

SHIKEN

A Journal of Language Testing and Evaluation in Japan

Volume 30 • Number 1 • May 2026

<https://doi.org/10.37546/JALTSIG.TEVAL30.1>

Contents

1. In Memoriam: Dr. Jim Sick

2. Validating the second stage of the EIKEN test
Natalie Correa and Todd Phillips

<https://doi.org/10.37546/JALTSIG.TEVAL30.1-1>



Testing and Evaluation SIG

ISSN 1881-5537

Shiken: A Journal of Language Testing and Evaluation in Japan

Volume 30 No. 1

May 2026

<https://doi.org/10.37546/JALTSIG.TEVAL30.1>

Editors

Heather Woodward

Rikkyo University

Benjamin Sanchez Murillo

Tsuru University

Reviewers

(see editorial board, plus additional reviewers)

Website Editor

Tophe Zuelke

English Language Education Council of Japan (ELEC)

Editorial Board

Edward Schaefer

Ochanomizu University

Trevor Holster

Kyushu High School

J. W. Lake

Tohoku University

In Memoriam: Dr. Jim Sick

It is with great sadness that we announce the passing of our colleague and friend Dr. Jim Sick on January 6. Like many EFL teachers in Japan, Jim came to the profession from other areas. Having gone to the New College of Florida and the Berklee College of Music, and playing guitar in San Francisco, he originally came to Japan as a jazz guitarist. He gradually got into EFL teaching, maybe to make ends meet, and decided to make that his career choice. He went on to earn his Master's and doctoral degrees from Temple University Japan (TUJ), becoming a recognized specialist in language assessment and statistical analysis, including Rasch measurement. Jim shared his expertise by teaching language testing courses at TUJ and served as a dedicated dissertation advisor to numerous doctoral students. A long-standing member of JALT and the Testing and Evaluation SIG (TEVAL), he also served as TEVAL President for four years. His scholarly contributions and leadership have left a lasting impact on the field of assessment in Japan. Jim will be remembered with deep gratitude by the many students he guided and by his colleagues for his leadership, expertise, and kindness. He will be sorely missed. Rest in peace, Jim. May your memory be a blessing.

Validating the second stage of the EIKEN test

Natalie Correia¹ and Todd Phillips²

natalie@h-lasalle.ed.jp

phillips_todd@asia-u.ac.jp

1. *Hakodate La Salle Academy*

2. *Asia University*

Abstract

Before using a test to make inferences about the abilities of test takers, it needs to be determined whether the test is effective and useful for its intended interpretations. With the Jitsuyo Eigo Gino Kentei (“the EIKEN test”) being among the most popular English proficiency tests for Japanese learners, there is a surprising lack of external research supporting the validity of the test. This small-scale study investigated the validity of the second grade EIKEN speaking test through analysis of data from 10 learners using many-facet Rasch measurement analysis. The results showed that the standards of evaluation for the speaking test could reliably differentiate examinee ability and demonstrate consistent inter-rater reliability.

Keywords: Eiken Test, Rasch analysis, test evaluation, validity

The scores of language tests are invaluable to teachers as we make inferences about the abilities of test takers which are often used to make decisions about people, placements and programs (Bachman, 2004). Before using a test for these purposes, it needs to be determined whether the test is effective and useful for these intended interpretations (Bachman, 2004; Bachman & Palmer, 2010; Hughes & Hughes, 2020). This can be determined by evaluating the validity of the test (Bachman & Palmer, 2010; Brown, 2007; Hughes & Hughes, 2020).

The Jitsuyo Eigo Gino Kentei (Test in Practical English Proficiency, hereafter referred to as “the EIKEN test”) is one of the most popular English proficiency tests among students in Japan with almost 4 million students taking part each year (EIKEN, 2022). Moreover, with several Japanese universities using a candidate's EIKEN proficiency level as a qualification for entrance application processes, it can be considered a high-stakes test. Although the Eiken Foundation claims high standards to ensure validity of their test levels from internal research, there has been insufficient outside research to support these claims, especially in recent years (Benson, 2013; Dunlea, 2008; Plumb & Watanabe, 2016). This paper will discuss the results of a small-scale study conducted with senior high school students in Japan to provide more insight into the validity of the EIKEN second stage’s speaking test.

Literature Review

The EIKEN Test

In the early 1960s, Japan’s Social Education Council had suggested to the Ministry of Education in Japan that Japanese learners may be more motivated by certificated proficiency tests (EIKEN, 2022). Subsequently, the Society for Testing English Proficiency (now called The Eiken Foundation) was established in 1963 and “the EIKEN test” was introduced to promote practical English in Japan. Currently, the EIKEN operates on a pass/fail system of which there are eight levels: Grade 5, 4, 3, Pre-2, Pre-2 Plus, 2, Pre-1, and 1. Each of the grades have sets of test items that are designed to act as goal posts to measure English ability with Grade 1 being the most difficult. Different versions of the test are used with each administration. The test is performed in two stages: 1) a written examination including a listening section,

and 2) a speaking test. With interest in communicative competence, this paper will mainly focus on the characteristics of the second stage.

The target language use (TLU) domain—the situations in which examinees will be using English outside of the test (Bachman & Palmer, 2010)—that the EIKEN speaking test attempts to assess is communicative competence in real-life situations (EIKEN, 2022). The speaking test administered at the second stage is designed to evaluate how well examinees can speak and interact in English. The test’s target population is Japanese students. Table 1 below depicts the target clientele for each grade based on their Japanese education level as well as a description of the English goal posts or benchmarks the level is designed to evaluate.

Table 1
EIKEN Grade Levels

EIKEN Grade	Target Clientele	English Goal Post
5	First-year junior high school students (ages 12~13)	simple words, phrases, and short sentences
4	Second-year junior high school students (ages 13~14)	simple words, phrases, and short sentences
3	Junior high school graduates (ages 14~15)	familiar, everyday topics such as likes and dislikes and basic personal and family information
Pre-2	First and second-year senior high school students (ages 15~17)	general aspects of daily life
Pre-2 Plus	Advanced high school students (ages 15~17)	detailed information, ideas, and feelings about familiar social topics
2	Senior high school graduates (ages 17~18)	social, professional, and educational situations
Pre-1	University students (ages 18+)	wider range of social, professional, and educational situations
1	University graduates (ages 18+)	

The first stage consists of reading and listening questions, with Grades 1 to 3 also including a writing section. Once the students are notified online if they have passed the first stage, a date is chosen for the speaking test of the second stage test. For Grades 1 to 3, examinees must take the speaking test of the second stage to pass the level, while for Grades 4 and 5 the speaking test is optional and occurs online. For Grades Pre-1 to 3, the speaking test is a face-to-face interview with just the examinee and one examiner in the room and Grade 1 (the highest stage) has two examiners. Examinees are evaluated based on their responses for the tasks, as well as pronunciation, use of vocabulary and grammar. The examinee’s “attitude toward actively engaging in conversation” is also a criterion for evaluation in Grades 1 to 3 (EIKEN, 2022). The aim of the current study is to evaluate this second stage and determine whether this assessment of oral communication proficiency is supported by validity evidence.

Validity of the EIKEN

Validity describes how well a test measures what it is intended to measure and judge the degree to which evidence “can support the adequacy and appropriateness of inferences and actions based on test scores” (Messick, 1989, p. 13 cited in Bachman, 2004, p. 259). With validity being specific to particular uses or

interpretations, measuring the validity of the EIKEN speaking test is difficult since it works on a level-based framework whereby each level may have a different definition of a given ability, purpose, and target clientele (EIKEN, 2022; Linn & Gronlund, 2000). The levels of the EIKEN have been compared to other norm-referenced English proficiency tests such as the TOEIC (Nagashima, 2001 cited in Marlowe & Asaba, 2017) and the TOEFL (Ishida, 2004). However, overall, very little outside research has been done on the validity of these levels of EIKEN with most declarations coming from in-house studies on the development of Can-do Lists designed for each level (Benson, 2013; Dunlea, 2008; Plumb & Watanabe, 2016). Internal research results have purported that while the Can-do lists are not exhaustive of all relevant real-world situations, the descriptors of situations tested at each level were shown to be applicable to the targeted test takers indicating favorable validity (Dunlea, 2008; Noguchi et al., 2007).

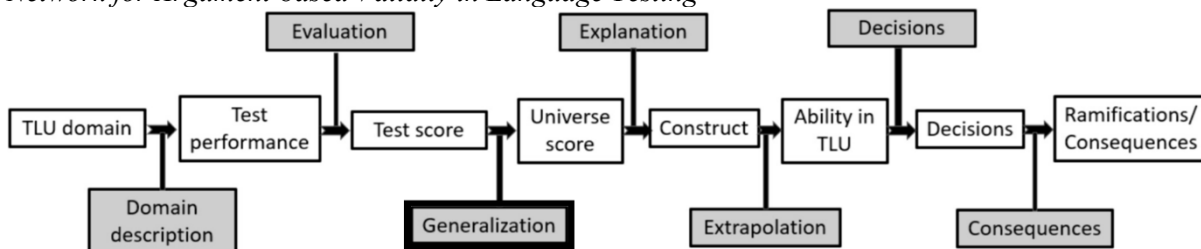
The Eiken Foundation states on their website that a majority of their validation resources have gone into assuring that the tests are relevant to learners at each level with focus on content validity and predictive validity. Recently, however, these validity models have been criticized for being inadequate in the delivery of supporting evidence since they can each only assess a limited amount of what is necessary for an overall view of validity of a measure (Cubilo, 2014). As such, the in-house studies of validity by the Eiken Foundation can be seen as needing to be updated both in research and theoretical framework. With the recognition of a need for a more unified logic-based validity argument, an argument-based approach to validity outlined by Kane (2006) and expanded by Chapelle et al. (2011) has come to the forefront as a tool for considering the interpretations of test scores.

Argument-based approach to validity

In the framework for an argument-based approach to validity, arguments are developed through a process of reasoning defined by a network of inferences and assumptions. The validity of the score's uses and interpretations can be evaluated based on the plausibility of these inferences and the thoroughness of the network as a whole. The building blocks of this network proposed by Kane (2013) and expanded on by recent researchers (Dai et al., 2024) are illustrated in white in Figure 1 below. Inferences made at each stage are shown in grey.

Figure 1

Network for Argument-based Validity in Language Testing



Note. Adapted from “Expanding Kane’s argument-based validity framework: What can validation practices in language assessment offer health professions education?” by D. W. Dai et al., 2024, *Medical Education*, 58(12), p. 1463.

Claims made by inferences require justification through the evaluation of a warrant or assumption (Cubilo, 2014). These warrants are evaluated by interpreting backing through theories, research, data, experience, etc. For instance, in the case of the EIKEN test, one may assume that a score received by an examinee is dependable and would be consistent with the score they would receive on other versions of the tests, regardless of the occasion or raters. In evaluation of assessments, this inference was commonly known as reliability. In the argument-based approach to validity framework, a score that would exist from a sample of measurements obtained under all admissible conditions would be considered a universe score. The

connection between a test score and this universe score is known as the generalization inference, boldly highlighted in Figure 1 above (Knoch & Elder, 2013; McCowan & McCowan, 1999). Evidence of test reliability which supports the claim that test scores are consistent can provide the backing to support the generalization inference (Schaefer & Martin, 2023). The present study evaluates whether there is enough reliability in EIKEN scores to have validity in generalization inferences.

Reliability of the EIKEN

The Eiken Foundation has stated that they have measures in place to ensure quality (EIKEN, 2022), however, measures of interrater reliability have not been stated. In terms of independent studies, there is little research which investigates the reliability of EIKEN test tasks or consistency across raters, and those that have been published only show moderate support for claims of reliability with coefficient estimates showing less than the minimum ideal 0.90. An investigation by MacGregor (1997) into reliability and validity of the Pre-2 level of the EIKEN found the Kuder-Richardson Formula reliability estimate to be only 0.82. Another small-scale study by Nielsen (2000) which explored the third level of the EIKEN test by first year students at a technical college found a reliability co-efficient of 0.86. With the EIKEN speaking test being a rater-mediated assessment, the speaking test is susceptible to lower reliability (Bachman, 2004). Thus, the lack of investigation into the EIKEN test to support validity claims is concerning (Benson, 2013; Dubien, 2023; Piggin, 2011).

The current study proposes a modern approach to evaluating reliability of the EIKEN test. To explore the generalization inference of validity, a method of research must be used which supports the claim that speaking test scores of examinees taking the EIKEN test are consistent and reliable. With raters possibly contributing to measurement error in the EIKEN speaking test, the approach the authors considered well-suited for its analysis is the psychometric modeling approach called many-facet Rasch measurement.

Many-facet Rasch measurement model

The Rasch model, developed by Georg Rasch (1960, 1980), provides a probabilistic framework for estimating item difficulty and examinee ability on a common measurement scale (Barkaoui, 2013). In this model, the probability of a particular response is described as a function of an examinee's location on a latent trait continuum relative to item difficulty (Wind, 2014, as cited in Bahrouni, 2016). The model can be extended to the many-Facet Rasch model (MFRM), which incorporates additional facets that may influence observed scores (Linacre, 1989). A facet is a variable of the scoring procedure which can affect the outcomes of student performances and therefore needs to be investigated (Bachman, 2004; Eckes, 2015). In rater-mediated assessments such as speaking tests, these facets may include candidate ability, task difficulty, and rater severity. By estimating these parameters simultaneously on a shared measurement scale, MFRM enables the identification of systematic differences among raters, making it a useful approach for examining interrater reliability and rating consistency (Eckes, 2015).

Previous studies evaluating the validity of oral performance-based assessment have used the MFRM as a means to analyze the reliability of a test's facets and its overall reliability as an L2 speaking assessment. A study by Bijani et al. (2022) used the model to investigate the difficulty in the scale categories of an oral proficiency test taken by 200 Iranian students. The study investigated the impact of rater training on severity and consistency on scale categories through a pre-post method design using the MFRM to analyze facets including test-takers, rater and rater group, scoring criterion, and interactions amongst them. It was found that although training did not have a large effect on raters' differences, the reliability was relatively high purportedly due to the high validity of the scoring rubric.

In the Japanese context, Bonk and Ockey (2003) utilized many-facet Rasch measurement analysis to examine a peer group discussion task with over 1000 Japanese English-major university students. The

facets analyzed included examinee, prompt, rater, and rating category items. The study found that although the variance in rater severity was large, the differences stabilized over time with returning raters becoming more severe and consistent. Moreover, general validity was found in rating scales, however, pronunciation and communicative skills were found to be difficult to rate when nearing the ends of scales. The reliability of peer discussion task activities for testing purposes has also been analyzed using Rasch analysis by Garside (2024). In this study, the facets of the examinee, rater, and category items were analyzed. The findings concluded that raters differed in severity but demonstrated sufficient internal consistency for the model to control for disparities. In another study, Koizumi et al. (2017) examined the rater reliability of the speaking section of the Global Test of English Communication Computer-Based Testing (GTEC CBT). MFRM was used to analyze the rating data of facets of test takers, criteria, and raters. The study found that rater severity and bias were low, and rater agreement and self-consistency were high. These studies demonstrate the applicability of using multifaceted Rasch measurement for analysis of aspects of oral-based performance assessment, including rater reliability.

Research questions

The literature demonstrates that although speaking tasks have been evaluated in the Japanese context, there are gaps in the research concerning one of the most popular oral-based performance assessments in Japan. Considering the foregoing discussion, the current study investigates the generalization validity of the EIKEN speaking test through analysis of measurement reliability across examinees, items, raters, and rating-scale categories. The research was guided by the following questions:

1. Do test items show consistency in their level of difficulty?
2. Does the test differentiate between examinees' level of proficiency for the target language use domain?
3. Do the rating-scale categories function as intended when applied by raters?
4. Do raters perform consistently, or do they differ in severity or leniency?

Method

Participants

To consider the validity of the EIKEN speaking test, a small-scale study was conducted at a senior high school in Japan. Ten recordings of a practice test for the Grade 2 EIKEN were evaluated by seven raters. The examinees were senior high school students taking a course in English proficiency tests. Ten students in the course volunteered to do recordings of the practice test for the study provided that they would be able to receive feedback on their performance after the session ended. Ages ranged from 16 to 18 years. All students were briefed on the format and expectations of the test as part of their course.

The raters were recruited based on their experience of evaluating EIKEN speaking test performance either in the role of a teacher practicing with a pupil, or as an official certified rater for EIKEN. All raters received training in the form of directives from the official EIKEN manual to (re)familiarize themselves with the scoring criteria. Raters 2, 4, 6, and 7 had prior official training through EIKEN, whereas Raters 1, 3, and 5 had not. Due to nondisclosure agreements, the officially trained raters were unable to specify the degree of their experience.

Speaking test procedure

Each of the ten examinees individually participated in the same practice test with the same interviewer (see Appendix). The test is divided into five sections. Table 2 details the sections of the test and the codes used for analysis.

Table 2
The Contents of the Second Stage of Eiken Grade 2

Items	Descriptions	Item Label	Criteria Label
Reading Out Loud	The examinee reads a passage out loud	ROL	CR
Question 1	The examinee is asked one question about the passage to assess their comprehension	RCC	C1
Question 2*	The examinee describes a sequence of three pictures; The examinee is given a prompt to start their response	Seq	C2C / C2G
Question 3	The examinee's opinion on a topic broadly related to the passage is elicited	PetOp	C3/4
Question 4	The examinee's opinion on a topic related to everyday life is elicited	EdOp	

Note. Examinees receive two separate scores for response to Question 2.

At the beginning of the test, the examinee was given a card which contained a passage of approximately 60 to 70 words. The card also had a three-panel sequence of illustrations that portrayed a story with a one-sentence prompt of how the story should begin. The examinee read the passage silently to themselves for 20 seconds then read the passage out loud (Reading Out Loud; ROL). The examiner then asked a comprehension check question based on the passage (Reading Comprehension Check; RCC). The examinee was then given 20 seconds to examine the sequence of pictures and prepare to narrate the story. For the second question, the examinee was asked to give narration of the story beginning with the prompt provided on the card (Sequence of Pictures; Seq). The examinee was then asked to turn the card over and was asked the third question which elicited the examinee's opinion about a topic broadly related to the passage. In the case of this test, the question was about pets (Pet-related Opinion; PetOp). The fourth question elicited the examinee's opinion about a topic related to everyday life. For this test, the question was about the educational importance of club activities (Education-related Opinion; EdOp). The structure of the Eiken Grade 2 test meets the Rasch model requirements, with all items being locally independent and measuring a single underlying construct (oral proficiency) (Eckes, 2015; Fulcher, 2014).

Each test was audio-recorded. Video recordings were not taken to protect the identity of the students, thus, points for "Attitude" were eliminated from the criteria usually used in the EIKEN speaking test. Recordings of the examinees were supplied to the raters according to a matrix to assure that every test was evaluated by five different raters. Each rater assessed seven to ten tests depending on personal time constraints; Raters 1, 2, and 3 evaluated all ten recordings, and Raters 4, 5, 6, and 7 evaluated seven recordings each.

Each rater evaluated the examinee's recording based on the EIKEN scoring criteria which includes the standards of evaluation for comprehension, phonological and linguistic accuracy, logicity of opinions, etc. The scoring criteria also details the means of rewarding or penalizing examinees for varieties of specific situations, such as particular insufficiencies and errors.

Data analysis

To maximize the degree of connectedness among facets, a rating design was devised based on the number of recordings assigned to each rater. The design satisfied the minimum requirements outlined by Linacre (2025b), who recommends at least 30 observations per test and 10 observations per rating-scale category to ensure reliable results in an MFRM analysis. The number of observations per examinee ranged from 30 to 42. In addition, across the first three items, the number of rating-scale categories meeting Linacre's

minimum threshold of 10 observations ranged from two categories (for ‘Reading Out Loud’) to four (for assessing the amount of content for the response to Question 2), with Question 1 providing three valid categories. Since Questions 3 and 4 shared the same scale, the combined total of scores produced valid categories for all five levels.

When scoring the Eiken Grade 2 test, raters use distinct rating scales for the first three tasks, whereas Questions 3 and 4 share a common scale. Raters were assigned identification numbers (1–7), and examinees were likewise assigned numbers (1–10). However, when the MFRM analysis was first conducted in the MINIFAC program (see Linacre, 2025a), the item facet was found to lack connectivity. The absence of connectivity was attributable to the use of distinct rating scales for three items, which eliminated cross-item overlap and prevented items from being placed on a shared measurement scale. To address this issue, group anchoring of the item facet was implemented with the average item difficulty fixed at a uniform mean (u-mean) of four (Linacre, 2025b). This adjustment aligned examinee, rater, and criteria measures on the same logit scale as the items. A relatively small u-mean was selected to facilitate clear visualization of all measures on the Wright map. Importantly, this transformation shifts the measurement scale upward without altering the relative positions of examinees or items, nor the probabilities of success, thereby preserving the interpretability of the measurement framework.

Results

Facet analysis of examinee, rater, and items

As illustrated by Figure 2 below, all four facets (examinee, rater, item, and criteria) are clearly shown on adjacent vertical columns, along with each rating scale displayed separately. As mentioned, the average difficulty of the item facet is now centered at 4 logits following the group anchoring procedure.

criteria measurements reflect that the comprehension question for the passage (C1; .52 logits) is the most challenging criterion for students to achieve a high score. The higher difficulty may indicate that students might struggle to demonstrate understanding beyond simple recitation. In contrast, simply reading the passage out loud (CR; -.40 logits) is much easier, indicating that students likely can read passages aloud with relative fluency, although they may read primarily for delivery rather than for meaning. The opinion-based questions (C3/4; .25 logits) targeting reasoning skills also seem relatively difficult for examinees, perhaps reflecting that Japanese students at this stage may not yet fully have developed the ability to express extended opinions. Slightly above average in difficulty is the category which evaluates the linguistic accuracy of the story narration (C2G; .10 logits), indicating that linguistic structures and systems could be moderately challenging for Japanese learners at this stage. Finally, C2C (-.47 logits), which evaluates the content of the narration, appears to be the easiest criterion. This may be attributable to the three-picture prompts providing substantial contextual support, including a starting prompt and speech bubbles which potentially help guide content generation. On the far right of the Wright map, each rating scale corresponding to a criterion is displayed separately. The three horizontal dots represent the rating scale category thresholds, or the logit points where a rater is equally likely to choose either of the two adjacent categories (Bond & Fox, 2015; Linacre, 2002). Moderate to wide gaps between thresholds generally indicate clear distinction for raters, whereas mid- to lower-level thresholds appear to be more compressed, making these scores harder for raters to differentiate. All criteria show some gaps in category usage, possibly due to small sample size or rater-calibration issues.

The MFRM analysis yielded several noteworthy findings regarding the functioning of the assessment. Referring to Table 3, examinee measures demonstrated a wide distribution, with high separation (3.92) and reliability (.94), suggesting that the instrument was effective in distinguishing performance levels within this sample. The Rater facet, by contrast, showed virtually no separation and zero reliability, which may reflect a high degree of consistency among raters and minimal evidence of systematic severity or leniency effects. Item separation was low (.84) with modest reliability (.41), which may suggest that the tasks did not vary substantially in difficulty, although this result should be interpreted with caution considering the small sample size. The Criteria facet showed stronger separation (2.51) and reliability (.86), potentially indicating that the rating scales captured meaningful distinctions overall.

Table 3

Statistics Summary for Each Facet

Facet	Mean	SE	Fixed Chi-Square	df	Separation	Reliability
Examinee	4.22	0.19	144.0	9	3.92	0.94
Rater	0	0.16	6.1	6	0	0
Item	4.00	0.14	8.3	4	0.84	0.41
Criteria	0	0.14	40.0	4	2.51	0.86

However, except for the C3/4 rating scale, between two and four categories of each criterion were consistently scored. Furthermore, the underuse of categories may be attributable to the limited number of examinees and raters, while the presence of narrow threshold gaps may indicate difficulties among raters in distinguishing between adjacent categories or possibly inconsistently interpreting the intended scale calibrations, such as penalizations.

To further confirm these findings, the Rasch-Thurstone thresholds were examined to assess how effectively each rating scale functions and whether raters can consistently distinguish between adjacent categories.

Category functioning

Due to the small sample size and the presence of underused categories across most criteria (particularly in CR and C1), the average measures appeared to show non-monotonic ordering. This disordering persisted even after the underused categories were removed and the analysis was repeated. As a result, category functioning for all criteria except those for the opinion-based questions (C3/4) was compromised by sparse and uneven category counts, limiting the stability of threshold estimates and interpretations of the results. In contrast, criterion C3/4 was used to rate two separate questions, yielding a sufficient number of observations for stable estimation; its category measures increased monotonically and met Linacre's (1999, 2002, 2025b) recommended standards for stable measurement.

Table 4

Category Performance for Criterion C3/4

Category Scores	Frequency (%)	Average Measures	Outfit MnSq	Rasch-Thurstone Thresholds	Category PEAK Probabilities%
1	14 (12)	-1.19	.6	low	100
2	21 (18)	-.20	1.9	-1.55	33
3	42 (36)	.01	.8	-.68	46
4	28 (24)	.44	.9	.52	40
5	13 (11)	.64	1.1	1.68	100

Analysis of the C3/4 rating scale showed generally acceptable category functioning. Category-level fit was acceptable overall, although Category 2 showed slightly elevated Outfit Mean-Square (MnSq) of 1.9, which may indicate greater variability in rater use at this level. Thresholds advanced in the expected order, but the spacing between adjacent thresholds was relatively narrow. Nevertheless, each category displayed a distinct peak probability, suggesting that all response options were actively and meaningfully used by raters. The results appear to indicate that the C3/4 scale operates reliably, though distinctions between adjacent categories—particularly in the mid-range—are relatively fine. The elevated variability observed for Category 2 may reflect how raters interpreted situations that necessitated penalization. For instance, in cases whereby “a maximum score of 2 points” is to be given to particular responses, some raters may have more strictly adhered to this standard than others. As such, this does not necessarily mean there is a substantive problem with the scale. Moreover, compressed thresholds are generally best interpreted as limited category separation rather than dysfunction, a pattern often associated with modest sample sizes and rater decision-making behavior. As all categories were sufficiently used and ordered as intended, there was no indication that category collapsing was necessary. Overall, the C3/4 criterion appears appropriate for continued use, with the expectation that clearer separation between categories may emerge with larger samples. The next section therefore turns to rater fit statistics to evaluate whether this variability could be associated with differences in rater behavior.

Rater fit and consistency

The Wright map indicated negligible separation and zero reliability for the Rater facet, suggesting overall consistency in rater severity; however, rater fit statistics were examined to assess whether any individual raters might exhibit misfit. Exact agreement percentages ranged from 38.9% to 47.7% across raters and exceeded the agreement expected by the model (approximately 33–34%) in all cases, indicating generally good interrater reliability. However, the uniformly elevated agreement also suggests that raters may have used the rating scale in a relatively similar and conservative way. As the EIKEN exam is a high-stakes assessment, more stringent control limits for rater fit statistics were applied, with acceptable Infit and Outfit mean-square values set between 0.80 and 1.20, as recommended by Myford and Wolfe (2003). In

Table 5, raters are arranged according to their severity measures, with higher severity appearing toward the top and greater leniency toward the bottom.

Table 5
Rater Severity, Fit, and Agreement Statistics

Rater	Severity Measure	SE	Infit MnSq	Outfit MnSq	Exact Agree.	
					Obs%	Exp%
4	0.13	0.17	0.83	0.81	47.70	33.30
2	0.13	0.14	1.43	1.49	44.60	33.70
1	0.09	0.14	1.12	1.22	44.90	33.80
5	0.04	0.16	0.60	0.56	46.70	33.70
7	0.03	0.17	0.66	0.69	38.90	33.30
6	-0.17	0.17	0.93	0.95	40.50	33.80
3	-0.25	0.14	1.15	1.10	44.60	33.60

Under the high-stakes fit criteria, three raters (Raters 3, 4, and 6) demonstrate acceptable severity and fit. Rater 1 shows marginal Outfit misfit suggesting isolated inconsistencies but remains within acceptable Infit limits indicating generally stable scoring behavior. Rater 5's overfit, together with high exact agreement, may suggest limited differentiation across performance levels. By comparison, Rater 7's overfit occurs without a high exact agreement, which may indicate consistent application of the rating criteria rather than a restricted use of the scoring scale. Rater 2, however, exhibited elevated Infit (1.43) and Outfit (1.49) MnSq values, exceeding the more stringent high-stakes control limits but remaining within generally acceptable Rasch thresholds of 1.50 (see, e.g., Eckes, 2015; Green, 2013; Linacre, 2025b; McNamara, 1996; Min & Aryadoust, 2021). This pattern appears to suggest some degree of unpredictability in scoring, though not to an extent necessarily indicative of random or dysfunctional rating behavior. While Rater 2's fit statistics do not warrant retraining, the elevated values may indicate that closer monitoring or focused calibration feedback may be beneficial in future administrations. Although overall rater fit and consistency were largely acceptable, fit statistics alone do not reveal whether individual raters interacted differently with specific examinees, items, or rating criteria. Therefore, bias and interaction analyses were conducted to examine whether any systematic rater-examinee, rater-item, or rater-criteria effects were present.

Rater bias and potential interaction effects

Three bias interactions were performed on the data: rater-examinee, rater-item, and rater-criteria. Raters 3, 5, 6, and 7 showed no evidence of potential bias across all examinees, items, or criteria ($p > .05$). However, Raters 1, 2, and 4 were all flagged for specific biases related to one or more of the targeted facets. Table 6 summarizes the types of statistically significant interactions indicating localized rather than systematic rater bias.

Table 6
Statistically Significant Rater-target Bias Interactions

Rater	Target Facets	Direction	Target-Joint Contrast	<i>t</i>	<i>p</i> < 0.05
1	Item ROL + Criterion CR	More lenient	2.54	3.10	.013
1	Item ROL+ Criterion CR	Harsher	-2.09	-2.37	.039
1	Criterion CR	More lenient	1.45	2.21	.046
1	Criterion CR	More lenient	1.44	2.37	.030
2	Examinee 2	Harsher	-1.62	-2.34	.044
4	Criterion C2G	Harsher	-1.76	-2.65	.023
4	Criterion C2G	Harsher	-1.55	-2.33	.040

Rater 1 accounted for the majority of flagged interactions, particularly in relation to criterion CR. Most of these interactions reflected more lenient-than-expected scoring, although one interaction indicated harsher-than-expected scoring involving item ROL and criterion CR. Taken together, this pattern may suggest a criterion-specific tendency for Rater 1 toward leniency on CR, with occasional context-dependent deviations in the opposite direction. Possible reasons for the sensitivity of criterion CR for Rater 1 are that the criterion may involve multiple components of speaking performance, which can be assigned different weights of importance when judgments are made. The rater may prioritize certain aspects of performance over others, potentially resulting in leniency or severity that deviates from the model's expectations. Another possible source of this bias is that the criterion descriptors may be broadly or ambiguously worded, allowing for greater variation in rater interpretation. In comparison, Rater 2 showed a single significant interaction with Examinee 2, which may indicate an isolated instance of harsher-than-expected scoring rather than a recurring pattern. Rater 4 demonstrated two significant interactions involving criterion C2G, both indicating harsher-than-expected scoring. While these effects were consistent in direction, they were limited in number and appear to reflect criterion-specific severity rather than global rater harshness.

Overall, the findings suggest that rater bias in this dataset is target-specific and context-dependent, with evidence supporting targeted recalibration for criterion CR, particularly for Rater 1, and continued monitoring rather than intervention for the other raters.

Discussion and Conclusions

This study addressed the gaps in research concerning the validity of the EIKEN test. It examined the generalization validity of the EIKEN speaking test through the analysis of reliability of the test items, examinees, raters, and method of evaluation. The EIKEN internal studies show test items to be relevant to the target language use (TLU) domain (oral communication) of each level (Benson, 2013; Dunlea, 2008; Noguchi et al., 2007; Plumb & Watanabe, 2016). The current study appears to confirm that individual test items of the speaking test showed limited variability in difficulty within the second-grade level. Moreover, the criteria seemed to capture meaningful distinctions in student performance. This may

suggest that the test effectively distinguished between high- and low-performing examinees. Moreover, the analysis was able to distinguish which aspects of the TLU may be more challenging for Japanese language learners based on the standards of evaluation for the EIKEN speaking test. For instance, facet analysis of categories appears to demonstrate that these particular students in this small-scale study struggled with reading comprehension and linguistic accuracy in sentence construction on the test. This seems to align with recent findings that reveal a decline in Japanese students' performance in these areas on nationwide academic achievement tests ("National Assessment of Academic Ability," 2024).

Regarding the raters, the analysis showed that the raters generally applied scoring criteria consistently, with only isolated deviations observed. Most raters displayed stable scoring patterns with a few showing evidence of overfit or minor unpredictability. In addition, the bias analyses largely confirmed that the vast majority of raters exhibited no systematic bias, with only a small number of isolated interactions possibly indicating localized deviations that did not compromise overall rating consistency. This may reflect that the rater panel maintained reliable and fair scoring practices overall. This congruous scoring behavior could suggest that the evaluation method used for the EIKEN speaking test is reliable enough for raters to consistently agree on examinee output. However, it is also possible that the lack of variability in scoring behavior may signify a restricted category range in the standards of evaluation.

Many criteria for the test items were affected by sparse or uneven category use, likely due to the small sample and rater decision patterns. The category functioning for the opinion-based question was generally acceptable enough to statistically evaluate. However, analysis showed that some decision-making behavior of the raters caused variability in category use, possibly due to interpretations of penalizations. One reason for this may be that standards for evaluating the content of the response were overlooked when the examinee demonstrated fluency in other areas, such as pronunciation and syntax. This corresponds to Bachman's (2004) concerns of interrater reliability in which there may be inconsistency in how responses are scored when there are different values of certain language features. While these deviations were not severe enough to make suggestions about rater training or retraining, targeted calibration for these specific scenarios could potentially support continued scoring quality.

The findings suggest several practical implications for rater calibration and scoring practices in the EIKEN speaking test. Although overall rater consistency appeared acceptable, the analysis identified some localized variability in scoring behavior related to specific criteria and rating scale use. In particular, the comprehension criterion (CR) appeared sensitive to rater interpretation, possibly because it requires raters to evaluate multiple aspects of performance simultaneously. Targeted calibration focusing on how comprehension should be interpreted, especially distinguishing between simple recitation and evidence of understanding, may therefore help support greater consistency in scoring.

In addition, variability in category use suggests that raters may sometimes interpret penalization rules differently when evaluating responses containing errors or insufficient content. Clarifying how and when score limitations should be applied may help reduce inconsistencies in these situations. Finally, compressed thresholds between adjacent score levels indicate that raters may find it challenging to differentiate between mid-level categories on the rating scale. Calibration activities involving borderline performances may help raters better recognize the levels of performance intended between adjacent score levels and may support continued scoring consistency in rater mediated speaking assessments such as the EIKEN interview.

In sum, the assessment appears to demonstrate effective differentiation of examinee ability, sound psychometric functioning, consistent rater performance, and generally reliable use of rating scales. While the restricted sample size for both examinees and raters limits definitive inferences that can be made about the EIKEN test in general, the findings of this study may bring us closer to assuming that the speaking test can claim generalization validity. Such observations may prove useful to educators and stakeholders

needing insight into the effectiveness of using a school-age EFL learner's EIKEN grade level to interpret their oral communication competency.

Further Research and Limitations

The small sample of examinees and raters was a limitation in the present study. A greater sample of test takers and raters would have allowed for more meaningful interpretations regarding the analysis of examinee, rater and item facets. In particular, further research on a larger scale could lend more insight into the rating-scale functioning to determine whether the underuse of categories was due to ambiguous thresholds or the small sample size. Moreover, while all raters received instructions based on the manual, they did not all have access to the training and opportunities to practice scoring received by official proctors to support evaluation quality. Ideally, using only raters with experience as official EIKEN interviewers would bolster claims of reliability.

References

- 2023 年度版 英検 2 級 過去 6 回全問題集 / 旺文社 [2023 Edition Eiken Grade 2 Past 6 Exam Questions] (pp. 142–143). Obunsha.
- Bachman, L. F. (2004). *Statistical analyses for language assessment book*. Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
- Bahrouni, F. (2016). Using multi-facet Rasch model (MFRM) in rater-mediated assessment. *The Journal of Teaching English for Specific and Academic Purposes*, 4(1), 195–212.
- Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 3, pp. 1301–1322). Wiley-Blackwell.
<https://doi.org/10.1002/9781118411360.wbcla070>
- Benson, S. (2013). A critical analysis of the STEP Eiken Test's validity and reliability. *The Journal of Kanda University of International Studies*, 25, 95–102.
- Bijani, H., Hashempour, B., & Said Bani Orabah, S. (2022). Development and validation of a training-embedded speaking assessment rating scale: A multifaceted Rasch analysis in speaking assessment. *International Journal of Research in English Education*, 7(3), 32–45.
<https://doi.org/10.52547/ijree.7.3.32>
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed). Routledge.
- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89–110.
- Brown, H. D. (2007). *Teaching by principles* (3rd edition). Longman.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2011). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.
- Cubilo, J. (2014). Argument-based validity in classroom and program contexts: Applications and considerations. *JALT Shiken Research Bulletin*, 18(1), 18–24.
- Dai, D. W., Vu, T., Knoch, U., Lim, A. S., Malone, D. T., & Mak, V. (2024). Expanding Kane's argument-based validity framework: What can validation practices in language assessment offer health professions education? *Medical Education*, 58(12), 1462–1468.
<https://doi.org/10.1111/medu.15452>
- Dubien, M. (2023). (PDF) Are Eiken tests really English proficiency exams? The case of the Eiken Grade 1 Test. 沖縄キリスト教学院大学論集 [Okinawa Christian University Journal], 21, 1–7.

- Dunlea, J. (2008). The EIKEN Can-do List: Improving feedback for an English proficiency test in Japan. *Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment*, 31, 245–262.
- Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd revised ed.). Peter Lang. <https://doi.org/10.3726/978-3-653-04844-5>
- EIKEN | Eiken Foundation of Japan. (2022). <https://www.eiken.or.jp/eiken/en/>
- Fulcher, G. (2014). *Testing second language speaking*. Routledge.
- Garside, P. (2024). Investigating the assessability of speaking proficiency in a group discussion context. *TEVAL - Shiken: A Journal of Language Testing and Evaluation in Japan*, 28(1), 1–18. <https://doi.org/10.37546/JALTSIG.TEVAL28.1-1>
- Green, R. (2013). *Statistical analyses for language testers*. Palgrave Macmillan.
- Hughes, A., & Hughes, J. (2020). *Testing for language teachers* (3rd ed.). Cambridge University Press.
- Ishida, M. (2004). 英語教員が備えておくべき英語力：英検準1級、TOEFL550点、TOEIC730点の目標値を中心に [English ability required for English teachers: With respect to the benchmarks of the EIKEN exam Level Pre-1, TOEFL score of 550, and TOEIC score of 730]. *ELEC Bulletin*, 111, 10–17.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th edition, Vol. 4, pp. 17–64). Greenwood Publishing.
- Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, 42(4), 448–457. <https://doi.org/10.1080/02796015.2013.12087465>
- Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessments (PELAs). *Papers in Language Testing and Assessment*, 2(2), 48–66. <https://doi.org/10.58379/YZLQ8816>
- Koizumi, R., Okabe, Y., & Kashimada, Y. (2017). A multifaceted Rasch analysis of rater reliability of the speaking section of the GTEC CBT. *ARELE: Annual Review of English Language Education in Japan*, 28, 241–256. https://doi.org/10.20581/arele.28.0_241
- Linacre, J. M. (1989). *Many-faceted Rasch measurement*. The University of Chicago.
- Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, 3, 103–122.
- Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3, 85–106.
- Linacre, J. M. (2025a). *Minifac Rasch measurement computer program (Version 4.4.4) [Computer software]*. Available at: <https://www.winsteps.com/minifac.htm> [Online]. [Accessed Aug 4th 2025].
- Linacre, J. M. (2025b). *A user's guide to FACETS: Rasch-model computer programs*. *Winsteps.com*. Available at: <https://www.winsteps.com/winman/copyright.htm> [Online]. [Accessed Aug 5th 2025].
- Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Merrill.
- MacGregor, L. (1997). The Eiken Test: An investigation. *JALT Journal*, 19(1), 24–42.
- Marlowe, J. P., & Asaba, M. (2017). Investigating the cognitive processes of translation writing tasks. In P. Clements, A. Krause, & H. Brown (Eds.), *Transformation in language education: JALT 2016 conference proceedings* (pp. 369–376). JALT.
- McCowan, R. J., & McCowan, S. C. (1999). *Item analysis for criterion-referenced tests* (No. ED501716; pp. 1–39). Center for Development of Human Services. <https://eric.ed.gov/?id=ED501716>
- McNamara, T.F. (1996). *Measuring second language performance*. Longman.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.
- Min, S., & Aryadoust, V. (2021). A systematic review of item response theory in language assessment: Implications for the dimensionality of language ability. *Studies in Educational Evaluation*, 68, 100963. <https://doi.org/10.1016/j.stueduc.2020.100963>

- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386–422.
- Nagashima, K. (2001). TOEIC、英検、中学校、高校で求められている英単語の段階的に分類 [Stepwise classification of English vocabulary required for TOEIC, the EIKEN exams, junior high schools, and high schools]. *STEP Bulletin*, 13, 184–201.
- National assessment of academic ability: Poor reading and writing skills cannot be left unaddressed. (2024, July 30). *The Yomiuri Shimbun*. <https://www.yomiuri.co.jp/editorial/20240730-OYT1T50001/>
- Nielsen, B. (2000). Determining test reliability and quality of Eiken Test items: A statistical analysis of first year kosen student responses to test items of an Eiken third level test. *釧路工業高等専門学校紀要 [Research Reports Kushiro National College of Technology]*, 34, 81–93.
- Noguchi, H., Kumagi, R., Wakita, T., & Wada, A. (2007). 日本語 Can-Do-statements における DIF 項目の検出 [Detecting DIF items in Japanese Can-do-statement]. *Japan Language Testing Association Journal*, 10, 106–118. https://doi.org/10.20622/jltaj.10.0_106
- Piggin, G. (2011). An evaluative commentary of the Grade 1 EIKEN Test. *Language Testing in Asia*, 1(4), 145–167. <https://doi.org/10.1186/2229-0443-1-4-144>
- Plumb, C., & Watanabe, D. (2016). A critique of the Grade 2 EIKEN test reading section: Analysis and suggestions. *Shiken*, 20(1), 12–17.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Original work published 1960). Danish Institute for Educational Research.
- Schaefer, E., & Martin, J. (2023). Language testing in changing times: An interview with Professor Daniel Isbell. *Shiken*, 27(2), 1-5. <https://doi.org/10.37546/JALTSIG.TEVAL27.2-1>
- Wind, S. A. (2014). *Evaluating rater-mediated assessments with Rasch measurement theory and Mokken scaling* [Dissertation, Laney Graduate School].

Appendix: Interview Card

Reprinted from 2023 年度版 英検 2 級 過去 6 回全問題集 / 旺文社 [2023 Edition Eiken Grade 2 Past 6 Exam Questions], pp. 142–143.

二次試験
面接

問題カード (A日程) ▶ MP3 ▶ アプリ ▶ CD 3 55~57

A New Type of Service

Nowadays, many people are interested in having a pet. However, when pets become sick or get hurt, the cost of medical treatment can be high. To deal with this, it is a good idea for pet owners to get insurance that covers pets' medical costs. Some companies offer such insurance, and by doing so they try to meet the needs of pet owners.

Your story should begin with this sentence: **One day, Yumi and her father were talking in their car.**



Questions

- No. 1 According to the passage, how do some companies try to meet the needs of pet owners?
- No. 2 Now, please look at the picture and describe the situation. You have 20 seconds to prepare. Your story should begin with the sentence on the card.
<20 seconds>
Please begin.
- Now, Mr. / Ms. —, please turn over the card and put it down.
- No. 3 Some people say that having a pet can help people reduce their stress. What do you think about that?
- No. 4 Today, many students take part in club activities at school. Do you think club activities are an important part of school education?
Yes. → Why?
No. → Why not?

Call for Papers

Shiken: A Journal of Language Testing and Evaluation in Japan is seeking submissions for publication in the December 2026 issue. Submissions received by 1 September, 2026 will be considered, although earlier submission is encouraged to allow time for revision. Shiken: A Journal of Language Testing and Evaluation in Japan aims to publish articles concerning language assessment issues relevant to classroom practitioners and language program administrators. This includes, but is not limited to, research papers, replication studies, review articles, informed opinion pieces, technical advice articles, and qualitative descriptions of classroom testing issues. Article length should reflect the purpose of the article. Short, focused articles that are accessible to non-specialists are preferred and we reserve the right to edit submissions for relevance and length. Research papers should range from 4000 to 8000 words, but longer articles are acceptable provided they are clearly focused and relevant. Novice researchers are encouraged to submit, but should aim for short papers that address a single research question. Longer articles will generally only be accepted from established researchers with publication experience. Opinion pieces should be of 3000 words or less and focus on a single main issue. Many aspects of language testing draw justified criticism and we welcome articles critical of existing practices, but authors must provide evidence to support any empirical claims made. Isolated anecdotes or claims based on "commonsense" are not a sufficient evidential basis for publication.

Submissions should be formatted as a Microsoft Word (.doc or docx format) using 12 point Times New Roman font, although plain texts files (.txt format) without formatting are also acceptable. The page size should be set to A4, with a 2.5 cm margin. Separate sections for tables and figures should be appended to the end of the document following any appendices, using the section headings "Tables" and "Figures". Tables and figures should be numbered and titled following the guideline of the *Publication Manual of the American Psychological Association, Seventh Edition*. Within the body of the text, indicate approximately where each table or figure should appear by typing "Insert Table x" or "Insert Figure x" centered on a new line, with "x" replaced by the number of the table or figure.

The body text should be left justified, with single spacing for the text within a paragraph. Each paragraph should be separated by a double line space, either by specifying a double line from the Microsoft Office paragraph formatting menu, or by manually typing two carriage returns in a plain text file. Do not manually type a carriage return at the end of each line of text within a paragraph.

Each section of the paper should have a section heading, following the guidelines of the *Publication Manual of the American Psychological Association, Seventh Edition*. Each section heading should be preceded by a double line space as for a regular paragraph, but followed by a single line space.

The reference section should begin on a new page immediately after the end of the body text (i.e. before any appendices, tables, and figures), with the heading "References". Referencing should strictly follow the guidelines of the *Publication Manual of the American Psychological Association, Seventh Edition*.

