neofilolog

Czasopismo Polskiego Towarzystwa Neofilologicznego
ISSN 1429-2173, eISSN 2545-3971, 2025, NR 64/1, 27–44
https://doi.org/10.14746/n.2025.64.1.3
http://poltowneo.org/

Yuki Otsuki
Meisei University, Japan
https://orcid.org/0000-0001-7331-0198
yuki.otsuki@h.meisei-u.ac.jp

Assessment of speaking interaction: Rater perception and testing criteria in Japanese elementary English education

This study aimed to explore assessment criteria for speaking interactions in performance tests conducted under the curriculum of formal elementary school English education in Japan, that are intended to foster students' communicative competence in an authentic language use environment. Building on the framework of Sato and McNamara (2018), a qualitative analysis of rater perceptions of communicative elements of young learner oral interaction was conducted. The current study examined the perception of nine raters. The material for analysis consisted of transcripts from video recordings of the elementary school students' role-play test. The findings indicated that the *flow of interaction* had a particularly strong impact on raters, along with five other factors. Perceptions of the use of *L1* and *nonverbal behaviors* were nuanced, and only sometimes viewed favorably. Furthermore, in interaction involving limited vocabulary and use of formulaic expressions, speed was not always perceived positively and was occasionally seen as rote memorization. These insights could be applied in the holistic assessment of speaking tests in young beginner learners.

Keywords: oral communication assessment, elementary English in Japan, factors affecting rater perception

Słowa kluczowe: ocena komunikacji ustnej, językangielski na poziomie podstawowym w Japonii, czynniki wpływające na percepcję oceniających

1. Introduction

Foreign language (FL) education in Japan is governed at the national level by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) through the *Course of Study*, which operates as a guideline (MEXT, 2017). The objective of FL education is to nurture communication skills among third and fourth grade students, who participate in 35 hours of FL activities per year and among fifth and sixth grade students learning an FL as a compulsory subject, involving 70 hours per year. Compulsory subjects require assessments each term, and MEXT outlines the assessment criteria which guide and maintain educational standards across the country. Aligned with this national framework, 47 prefectural boards of education operate in Japan, each adapting the guidelines to best serve the distinctive needs of their communities. They select textbooks for compulsory subject courses and secure additional teaching staff, or assistant language teachers (ALTs). While ALTs are sometimes placed in classrooms, English is mainly taught by generalist elementary school teachers many of whom have not studied methods of teaching English in their pre-service teacher training. Consequently, further support is needed, as only 6.7% are certified English teachers (MEXT, 2022). In this context, well-structured and supportive textbooks may play a vital role in ensuring more uniform instruction nationwide. However, each textbook includes approximately eight to nine opportunities for oral performance assessments throughout the year. Teachers, who may lack specialized knowledge in English teaching methods, are required to conduct these performance tests. Given this background, research and support in this area are particularly needed in Japan. As also highlighted by Nikolov and Timpe-Laughlin (2020), despite the clear focus on enhancing listening-comprehension, speaking, and interaction in age-appropriate teaching methodology and achievement goals for young learners, research on assessing young learners' oral and aural language abilities remains insufficient.

Given this need, this study aimed to explore the criteria for assessing the elementary school students' oral interaction, specifically within the framework of Japan's formal education system, where *the Course of Study* emphasizes the development of communication skills as the core educational objective. Drawing on Sato and McNamara (2018), which holistically analyzed listener perception to assess communication ability, I conducted a qualitative study to identify assessment criteria based on listener-as-rater perception. This study utilized audio data from classroom instruction, using the textbook provided by the local board of education, and having raters assess the students' role-play interaction to identify assessment

criteria. The following section reviews previous studies on children's oral interaction and assessment.

2. Literature review

2.1. Oral interactions

Children begin FL learning through auditory reception activities, such as songs and games, while engaging in play. Concerning cognitive processes, Hulstijn (2015) proposed basic and higher language cognition (BLC–HLC) theory, stating that BLC primarily consists of largely implicit knowledge in the domains of phonetics, prosody, phonology, morphology, and syntax, which is combined with explicit lexical knowledge. BLC is limited to the processing of oral language skills in utterances and contains high-frequency lexical, grammatical, phonotactic, and prosodic elements (Tracy-Ventura, et al., 2014). Johnstone (2009) contended that young learners of a second language (L2) initially receive sounds in whole chunks and that their reception becomes more analytical as they develop. From both the cognitive aspect of language learners' skill levels and the perspective of children's cognitive development stages, the approach of connecting auditory input to interaction is appropriate and aligns with second-language acquisition (SLA) theory (e.g., Ellis, 2015; Gass, et al., 2022; Lightbown, Spada, 2022). From the perspective of SLA theory, which emphasizes the shift from target language (TL) input to output, communication pressure promotes language acquisition. Long's (1980) interaction hypothesis posits that input is processed through negotiating meaning during interactions, providing opportunities for speech production. Mackey (2020) reviewed the cognitive-interactionist paradigm, defining how interaction and corrective feedback facilitate language acquisition through input, output, and feedback processes. This perspective highlights how interactional processes offer learning opportunities and how corrective feedback can enhance learners' linguistic outputs. However, assessments of oral interactions, especially among young learners, remain limited (Mora, 2006).

2.2. Assessing oral interactions in young learners

In terms of assessing oral interactions, research focusing on adults typically quantifies speech data from the perspectives of complexity, accuracy, and fluency, or develops rubrics appropriate to the context. In young leaners'

classroom, Genesee and Upshur (1996) advocated continuous and formative assessments that include supportive feedback to aid learning. Similarly, Butler (2005) introduced the concept of classroom harmonization to link learning and teaching in the classroom context: "It includes both the arrangement of the physical conditions of a given classroom as well as the integration of various psychological variables pertaining to both students and teachers" (Butler, 2005: 438). Likewise, Britton (2021) emphasizes the need for assessment practices tailored to the unique context of each classroom, a concept referred to as *assessment for learning,* which aims to effectively monitor and support students' progress. Furthermore, in young learners' classrooms, the importance of *assessment for learning* has been widely recognized (e.g., Brown, 2005; Butler, 2022; 2024; Carless, 2005; Nikolov, 2016; William, 2011; William, Thompson, 2017). Rather than focusing solely on assessing specific skill elements, this approach emphasizes the importance of a comprehensive, holistic understanding in assessing students' progress, which reflects the way children learn, as argued by Pinter (2017). Children are more holistic learners who focus on meaning and the whole message delivered instead of analyzing the structure of language. They are eager to use language immediately and tend to use it before learning the rules (e.g., Cameron, 2003; Rich, 2019; Rixon, 2016). To incorporate this holistic learning into assessment, understanding the perceptions of raters seemed essential. Therefore, identifying what elements of children's interactions are rated highly and what are rated lower became necessary. Aiming to gain a deeper understanding of the factors influencing assessment, I sought to thoroughly capture the raters' perception through a series of dialogues, based on Sociocultural Theory (SCT). SCT, which originated from Vygotsky's psychology on children's learning, emphasizes the importance of interactions with environmental factors, such as teachers' scaffolding, in supporting cognitive development, particularly in language learning. According to this theory, language mediates thought, and the social environment is not merely a setting for learning, but a source of development itself (Vygotsky, 1987; Lantolf, Poehner, 2014). Therefore, when assessing the language of young second language learners, it became crucial to understand how listeners, often teachers and raters in the classroom, interpret the development of interactions that take place uniquely within this specific context.

2.3. Layperson perspectives on L2 communicative competence

Sato and McNamara (2018) argued that the assessment of L2 communication ability has rarely been conducted from the perspectives of non-native speakers, even though interlocutors in real-world communication are not al-

ways trained language professionals, or native speakers. Their study focused on layperson views at a Japanese university, where 23 individuals provided intuitive ratings of their L2 communication ability after watching video performances. They rated performances on a 7-point scale and provided open-ended feedback without predetermined criteria, from which seven categories affecting perceptions of competence were identified: (a) English-language features, (b) overall communicative success, (c) content, (d) interaction quality, (e) nonverbal behaviors (NVBs), (f) speaker composure/attitude, and (g) other. These categories may serve as potential assessment criteria. However, the participants in that study were adults. Therefore, the current study aimed to apply Sato and McNamara's (2018) research method to investigate which elements of oral interaction in the FL education of Japanese elementary school beginners are perceived by listeners as demonstrating high communication ability. The results are likely to be useful for assessing oral interactions.

3. Research design and methodology

3.1. Purpose of the study

The purpose of the study was to investigate which elements of oral interaction in FL education for young beginner learners in Japanese elementary schools were perceived by raters as demonstrating high, or low, communicative ability. Drawing on Sato and McNamara (2018), this study sought to determine which elements affect the ease, or difficulty, of understanding oral interactions among elementary students, by adopting a partial replication study approach. While Marsden and colleagues (2018, p. 328) argue that "direct or close replication is not appropriate" for qualitative research involving multi-faceted interpretation, this study is arguably justified in replicating the framework of the original study, as it does not focus on ideologically-driven exploration. One of the notably distinctive features of this study is the selection of the raters. In contrast to Sato and McNamara's (2018) study, which used layperson raters, this study selected raters with some understanding of the context and developmental stages of children.

3.2. Research questions

Two research questions were established:

1) What elements of spoken interaction by Japanese fifth-grade students contribute positively to raters' impressions during the English role-play test?

2) What elements of spoken interaction by Japanese fifth-grade students negatively impact raters' impressions during the English role-play test?

3.3. Research paradigm

This study is grounded in the constructivist paradigm (Crotty, 1998), which emphasizes the co-construction of knowledge through interaction. The themes explored were developed through a process in which both the raters and the researcher (author) collaboratively deepened their understanding of raters' value perceptions.

3.4. Methodology

This research used a case-study methodology (Merriam, 1998), focusing on fifth-grade students in Japan. This approach was chosen to facilitate an in-depth exploration of the phenomena within the real-life context of classroom interactions, using role-play tasks from textbooks, aiming to simulate how students use language in real-world situations.

3.5. Participants

Nine raters participated in the study. The raters' demographics and background knowledge of Japanese elementary school English policies are summarized in Table 1.

Four men and five women participated (age range: 20–60 years). Seven participants were native speakers of Japanese, while one spoke Taiwanese Mandarin and another Tagalog as their L1. Additionally, information was acquired concerning their occupational backgrounds and knowledge of elementary school education of English.

Table 1. Summary of the raters' characteristics

Name	Age	Gender	Knowledge of English policy	Occupational background
A	50s	F	5	Junior high school English teacher (29 years) and Elementary school English advisor (1 year)
B	60s	F	5	Elementary school English teacher (3 years)

Table 1 – cont.

Name	Age	Gender	Knowledge of English policy	Occupational background
C	20s	F	3	Japanese lecturer in Chinese university (2 years)
D	30s	M	3	Senior high school English teacher (5 years)
E	30s	F	4	University lecturer (3 years)
F	30s	M	4	University lecturer (1 year) and cram school teacher (10 years)
G	20s	M	3	Cram school teacher (4 years) and Japanese language institute instructor (4 years)
H	30s	F	5	Assistant language teacher at an elementary school (3 years)
I	20s	M	3	Senior high school English teacher (1 year)

Note. Knowledge of English policies of Japanese elementary schools was self-rated on a 5-point scale (5 = well known, 3 = somewhat familiar, 1 = no knowledge)

3.6. Materials

Video recordings of children's interactions were collected, and transcriptions were prepared. These materials allowed participants to listen to authentic children's interactions. In a typical fifth-grade textbook, speaking- performance tests are included, such as tests that ask students to "introduce your friend" or "give a friend directions." The former is an example of speaking–presentation, while the latter is an example of speaking–interaction. For this study, the speaking-interaction test on "restaurant role-play" was chosen.

Fifteen transcripts were randomly selected from video recordings of one-on-one interactions between a total of 82 fifth-grade students and a Filipino female ALT who has been teaching throughout the year.

As the textbook suggests, seven class hours were used in this study to introduce and practice the following formulaic expressions:

A: "May I help you? What would you like?"
B: "I'd like a hot dog."
A: "Here you are."
B: "How much is it?"
A: "It's 300 yen."

The one-on-one role-play interactions with the ALT used phrases from their textbook in simulated restaurant scenes, as referenced in Szpotowicz and Lindgren (2011). Additionally, general greetings were also incorporated before the role-playing began. These role-play interactions were

video-recorded by the students themselves as routine work, and the recordings submitted to the teacher were transcribed.

The four transcriptions (out of 15) were specifically selected for their notable characteristics: one with the most interaction turns, one with the fewest interaction turns, one with the longest periods of silence, and one with the shortest total interaction time, indicating the fastest speech.

Supplemental material in the form of video-recordings was also collected from an out-of-school club. Five fifth-grade students who were familiar with the same textbook, and who belonged to this club, also performed the same role-play interaction while being video-recorded. This enabled the use of video materials, in addition to transcribed data, as only transcribed data were allowed to be used in public schools.

3.7. Data collection methods and procedure

Data were collected using open-ended questionnaires and semi-structured interviews. Raters were asked to intuitively judge the level of communicative competence, such as English proficiency, of speakers, assigning a four-point scale rating and then providing open-ended explanations for their assessment during the period from March to April 2022. All raters received 15 transcriptions and five videos of Japanese fifth-grade students' interaction data, in total approximately 40 minutes. Initially, they were asked to provide an overall rating (A = excellent, B = good, C = fair, and D = poor) to judge the speakers' communication abilities. Along with the scoring sheet, a questionnaire containing two open-ended questions regarding points and reasons for assigning high and low scores was also included. The questionnaire was distributed via email, and respondents had the option of returning their answers either by email, or by handing in a paper copy in person.

After the data were collected, four raters were asked to participate in further interviews to clarify written answers that the author had difficulty understanding. Each participant met the author individually face-to-face. Each participant was interviewed for 10 minutes to 1 hour. First, the author asked the participants whether they would like to add to the answers they had written in March. They were then asked to explain the meanings of expressions in their written answers that were difficult for the author to understand, such as "parroting" and "L1 content words," including the interpretation of filler use, and L1 use that appeared to be evaluated as both positive and negative. After the themes were summarized, a report was sent to each participant for member checking in May 2022.

3.8. Data analysis

The data were analyzed using thematic analysis (Braun, Clarke, 2006; Braun et al., 2019), which involved identifying, analyzing, and extracting themes from the raters' open-ended answers. The study aimed to inductively identify the reasons for high and low evaluations in student interactions. Additionally, a deductive approach was used to compare the findings with those from the initial evaluations.

3.9. Ethical consideration

In conducting this study, strict adherence to ethical guidelines was prioritized to ensure the protection of participants' privacy and integrity. Approval was obtained from the school principal to use verbatim transcripts of audio recordings of classroom interactions, with a clear condition that all data would be treated in accordance with children's privacy protection standards. Additionally, for data involving five children from an out-of-school international exchange organization, written consent was obtained from their parents or guardians affirming their agreement to participate and their understanding of how the data would be used.

For the raters involved in evaluating the interactions, written consent was obtained. Raters were informed of the privacy measures in place and agreed to participate by submitting their responses via an online form. This process ensured that all participants were fully aware of the study's aims and their rights, including confidentiality and the use to be made of the data they provided.

4. Results

4.1. Thematic analysis of positive elements

Regarding high evaluations, thematic analysis generated four themes from 14 codes. The highest percentage was accounted for by statements related to *interactions* (48.57%). *Interactions* included *flow*, *negotiation of meaning*, use of *communication strategies (CSs)*, *repetition*, and *no silence.* For example, "the students who ask for *clarification* and use *active responses* keep conversation *flowing*" (Rater A; B; G; E; I). Rater B specifically stated, "Using *active responses* and *fillers* to show that the student is thinking was more favorable than providing articulated quick responses in this interaction."

The second highest percentage was for *English language knowledge* (22.85%). Basic knowledge of *formulaic expressions* and *listening comprehension* ability, which was also expressed as an understanding of what the interlocutor asked, seemed to be the baseline for evaluation. Rater C pointed out that one of the primary causes of silence is a lack of listening skills and emphasized the importance of listening comprehension in English. A further 11.42% of the participant explanations highlighted *attitudes* such as *confidence*, *willingness to communicate*, and *listenership*. Rater E also mentioned affective factors such as "speaking without anxiety" and "students should not worry about making mistakes." McCarthy and McCarten (2018) identified four core concepts of conversational behavior: (a) organizing one's own talk, (b) taking into account other speakers, (c) listenership, and (d) organizing the conversation as a whole, with listenership involving showing positive attitudes to interlocutors and using expressions such as "right" and "Uh huh." These *active responses* form of *CSs*. Such strategies were included in the theme of *interactions*, when speakers physically expressed their intent both verbally and nonverbally. *Overall communicative success,* accounting for 8.57%, encompassed features such as *goal-oriented* and *task-related behaviors*; for example, "the goal of this interaction is to convey meaning using English, so being able to place an order and make a payment is a primary indicator" (Rater D; E). Only one code, *NVB*, was presented in one video sample, which showed a student nodding and using hand gestures to fill the interaction gaps. *L1 use* was not favored, except for *fillers*. Some questionnaire answers were vague and elicited mixed interpretations; therefore, further explanation of these answers

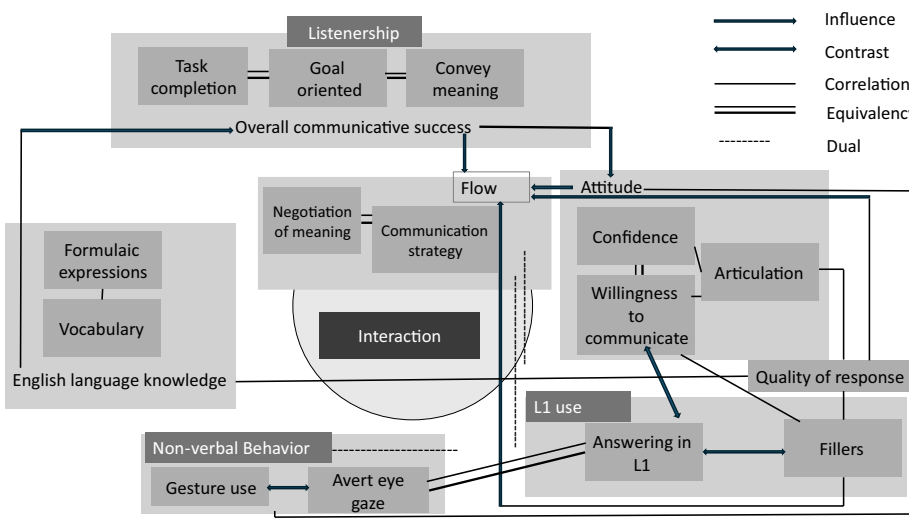

Figure 1. Themes concerning reasons for high initial evaluations

was requested later in the individual interviews. Figure 1 thematically maps the raters' initial high evaluation points. As shown, the main theme, *interactions*, overlapped with *English language knowledge*, *overall communicative success*, and *attitude*. The *L1* category emerged distinctively as negative in terms of *L1 use* and positive for *fillers*.

4.2. Thematic analysis of negative elements

The reasons for low ratings were generally the opposite of those for high ratings; for example, "a student stopped the flow of interactions" (Raters F; G; A) or "a student did not solve problems in conversations" (Raters B; I). As with the findings of Sato and McNamara's (2018) study, no codes related to *content* and *cohesion* were found among the reasons for high or low ratings in this study. Most of the themes that emerged corresponded to *interactions* (48.48%) and *English language knowledge* (18.18%). Two themes that markedly differed from the high-ratings findings were the number of *NVBs* (12.12%) used and the *quality of response*. *NVBs*, especially averting eye gaze, resulted in low ratings. *Quality of response* was variously specified as follows: (a) excessive *L1 use*, (b) *L1 use* for content words, (c) too many *fillers*, and (d) *mimicking* or *parroting*. Excessive use of *CSs* such as *fillers* and *repetitions* yielded a negative impression. Moreover, two answers indicated that some students pretended to understand the questions as, even if the students answered immediately, the dialogue was somewhat disconnected. Ambiguity was present in situations where both high and low ratings were given for *L1 use* and *quality of response*. Thus, follow-up interviews were conducted to clarify some answers.

4.3. Follow-up interviews

None of the respondents indicated that they wished to provide additional remarks. However, they were prepared to provide further examples from their teaching contexts that helped confirm that there were no discrepancies in the results.

Concerning unclear responses, clarification was sought regarding *parroting* and *L1 content words*.

Parroting was indicated by Rater B: "*Parroting* or *mimicking* is not a strategy intended to be a *filler* to connect interactional turns nor finding one's own words, but rather to pass the moment by responding immediately by repeating the word used in the question." Rater A noted: "The immediate

response made the role-play like a rote memorization test of *formulaic expressions*. In real-world interactions, *hesitation* and *rephrasing* seemed more natural than quick responses."

L1 use was explained by two participants, Raters G and H, as meaning that L1 was used not as a *filler* but as a response to the question, such instances were not seen as using CSs. Rater G stated that using *content words* not only revealed a lack of knowledge of the TL but also a lack of imagination, reflecting that the listener may not be able to understand the L1 that the students used. Rater H added that frequent *L1 use*, not only for *content words*, would make her feel as if her Japanese proficiency was dismissed and that she perceived the students' lack of imagination regarding the listeners' feelings. Such a response would involve an overt attitude of *good listenership*. However, a few *fillers* used in the L1 were favorably received in terms of playing a crucial role in connecting the *flow* of interactions. It would also appear that students' low English proficiency led them translate entire interactions word-for-word, giving the impression that they were far from the level at which they could use formulaic expressions naturally. Overall, *continuity* and *flow* of conversations, even in short exchanges, appeared to be the key features of oral interaction.

4.4. Additional emerging themes

In this section, themes indicating positive and negative elements, as well as the final themes that emerged from interpretations as clarified in interviews, are presented and briefly discussed.

Interactional flow. Strategies necessary to continue *interactional flow* are important. These include *negotiation of meaning,* such as *clarification requests,* and use of *CSs*, along with *NVBs*. These should be used to avoid *disfluency features* such as s*ilent pauses* and *L1 use* for *content words*.

Task completion. Whether *communicative success* is achieved is highly important for accomplishing simple tasks such as role-playing tasks using textbooks.

English-language features. *Formulaic expressions* need to be memorized and used simultaneously. Simultaneously, basic *listening skills* for the task must be learned.

Attitude. *Confidence* and *articulation of speech* have the potential to produce a positive atmosphere with interlocutors, which enhances good listenership during communication.

NVBs. Using physical gestures to convey messages may help achieve communicative success, such as nodding intended to mean "yes." However,

averting eye gaze and other negative *NVBs* can negatively affect communication.

Quality of response. Mimicking and *parroting* to respond quickly in order to maintain *flow*, or "pass the time," during the interaction resulted in a negative impression for raters. Excessive *L1 use* and of L1 *fillers* exerted both positive and negative influences. A summary is presented in Table 2.

Table 2. Summary of speaking-interaction criteria for fifth-graders in Japan

Element	Positive	Negative
Interactional flow	Smooth, flow	Too long and too many silences that make listeners uncomfortable
Task completion	Task was accomplished	Task was not accomplished.
	(Task example)	
English language feature	Attainment of target formulaic expressions	
Attitude	Showing to the listeners his/her understanding. Speaking clearly enough to be heard.	Not ready or willing to interact.
Quality of response	Clear voice to be heard by the interlocutor. Controlling the speed and placing an emphasis to clearly deliver the message.	Not too fast. Not overuse of L1. Not answering in L1.
NBV	Nodding and use of hand gestures to implement English words.	Avert eye gaze

5. Discussion

Through dialogue with the raters, I reflexively learned from the initial responses on the questionnaire as well as from the oral interviews. This corresponds to the concepts illustrated by Yagi and Nakayama (2021). The raters indicated what is appropriate and effective for developing oral interaction skills in Japanese elementary school students.

In addressing the research question regarding what elements in spoken interactions positively influenced listeners, six themes were identified from the intuitive responses of the raters: *interactional flow, task completion*, *English-language features, attitude*, *NVBs*, and *quality of response*. These findings suggest that the use of formulaic expressions learned in class during role-play tasks notably facilitated listener perception in terms of promoting effective achievement in these areas.

These themes largely align with the findings of Sato and McNamara (2018), who focused on adults. However, there were notable differences in that issues related to *cohesion* or *content* were not evident in this study. This absence may reflect the developmental and educational contexts of elementary students, who are naturally more focused on basic language functions and interactional dynamics. Therefore, it is clear from the results that applying adult assessment criteria directly to children is not appropriate. Teachers developing assessment criteria should be mindful of these differences and modify them accordingly.

Concerning the second research question, pertaining to the elements that pose difficulties for listeners in the spoken interactions of fifth-grade elementary students in Japan, most of the negative factors identified were inversely related to the positive elements from the first question. For example, issues such as "insufficient basic listening comprehension" and "inability to complete tasks" were frequently mentioned. These difficulties mirror the challenges faced when elementary students fail to apply or understand the formulaic expressions taught in class.

Sato and McNamara (2018) focused on adult learners and highlighted *English-language features* primarily related to aspects of fluency, such as *speech rate*, *pause phenomena*, and *repair phenomena*. In contrast, the participants in this study specifically noted that the elementary students often struggled with the fundamental skills of *listening comprehension* and in using *formulaic expressions* appropriately. These basic competencies formed the core of the English-language features assessed in this study.

Fluency components such as *repair* and *breakdown* yielded a complex mixture of both positive and negative responses. For instance, *parroting* or *mimicking* was not regarded as a strategy merely to fill gaps or connect conversational turns to find one's own words. Instead, such usage was seen as a way to respond quickly by immediately repeating the word used in the question, which made the role-play resemble a rote memorization test of formulaic expressions. This approach contrasts with real-world interactions in which hesitation and rephrasing are often perceived as more natural than quick responses.

Furthermore, when the L1 was used not only for *fillers* but also for actual *content words* or excessively in responses, it was typically perceived negatively. This type of response, however, could signify an overt demonstration of good listenership (McCarthy, McCarten, 2018). When the L1 was frequently used, it often gave the impression of a basic deficiency in English knowledge, leading to negative perceptions among the listeners. However, when the L1 served to provide *fillers*, it reduced silence and conveyed to raters that the speakers were actively trying to formulate responses, which

could be seen as a positive aspect of engagement. Overall, the way the L1 was utilized played a significant role in how listeners judged the continuity and flow of conversations, which are crucial factors for assessing students' L2 fluency and comprehension.

These reconsidered responses clearly highlight the difficulties faced by elementary students and contrast with findings involving adult learners, thereby providing a more comprehensive understanding of the factors that hinder effective communication in this educational context.

6. Implications and limitations

The summarized criteria for elementary school students' speaking interactions highlight important aspects that can be adopted for speaking assessments. These criteria, when aligned with classroom-based tests, could potentially be applied to the holistic component of speaking interaction. Furthermore, it is suggested that the flow of interaction may have the most substantial influence on rater perception.

The limitations of this study include its small sample size and the very narrow lenses of the scope. Further studies are recommended to deepen the understanding of these findings by exploring similar research.

7. Conclusion

This study yielded six important themes relevant to the assessment of fifth grade students in English classroom-based role-play tasks in Japan. The analysis focused on the reasons underlying high and low ratings given to children's English performances. Each of the six themes was intricately related to one another, with most of them showing a connection to *interactional flow.* However, the presence of *L1 usag*e and *NVB* may act as a double-edged sword and could potentially be considered as reverse-coded items, making it necessary to further investigate how specific elements impact a holistic perspective.

In speaking research on adults, where criteria based on CAF are commonly used, Sato and McNamara (2018) employed a qualitative approach to extract assessment criteria based on raters' perceptions. This method was partially replicated with this study. The current study focused on Japanese children revealed some alignment with adult criteria, however, there were differences. These were interpreted as due to children's limited vocabulary knowledge and the use of formulaic expressions in classroom-based

tests. Specifically, temporal aspects were not always interpreted in the same way. Fast responses, often relying on memorized formulaic expressions, may not reflect natural spontaneous interaction. Additionally, fillers, which were often categorized as disfluencies, were not necessarily perceived as negative as in adult interactions, but rather viewed as strategies to avoid silence.

Ultimately, while assessing interactional flow provides valuable insights into children's speaking interactions, it may also be worthwhile to distinguish between formulaic expressions and spontaneous interaction in the assessment process.

Acknowledgements
I would like to express my deepest gratitude to Prof. Marianne Nikolov for her invaluable advice during the writing of this study.

BIBLIOGRAPHY

Braun V., Clarke V. (2006), *Using thematic analysis in psychology.* "Qualitative Research in Psychology", No. 3(2), pp. 77–101.

Braun V., Clarke V., Hayfield N., Frith H., Malson H., Moller N., Shah-Beckley I. (2019), *Qualitative story completion: Possibilities and potential pitfalls.* "Qualitative Research", No. 16, pp. 136–155.

Britton M. (2021), *Assessment for learning in primary language learning and teaching,* Cambridge: Multilingual Matters.

Brown S. (2005), *Assessment for learning.* "Learning and Teaching in Higher Education", No. 1, pp. 81–89.

Butler Y.G. (2005), *Comparative perspectives towards communicative activities among elementary school teachers in South Korea, Japan and Taiwan.* "Language Teaching Research", No. 9, pp. 423–446.

Butler Y.G. (2022), *Assessing young learners,* (in:) Fulcher G., Harding L. (eds.), The Routledge handbook of language testing. London: Routledge, pp. 153–170.

Butler Y.G. (2024), *Self-assessment in second language learning*. "Language Teaching", No. 57, pp. 42–56.

Cameron L. (2003), *Challenges for ELT from the expansion in teaching children*. "ELT Journal", No. 57(2), pp. 105–112.

Carless D. (2005), *Prospects for the implementation of assessment for learning.* "Assessment in Education: Principles, Policy & Practice", No. 12, pp. 39–54.

Crotty M.J. (1998), *The foundations of social research: Meaning and perspective in the research process*. London: Sage.

Ellis R. (2015), *Understanding second language acquisition.* Oxford: Oxford University Press.

Gass S.M., Behney J., Plonsky L. (2022), *Second language acquisition, an introductory cours*e. New York: Routledge.

Genesee F., Upshur J.A. (1996), *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.

Hulstijn J.H. (2015), *Language proficiency in native and non-native speakers: Theory and research*. Amsterdam: John Benjamins.

Johnstone R. (2009), *An early start: what are the key conditions for generalized success?,* (in:) Enever J., Moon J., Raman U. (eds.), Young learner English language policy and implementation: International perspectives. Reading: Garnet Education Publishing, pp. 31–42.

Lantolf J. P., Poehner, M.E. (2014), *Sociocultural theory and the pedagogical imperative in L2 education: Vygotskian praxis and the research/practice divide.* Routledge.

Lightbown P.M., Spada N. (2022), *How languages are learned.* Oxford: Oxford University Press.

Long M. (1980), *Input, interaction, and second language acquisition*. University Microfilms International.

Mackey A. (2020), *Interaction, feedback and task research in second language learning*. Cambridge: Cambridge University Press.

Marsden E., Morgan-Short K., Thompson S., Abugaber D. (2018), *Replication in second language research: Narrative and systematic reviews and recommendations for the field.* "Language Learning", No. 68.2, pp. 321–391.

McCarthy M.J., McCarten J. (2018), *Now you're talking! Practicing conversation in second language learning,* (in:) Jones C. (ed.), Practice in second language learning. Cambridge: Cambridge University Press, pp. 7–29.

Merriam S.B. (1998), *Qualitative research and case study applications in education. Revised and expanded from "Case study research in education."* New Jersey: Jossey-Bass Publishers.

MEXT (2017), Shogakko gakushu shido yoryo gaikokugo katsudou, gaikokugo hen [The elementary school course of study for foreign languages]. https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/03/18/1387017_011.pdf

MEXT (2022), Reiwa 4nenndo "Eigokyoiku jisshi jokyo chosa" no kekka nit suite shogakko [Results of the 2022 Survey on the implementation of English education in elementary schools]. https://www.mext.go.jp/a_menu/kokusai/gaikokugo/1415043_00004.htm

Mora J.C. (2006), *Age effects on oral fluency development,* (in:) Muñoz C. (ed.), Age and the rate of foreign language learning. Clevedon: Multilingual Matters, pp. 65–88.

Nikolov M. (2016), *A framework for young EFL learners' diagnostic assessment: 'Can do statements' and task types*, (in:) Nikolov M. (ed.), Assessing young learners of English: Global and local perspectives. New York: Springer Publishing, pp. 65–92.

Nikolov M., Timpe-Laughlin V. (2020), *Assessing young learners' foreign language abilities.* "Language Teaching", No. 54, pp. 1–37.

Pinter A. (2017), *Teaching young language learners*. Oxford: Oxford University Press.

Rich S. (2019), *Early language learning teacher education*, (in:) Garton S., Copland F. (eds.), The Routledge handbook of teaching English to young learners. London: Routledge, pp. 44–59.

Rixon S. (2016), *Do developments in assessment represent the 'coming of age' of young learners' English language teaching initiatives? The international picture,* (in:) Nikolov M. (ed.), Assessing young learners of English: Global and local perspectives. Heidelberg, Germany: Springer, pp. 19–41.

Sato T., McNamara T. (2018), *What counts in second language oral communication ability? The perspective of linguistic laypersons.* "Applied Linguistics", No. 40(6), pp. 894–916.

Szpotowicz M., Lindgren E. (2011), *Language achievements: a longitudinal perspective*, (in:) Enever J., (ed.), Early Learning in Europe. ELLiE study. London: British Council, pp. 125–144.

Tracy-Ventura N., McManus K., Norris J.M., Ortega L. (2014), *'Repeat as much as you can': elicited imitation as a measure of oral proficiency in L2 French*, (in:) Leclercq, P., Edmonds, A., & Hilton, H. (eds.), Measuring L2 proficiency perspectives from SLA. Bristol, United Kingdom: Multilingual Matters, pp. 143-166.

Vygotsky L.S. (1987), *Thinking and speech.* (in:) Rieber R. W., Carton A. S (eds.), The collected works of L. S. Vygotsky. Volume 1: Problems of general psychology. Plenum, pp. 39-288.

Wiliam D. (2011), *What is assessment for learning?.* "Studies in Educational Evaluation", No. 37(1), pp. 3–14.

Wiliam D., Thompson M. (2017), *Integrating assessment with learning: what will it take to make it work?,* (in:) Dwyer C.A. (ed.), The future of assessment. London: Routledge, pp. 53–82.

Yagi M., Nakayama A., Nakai Y. (eds.), Shitsuteki gengo kyoiku wo kangaeyo. [Doing qualitative research in language education]. (2021). Hitsuji shobo.

Received: 22.08.2024
Revised: 25.11.2024