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

Investigating the assessability of speaking proficiency in a group discussion context

Paul Garside
garsidepaul@hotmail.com
Meiji University

Abstract

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

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

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

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

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

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

Literature Review

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

Defining Speaking Proficiency

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

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

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

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

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

Assessing Speaking Proficiency

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

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

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

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

Many-Facet Rasch Measurement

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

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

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

Gaps and Research Questions

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

The research questions (RQs) are stated as follows:

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

Methods

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

Participants

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

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

Recorded Discussions

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

For all groups, the question prompts were:

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

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

Table 1

Rating Plan

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

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

The Scoring Rubric

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

Fluency

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

Accuracy

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

Strategy Use

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

Active Listening

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

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

Content

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

Item Fit in MFRM

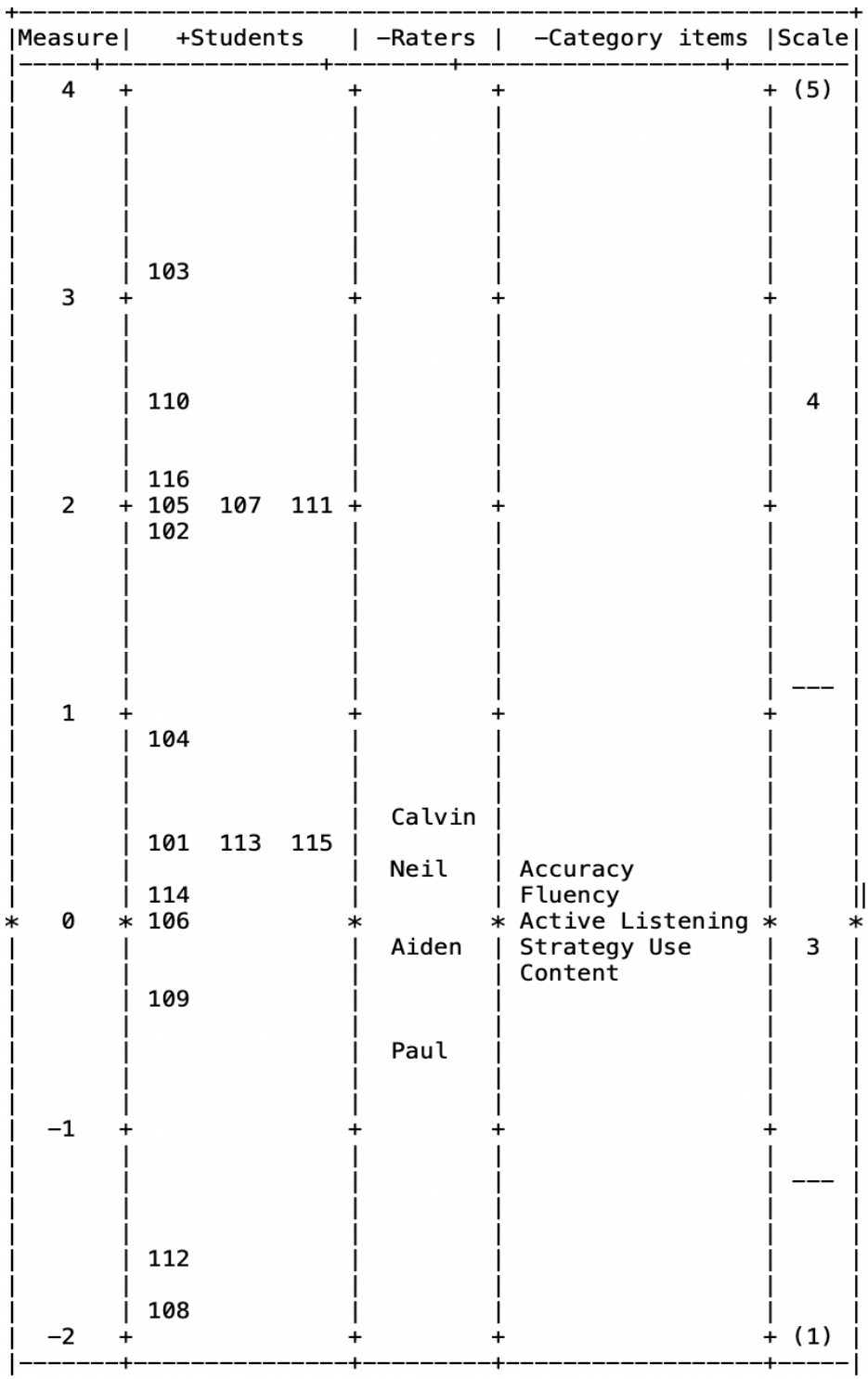
Item fit is an important assumption of Rasch modeling. Items that fit the model should have infit and outfit mean square (MNSQ) values close to the expected 1.0, although Linacre (2002) has argued that a range of .5 to 1.5 is productive for measurement purposes, and therefore acceptable in lower stakes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

Results and Discussion

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

Figure 1

All-facet Wright Map for the MFRM Analysis



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

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

Table 2

Rating Scores Based on Recorded Group Discussions

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

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

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

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

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

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

Raters

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

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

Table 3

Rater Severity and Model Fit

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

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

Category Items

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

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

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

Table 4

Model Fit of Rubric Category Items

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

Note. All statistics are based on Rasch logits.

Conclusion

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix A

Scoring Rubric with Level Descriptors

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