



Assessing functional adequacy in oral dialogic speech: Validating a rating scale for L2 interaction

RESEARCH

INEKE VEDDER

FOLKERT KUIKEN

KRISTY SUET-SIN CHEUNG

ASTRID MORRISON

TONGZHOU XIAO

PAOLO ORRÙ

**Author affiliations can be found in the back matter of this article*

WHITE ROSE
UNIVERSITY PRESS
Universities of Leeds, Sheffield & York

ABSTRACT

Functional Adequacy (FA), a multidimensional construct emphasising successful task completion, has developed into an established line of research in second language acquisition (SLA). Kuiken and Vedder's (2017, 2018) FA rating scale, which operationalises FA across four dimensions (Task Requirements, Content, Comprehensibility, Coherence & Cohesion), has been shown to provide a reliable and valid instrument for assessing second language (L2) performance in monologic tasks. Contrary to monologue performance, few studies have, however, investigated L2 interaction due to its inherent complexity (Galaczi & Taylor, 2020). Recent research has called for adaptations of the FA monologue scale for dialogic data (González-Lloret, 2022). To this end, Kuiken and Vedder (2025) substituted the Coherence & Cohesion dimension of the monologue scale with a new dimension, *Interaction*, focusing on the adequacy of speakers' engagement in situated communicative interaction. Our study explores the reliability, validity, and applicability of the adapted FA scale for interactional speech. Through three pilot studies involving L2 learners of Italian and English with three groups of raters, three issues were addressed: the correlations of raters' judgments on the Interaction dimension, the associations among the four FA dimensions, and the raters' experiences and challenges in applying the adapted scale. The three pilots showed a largely sufficient to high interrater reliability for Interaction. Moderate to high correlations were found between Interaction and Task Requirements, Content, and Comprehensibility. The raters' feedback confirmed the applicability of the scale and the relevance and clarity of the descriptors for assessing dialogic performance. With some caution, we may conclude that the adapted FA scale constitutes a promising instrument for assessing L2 oral interaction.

CORRESPONDING AUTHOR:

Ineke Vedder

University of Amsterdam, NL
s.c.vedder2@contact.uva.nl

KEYWORDS:

Functional Adequacy; rating scale; oral interaction; L2 Italian; L2 English; rater training

TO CITE THIS ARTICLE:

Vedder, I., Kuiken, F., Cheung, K. S.-S., Morrison, A., Xiao, T., & Orrù, P. (2026). Assessing functional adequacy in oral dialogic speech: Validating a rating scale for L2 interaction. *Journal of the European Second Language Association*, 10(1), 99–112. DOI: <https://doi.org/10.22599/jesla.180>

Since the 1990s, linguistic performance elicited by language tasks has often been assessed in terms of complexity, accuracy, and fluency (CAF; [Housen & Kuiken, 2009](#); [Skehan, 2009](#)). More recently, in addition to CAF, functional adequacy (hereinafter referred to as FA), a multi-layered concept referring to the communicative components of second language (L2) performance, has emerged as a key construct in the field of Second Language Acquisition (SLA; [Pallotti, 2009](#); [Pallotti, 2022](#); [Révész et al., 2016](#)). Within the framework of task-based language teaching (TBLT) and task-based language assessment (TBLA), Kuiken and Vedder ([2017, 2018](#)) developed a rating scale for the assessment of FA, defined as successful task completion in relation to the task and its context, for oral and written monologue tasks.

Kuiken and Vedder's ([2017, 2018](#)) scale for FA is a six-level Likert scale, comprising four dimensions: Task Requirements, Content, Comprehensibility, and Coherence & Cohesion, inspired by Grice's ([1975](#)) conversational maxims of quantity, relation, manner, and quality. Based on the can-do statements and descriptors for oral and written production of the Common European Framework of Reference (CEFR; [Council of Europe, 2001, 2020](#)), the FA scale descriptors were formulated for each of the six levels of these four dimensions. Task Requirements focuses on the extent to which the task is completed in accordance with genre, task, addressee, speech act, and register. Content considers the adequacy of the number of information units or concepts expressed in the text and their thematic elaboration. Comprehensibility refers to the amount of effort required from the listener/reader to understand and process the message conveyed by the speaker/writer. Coherence & Cohesion addresses the adequacy of the text in terms of the occurrence of strategies for coherence (e.g., use of discourse markers) and cohesive ties (e.g., deictic elements; see Appendix A). Several experimental studies, involving different educational settings, learners, raters (experts and non-experts), proficiency levels, task types, and target and source languages, have since been conducted to establish the applicability of the FA scale. These studies have demonstrated that the scale is a reliable, valid, and user-friendly tool in various learning contexts ([Ekiert et al., 2018](#); [Lu & Suzukida, 2026](#); [Strobl & Baten, 2022](#); see [Kuiken & Vedder, 2022](#), for an overview).

An important issue for further investigation concerns, however, the applicability of the FA scale for oral interactional tasks. In research on oral L2 performance, the emphasis has been so far on monologic rather than on interactional speech. One of the reasons for this paucity of studies is that interactional speech involves multiple factors, making it complex and challenging to assess ([Galaczi & Taylor, 2018](#); [Grabowski, 2013](#); [Sato, 2014](#)). As argued, among others, by González-Lloret ([2022](#)), Pallotti ([2019](#)), and Youn ([2018, 2020](#)), the adequacy of interactional speech may be affected by multiple artefacts (e.g., videochat, WhatsApp, mobile phone) and variables (e.g., socio-pragmatic, cultural, and situational factors; individual participant characteristics) mediating the interaction. A next step for research on FA (and TBLT and TBLA) is thus to explore the applicability of the FA scale for interactional tasks. To that end, in response to calls for expanding the use of the FA rating scale for oral interaction, Kuiken and Vedder ([2025](#)) reformulated the Coherence & Cohesion dimension of the original monologue scale into a new dimension, Interaction, focusing on the adequacy of speakers' engagement in situated communicative interaction. To test the adapted FA scale for oral interactional speech, three pilots were conducted among two groups of L2 learners of Italian and one group of L2 learners of English (see Section 3.2 below).

The study at hand aims to explore the reliability, validity, and applicability of the adapted FA scale. An additional purpose of our research is to investigate the experiences of the raters during rater training and while they individually assessed the interactional language samples. More specifically, we analyse the correlations of raters' judgments on the new dimension Interaction, and the associations between Interaction and the other three scale dimensions (i.e., Task Requirements, Content, Comprehensibility). Furthermore, the paper examines the challenges raters encountered when assessing interactional performance, together with the strategies they employed to overcome these difficulties.

2.1. STUDIES ON THE ASSESSMENT OF L2 INTERACTION

The ability to interact with other speakers has long been a central component in classic theories of L2 communicative competence (Bachman & Palmer, 2010; Canale & Swain, 1980; Celce-Murcia et al., 1995; Hymes, 1967, 1972). As mentioned in the Introduction, the specific field where FA is situated, TBLT and TBLA (e.g., Kuiken & Vedder, 2022), has predominantly focused on monologic tasks rather than dialogic speaking tasks. This emphasis on monologues may stem from their methodological advantages, as monologic speech tends to be easier to control and displays more predictable linguistic patterns compared to dialogic performance, where researchers face challenges in terms of the operationalisation of interactional features, such as between-speaker pauses, overlaps, and turn-taking (Tavakoli, 2016, 2018; Tavakoli & Wright, 2020).

Only a few studies have looked at the impact of interaction as an independent variable on linguistic performance, comparing tasks that were similar in all respects except for the +/- monologue condition (Gilabert et al., 2011; Michel et al., 2007; Michel, 2011). The main outcome of these studies was that in dialogues, both native and non-native speakers tended to produce more accurate and fluent, but syntactically less complex utterances, probably because learners rely on each other's turns in constructing their sentences.

In a conceptual and methodological study among native (L1) speakers of Italian, Pallotti (2019) explored the notion of “interactional difficulty,” suggesting an operationalisation of the construct in terms of three parameters: (a) number of turn exchanges, (b) number of initiating moves, and (c) visual access among participants. Turn-taking in an interaction may not be easy, as speakers need to identify “transition-relevant places” (Sacks et al., 1974, p. 703) and exploit them to insert their contribution in the flow of a conversation. Research has additionally shown that L2 learners' ability to initiate and yield turns develops slowly. In the early stages of L2 acquisition, learners speak only when a more competent interlocutor invites them to do so (Pekarek Doehler & Pochon-Berger, 2015). Initiating turns progresses the conversation, having a pro-active function which makes a reaction from the interlocutor more relevant. Visual access, the third factor distinguished by Pallotti, is different from the other two parameters (i.e., number of turn exchanges and initiating turns), as it does not require the observation of participants' speaking performance. For this reason, we have not included it in our operationalisation of interactional competence in the study at hand.

Other studies, focusing on specific dimensions of L2 performance in interactions, such as comprehensibility (Nagle et al., 2022) or fluency (Tavakoli, 2016, 2018), have also emphasised the different nature of dialogues compared to monologues. The goal of the study by Nagle et al. (2022) was to compare how comprehensibility was assessed by external listeners, speakers, and their interaction partners. Twenty pairs of L2 learners of English rated themselves and their partners on several occasions, and the recordings of the interaction were subsequently presented to external raters for evaluation. The results demonstrated that while self-assessed and peer-assessed scores of the interlocutors aligned with each other, external raters assigned significantly lower scores. The study by Tavakoli (2016), examining interactive aspects of fluency in dialogic performance (e.g., interruptions, overlaps, between-turn pauses) explored how fluency measurement differs between monologues and dialogues. Data were derived from 35 monologic and dialogic task performances collected among L2 learners of English. Highlighting the importance of between-turn pauses, the author argues that L2 fluency research has not appropriately described how such features can be measured in dialogic contexts and concludes that fluency measurement in dialogues remains less systematic and research-driven than in monologues. Sato (2014) investigated the construct of oral fluency in L2 English during peer interaction. Participants were 56 Japanese university students, and four native English speakers who rated the learners' oral fluency. The results showed that learners' individual performance did not necessarily predict the quality of their performance in an interactional context.

2.2. TOWARDS AN FA RATING SCALE FOR ORAL INTERACTION

Developing a rating scale which can also be employed for interactional tasks requires careful attention to the conceptualisation and operationalisation of interaction, the design and validation of the rating tool, and the implementation of the rating process, including

rating training. Our definition of FA in dialogic speech is based on the can-do statements as formulated in the CEFR (Council of Europe, 2001, 2020), referring to a speaker's ability of taking and keeping the floor, cooperating to the development of the interaction, and asking and providing clarification. We operationalised interactional FA in terms of such phenomena as (a) turn-taking, opening, and closing; (b) sequencing, alignment, and pausing behaviour; (c) asking for and providing relevant and detailed clarification.

To investigate the potential use of the FA scale for oral interaction, the first dilemma faced was the decision between the expansion of the monologue scale with a new scale dimension, and its adaptation by substituting one of the four scale dimensions (i.e., Task Requirements, Content, Comprehensibility, Coherence & Cohesion). Kuiken and Vedder (2025) chose to reformulate the Coherence & Cohesion dimension in the original scale ("less is better") in terms of the adequacy of speakers' engagement in situated communicative interaction. The reasons why Coherence & Cohesion was substituted with a new dimension labelled Interaction were that in an interactional exchange, coherence is mainly realised by mutual alignment of conversational turns. Speakers ideally connect their contribution to what has just been said by their interlocutor. Cohesion, furthermore, is often problematic to assess in dialogic speech, as utterances in interactional turns, comprising back-channels ("yes," "okay"), are generally short and contain a relatively lower number of cohesive and anaphoric devices compared to monologic turns.

The adapted FA scale for oral interaction thus comprises four dimensions: Task Requirements, Content, Comprehensibility, and the new Interaction dimension (see Appendix B). The leading question underlying Interaction is: "To what extent can the speaker take and keep the floor, contribute to the progress of the interaction, and ask for and provide clarification?" The scale descriptors of the six proficiency levels of Interaction are based on the descriptors for interactional competence of the CEFR (Council of Europe, 2001, 2020) and range from level 1 ("not at all adequate") to level 6 ("absolutely adequate"). The descriptors of each level refer to the interlocutor's ability of taking/keeping the floor, mutual co-operation, and asking for/providing clarification, e.g., "the interaction is carried out only thanks to the contribution of the interlocutor" (level 1); "the speaker needs a lot of help from the interlocutor to start, maintain, or end the interaction" (level 2); "utterances are partly connected to the interlocutor's contribution" (level 3); "the speaker gives and asks for clarification and repetition, although not always appropriately" (level 4); "the speaker contributes to the development of the interaction with follow-up statements and inferences showing comprehension" (level 5); "the speaker can select suitable and appropriate phrases to invite the interlocutor to contribute" (level 6). For full descriptors of each level, see Appendix B.

3. PILOTING THE FA SCALE FOR ORAL INTERACTIONAL TASKS

3.1. RESEARCH QUESTIONS

To establish the reliability, validity and applicability of the FA scale for oral interactional tasks, three pilots were set up. We formulated the following research questions (RQ1-3):

RQ1: How do raters' judgments of FA correlate on the Interaction dimension?

RQ2: How are raters' judgments on the four dimensions of FA associated?

RQ3: What are the main challenges that raters may encounter when assessing interactional data?

3.2. PARTICIPANTS

We collected interactional data from 52 university students of L2 Italian (Pilot 1: 13 dyads, Pilot 2: 13 dyads) and 42 students of L2 English (Pilot 3: 21 dyads). The reason to split up the Italian data and to treat them separately was that different teams of raters were involved in the two pilots. The self-reported proficiency level of the L2 speakers of both Italian and English, gathered via intake forms, ranged between B1 and C1. Data in Pilot 1 were obtained from 26 students in two distinct contexts: 10 study-abroad students with various native languages in Italy and 16 L2 learners of Italian in Hungary, all of whom were native speakers of Hungarian. Although these two groups differed in terms of learning context and L1 background, they

were included in the same pilot because data were collected using the same study protocol and interactional task. Thus, the distinction between the two groups reflects differences in participants' learning contexts rather than differences in the procedures used to generate the interactional data. The group in Italy was in an L2 immersion context, whereas the Hungarian group represented a foreign-language learning context. Data in Pilot 2 were collected in Poland (Krakow) among 26 university students of L2 Italian, with Polish as their L1. Finally, the L2 English data in Pilot 3 were collected in Italy, from Italian students of Modern Languages across different universities (Table 1).

PILOT	PARTICIPANTS	DATA COLLECTION
1. L2 Italian	N = 26 (13 dyads) Level B1-C1 Various L1s (Italy); L1 Hungarian (Hungary)	Italy (Cagliari) Hungary (Debrecen)
2. L2 Italian	N = 26 (13 dyads) Level B1-C1 L1 Polish	Poland (Krakow)
3. L2 English	N = 42 (21 dyads) Level B1-C1 L1 Italian	Italy (various universities)

Table 1 Participants and data collection.

3.3. TASK AND PROCEDURE

Oral interactional data were elicited through a problem-solving task with an information gap. The task requires pairs of participants to negotiate and to choose together a Bed & Breakfast (B&B) for a two-week holiday based on six descriptions of B&B establishments in various regions in Italy. There were six (partly contrasting) criteria that the participants had to consider (e.g., presence of a garden, vicinity of the city centre, possibility to swim). To trigger negotiation, there was no "ideal" solution, as none of the six B&Bs fully met all criteria.

All students were asked to sign a consent form and complete a questionnaire comprising personal data (e.g., L1, instructional background, self-perceived level of L2 Italian or L2 English). Data for Italian, as mentioned above, were divided into two pilots, because two different teams of raters were involved in the two experiments (nine and four, respectively), with differences in prior knowledge regarding TBLT and TBLA (see Section 3.4). The researchers first explained the purpose of the task (i.e., pair work on a communicative language task); each dyad then received two different sets of written prompts (one for each participant) containing a description of three of the six B&Bs. The students were asked to focus on communication, negotiation, and interaction, rather than on accuracy. Participants were matched arbitrarily, and there was no time limit for completing the task. Dyads typically completed the task in 8 to 10 minutes, though a few pairs finished in less than 5 or took longer than 15 minutes. Data were audio-recorded; recordings took place in a language lab or in a classroom. In a 5-minute post-task interview, we asked each dyad about their experiences during task performance, opinions about the target task and procedure, emotional responses (feelings of enjoyment or anxiety), difficulties encountered, and strategies used during the negotiation.

3.4. RATERS AND RATER TRAINING

The first decision we had to make was if the interactional phenomena we aimed to investigate should be measured jointly or individually. The main reason why we chose to assess each participant separately was that individual learner characteristics, such as speaking style, extraversion vs. introversion, anxiety, and pausing behaviour, may have an impact on interactional competence. What is more, joint assessment, per dyad, resulted to be more challenging, as participants may often show asymmetries in linguistic, pragmatic, and interactional skills.

As has often been emphasised in the literature, rater training is crucial to familiarise raters with the rating instrument, scale descriptors, and levels (Pill & Smart, 2020; Rezaei & Lovorn, 2010). Furthermore, for establishing the reliability and validity of the rating tool, oral (and

written) samples need to be rated by more than one rater (Van den Branden, 2020). In the current study, the oral data of the 26 dyads of Pilot 1 and 2 (L2 Italian) and the 21 dyads of Pilot 3 (L2 English) were assessed by three groups of raters. Pilot 1 involved nine raters. All of them were native speakers of Italian and master's students of Modern Languages, who had just finished a one-month course on TBLT and TBLA, and had been asked to volunteer in the rating experiment. The four raters involved in Pilot 2 were also master's students of Modern Languages, but differently from the raters in Pilot 1, they were less familiar with TBLT and TBLA. All raters of Pilots 1 and 2 were non-experts, insofar as they had very little or no prior rating experience. The three raters of Pilot 3, PhD researchers in Linguistics, were near-native speakers of English, with Chinese or Spanish as their L1. All of them had experience in language teaching. However, as they had little experience in language assessment, in the present study they were considered non-expert raters too. None of the raters of Pilots 1, 2, and 3 were familiar with accented Italian or accented English. Table 2 shows an overview of the raters and the training sessions they received.

PILOT	RATERS (NON-EXPERTS)	TRAINING DURATION	TRAINING LOCATION
1. L2 Italian	N = 9 MA students L1 Italian	3–4 hours (two sessions)	On site
2. L2 Italian	N = 4 MA students L1 Italian	3–4 hours (two sessions)	Online
3. L2 English	N = 3 PhD L2 English L1 Chinese, L1 Spanish	3–4 hours (two sessions) 1-hour discussion Possibility to reconsider scores	Online

Table 2 Raters' information and training.

The duration of the rater training sessions was three to four hours in total for the two training sessions conducted in Pilots 1 and 2. In Pilot 3, training comprised two sessions of three to four hours in total, followed by a separate one-hour discussion session during which raters had the possibility to reconsider their scores. In Pilot 1, rater training took place on site; in Pilots 2 and 3, due to the geographical distribution of the raters, who came from different universities, training sessions were held online.

In Session 1, the researchers introduced the construct of FA and the dimensions and scale levels of the rating scale. Trial materials were then presented and discussed. In preparation for Session 2, we asked the raters to assess the speech samples of two dyads. In Session 2, guidelines and feedback were provided based on the ratings of these samples, and instructions for the individual scoring were presented. In addition, two other dyads were assigned to be individually assessed at home. The raters were encouraged to rate the recordings within a time span of one to two weeks. Speech samples were randomly assigned to avoid a possible rater bias due to order effects. In the case of Pilot 3, despite overall consistency, some discrepancies emerged among the three raters' scoring patterns, particularly in the dimensions of Content and Comprehensibility. For this reason, the three raters organised an additional session (this time without the researchers). During this session, they revisited a set of four differing cases and discussed and shared their understanding of performance descriptors. Data of these pairs were excluded from the scoring of the remaining 21 dyads. This was done because the three raters rated the excluded pairs together and reached a consensus in this additional session. The extra meeting proved essential for reinforcing interrater reliability and aligning evaluative standards across different L2 teaching backgrounds. Finally, a retrospective group meeting with each of the three rating teams of Pilots 1–3 was organised a few weeks after all speech samples had been scored. In this meeting, we discussed raters' experiences during and after the training sessions, challenges they had come across, and strategies they had employed to overcome these difficulties.

4. RESULTS

4.1. INTERRATER RELIABILITY

To assess interrater reliability for both the Italian and English texts (RQ1), intraclass correlation coefficient (ICC) and Cronbach’s alpha (Bachman, 2004) were calculated among the nine raters of Italian in Pilot 1, the four raters of Italian in Pilot 2, and the three raters of English in Pilot 3 (see Table 3). ICC estimates and their 95% confident intervals were calculated using SPSS statistical package version 30.0 (IBM Corp., 2024) based on a mean-rating, absolute agreement, two-way random model. The interrater reliability assessment of the spoken texts in terms of Task Requirements and Content showed a moderate to high ICC across all tasks, ranging from .551 (for Content in Pilot 3) to .867 (for Content in Pilot 1). Regarding Comprehensibility, raters’ scores in Pilot 1 and Pilot 3 showed good reliability in both ICC and Cronbach’s alpha, whereas these two indices were not satisfactory in some dimensions for Pilot 2; see Section 5 for a possible explanation of this finding.

In terms of interrater reliability of the Interaction dimension, both ICC and Cronbach’s alpha were high to excellent across all raters in the three pilot studies. Results of ICC ranged between .762 (for Interaction in Pilot 3) to .905 (for Interaction in Pilot 1); Cronbach’s alpha ranged between .809 (for Interaction in Pilot 3) to .920 (for Interaction in Pilot 1).

DIMENSION	PILOT 1: L2 ITALIAN		PILOT 2: L2 ITALIAN		PILOT 3: L2 ENGLISH	
	ICC	α	ICC	α	ICC	α
Task Requirements	.828	.841	.832	.879	.795	.833
Content	.867	.875	.753	.778	.551	.653
Comprehensibility	.845	.872	.325	.465	.614	.686
Interaction	.905	.920	.793	.833	.762	.809

Table 3 Intraclass correlations (ICC) and Cronbach’s alpha (α) in all pilot studies.

4.2. CORRELATIONS BETWEEN DIMENSIONS OF FUNCTIONAL ADEQUACY

In terms of how raters’ judgments on the four dimensions of FA are associated (RQ2), we found that all three pilot groups demonstrated high average scores across the four dimensions, with mean scores generally ranging from 4.13 to 5.02. The L2 English participants (Pilot 3) achieved the highest mean scores in Task Requirements (M = 5.02) and Content (M = 4.77), while the L2 Italian participants (Pilot 1) scored highest in Comprehensibility (M = 4.89). In the Interaction dimension, the L2 English group (Pilot 3) also received the highest mean score (M = 4.75; see Table 4).

DIMENSION	PILOT 1: L2 ITALIAN (N = 26)		PILOT 2: L2 ITALIAN (N = 26)		PILOT 3: L2 ENGLISH (N = 32)	
	MEAN	SD	MEAN	SD	MEAN	SD
Task Requirements	4.72	.93	4.49	.72	5.02	.75
Content	4.13	.83	4.72	.83	4.77	.58
Comprehensibility	4.89	.77	4.72	.85	4.86	.67
Interaction	4.53	.68	4.48	.68	4.75	.70

Table 4 Mean rating scores and standard deviations (SDs) for all pilot studies.

Pearson’s product-moment correlations were calculated to examine the extent to which raters’ judgements of the four FA dimensions were related to one another (see Table 5). All correlations were statistically significant and ranged from moderate to high, from $r = .518$ between Content and Comprehensibility in Pilot 2 to $r = .931$ between Content and Interaction in Pilot 1. In Pilot 1, the strongest correlation was between Content and Interaction, $r = .931$, followed by Content and Task Requirements, $r = .907$, and Comprehensibility and Interaction, $r = .804$. In Pilot 2, correlations ranged from $r = .518$ between Content and Comprehensibility to $r = .889$ between Content and Task Requirements. The correlations involving Interaction in Pilot 2 were $r = .882$ with Task Requirements, $r = .858$ with Content, and $r = .622$ with Comprehensibility. In Pilot 3, correlations ranged from $r = .713$ between Content and Comprehensibility to $r = .857$ between Interaction and Task Requirements. The correlations involving Interaction in Pilot 3 were $r = .857$ with Task Requirements, $r = .773$ with Content, and $r = .829$ with Comprehensibility. Overall, the

correlation coefficients varied across the three pilots and across different pairs of dimensions. This indicates that in different studies, the new dimension is distinctively differentiable from other domains on the FA scale.

DIMENSION	PILOT 1 (L2 ITALIAN)			PILOT 2 (L2 ITALIAN)			PILOT 3 (L2 ENGLISH)		
	TASK REQ.	CONT.	COMP.	TASK REQ.	CONT.	COMP.	TASK REQ.	CONT.	COMP.
Cont.	.907**			.889**			.722**		
Comp.	.690**	.818**		.610**	.518**		.794**	.713**	
Int.	.786**	.931**	.804**	.882**	.858**	.622**	.857**	.773**	.829**

Table 5 Pearson's product-moment correlations between dimensions of functional adequacy for English spoken interaction texts.

Note. Task req. = Task requirements; Cont. = Content; Comp. = Comprehensibility; Int. = Interaction; ** $p < .01$.

5. RATERS' PERCEPTIONS AND STRATEGIES

Overall, both L2 Italian and L2 English raters' reflections on the applicability of the interactive FA scale revealed a positive reception toward the instrument's clarity, construct coverage, and relevance for assessing dialogic performance. Raters agreed that the four dimensions of the adapted FA scale were clearly distinguished from one another, facilitating consistent judgments across samples. The keyword descriptors within each level (e.g., "not at all adequate," "scarcely," and "somewhat") were considered comprehensible and aligned with performance features observable in the data.

During the retrospective interview, raters' responses suggested that ratings of Comprehensibility may be more strongly influenced by individual listeners' perceptions of the effort required to interpret the message ("If I can understand them, I always assign a 6"). The comment also illustrates how raters may rely on different interpretive strategies when evaluating comprehensibility in dialogic performance.

In the case of Pilot 3, despite overall consistency, some discrepancies in the dimensions of Content and Comprehensibility came up among the L2 English raters (see also Section 3.4). To understand their perception of the scale dimensions, the raters were invited to hold an extra meeting to discuss their scores after they completed the individual rating. Four dyads were re-rated and discussed among the three raters; these samples were then removed from the data set. In the meeting, the raters also discussed the applicability and usefulness of the scale.

While all the raters from the three pilots perceived the scale as a useful analytical tool for dialogic assessment, they also identified some difficulties. In the case of L2 English raters, subtle differences were detected when distinguishing between the upper levels of the scale which were then discussed in later sessions. This posed interpretive challenges, as raters struggled to consistently differentiate high-level adequacy across dimensions. This finding aligns with previous studies indicating that higher proficiency levels demand greater calibration due to finer performance gradations (Davis, 2016; Kuiken & Vedder, 2017).

Dimensions beyond Interaction, particularly Content and Comprehensibility, had greater variability in scores. A possible explanation of this variability, contrary to earlier findings (Kuiken & Vedder, 2022), is the major emphasis during rater training on the new Interaction dimension—in which the researchers/trainers were particularly interested—compared to the attention devoted to the three original dimensions of the monologue scale (i.e., Task Requirements, Content, and Comprehensibility).

Differences in the length and structure of interlocutors' turns, ranging from brief exchanges to extended monologic contributions, appeared to add difficulty to the evaluation of the adequacy of the interaction. Differences in raters' L1 backgrounds across pilots—L1 Italian in Pilots 1 and 2, and L1 Chinese and L1 Spanish in Pilot 3—may also have influenced the perception of interactional adequacy, underscoring the potential interplay between raters' sociolinguistic backgrounds and their interpretive frameworks. As noted by Davis (2021), such factors may contribute to differential rater severity, consequently highlighting the need for sustained reflection, calibration sessions, and reiterative training during the rating process.

The raters also reported several strategies to enhance reliability and to mitigate bias during the assessment process. These included the systematic provision of benchmark samples, and the discussion of challenging cases. They also suggested that the reliability and applicability

of the interactive FA scale could benefit from integrating in rater training more task-specific examples and content- and culture-relevant criteria, such as turn-taking conventions, to anchor interpretations more firmly.

All in all, the findings indicate that raters across the three pilot studies regarded the FA scale as an effective and adaptable instrument for evaluating interactional language performance. At the same time, its successful implementation appears to benefit from sustained rater training. Consistent with Fulcher (2003) and Brown (2005), continuous feedback and group discussions are key in developing shared rating practices and minimising subjectivity.

6. DISCUSSION

6.1. THEORETICAL IMPLICATIONS FOR SLA, TBLT AND TBLA

The present study provides empirical support for adapting the FA scale, originally considered for monologic performance (Kuiken & Vedder, 2017, 2018), for dialogic interaction, a mode central to real-world communication, yet underrepresented in assessment research (Galaczi & Taylor, 2020; González-Lloret, 2022). By reformulating the Coherence & Cohesion dimension into Interaction, the study operationalises FA through observable interactional behaviours such as turn-taking, clarification practices, and collaborative management of interaction. The generally high interrater reliability observed for this new dimension across the three pilot studies indicates that key aspects of interaction can be assessed when supported by clearly defined descriptors and structured rater training. The results therefore are consistent with the proposition that FA is not limited to monological performance, but also applies to the dynamic, reciprocal processes that shape spoken interaction.

In addition, the results further contribute to the ongoing effort to operationalise FA within task-based frameworks. The moderate to high correlations between Interaction and the other FA dimensions suggest that, although conceptually distinct, these dimensions are closely related in the assessment of task-based performance. The high association also reflects the fact that successful interaction depends on several closely interconnected abilities, such as developing content, maintaining comprehensibility, and responding appropriately to an interlocutor, which communicative competence models often conceptualise as integrated components of language use (Canale & Swain, 1980; Bachman & Palmer, 2010; Hall et al., 2011). Future research should further investigate the extent to which the four dimensions can reliably distinguish different profiles of performance, for example learners showing strong interactional engagement despite limited content or successful task fulfilment despite relatively weak interactional initiative. The qualitative findings highlight that raters of the three pilot studies perceived the descriptors as clear and representative of observable interactive behaviour, but sometimes struggled to distinguish performance at higher proficiency levels, echoing previous research (Davis, 2016; Kuiken & Vedder, 2017).

From an SLA perspective, the results support the claim that interactional performance is associated with several communicative resources, including pragmatic, discourse, and strategic abilities involved in sustaining interaction (Roever & Ikeda, 2022). This interconnectedness is compatible with communicative competence frameworks that conceptualise interaction as involving several closely related pragmatic, discourse, and strategic resources operating simultaneously during communication. In this sense, the findings align with interactionist SLA approaches (e.g., Gass & Mackey, 2007; Long, 1996), but also with research in Conversation Analysis for SLA, which views interactional competence as developing through participants' management of conversational practices, including turn-taking, repair, alignment, and sequencing practices (Hall et al., 2011; Hellermann, 2008; Markee, 2008; Pekarek Doehler & Pochon-Berger, 2015; Wagner et al., 2018). In this respect, the study should be considered primarily as an exploratory attempt to operationalise interaction-related features, such as turn management, clarification behaviour, and collaborative task completion, within a task-based assessment framework.

6.2. PEDAGOGICAL IMPLICATIONS

Beyond theoretical and assessment-oriented contributions, the findings have several pedagogical implications for classroom practice, rater training, and curriculum design in TBLT/TBLA contexts.

Firstly, the generally high reliability among raters in the Interaction dimension suggests that the adapted FA scale could serve as a formative assessment instrument for evaluating and supporting learners' interactional skills. Similar to other frameworks developed for assessing oral interactional competence (e.g., Galaczi & Taylor, 2018; Youn, 2020; Young, 2011), the scale operationalises features such as turn-taking, topic management, clarification requests, and collaborative task fulfilment through explicit CEFR-derived descriptors. In this respect, the adapted scale extends earlier work on Functional Adequacy by incorporating interaction-related criteria into the FA framework. This supports classroom assessment and feedback practices that foreground communicative effectiveness rather than exclusively formal linguistic accuracy.

In addition, the findings from the raters' reflections and experiences across the three pilot studies emphasise the central role of training and calibration in achieving consistent judgments. Teachers and raters benefited from shared calibration sessions, benchmark samples, and reflective discussion of difficult cases, which align with established practices in language assessment (Brown, 2005; Pill & Smart, 2020). At the same time, the findings indicate that the successful implementation of the scale depends substantially on sustained training and shared interpretation of descriptors. This raises practical questions concerning the feasibility of applying the scale in classroom contexts where teachers may have limited experience with performance-based assessment. Incorporating collaborative calibration activities into teacher education programmes could nevertheless strengthen assessment literacy by helping practitioners interpret descriptors more consistently and better understand how interactional behaviours, task requirements, and communicative goals may shape assessment decisions.

More specifically, both the original and the adapted FA scales might encourage greater attention to communicative aspects of performance as goal-oriented language use rather than solely formal linguistic accuracy. In the case of the adapted interaction scale specifically, the explicit focus on engagement and cooperation may additionally draw attention to interaction-related behaviours such as turn management, building on interlocutors' contributions, and requesting clarification. Embedding FA descriptors into classroom activities, peer assessment, or reflective journals may support learners' ability to reflect on and monitor their communicative behaviour over time. Furthermore, the emphasis on adequacy rather than accuracy aligns with communicative principles of TBLT and TBLA by encouraging learners to focus on meaning, negotiation, and successful task completion during interactional tasks.

7. CONCLUSION

This study explored the reliability, validity, and applicability of Kuiken and Vedder's (2017, 2018) adapted FA rating scale for the assessment of L2 oral interactional performance. The findings of three pilot studies with learners of L2 Italian and English suggest that the FA framework can be adjusted from monologic to dialogic task performance, while also highlighting the importance of rater training and calibration in the assessment of spoken interaction.

Quantitatively, the results provided preliminary evidence for the reliability and internal consistency of the adapted scale. The Interaction dimension achieved generally high interrater reliability across the three pilots, suggesting that the revised descriptors can support consistent rater judgments when accompanied by structured training. Correlations between Interaction and the other FA dimensions suggest a close connection among different aspects of dialogic task performance, while still pointing to conceptually distinguishable components of Functional Adequacy, such as task fulfilment, content development, comprehensibility, and interactional management. Overall, the study suggests that the FA framework can be feasibly adapted and applied to interactive speech within a task-based assessment context, thereby extending the applicability of the original monologue FA scale to dialogic performance.

Qualitative data complemented these results by providing insight into how raters interpreted and applied the scale. The non-expert raters reported that the descriptors were sufficiently clear and applicable to the assessment of dialogic performance. Their reflections also highlighted the important role of iterative training, benchmarking, and collaborative moderation. These findings suggest that, with appropriate support, non-expert raters may be able to assess dialogic performance, meaning that such assessment need not be restricted to professional raters. At the same time, the study also shows that the scale does not function entirely "out of the box," as consistency in scoring depends substantially on the amount of training and discussion provided during the rating process.

Taken together, the findings suggest that the FA scale, adapted from monologic to interactive tasks, may constitute a promising instrument for assessing L2 oral interaction. Additionally, the findings indicate that the scale can become a potential pedagogical and formative assessment tool to support teachers and learners in understanding and improving interactional tasks.

The study also has a number of limitations, which future studies should address. The first limitation, which has already been discussed, concerns the strong focus during rater training on the new Interaction dimension, whereas comparatively less attention was devoted to the three dimensions from the original monologue scale, particularly Content and Comprehensibility, which were equally unfamiliar to the raters. Differences in training modality (on site in Pilot 1; online in Pilots 2 and 3) and data collection may have also influenced the results. Although all raters were initially considered non-experts, the findings further highlight the extent to which rater reliability might depend on prior knowledge of concepts related to TBLT, assessment, and task performance, as well as the amount of training provided during the rating process.

The data discussed in the present study are based on two target languages and relatively small samples of learners and raters. Future studies should therefore focus on validating the scale with larger and more diverse samples of raters and learners in different learning contexts, with different source and target languages and interactional task types. Further work could also explore its use in classroom-based assessment. In addition, further research might investigate how FA increases over time, how learners' interactional skills develop at different stages, and how the FA framework can be employed for technology-mediated interaction, thereby extending its relevance to contemporary communicative practices.

In sum, this study offers an initial exploration of how the FA framework, originally developed for monologic tasks, can be adapted to oral interaction while also highlighting the methodological and conceptual challenges involved in assessing dialogic performance within task-based assessment contexts.

DATA ACCESSIBILITY STATEMENT

The anonymised rating data and analysis materials supporting the findings of this study are available from the corresponding author upon request.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Appendices.** Appendix A and B. DOI: <https://doi.org/10.22599/jesla.180.s1>

ETHICS AND CONSENT

The study protocol and procedures of the project “Assessment of Functional Adequacy in L2 speech” (FGW-565) were reviewed and approved by the Amsterdam Institute for Humanities Research Ethics Committee on March 30, 2023. The study was conducted in accordance with the ethical standards of the Declaration of Helsinki. Written informed consent was obtained from all individual participants prior to their enrolment in the study.

AUTHOR CONTRIBUTIONS

Conceptualisation, Methodology: Ineke Vedder, Folkert Kuiken; Formal coding and analysis: Kristy Suet-Sin Cheung, Astrid Morrison, Tongzhou Xiao, Ineke Vedder, Paolo Orrù; Writing (original draft): Ineke Vedder, Kristy Suet-Sin Cheung, Astrid Morrison, Tongzhou Xiao; Writing (review & editing): Ineke Vedder, Kristy Suet-Sin Cheung, Astrid Morrison, Paolo Orrù.

AUTHOR AFFILIATIONS

Ineke Vedder  orcid.org/0000-0002-2677-0228
University of Amsterdam, NL

Folkert Kuiken  orcid.org/0000-0002-1667-2501
University of Amsterdam, NL

REFERENCES

- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press. <https://doi.org/10.1017/CBO9780511667350>
- Brown, A. (2005). Interviewer variability in oral proficiency interviews. Peter Lang.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47. <https://doi.org/10.1093/applin/1.1.1>
- Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics*, 6(2), 5–35. <https://doi.org/10.5070/L462005216>
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.
- Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment—Companion volume*. Council of Europe Publishing.
- Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, 33(1), 117–135. <https://doi.org/10.1177/0265532215582282>
- Davis, L. (2021). Rater and interlocutor training. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 322–338). Routledge. <https://doi.org/10.4324/9781003220756-25>
- Ekiert, M., Lampropoulou, S., Révész, A., & Torgersen, E. (2018). The effects of task type and L2 proficiency on discourse appropriacy in oral task performance. In N. Taguchi & Y. Kim (Eds.), *Task-based approaches to teaching and assessing pragmatics* (pp. 247–263). John Benjamins. <https://doi.org/10.1075/tblt.10.10eki>
- Fulcher, G. (2003). *Testing Second Language Speaking* (1st ed.). Routledge. <https://doi.org/10.4324/9781315837376>
- Galaczi, E. D., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, 15(3), 219–236. <https://doi.org/10.1080/15434303.2018.1453816>
- Galaczi, E. D., & Taylor, L. (2020). Measuring interactional competence. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 338–348). Routledge. <https://doi.org/10.4324/9781351034784-36>
- Gass, S. M., & Mackey, A. (2007). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 175–199). Lawrence Erlbaum Associates.
- Gilbert, R., Barón, J., & Levkina, M. (2011). Manipulating task complexity across task types and modes. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 105–138). John Benjamins. <https://doi.org/10.1075/tblt.2.10ch5>
- González-Lloret, M. (2022). The present and future of functional adequacy. *TASK*, 2(1), 146–158. <https://doi.org/10.1075/task.21008.gon>
- Grabowski, K. C. (2013). Investigating the construct validity of a role-play test designed to measure grammatical and pragmatic knowledge at multiple proficiency levels. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 149–171). Palgrave Macmillan. https://doi.org/10.1057/9781137003522_6
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41–58). Academic Press. https://doi.org/10.1163/9789004368811_003
- Hall, J. K., Hellermann, J., & Pekarek Doehler, S. (Eds.). (2011). *L2 interactional competence and development*. Multilingual Matters. <https://doi.org/10.21832/9781847694072>
- Hellermann, J. (2008). *Social actions for classroom language learning*. Multilingual Matters. <https://doi.org/10.21832/9781847690272>
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473. <https://doi.org/10.1093/applin/amp048>
- Hymes, D. (1967). Models of the interaction of language and social setting. *Journal of Social Issues*, 23(2), 8–28. <https://doi.org/10.1111/j.1540-4560.1967.tb00572.x>

- Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Penguin Books.
- IBM Corp. (2024). *IBM SPSS Statistics for Windows (Version 30.0)* [Computer software]. IBM Corp.
- Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing*, 34(3), 321–336. <https://doi.org/10.1177/0265532216663991>
- Kuiken, F., & Vedder, I. (2018). Assessing functional adequacy of L2 performance in a task-based approach. In N. Taguchi & Y. Kim (Eds.), *Task-based approaches to teaching and assessing pragmatics* (pp. 265–285). John Benjamins. <https://doi.org/10.1075/tblt.10.11kui>
- Kuiken, F., & Vedder, I. (2022). Measurement of functional adequacy in different learning contexts: Rationale, key issues, and future perspectives. *TASK*, 2(1), 8–32. <https://doi.org/10.1075/task.00013.kui>
- Kuiken, F., & Vedder, I. (2025). From CAF to CAFFA: Measuring linguistic performance and functional adequacy in task-based language teaching. In M. East (Ed.), *Broadening the horizon of TBLT: Plenary addresses from the second decade of the International Conference on Task-Based Language Teaching* (pp. 147–161). John Benjamins. <https://doi.org/10.1075/tblt.17.08kui>
- Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press. <https://doi.org/10.1016/B978-0-12589042-7/50015-3>
- Lu, X., & Suzukida, Y. (2026). Relationships of linguistic complexity and accuracy to functional adequacy across L2 English speaking and writing: Roles of L2 proficiency and task type. *International Journal of Applied Linguistics*, 36(1), 322–335. <https://doi.org/10.1111/ijal.12775>
- Markee, N. (2008). Toward a learning behavior tracking methodology for CA-for-SLA. *Applied Linguistics*, 29(3), 404–427. <https://doi.org/10.1093/applin/amm052>
- Michel, M. C. (2011). Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 141–174). John Benjamins. <https://doi.org/10.1075/tblt.2.12ch6>
- Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, 45(3), 241–259. <https://doi.org/10.1515/iral.2007.011>
- Nagle, C. L., Trofimovich, P., O'Brien, M. G., & Kennedy, S. (2022). Comprehensible to whom? Examining rater, speaker, and interlocutor perspectives on comprehensibility in an interactive context. *The Modern Language Journal*, 106(4), 675–693. <https://doi.org/10.1111/modl.12809>
- Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590–601. <https://doi.org/10.1093/applin/amp045>
- Pallotti, G. (2019). Assessing tasks: The case of interactional difficulty. *Applied Linguistics*, 40(1), 176–197. <https://doi.org/10.1093/applin/amx020>
- Pallotti, G. (2022). Holistic and analytic assessment of functional adequacy. *TASK*, 2(1), 85–114. <https://doi.org/10.1075/task.21014.pal>
- Pekarek Doehler, S., & Pochon-Berger, E. (2015). The development of L2 interactional competence: Evidence from turn-taking organization, sequence organization, repair organization and preference organization. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-based perspectives on second language learning* (pp. 233–268). De Gruyter Mouton. <https://doi.org/10.1515/9783110378528-012>
- Pill, J., & Smart, C. (2020). Raters: Behavior and training. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 135–144). Routledge. <https://doi.org/10.4324/9781351034784-15>
- Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, 37(6), 828–848. <https://doi.org/10.1093/applin/amu069>
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18–39. <https://doi.org/10.1016/j.asw.2010.01.003>
- Roever, C., & Ikeda, N. (2022). What scores from monologic speaking tests can(not) tell us about interactional competence. *Language Testing*, 39(1), 7–29. <https://doi.org/10.1177/02655322211003332>
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735. <https://doi.org/10.2307/412243>
- Sato, M. (2014). Exploring the construct of interactional oral fluency: Second language acquisition and language testing approaches. *System*, 45, 79–91. <https://doi.org/10.1016/j.system.2014.05.004>
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532. <https://doi.org/10.1093/applin/amp047>
- Strobl, C., & Baten, K. (2022). Assessing writing development during study abroad: The role of task and measures of linguistic and communicative performance. *TASK*, 2(1), 60–84. <https://doi.org/10.1075/task.21010.str>
- Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, 54(2), 133–150. <https://doi.org/10.1515/iral-2016-9994>

- Tavakoli, P.** (2018). L2 development in an intensive study abroad EAP context. *System*, 72, 62–74. <https://doi.org/10.1016/j.system.2017.10.009>
- Tavakoli, P., & Wright, C.** (2020). *Second language speech fluency: From research to practice*. Cambridge University Press. <https://doi.org/10.1017/9781108589109>
- Van den Branden, K.** (2020). Measuring task-based performance. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 316–325). Routledge. <https://doi.org/10.4324/9781351034784-34>
- Wagner, J., Pekarek Doehler, S., & González-Martínez, E.** (2018). Longitudinal research on the organization of social interaction: Current developments and methodological challenges. In S. Pekarek Doehler, J. Wagner, & E. González-Martínez (Eds.), *Longitudinal studies on the organization of social interaction* (pp. 3–35). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-57007-9_1
- Youn, S. J.** (2018). Rater variability across examinees and rating criteria in paired speaking assessment. *Papers in Language Testing and Assessment*, 7(1), 32–60. <https://doi.org/10.58379/YVWQ3768>
- Youn, S. J.** (2020). Interactional features of L2 pragmatic interaction in role-play speaking assessment. *TESOL Quarterly*, 54(1), 201–233. <https://doi.org/10.1002/tesq.542>
- Young, R. F.** (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 426–443). Routledge.

Vedder et al.
*Journal of the European
 Second Language
 Association*
 DOI: 10.22599/jesla.180

112

TO CITE THIS ARTICLE:

Vedder, I., Kuiken, F., Cheung, K. S.-S., Morrison, A., Xiao, T., & Orrù, P. (2026). Assessing functional adequacy in oral dialogic speech: Validating a rating scale for L2 interaction. *Journal of the European Second Language Association*, 10(1), 99–112. DOI: <https://doi.org/10.22599/jesla.180>

Submitted: 07 April 2026

Accepted: 25 August 2026

Published: 22 September 2026

COPYRIGHT:

© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <https://creativecommons.org/licenses/by/4.0/>.

Journal of the European Second Language Association, is a peer-reviewed open access journal published by White Rose University Press.

