (2022). Exploring perceived difficulty of graded reader texts. *Reading in a Foreign Language, 34*(2), 249-270. http://hdl.handle.net/10125/67425

Browne, C. (1998). Japanese high school textbooks: How readable are they? *Temple University Japan Working Papers in Applied Linguistics, 12,* 1-13.

Browne, C., Culligan, B., & Phillips, J. (2013). The new general service list: A core vocabulary for EFL students and teachers. *JALTs The Language Teacher, 34*(7), 13-15.

Chujo, K., & Hasegawa, S. (2004). Goi no cover ritsu to readability kara mita daigaku eigo nyushi mondai no nanido [Assessing Japanese college qualification tests using JSH text coverage and readability indices]. Nihon University Student Faculty of Engineering Research Report B, 37, 45-55. http://www.cit.nihon-u.ac.jp/laboratorydata/kenkyu/publication/journal_b/b37.5.pdfCVLA: CEFR-based Vocabulary Level Analyzer (ver. 2.0). (2023). *Home*. https://cvla.langedu.jp/

Dunlea, J., & Matsudaira, T. (2009). Investigating the relationship between the EIKEN tests and the CEFR. *Linking to the CEFR levels: Research perspectives,* 103-110.

EIKEN. (2023). Research. https://www.eiken.or.jp/eiken/en/recognition/

English Profile. (n.d.). *Wordlists.* https://www.englishprofile.org/wordlists

Flesch, R. (n.d.). *How to Write Plain English.* University of Canterbury: Management, Marketing, and Entrepreneurship. https://web.archive.org/web/20160712094308/http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221-233. https://doi.org/10.1037/h0057532

Grabe, W., & Yamashita, J. (2022). Cambridge Applied Linguistics. Cambridge University Press.

Haberlandt, K. and Graesser, A. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General, 114*(3), 357-374. https://doi.org/10.1037/0096-3445.114.3.357

Hermosa, N. (2002). *The Psychology of Reading.* University of the Philippines -Open University.

Hirsh, D. and Nation, I.S.P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language 8(*2), 689-696.

Holster, T. A., Lake, J. W., & Pellowe, W. R. (2017). Measuring and predicting graded reader difficulty. *Reading in a Foreign Language, 29*(2), 218–244. https://nflrc.hawaii.edu/rfl/item/377

Hong, J. F., Tseng, H. C., Peng, C. Y., & Sung, Y. T. (2020). Linguistic Feature Analysis of CEFR Labeling Reliability and Validity in Language Textbooks. Journal of Technology & Chinese Language Teaching, 11(1).

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

Kukita, D., & Fukuda, M. (2015). On the Readability of the English Technical Writing Test: With Special Reference to the Textbooks Used in Technical High Schools and Colleges of Technology in Japan. *Bulletin of Miyazaki Municipal University Faculty of Humanities, 22*(1), 251-260.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans to thinking machines* (pp. 316–323). Multilingual Matters.

Linguapress. (n.d.). *Flesch-Kincaid Readability Test.* Linguapress. https://linguapress.com/teachers/flesch-kincaid.htm

MetaMetrics. (2016). *Lexile® Measurement of Tests: EIKEN & Test of English for Academic Purposes.* MetaMetrics.https://metametrics.s3.amazonaws.com/public/dynamic/international/pdfs/EikenText_Measurement_Report_Digital.pdf

MetaMetrics. (2022). *Lexile® Framework for Reading: Development and Validity Evidence.* MetaMetrics. https://hubsupport.lexile.com/Images/Lexile%20Framework%20for%20Reading%20Validity%20Evidence_2022.pdf

MetaMetrics. (2024). *The Lexile Framework of Reading.* https://hub.lexile.com/analyzer

MEXT: Ministry of Education, Culture, Sports, Science and Technology. (2012*). CEFR-based framework for ELT in Japan.* https://www.mext.go.jp/b_menu/shingi/chousa/shotou/092/shiryo/__icsFiles/afieldfile/2012/09/24/1325972_2_1.pdf

Nation, I. S. P. (2013). *Teaching & learning vocabulary.* Heinle Cengage Learning.

Nuttall, C. (1996). *Teaching reading skills in a foreign language*. Heinemann.

Ocampo, S. (1997). *Trends in Reading Instruction.* University of the Philippines- Open University.

Runnels, J. (2014). An exploratory reliability and content analysis of the CEFR-Japan's A-level can-do statements. *JALT journal, 36*(1), 69-89.

Štajner, S., Evans, R., Orasan, C., & Mitkov, R. (2012). What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility* (pp. 14-22).

Strauss, U., Grzybek, P., & Altmann, G. (2007). Word length and word frequency (pp. 277-294). Springer Netherlands.

Sugiura, R., Imai, N., Hamilton, M., Dean, E., & Ashcroft, R. (2020). Input and Output in Japanese High School Government-Approved English Textbooks. *Journal of Higher Education Tokai University, 21*, 1-16.

Text Inspector. (n.d.). *Analyse*. https://textinspector.com

Tono, Y. (2022). *CEFR-J wordlist.* http://www.cefr-j.org/download.html

Tono, Y. (n.d.). *Using corpora for reference level descriptions of the CEFR and the CEFR-J* https://languages-cultures.uq.edu.au/files/28901/Using-corpora-for-reference-level-descriptions.pdf

Uchida, S. (n.d.) CVLA: CEFR-based Vocabulary Level Analyzer (ver. 2.0).

Uchida, S., & Negishi, M. (2018). Assigning CEFR-J levels to English texts based on textual features. In Y. Tono & H. Isahara (Eds.), *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference* (pp. 463-467).

Wallace, C. (1992). *Reading.* Oxford University Press.

Waring, R. (n.d.). Writing a graded reader. Retrieved June 29, 2024, from https://www.er-central.com/authors/writing-a-graded-reader/writing-graded-readers-rob-waring/#google_vignette

Appendix A

Lexile Ranges of Texts Per Test

Texts	Lexile Range Min	Lexile Range Max
2018-1	810.00	1200.00
2018-2	810.00	1200.00
2018-3	810.00	1200.00
2019-1	1010.00	1200.00
2019-2	810.00	1200.00
2019-3	810.00	1400.00
2020-1	810.00	1200.00
2020-2	810.00	1200.00
2020-3	810.00	1200.00
2021-1	810.00	1200.00
2021-2	810.00	1200.00
2021-3	810.00	1200.00
2022-1	810.00	1200.00
2022-2	810.00	1200.00
2022-3	810.00	1200.00

Appendix B

Text Scores Derived from Index Calculations

1	2	3	4	5	6	7	8	9	10
Texts	Text Inspector	Text Inspector	Text Inspector	Lexile Analyzer	Text Inspector	Flesch	AntWord Profiler	CVLA	Text Inspector
	Word count	Average sentence length	Flesch Kincaid grade levels	Lexile grades	Flesch Reading Ease scores	FRE grade levels	Total coverage	CEFR-J levels	CEFR levels
2018-1A	270	15.88	8.26	3-7	64.13	8-9	96.3	B2.1	B2
2018-1B	274	15.22	7.79	3-7	66.34	8-9	93.0	B1.2	B1
2018-1C	201	18.27	7.03	3-7	77.17	7	98.6	B1.2	B1
2018-1D	344	18.11	8.93	5-12	63.28	8-9	92.1	B2.1	B1+
2018-1E	373	16.2	8.45	5-12	63.36	8-9	93.7	B2.2	B2+
2018-2A	265	15.59	10.08	5-12	50.55	10-12	83.5	B1.2	B2
2018-2B	260	18.64	9.85	5-12	57.51	10-12	94.7	C1	B2
2018-2C	221	14.73	6.12	3-7	77.42	7	95.1	A2.2	B1
2018-2D	354	16.86	7.62	3-7	70.47	7	96.1	B2.1	B1
2018-2E	356	18.74	10.25	5-12	54.98	10-12	86.3	C1	B1+

1	2	3	4	5	6	7	8	9	10
Texts	Text Inspector	Text Inspector	Text Inspector	Lexile Analyzer	Text Inspector	Flesch	AntWord Profiler	CVLA	Text Inspector
	Word count	Average sentence length	Flesch Kincaid grade levels	Lexile grades	Flesch Reading Ease scores	FRE grade levels	Total coverage	CEFR-J levels	CEFR levels
2018-3A	264	18.86	10.18	5-12	55.67	10-12	94.0	C1	B2+
2018-3B	274	16.12	7.97	5-12	66.66	8-9	90.5	B2.1	B1
2018-3C	227	17.46	7.13	3-7	75.07	7	94.0	B2.1	B1
2018-3D	362	17.24	8.34	5-12	65.94	8-9	93.1	B2.1	B1+
2018-3E	372	16.91	10.45	5-12	50.26	10-12	92.0	B2.1	B1+
2019-1A	267	15.71	7.82	5-12	67.00	8-9	87.5	B2.1	B2+
2019-1B	274	17.12	8.23	5-12	66.57	8-9	92.2	B2.1	B1
2019-1C	226	18.83	8.25	5-12	69.43	8-9	96.3	B1.2	B1
2019-1D	334	16.70	9.01	5-12	60.20	8-9	94.1	B2.1	B1+
2019-1E	365	18.25	11.25	5-12	46.92	college	97.3	C1	B2
2019-2A	276	18.40	9.37	5-12	60.65	8-9	94.9	B2.1	B1+
2019-2B	255	18.21	8.73	5-12	64.93	8-9	90.1	B2.1	B1+
2019-2C	229	15.27	5.41	3-7	83.47	6	95.9	A2.1	A2+
2019-2D	358	17.90	9.16	5-12	61.29	8-9	96.1	B1.1	B1+
2019-2E	361	18.05	10.54	5-12	*51.65*	10-12	93.6	C1	B2
2019-3A	267	19.07	11.82	10-12	44.26	college	94.7	C1	B2+

1	2	3	4	5	6	7	8	9	10
Texts	Text Inspector	Text Inspector	Text Inspector	Lexile Analyzer	Text Inspector	Flesch	AntWord Profiler	CVLA	Text Inspector
	Word count	Average sentence length	Flesch Kincaid grade levels	Lexile grades	Flesch Reading Ease scores	FRE grade levels	Total coverage	CEFR-J levels	CEFR levels
2019-3B	271	18.07	10.09	10-12	54.89	10-12	91.6	C1	B2+
2019-3C	230	14.38	5.46	3-7	81.53	6	90.7	B2.1	A2+
2019-3D	359	18.89	10.22	5-12	55.45	10-12	96.4	B2.2	B2
2019-3E	356	22.25	10.36	5-12	60.44	8-9	98.0	B2.2	B2
2020-1A	260	16.25	8.31	3-7	64.42	8-9	89.9	B1.1	B1
2020-1B	261	17.40	9.33	5-12	59.19	10-12	94.7	B2.2	B1+
2020-1C	213	16.38	8.03	3-7	66.68	8-9	97.7	B2.1	B1
2020-1D	362	18.10	10.18	5-12	54.32	10-12	95.7	B1.2	B1+
2020-1E	356	18.74	8.85	5-12	64.96	8-9	96.0	B1.2	B2
2020-2A	267	16.69	9.39	5-12	57.45	10-12	92.2	B2.2	B2
2020-2B	265	18.93	10.27	5-12	55.14	10-12	95.9	C1	B2
2020-2C	210	16.15	7.79	3-7	67.97	8-9	93.3	B2.1	B1
2020-2D	367	17.48	8.46	5-12	65.54	8-9	91.8	B1.1	B2
2020-2E	357	14.88	9.15	3-7	*55.95*	10-12	88.6	B2.1	B2+

1	2	3	4	5	6	7	8	9	10
Texts	Text Inspector	Text Inspector	Text Inspector	Lexile Analyzer	Text Inspector	Flesch	AntWord Profiler	CVLA	Text Inspector
	Word count	Average sentence length	Flesch Kincaid grade levels	Lexile grades	Flesch reading ease scores	FRE grade levels	Total coverage	CEFR-J levels	CEFR levels
2020-3A	268	14.11	7.61	3-7	65.62	8-9	92.5	B2.1	B2
2020-3B	266	17.73	8.54	5-12	65.43	8-9	97.4	C1	B2
2020-3C	222	15.86	7.23	3-7	71.46	7	94.2	A2.1	B1
2020-3D	349	14.54	7.56	3-7	66.75	8-9	90.2	B2.2	B2
2020-3E	343	17.15	10.50	5-12	50.32	10-12	90.6	C1	B2
2021-1A	275	15.28	9.08	3-7	57.20	10-12	96.5	B2.1	B1+
2021-1B	260	16.25	8.77	5-12	61.16	8-9	92.3	B2.1	B1+
2021-1C	214	19.45	9.42	5-12	62.17	8-9	97.2	B2.1	B1+
2021-1D	363	18.15	8.65	5-12	65.36	8-9	91.5	B1.2	B1+
2021-1E	362	16.45	8.89	5-12	60.66	8-9	93.9	C1	B1+
2021-2A	265	18.93	8.85	5-12	65.35	8-9	92.1	B1.1	B1
2021-2B	271	15.06	7.26	3-7	69.80	8-9	98.9	B1.2	B1
2021-2C	221	13.81	6.94	3-7	69.93	8-9	95.3	A2.1	B1
2021-2D	354	16.09	8.72	5-12	61.21	8-9	85.5	B2.2	B1+
2021-2E	359	14.36	8.22	3-7	61.71	8-9	95.5	C1	B2

1	2	3	4	5	6	7	8	9	10
Texts	Text Inspector	Text Inspector	Text Inspector	Lexile Analyzer	Text Inspector	Flesch	AntWord Profiler	CVLA	Text Inspector
	Word count	Average sentence length	Flesch Kincaid grade levels	Lexile grades	Flesch Reading Ease scores	FRE grade levels	Total coverage	CEFR-J levels	CEFR levels
2021-3A	274	17.12	8.62	5-12	63.79	8-9	89.7	B2.1	B1
2021-3B	270	19.29	9.02	5-12	64.75	8-9	94.0	B2.2	B2+
2021-3C	234	15.60	7.14	3-7	71.69	7	92.7	B1.1	A2+
2021-3D	362	18.10	8.78	5-12	64.37	8-9	93.2	B1.2	B1+
2021-3E	358	15.57	9.33	3-7	55.87	10-12	86.0	B2.1	B2
2022-1A	263	16.44	8.81	5-12	61.16	8-9	90.8	B2.2	B1+
2022-1B	272	14.32	6.91	3-7	71.00	7	88.9	B2.1	B2
2022-1C	234	15.60	6.58	3-7	75.67	7	97.4	A2.2	A2+
2022-1D	359	19.94	9.58	5-12	61.93	8-9	87.7	B2.1	B2+
2022-1E	352	18.53	9.44	5-12	60.41	8-9	94.5	C1	B2
2022-2A	260	15.29	8.39	3-7	62.13	8-9	97.5	C1	B2
2022-2B	276	17.25	8.62	5-12	63.96	8-9	92.6	B1.1	B1
2022-2C	221	18.42	8.63	5-12	66.03	8-9	93.2	B2.1	B1+
2022-2D	362	17.24	8.60	5-12	64.07	8-9	83.6	B2.2	B1+
2022-2E	365	16.59	8.18	5-12	65.99	8-9	88.2	B1.1	B2
2022-3A	269	14.16	6.21	3-7	75.79	7	93.0	B1.1	A2+
2022-3B	262	13.10	5.78	3-7	76.97	7	91.9	A2.2	A2+

1	2	3	4	5	6	7	8	9	10
Texts	Text Inspector	Text Inspector	Text Inspector	Lexile Analyzer	Text Inspector	Flesch	AntWord Profiler	CVLA	Text Inspector
	Word count	Average sentence length	Flesch Kincaid grade levels	Lexile grades	Flesch Reading Ease scores	FRE grade levels	Total coverage	CEFR-J levels	CEFR levels
2022-3C	240	16.00	7.61	3-7	68.98	8-9	93.7	B2.1	B2
2022-3D	361	21.24	10.08	5-12	60.61	8-9	88.5	B2.1	B1+
2022-3E	362	22.62	11.06	5-12	56.04	10-12	96.1	C1	B2+

Call for Papers

Shiken: A Journal of Language Testing and Evaluation in Japan 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 the most 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 including adding DOI or URL information wherever possible. 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 two expert reviewers. Initial evaluation is usually completed within four weeks. The whole process from submission to publication is normally completed within six months.