Articles

Developing a Rubric for Interactional Competence Using Many-Facet Rasch Measurement

Thomas Stones
Kwansei Gakuin University

The teaching and assessment of interactive speaking skills is a key aim in many English-language programs, and the right assessment rubric is a key component of any effective course. There is no one-size-fits-all approach and rubrics need to be validated, feedback collected, and revisions made. This paper reports on a small-scale, exploratory study undertaken to develop and improve a rubric for assessing interactive discussion skills. Utilizing many-facet Rasch measurement (MFRM), the paper reports on an analysis of the rubric originally used on the course. Based on these findings, the rubric was revised and subject to a second round of analysis. The findings indicate that the revisions led to considerably improved rubric function, due primarily to a reduced number of scale points and more clearly defined rubric categories. Finally, the paper suggests a number of recommendations for rubric creation that can be applied to a range of assessment contexts.

概要インタラクティブなスピーキングスキルの指導と評価は、多くの英語学習プログラムにおいて重要な目標であり、適切な評価ルーブリックは効果的な授業の重要な要素である。万能なアプローチは存在しない為、フィードバックを収集し、ルーブリックは検証、改善される必要がある。本稿では、対話型ディスカッションのスキルを評価するためのルーブリックを開発・改善するために行われた、小規模で探索的な研究について考察する。多相ラッシュ測定（MFRM）を活用し、当初コースで使用されていたルーブリックの分析について報告する。その結果に基づき、ルーブリックが改訂され2回目の分析が行われたところ、ルーブリックの機能に大幅な改善が見られた。その主な理由は、尺度点数を減らし、ルーブリックのカテゴリーをより明確に定義したことに

https://doi.org/10.37546/JALTJJ46.1-1

JALT Journal, Vol. 46, No. 1, May 2024

あることがわかった。最後に、本論文は様々な評価の文脈に適用可能なルーブリック作成に関する多くの提言を提示している。

Keywords: interactional competence; many-facet Rasch measurement; rubric development; speaking assessment; test validation

The effective development of speaking skills has become a core component of many English-language programs, and the range of approaches to assessing speaking has grown concurrently. Speaking assessments themselves can range in orientation from individual, paired and group and include tasks that test global speaking proficiency, interaction in specific scenarios or the use of pre-defined language forms or discourse functions (Luoma, 2004). In most cases, the speaking test represents a performance assessment (Johnson et al., 2009) where learners demonstrate the spoken language skills in authentic or semi-authentic ways. Central to the assessment of speaking is the use of a rubric to base judgements on participant performance. Any rubric should reflect the performances of the test participants and ability with the target skill as accurately as possible (Green, 2013). However, in many contexts various pressures mean that it is not often possible to investigate the effective functioning of rubrics or scoring systems, with teachers frequently left to trust in their own professional judgement as to how well these documents are functioning. Indeed, Janssen et al., (2015) note that in-depth studies of rubric development are relatively few and far between. Thus, in small part, this study aims to address this by exploring the validity of a rubric for a paired speaking test, primarily using many-facet Rasch measurement (MFRM).

Paired Speaking Tests and Interactional Competence

In recent years, the use of paired or group discussion tasks in universities has increased (e.g., Bonk & Ockey, 2003; Leaper & Brawn, 2019; Nitta & Nakatsuhara, 2014), as well as in a number of the higher-level Cambridge assessments (e.g., Cambridge 2008). The use of group discussions necessarily requires the incorporation of rubric categories that deal with interactional competence, which is the ability to effectively co-construct an interaction with an interlocutor within a specific context (Kramsch, 1986). High-profile examples are the interaction component in the CEFR (Council of Europe, 2001), which formed the basis of the similarly named ‘interactive communication’ category in the Cambridge exams (Galaczi et al., 2011). The inclusion of interactive competence measures is necessary to represent the

co-constructed nature of dialogic speech in group speaking tasks (Nitta & Nakatsuhara, 2014). Indeed, its inclusion in speaking tests has also been persuasively argued for by Roever and Kasper (2018) who note that it can lead to a far richer range of information on participants' ability to engage in the type of interactive, group-based talk that is common to many academic and professional contexts than potentially unstable predictions from assessments focused primarily on monologic production. Galaczi (2014) highlights several key areas of interaction that are central to the maintenance of an effective, co-constructed discussion which are topic development, listener support & turn-taking management, with greater topic development across turns and speakers particularly noticeable at higher CEFR levels. Similarly, Leaper and Brawn (2019) analysed the progression of learners over a two-year period focusing on four main areas of interaction: initiating, responding, developing and collaborating. Therefore, including interactional components into a rubric can help educators promote essential interactive skills for use beyond the classroom, as well as provide valuable reference information that enhances the validity and transparency of the awarding of scores (Jeong, 2015).

Many-Facet Rasch Measurement & Rating Scales

Once a suitable rubric has been created it is important to investigate how well it is functioning, which is where MFRM is of great use. MFRM is a statistical technique devised by Linacre (1994) that provides a 'rich set of highly efficient tools' (Eckes, 2015, p. 19) to examine how various facets of assessments interact to contribute to the assignment of scores. MFRM, therefore, can contribute to test development and administration due to the range of facets that can be compared and analysed (McNamara, 1996) and can inform revisions to rating scales for more meaningful and accurate scoring (Bond & Fox, 2015). MFRM has been gradually adopted in a variety of fields but has also become increasingly influential within applied linguistics and language teaching over the last 20 – 30 years (McNamara & Knoch, 2012). It has featured in a range of journals and has been used to investigate a variety of assessment types (Aryadoust et al., 2021) and was in fact foundational in the formation of the 6 CEFR levels (Council of Europe, 2001). More specifically, MFRM can be used to detect a variety of rater effects, such as leniency/severity, central tendency, randomness, halo, and restriction of range (Myford & Wolfe, 2003) as well as other demographic factors including gender, age, or attractiveness (Murphy & DeShon, 2000), format of test delivery (Nakatsuhara et al., 2020) or difficulty of assessment topics (Engel-

hard, 1992). In terms of research on scale function, Chen and Liu (2016) found that a 5-point rather than a 10-point scale functioned more effectively when evaluating written discourse completion for an email task. Janssen et al. (2015) similarly found that reducing the number of points on scales of a variety of sizes, some up to 20 points per rubric section, led to far more reliable scoring. Further, McDonald (2018) was able to considerably improve the functioning of a 9-point rubric to assess speaking skills by adopting a 5-point scale. Bonk and Ockey (2003) also utilised MFRM to explore the functioning of their group oral assessment in a Japanese University. They found that raters varied considerably in terms of the severity of scoring by as many as 2 points on a 9-point scale, despite training and practice on rubric use. Thus, MFRM is a highly flexible tool that can bring focus to areas of rubric and assessment performance that are difficult to obtain through other methods.

Rubrics & Raters

In addition to the rubric, raters can also introduce a large amount of unwanted variability to any score. The assessment of any spoken performance necessitates a subjective judgement on the part of the rater (McNamara, 1996), and human raters are inevitably fallible and may imperfectly represent any given performance (Eckes, 2015). This can add levels of construct-irrelevant variance, known generally as 'rater effects', which are a consequence of rater and not candidate performance (Scullen et al., 2000). There are myriad ways in which raters can differ in their application and interpretation of scoring rubrics and learner performances as well as potentially exhibiting other biases based on length of experience, pedagogical preferences, and educational background (Eckes, 2015). Furthermore, teachers can incorporate external additional factors when assigning grades, such as effort and behaviour throughout the course (Randall & Engelhard, 2009), or including factors such as body language and gaze despite them not being part of the scale (Orr, 2002). Rater training does help improve accuracy and eliminate extreme scoring phenomenon (Davis, 2016; Yan & Chuang, 2022), but despite even substantial and sustained attempts at rater training, some errors and biases can persist (McNamara, 1996; Myford & Wolfe, 2003). Therefore, it is essential to take remedial action where appropriate, as such running a MFRM analysis and providing the results to the teachers themselves (Myford & Wolfe, 2003). Validation of rater performance is also central to rubric development as it can lead to unreliable scoring or can indicate that rubric wording is not providing sufficient clarity to raters.

Research Context

This research took place at a Japanese university in the Kansai region. The speaking test used is part of the seminar-skills course, which is, in turn, part of a two-year (four semester), compulsory English-language program aimed at developing basic EAP skills. Student proficiency levels vary widely from A2 up to B2 and in some rarer cases C1. This seminar-skills course is the level 3 (of 6) course. The speaking test is a 5-minute paired discussion on a topic that participants had been studying for the previous 3 weeks. The weeks prior to the tests also introduced various discussion skills to be used in the test. As such, the test represented a summative assessment of content covered. At this level, the discussions skills are introductory and are closer to more general interactive competence than a full academic discussion. The original rubric included three categories: Discussion Skills, Discussion Questions, and Delivery and Effectiveness and could be described as a hybrid checklist-rating scale model whereby two of the rubric sections specify language to be used and checked off, but those sections were scored on a scale. The third section is a more typical rating scale with only the relevant constructs listed within it. The three rubric categories have scales of 10, 20 and 20 points respectively (see Appendix A). The scoring system was initially this single sheet but based on feedback received that it was difficult to discriminate between points on this scale, a more detailed explanation of the different bands was added to the primary rubric (Appendix B). Thus, the teachers had a 'live' rubric (Appendix A) for use while scoring participants as well as a 'detailed rubric' (Appendix B) that was also used as a reference to offer more guidance on the requirements for each category. In addition, the university-wide scoring policy sets 60% as a pass and states that the average score should be around 70 – 75%. The rubric was pre-existing within the program and had undergone various minor adjustments over a number of years and frequently received a range of feedback, from very positive to very negative. This feedback acted as the trigger for this research which intends to more deeply explore where the strengths and weaknesses in the rubric lie and find ways to improve it with the use of MFRM.

Research Aims

This research aims to explore the effective functioning of a rubric used for a paired-speaking test on a discussion skills course at a Japanese university. The aims of the research are as follows:

1. Does a many-facet Rasch measurement analysis reveal any issues in rubric functioning in the original rubric?

2. Based on question 1, what revisions should be made to the rubric?
3. Does a many-facet Rasch measurement analysis reveal any improvements in rubric functioning in the revised rubric?

Methods

For the examination of the original rubric, three speaking tests consisting of two participants each (total 6 participants) were video recorded and graded by 11 raters. The discussion topic for this test was 'Accommodation Options for University Students' and the problems and benefits associated with the various options. The participants were drawn from classes that covered the range of levels represented in the course with participants from the lowest, highest classes and mid-level classes selected. Students were originally assigned to classes based on TOEFL ITP scores taken at the program entry point. All raters were teachers that have had some experience on the course, work at the featured institution and held at least master's level-qualifications in TESOL or a related field and/or TESOL teaching certificates. All raters rated all speaking tests, meaning that the data was fully crossed. After teachers graded the speaking tests, they were asked to respond to a short questionnaire that focused on the validity and usability of the rubrics.

Thus, this is a small-scale, exploratory study taking an investigative approach to instrument development, looking to explore the functioning of the rubric and the teachers views thereof without aiming to prove or disprove a predetermined hypothesis (Singh, 2007). It is part of a broader study in which the primary area of data-collection is quantitative, with qualitative data used to supplement the quantitative findings (Morgan, 1998). The qualitative data would serve to add completeness to the picture gained from the Rasch analysis as well as provide methodological triangulation and facilitate instrument development (Bryman, 2006). This aims to align with the view of teachers and assessors as a community of professional practitioners who have the responsibility to uphold standards and contribute to a dialogue of continual, iterative improvement (Fulcher & Davidson, 2007) that should be mutually developed by key stakeholders in the local context (Hamp-Lyons, 1991; Ockey et al., 2013). This mixed methods approach is also in line with the Common European Framework (2001) recommendations on rubric development which suggests the use of intuitive, qualitative, and quantitative methods, where intuitive and qualitative elements can include informed, experience-based contributions with opportunities for feedback and review.

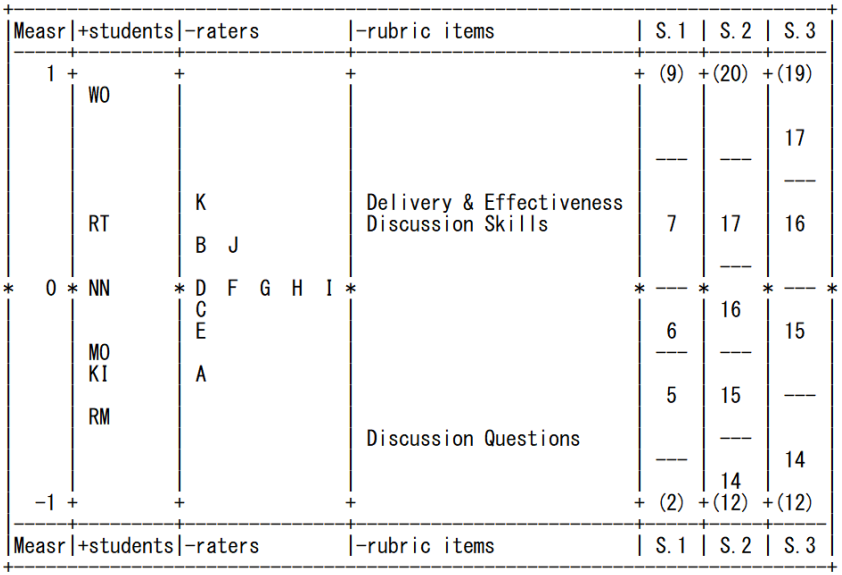
However, although the data collected from the questionnaire provided valuable insights, space restrictions preclude a detailed analysis of the responses. Also, despite the small sample size, it should be noted that MFRM does not necessarily need a large data set if the data fits the Rasch model. Indeed, Linacre's foundational work on MFRM (1994) reanalysed Guilford's (1954) data that featured only three raters and seven participants.

Findings from the Many-Facet Rasch Measurement Analysis

The raters' scores were input to the *Facets* (Linacre**,** 2001) program for performing MFRM. The Partial Credit Model was used to better compare the functioning of the three rubric categories. The Partial Credit Model analyses the rubric categories separately, so allows for more precision compared to analysing the rubric as a whole (Bond & Fox, 2015). This is also helpful where the rubric categories have different length scales as is the case here.

Figure 1 shows the Wright Map for the MFRM analysis. The Wright Map displays all facets ordered along a vertical logit scale (the leftmost column). The first column, students, orders the participants' ability from higher ability at the top to lower ability at the bottom. The next column gives the rater severity/leniency information with more severe raters placed at the top. 'Rubric items' orders the difficulty of the three rubric categories from easier at the top to harder at the bottom. The final three columns compare the relative difficulty of individual scale points. An initial analysis yields several interesting findings. Firstly, there is a narrow range of ability indicated on the logit scale, covering only 1.53 logits. The Rubric Items column shows the Discussion Questions section is the hardest, with the remaining two categories, Discussion Skills and Delivery and Effectiveness, exhibiting similar levels of challenge to each other. Rubric categories varying in difficulty is not necessarily an indicator of a misfunctioning rubric and can be desirable as different subskills may pose differing levels of challenge. In fact, the relative difficulty of the Questions section is likely a factor of poor assessment design. To score well on this section, most of the pre-determined discussion questions need to be asked. Within a short discussion, it is virtually impossible for both participants to ask all 4 questions, especially as some become redundant after one person has used them. Furthermore, the three columns on the far right raise some concern as they are considerably misaligned. Ideally, a score of 16, for example, on one component should align with a score of 16 on another if they are of similar difficulty. However, there is considerable misalignment, which will be further explored below.

Figure 1
Wright Map for Student, Rater & Rubric items



Statistical Findings from the Rasch Analysis

Tables 1–3 provide more detailed statistical information on rubric function. Examination of the student facet in Table 1 shows a separation index of 3.77 with a high reliability coefficient (0.93). This figure indicates how many different levels of ability were found among the learners, and a figure of 3.77 suggests that there were just under 4 ability levels across the cohort. The rater facet had a separation index of 0.76 and a reliability coefficient of 0.36. A separation index closer to 1.00 is ideal as it indicates that the raters are 'functioning as one' and producing similar scores for similar performances (Eckes, 2015). A higher separation index would indicate large differences in scores awarded for the same performance and would thus be undesirable. A reliability coefficient for raters closer to 0 rather than 1 is preferable (Eckes, 2015), so 0.36 is relatively good. These are encouraging findings in terms of rater reliability and suggest that a high level of consistency among raters. However, this could be partly due to the university's grading policy, which can have a narrowing effect on scoring. The rubric facet has a separation statistic of 5.19 with a high reliability coefficient (0.96). This is somewhat

problematic as it shows the rubric can only distinguish 5 ability levels. In and of itself, that is not a problem, but given that two of the category scales have 20 points available, it implies that only 25% of the scale points are being used, leading to a significant amount of redundancy.

Table 1

Separation Statistics for the Three Facets

	Root-mean Square Error	Separation Index	Reliability Coefficient	x^2
Student facet	0.13	3.77	0.93	0.00
Rater facet	0.18	0.76	0.36	0.07
Rubric items	0.09	5.19	0.96	0.00

Table 2

Measures and Fit Statistics for Raters and Rubric Categories

	Measure	SE	Infit MNSQ	Outfit MNSQ
Rater Facet				
A	-0.44	0.19	1.11	1.00
B	0.20	0.17	1.34	1.28
C	-0.08	0.18	0.93	0.98
D	-0.05	0.18	1.42	1.45
E	-0.24	0.18	1.28	1.20
F	-0.01	0.18	1.12	0.97
G	-0.05	0.18	0.56	0.57
H	0.05	0.18	0.32	0.32
I	-0.05	0.18	1.40	1.39
J	0.23	0.17	0.70	0.86
K	0.43	0.17	0.72	0.71
Rubric Facet				
Discussion Questions	-0.70	0.10	1.12	1.06
Discussion Skills	0.29	0.08	1.02	1.04
Delivery and Effectiveness	0.41	0.10	0.82	0.82

Table 2 gives more details on the raters' performances, with Infit MNSQ statistics largely falling within acceptable ranges. Infit MNSQ square statistics detail the extent to which the data matches the Rasch model and can serve to highlight various phenomenon among individual raters, such as erratic or conservative scoring (Myford & Wolfe, 2003). The acceptable range for this statistic varies depending on the purposes of the instrument with tighter ranges, for example between 0.8 and 1.2, preferred for higher-stakes situations and 0.7 – 1.3 for 'run of the mill' situations (Wright & Linacre, 1994), although often 0.5 – 1.5 is used, especially as small samples can widen the range of fit statistics (Wu & Adams, 2013). The Infit MNSQ statistics of the data analysed for this study generally fall between 0.7 – 1.3, suggesting good model fit and no erratic scoring, with no raters above 1.5. Two raters fell below the lower threshold, with Rater H at 0.32 and Rater G at 0.56. These indicate 'overfit' meaning that the raters more conservatively stuck to a narrow range of scores.

Overall, these figures would generally suggest fairly good rubric functioning and good model fit, with raters scoring in a fairly consistent manner. However, the narrow range of difficulties the rubric can discriminate and the misalignment of the scoring thresholds warrant further investigation.

Table 3

Rubric Functioning

Scale Point	Number of Observations	Av. Measure	Outfit MNSQ	Rasch-Andrich Threshold	Standard Error
Discussion Questions					
2	1	0.14	1.0		
3	0	N/A	N/A	N/A	N/A
4	2	0.16	0.8	-0.59	1.02
5	5	0.22	1.0	-0.73*	0.61
6	4	0.49	1.5	0.52	0.41
7	16	0.62	1.2	-0.95*	0.35
8	25	0.68	1.2	0.19*	0.27
9	13	1.06	1.0	1.56	0.33

Scale Point	Number of Observations	Av. Measure	Outfit MNSQ	Rasch-Andrich Threshold	Standard Error
Discussion Skills					
12	5	-0.76	1.0		
13	8	-0.83*	0.6	-1.24	0.49
14	8	-0.41	1.9	-0.67	0.34
15	9	-0.62*	1.4	-0.67	0.31
16	20	-0.28	1.1	-1.18*	0.29
17	2	0.22	0.5	2.13	0.35
18	12	0.17*	1.0	-1.71*	0.36
19	1	0.42	0.9	2.82	0.75
20	1	0.68	0.9	0.53*	1.04
Delivery and Effectiveness					
12	6	-0.87	1.0		
13	7	-0.98*	0.5	-0.99	0.45
14	23	-0.61	0.9	-1.89*	0.33
15	10	-0.50	1.2	0.3	0.29
16	11	-0.11	0.7	-0.40*	0.32
17	4	0.27	0.7	0.97	0.42
18	4	0.38	0.8	0.21*	0.52
19	1	0.56	0.8	1.79	1.03

* indicates where scale points do not advance in a linear fashion.

A Closer Look at Rubric Functioning

It is in the analysis of how individual points on the rating scales were awarded that the issues implied from the misalignment in the Wright Map and narrow separation index of 5.19 are fully explained. We can see that for all rubric categories a profound clustering of scores around particular scale points occurred. For example, the Number of Observations column shows that the majority of scores for Discussion Questions fall at scale points 7, 8, and 9 and at 14, 15, and 16 for Delivery and Effectiveness. Discussion Skills showed a greater spread, but a number of scale points were seldom selected,

most notably 17, 19, and 20. Furthermore, even though Discussion Skills and Delivery and Effectiveness are 20-point scales, less than half of these scale-points were actually used, with no scores awarded below 12 for either Discussion Skills or Delivery and Effectiveness. Even when considering that the university's scoring policy, and the small sample are likely having a narrowing effect, these results still suggest there are more scale points than there are levels of ability. Linacre's (2002) recommendation is that at least 10 observations per scale-point is needed for reliable analysis. With a total of only 66 ratings collected for this study (from 11 raters scoring 6 students across each of the 3 criteria), this is mathematically impossible with a 20-point scale, but the trends that are visible here are likely to be repeated with a greater number of raters and test takers, although a larger sample would be needed to confirm this.

Further issues can be seen in the average measures and Rasch-Andrich Threshold scores. In both cases, these should increase in line with the increase in the rating-scale points to suggest that a higher score on the rubric represents a higher-level of ability on the latent variable. Disordered categories, where the average measure and Rasch-Andrich threshold for a higher scale-point are below that of a lower scale point, reveal instances when the thresholds do not advance in a step-by-step manner and indicate that the rubric scale points are overlapping in the minds of the raters and, therefore, do not represent a distinct level of ability on the latent variable (Linacre, 2020). These points are marked with an asterisk in Table 3. The recommended distance between scale points is 1.4 – 5 logits (Linacre, 2002). Again, with a total range of 1.5 logits, this is clearly impossible in this data set and is a function of the extremely narrow range of scores awarded relative to the far wider span of the rubric. Another problem with some of the Rasch-Andrich thresholds is the large standard error figures associated with some of them. This is caused by the very low number of observations for several scale points, thus reducing their precision.

A visual representation of the trends indicated in Table 3 can be seen in Figures 2 – 4. These graphs display the probability of a particular score on the scale being awarded as difficulty increases. With a well-functioning rubric, the graph should appear as several distinct curves, similar in appearance to bell-curves, with the peak of each clearly separate from its neighbour, thus indicating that at each point on the latent variable, that score is the most likely. No lines should be subsumed by others, and curves should cross around their mid-points. Figures 2, 3, and 4 are obviously some distance from such a pattern.

Figure 2

Category Probability Scores: Discussion Questions

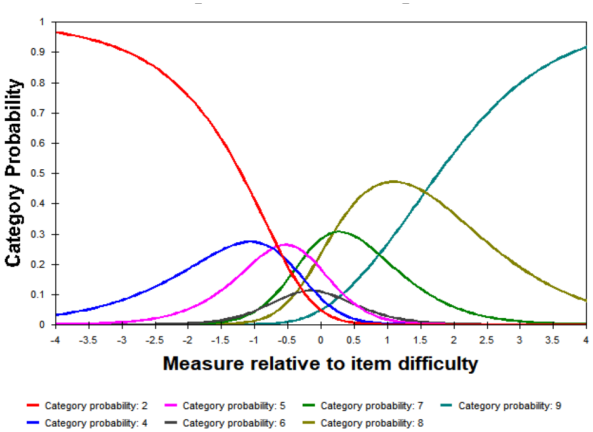


Figure 3

Category Probability Scores: Discussion Questions

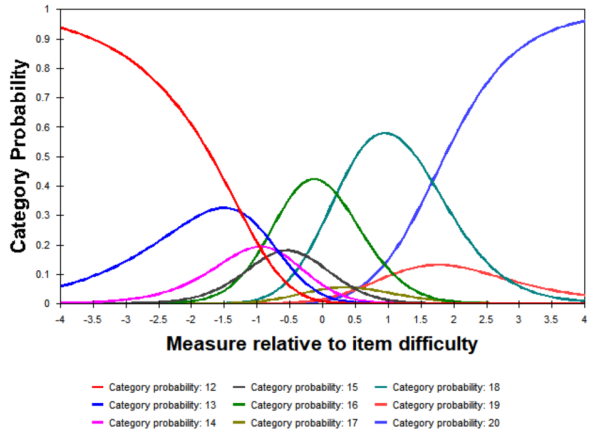
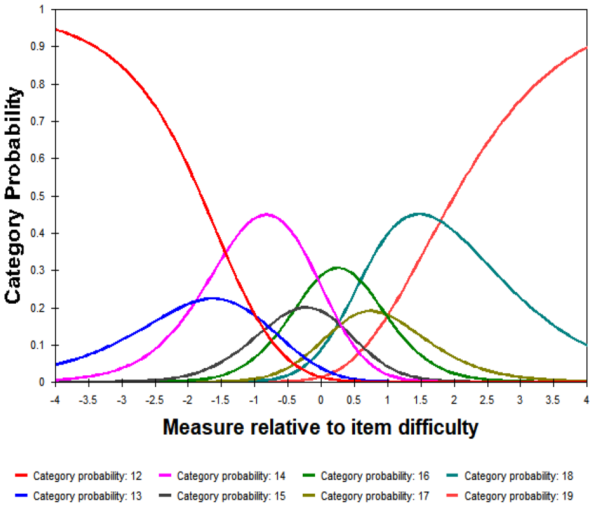


Figure 4

Category Probability Scores: Delivery and Effectiveness



Each of the three graphs show an amount of chaos in their alignment, with very few peaks distinct from the next and with a number subsumed by others. This suggests that raters do not have a clear idea of what level of performance is reflected by each point on the scale and indicates inconsistency in how points are awarded. The typical recommendation in such cases is to collapse the scale-points (Bond & Fox 2015; Eckes, 2015; Linacre, 2002).

Rubric Development Process

In light of the Rasch findings and feedback from teachers, a rubric revision process was undertaken that involved extensive discussions and multiple stages of drafting and redrafting. The major changes are summarized below.

Reduced number of scale points. In line with other studies where fewer scale-points improved functioning (Bonk & Ockey, 2003; Janssen et al., 2015; McDonald, 2018) recommendations for interpreting the output of an analysis using MFRM (Bond & Fox, 2015; Eckes, 2015; Linacre, 2002) and the results of the statistical analysis that the rubric distinguishes five levels of ability, the number of rubric categories was reduced. The revised scale goes from 1 – 5, with half scores at 3.5 and 4.5, so ultimately contains seven points. This is also the suggested maximum of seven that human raters can

deal with in short-term memory (Miller, 1956). As there were virtually no failing scores in the analysis of the original rubric, only two were awarded for Discussion Questions, it was thought that there needed to be some options for poor performances not fully represented in the original sample. Likewise, in creating the descriptors, it was felt that teachers would want more than three options for passing scores, especially as a maximum score is rarely awarded.

Move to a more general assessment of interactional goals. The original rubrics required learners to produce specific language, but this created several issues, so the descriptors will focus on interaction in general, rather than specific phrase production. Specific language should be taught in the course, but not mandated to be used within the rubric itself. Wiliam (2011) suggests including course language in the rubric itself to provide a connection to the course content, but its use should be subordinate to the achievement of interactional goals and avoid construct reductionalism (Green, 2013). Therefore, the Discussion Questions and Skills will be merged into a general 'Interaction' category, with the descriptors drawn from the interactional competence rubric developed by May et al. (2020) and the findings of Galaczi (2014).

Separate and reduce the constructs in the Delivery and Effectiveness category. This section was divided into two categories: Fluency and Language Use. The fluency category is based on that used by Nitta & Nakatsuhara, (2014), Iwashita et al. (2001), and later incorporated by McDonald (2018), as well as the criteria for the IELTS Speaking Test (IELTS, n.d.). Similarly, the Language Use category aims to incorporate the constructs of complexity and fluency and drew heavily on the IELTS criteria (IELTS, n.d.). This replaced the 'unit language' section as it was felt that the load placed on raters to reliably track the usage of 15 or so words that were included in each unit added to the already heavy cognitive burden that is often characteristic of scoring a performance with multiple traits (Hamp-Lyons & Henning, 1991).

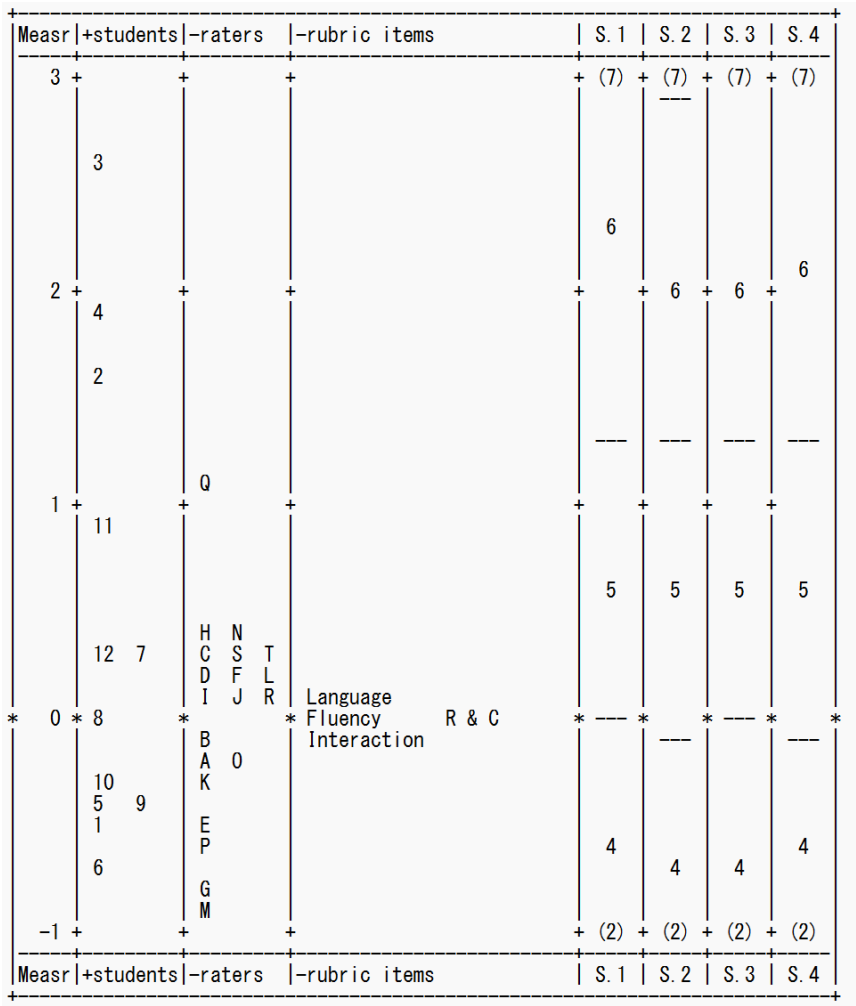
Add a Relevance & Content Category. This category was added as one intended outcome of the course is that the students engage with articles on the unit topics. Thus, this category was added to provide performance-based evidence that this goal has been achieved and thus the assessments align with the intended learning outcomes (Wiggins & McTighe, 2005). It was partly based on the descriptors in the Discourse Management Category for the Cambridge First Certificate (Cambridge, 2008).

Analysis of Revised Rubric

The redesigned rubric (see Appendix C) was tested with an expanded group of raters. In total, 20 raters scored the videos, all of whom had master's degrees in TESOL and/or extensive experience of teaching language. Some raters were from recruited from outside of the program; this was deemed important as the program coordinators suggested that teachers within the program tended to award most scores of around 70%, regardless of the wording of the rubrics. Using raters without such preconceptions should prove a better test of the validity of the descriptors. No mention was made in the instructions that a passing score was 60% and that an average of around 70 – 75% is expected by the institution. Additional speaking test videos were also added to provide a better spread of performances. All videos from the original rating session were included, plus three more, making a total of 12 performances. These additions now mean that this is not now a direct A to B comparison, but it was felt that expanding the sample was more important to explore rubric function more fully. In fact, it would have been ideal to have a greater number of performances to be rated, but logistical issues limited this. Further, standardisation training was not conducted before scoring. This is clearly less than ideal but does reflect the reality of the program, where standardization cannot always occur, and, therefore, provides a robust test of rubric performance in context-realistic conditions. Additionally, it was necessary to recode the scale as *Facets* cannot take decimals. Therefore, for the purposes of the Rasch analysis the scale points were recoded as follows: 3 = 3, 3.5 = 4, 4 = 5, 4.5= 6, 5 = 7.

Overall, the Wright Map (Figure 5) and Tables 4 – 6 shows several interesting findings. Regarding the spread of student abilities, the rubric identified a wider range of abilities as the separation statistics stood at 8.08, as shown in Table 4, slightly wider than the 7-point scale. Also, it is clear that the sample is skewing positively as no learners displayed abilities below -1 logits, but three above +1 logits. This is an artifact of the university policy of setting a passing grade at 60%, so the scale points 0 to 2 are less extensively used, as a pass should be achievable for most. The wording of the rubric was deliberately chosen such that the majority of scores would fall above this threshold. Also, given that a number of raters were unaware of this, it suggests the descriptor wording is targeting a suitable difficulty level for this cohort and the resulting skew in fact aligns the assessment institutional expectations, while still maintaining the ability to distinguish differing ability levels.

Figure 5
Wright Map for Revised Rubric



The spread of rater severity, column 2, now ranges from -0.94 to 1.10, a total range of 2.04 logits, an increase from 0.87 from the original rubric. Also, the separation statistic of 2.58 and reliability at 0.87 now suggest at least two statistically significant different levels of severity and a lower likelihood

of repeatability. These figures have increased from the original separation of 0.76 and reliability of 0.36. This is clearly worse than the original rubric and could be due to the novelty of the rubric, which was new to all raters. Rater training, ideally over a period of time, should bring scores closer into alignment. Indeed, it has been found that experienced teacher-raters can provide more-or-less reliable scores using their background and experience, as is likely the case here, but specific training with a given rubric can lead to considerable improvement in reliability and reduced severity ranges (Yan & Chuang, 2022).

The data on the rubric categories in column three now show the rubric categories as bunching very tightly together, with a separation statistic of 0.60, suggesting similar difficulty levels. The low reliability coefficient (0.26) supports this and demonstrates that the different rubric categories are similarly difficult. This may or may not be an improvement, as it could be indicative of halo effects (Myford and Wolfe, 2004).

Table 5 shows the fit statistics for the rubric, as all fall very close to 1 and within the narrower range of 0.7 – 1.3 (Wright & Linacre, 1994) suggesting good fit to the Rasch model.

Table 4
Separation Statistics for Revised Rubric

	Root-mean Square Error	Separation Index	Reliability Coefficient	x^2
Student facet	0.13	8.08	0.98	0.00
Rater facet	0.17	2.58	0.87	0.00
Rubric items	0.08	0.60	0.26	0.26

Table 5
Rubric Categories in Fit Order

Category	Measure	SE	Infit MNSQ	Outfit MNSQ
Interaction	-0.11	0.08	1.26	1.30
R & C	0.01	0.07	0.96	0.96
Language	0.10	0.08	0.94	0.94
Fluency	-0.01	0.08	0.80	0.79

Table 6 gives details on the raters, and overall, the raters fit the model well. Almost all fall between 0.5 – 1.5, with three underfitting with Infit MNSQ between 1.5 and 2.0. Two raters, R and F, are relatively close to the 1.5 threshold; however, rater N is somewhat higher. Only two raters, C and L, exhibited overfit, but overfit rarely causes any validity issues for measurement, especially when rater agreement is encouraged (Linacre, 2020). However, it could be indicative of halo effects where examiners show less variance than expected and assign identical scores across categories despite differing performances within each category (Myford & Wolfe, 2004). One simple method for investigating this suggested by Myford and Wolfe is to calculate the percentage of grades awarded by each rater that are identical across categories. This is shown in the rightmost column, and there appears not numerous incidences of halo effect. The two most overfitting raters, perhaps unsurprisingly, had the highest percentages of identical scores, but the 3rd lowest had none. However, further training would likely be beneficial (Linacre, 2012), especially for the three that underfit. The underfit exhibited here does not appear large enough to invalidate the measures, and so for the purposes of this study, it is not necessary to remove these ratings. Overall, without any formal training on the use of the rubric, these figures are encouraging and would improve with a standardisation session. Further encouraging statistical support is the close match of exact agreements, the Rasch Model expect this to be 31.1%, and the data yields a score of 31.2%.

Table 6
Raters in Fit Order

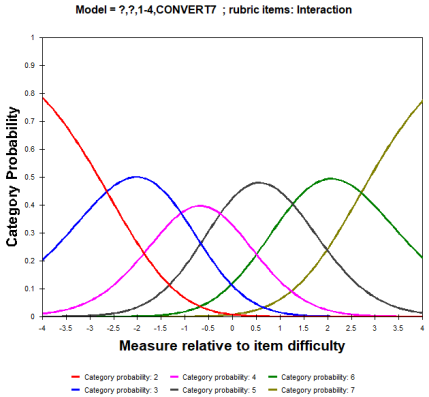
Rater	Measure	SE	Infit MNSQ	Outfit MNSQ	% Identical
N	0.40	0.16	1.87	1.91	8.33
R	0.10	0.17	1.65	1.63	16.67
F	0.15	0.17	1.59	1.58	8.33
J	0.10	0.17	1.28	1.39	16.67
Q	1.10	0.17	1.31	1.33	8.33
K	-0.30	0.17	1.27	1.27	33.33
G	-0.78	0.18	1.08	1.15	0.00
P	-0.56	0.17	1.06	1.07	8.33
O	-0.24	0.17	1.05	0.98	8.33

Rater	Measure	SE	Infit MNSQ	Outfit MNSQ	% Identical
T	0.26	0.17	0.99	1.02	8.33
M	-0.94	0.18	0.89	0.86	16.67
S	0.31	0.17	0.82	0.81	0.00
A	-0.21	0.17	0.77	0.77	0.00
B	-0.07	0.17	0.73	0.76	16.67
D	0.21	0.17	0.71	0.70	16.67
H	0.37	0.16	0.68	0.69	8.33
E	-0.47	0.17	0.61	0.62	8.33
I	0.07	0.17	0.59	0.57	0.00
C	0.31	0.17	0.45	0.46	25.00
L	0.21	0.17	0.39	0.38	33.33

Figure 6 gives the combined Category Probability Curves for the revised rubric overall. In general, the results here are very positive as each scale point is relatively distinct from its neighbour, the peaks are even and are not overlapping, and the peaks are not subsumed by others. In general, this points to a well-functioning rubric and is largely what could be hoped for in this context.

Figure 6

Overall Category Probability Curves



In addition to the overall rubric performance, it is also important to look at the individual category response curves (Andrich, 1996), as shown in Figures 7 – 10. Similarly positive results to the overall category curves are evident; however, some areas where further progress could be made. On the positive side, most peaks occupy their own space along the latent variable, but there are also clear exceptions to this, especially scale-point 4 in the Interaction and R and C categories, and to a lesser extent point 6 for fluency, where peaks are subsumed. Despite this, the improvement from version 1 is clear and substantial.

Figure 7

Category Probability Curves: Interaction

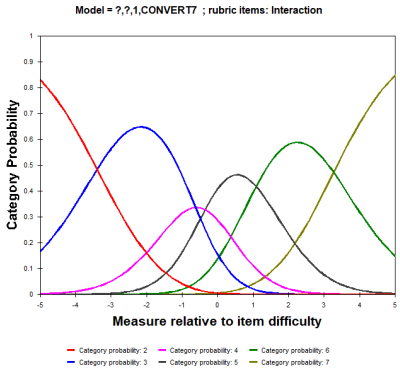


Figure 8

Category Probability Curves: Fluency

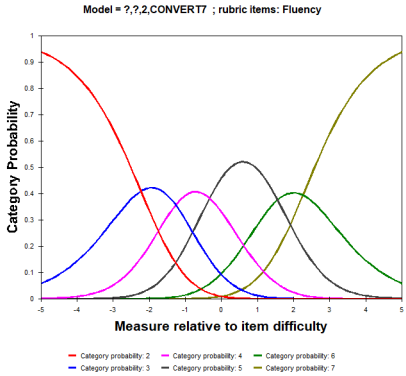


Figure 9

Category Probability Curves: Language

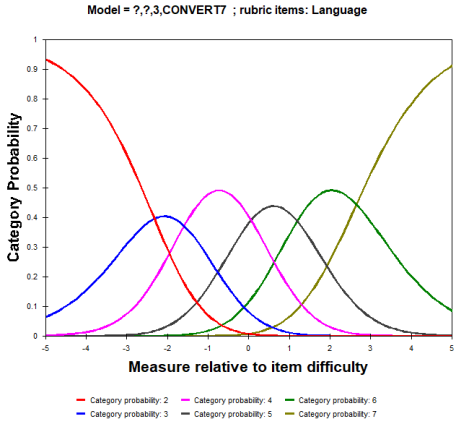


Figure 10

Category Probability Curves: Relevance and Content

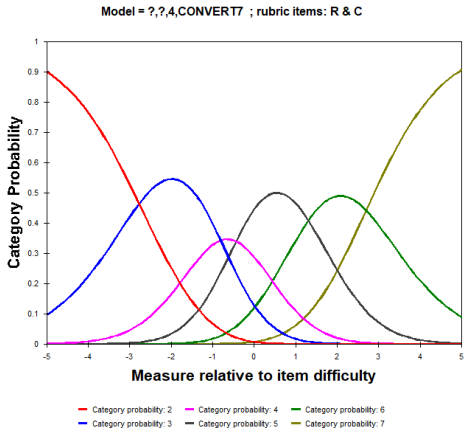


Table 7 gives specific statistical information for the four rubric categories. For all categories, the step calibrations advance monotonically, as per Linacre's recommendation (2002), a clear contrast from the first rubric iteration. Also, all passing grades, scores 3 to 7, have more than 10 observations, and so similarly meet Linacre's (2002) minimum requirement for stability.

However, Linacre (2002) also recommends that there be a minimum of 1.4 logits between the category thresholds, as can be seen in the Rasch-Andrich Threshold column, but this is not the case as a number of places where the spacing is below these recommendations can be seen. Instances of where the distance is below 1.4 logits are shown in bold in Table 7. This implies these scale points do not represent a suitably distinct level of ability on the latent variable; however, all scales do advance monotonically, which represents significant progress.

Table 7
Revised Rubric Step Calibrations

Scale Point	Number of Observations	Av. Measure	Rasch-Andrich Threshold	Distance to Next Category	Standard Error
			Interaction		
1	0	N/A	N/A	N/A	N/A
2	2	-1.12	N/A	N/A	N/A
3	34	-0.40	-3.39	2.77	0.72
4	48	0.28	-0.62	0.27	0.20
5	78	0.39	-0.35	1.44	0.16
6	63	1.08	1.09	2.18	0.17
7	15	2.48	3.27	N/A	0.30
			Fluency		
1	0	N/A	N/A	N/A	N/A
2	5	-1.23	N/A	N/A	N/A
3	24	-0.47	-2.23	0.96	0.47
4	59	-0.25	-1.27	0.95	0.22
5	87	0.27	-0.32	1.88	0.16
6	41	1.52	1.56	0.70	0.19
7	24	2.27	2.26	N/A	0.26

Scale Point	Number of Observations	Av. Measure	Rasch-Andrich Threshold	Distance to Next Category	Standard Error
			Language		
1	0	N/A	N/A	N/A	N/A
2	5	-1.29	N/A	N/A	N/A
3	24	-0.61	-2.32	0.73	0.47
4	75	-0.26	-1.59	1.65	0.22
5	71	0.37	0.06	1.13	0.16
6	47	1.20	1.19	1.46	0.19
7	18	2.20	2.65	N/A	0.28
			R & C		
1	0	N/A	N/A	N/A	N/A
2	4	-0.95	N/A	N/A	N/A
3	33	-0.51	-2.77	1.96	0.52
4	51	-0.30	-0.81	0.38	0.20
5	83	0.40	-0.43	1.74	0.16
6	50	1.42	1.31	1.39	0.18
7	19	1.89	2.70	N/A	0.28

Conclusion & Reflections

Overall, the use of a Rasch analysis has led to considerable improvements in the rubric functioning, with scale points and categories far more clearly delimitated, leading to far more reliable scoring. However, more work needs to be done in terms of validation as the small sample of test takers mean there could be more clarity in terms of the number of levels of ability the rubric can identify. Also, several scale points still have relatively narrow logit distances between them, so closer attention to the wording of the descriptors or a merging of some scale points could be areas that would improve functioning still further. Indeed, it has been argued that adhering to a consistent number of scale points across categories, although the norm and appearing 'neat' on the surface, may come with validity issues (Humphry & Heldsinger, 2014) as unnecessary scale points may be added for the sake of appearances.

Furthermore, although the categories now appear to be better matched in terms of overall difficulty, this can in fact provide less information on the sub-skills that make up the assessment, making it potentially less valuable. In our case, it appears that the apparent lack of halo effect means that the categories are of a similar level of difficulty, but care needs to be taken when interpreting such trends.

Through the process of developing this rubric there emerged some general principles that could be generally applied to rubric development, namely:

- **Less is more regarding scale points**. Frequently, a small number of scale points have been found to perform better than a larger number (Bonk & Ockey, 2003; Janssen et al., 2015; McDonald, 2018). This increases clarity as to what a particular score means and therefore allows for better feedback and clearer performance expectations. Although it may be tempting to allow a large range of points to be awarded for greater flexibility; in reality, this can lead to inconsistent scoring across raters and so should be avoided.
- **Separate constructs into clear categories.** In the original version of the rubric, there was some confusion arising from indistinctly defined constructs. By separating these into categories with clearly defined boundaries, raters and test takers alike will have clearer expectations as to what any rubric category is trying to target. This also helps to add to the granularity of the assessment as specific information can be provided about sub-dimensions of an overarching skill. This can reveal information on which aspects of performance pose differing levels of challenge to learners and action can be taken accordingly. Of course, this is assuming raters are scoring each category distinctly from the others and that halo effects are not evident. This is important as categories that align well on the Wright Map may look tidy but could indicate other issues.
- **Look to the bigger picture, avoid a check box approach.** The original rubric included individual phrases that were checked when used. Such an approach can be appropriate in some cases, but it has been argued that it can be reductionalist (Green, 2013) as it ignores certain aspects of performance. Some teachers commented, for example, that it is unclear if any phrases need to be pronounced perfectly or with 100% grammatical accuracy for points to be awarded. As such, seemingly simple checkbox approaches can in fact add complexity and reduce reliability if expectations are not clearly set.

- **Carefully word the descriptors based on the performance expectations of the cohort.** If an institution, as was the case here, has guidelines in terms of the passing score, then descriptors need to be written such that the minimum expected performance is 'set' to this benchmark. Knowledge of cohort ability and the general levels of performance they are capable of is essential here, as is teacher and assessor input.
- **Involve colleagues in the process of rubric development.** Despite teacher comments not featuring in this paper, they did play a significant role in the development process and provided valuable insights into teacher perceptions of rubric function and its usability. Adding a learner perspective in any future studies would strengthen any future research findings and involve more key stakeholders, as suggested for the development of any well-rounded testing instrument (Fulcher & Davidson, 2007; Hamp-Lyons, 1991; Ockey et al., 2013).

These recommendations need to be caveated with the proviso that the needs of all stakeholders in the local context need to be considered in the design of assessment instruments, but MFRM would likely be a useful tool where rater-mediated assessment is employed, regardless of the form of the rubric.

Thomas Stones has been working in language teaching for more than 15 years and currently works at the Department of Economics at Kwansei Gakuin University. He has a range of research interests including developing skills in interactional competence, assessment validation using Rasch-based methods, the teaching and assessment of listening skills as well as developing skills in self-directed learning. He has presented and published on all of these topics.

Appendices

All appendices are available from the online version of this article at https://jalt-publications.org/jj.

References

Andrich, D. A. (1996). Measurement criteria for choosing among models for graded responses. In A. von Eye & C. C. Clogg (Eds.), *Analysis of categorical variables in developmental research* (pp. 3–35). Academic Press. https://doi.org/10.1016/B978-012724965-0/50004-3

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 1–35. https://doi.org/10.1177/0265532220927487

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

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

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research, 6*(1), 97–113. https://doi.org/10.1177/1468794106058877

Cambridge. (2008). Assessing speaking performance – level B2. https://www.cambridgeenglish.org/images/168619-assessing-speaking-performance-at-level-b2.pdf

Chen, Y., & Liu, J. (2016). Constructing a scale to assess L2 written speech act performance: WDCT and e-mail tasks. *Language Assessment Quarterly, 13*(3), 231–250. https://doi.org/10.1080/15434303.2016.1213844

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* https://rm.coe.int/16802fc1bf

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

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement.* Peter Lang.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171–191. https://doi.org/10.1207/s15324818ame0503_1

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge. https://doi.org/10.4324/9780203449066

Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, *35*(5), 553–574. https://doi.org/10.1093/applin/amt017

Galaczi, E. D., ffrench, A., Hubbard, C., & Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach. *Assessment in Education: Principles, Policy & Practice, 18*(3), 217–237. https://doi.org/10.1080/0969594X.2011.574605

Green, A. (2013). *Exploring language assessment and testing: Language in action*. Routledge. https://doi.org/10.4324/9781315889627

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). McGraw-Hill.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Ablex.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41*(3), 337–373. https://doi.org/10.1111/j.1467-1770.1991.tb00610.x

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253–263. https://doi.org/10.3102/0013189X14542154

IELTS (n.d.). *Speaking: Band descriptors* (public version). https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx

Iwashita, N., Elder, C., & McNamara, T. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning, 51*(3), 401–436. https://doi.org/10.1111/0023-8333.00160

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing*, *26*, 51–66. https://doi.org/10.1016/j.asw.2015.07.002

Jeong, H. (2015). Rubrics in the classroom: Do teachers really follow them? *Language Testing in Asia*, *5*(6), 1–14. https://doi.org/10.1186/s40468-015-0013-5

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks.* Guilford.

Kramsch, C. (1986). From language proficiency to interactional competence, *The Modern Language Journal*, *70*(4), 366–372. https://doi.org/10.2307/326815

Leaper, D. A., & Brawn, J. R. (2019). Detecting development of speaking proficiency with a group oral test: A quantitative analysis. *Language Testing*, *36*(2) 181–206. https://doi.org/10.1177/0265532218779626

Linacre, J. M. (1994). *Many-Facet Rasch Measurement*. (2nd ed.). MESA Press.

Linacre, J. M. (2001). *FACETS* [Computer program, version 3.36.2]. MESA Press.

Linacre, J. M. (2002). Optimal rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85–106.

Linacre, J. M. (2012). *Many-Facet Rasch Measurement: Facets tutorial*. https://www.winsteps.com/a/ftutorial2.pdf

Linacre, J. M. (2020). *A user's guide to Facets*. https://www.winsteps.com/manuals.htm

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733017

May, L., Nakatsuhara, F., Lam, D., & Galaczi, E. (2020). Developing tools for learning oriented assessment of interactional competence: Bridging theory and practice. *Language Testing, 37*(2), 165–188. https://doi.org/10.1177/0265532219879044

McDonald, K. (2018). Post hoc evaluation of analytic rating scales for improved functioning in the assessment of interactive L2 speaking ability. *Language Testing in Asia*, *8*(19), 1–23. https://doi.org/10.1186/s40468-018-0074-3

McNamara**,** T. F. (1996). *Measuring second language performance.* Longman.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555–576. https://doi.org/10.1177/0265532211430367

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97. https://doi.org/10.1037/h0043158

Morgan, D. L. (1998). Practical strategies for combining qualitative and quantitative methods: Applications for health research, *Qualitative Health Research*, *8*(3), 362–376. https://doi.org/10.1177/104973239800800307

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*(4), 873–900. https://doi.org/10.1111/j.1744-6570.2000.tb02421.x

Myford, C. M., & Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Myford, C. M., & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

Nakatsuhara, F., Inoue, C., & Taylor, L. (2020). Comparing rating modes: Analysing live, audio, and video ratings of IELTS speaking test performances. *Language Assessment Quarterly*, *18*(2), 83–106. https://doi.org/10.1080/15434303.2020.1799222

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral task performance. *Language Testing*, *31*(2), 147–175. https://doi.org/10.1177/0265532213514401

Ockey, G. J., Koyama, D., & Setoguchi, E. (2013). Stakeholder input and test design: A case study on changing the interlocutor familiarity facet of the group oral discussion test. *Language Assessment Quarterly, 10*(3), 292–308. https://doi.org/10.1080/15434303.2013.769547

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores, *System 30*(2), 143–154. https://doi.org/10.1016/S0346-251X(02)00002-7

Randall, J., & Engelhard, G. (2009). Examining teacher grades using Rasch measurement theory. *Journal of Educational Measurement, 46*(1)*,* 1–18. https://doi.org/10.1111/j.1745-3984.2009.01066.x

Roever, C., & Kasper, G, (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing*, *35*(3), 331–355. https://doi.org/10.1177/0265532218758128

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*(6), 956–970. https://doi.org/10.1037/0021-9010.85.6.956

Singh, K. (2007). *Quantitative social research methods*. Sage. https://doi.org/10.4135/9789351507741

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). ASCD.

Wiliam, D. (2011). *Embedded formative assessment.* Solution Tree Press.

Wright, B., & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement, 14*, 339–355.

Yan, X., & Chuang, P. L. (2022). How do raters learn to rate? Many-facet Rasch modelling of rater performance over the course of a rater certification program. *Language Testing*, *40*(1), 153–179. https://doi.org/10.1177/02655322221074913

Appendix A

Live rubric

NAME:	Student Number:	Class:	/50

	EXCELLENT					PASS					
DISC. QUESTIONS	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**

1. What do you think about?

2. Should university students___?

3. What are the main benefits of___?

4. What are the main problems with ____ ?

Comments:

A score of 6 = basic use of two questions

	EXCELLENT					PASS				
DISCUSSION SKILLS	**20**	**18**	**16**	**14**	**12**	**10**	**8**	**6**	**4**	**2**	**0**

1. Introduce topics (let's talk about…)

2. Discuss benefits and/or problems

((one of) the main problems with ___ is…)

((one of)the main benefits of ___ is…)

3. Answer questions/give opinions with extra information

4. Deals with communication problems

Comments

A score of 12 = basic use of 2 skills.

	EXCELLENT					PASS					
DELIVERY & EFECTIVENESS	**20**	**18**	**16**	**14**	**12**	**10**	**8**	**6**	**4**	**2**	**0**

Student uses unit language and pronunciation appropriately (10)

Student has good control over pacing and hesitation (5)

Student actively participates to develop the discussion (5)

Comments:

Deduct up to 3 points if eye contact is absent/ineffective. Eye contact deduction: ☐

Appendix B

Detailed rubric

DISCUSSION QUESTIONS	
9 – 10	3 - 4 Questions are used with no form errors. All questions are used at completely appropriate points in the discussion and delivered effectively.
7 – 8	Two to three questions are used with very minor surface errors and/or slightly unnatural placement in the discussion. Delivery of the questions is fairly effective.
6	Basic use of two questions/effective use of one question. Two Questions may be produced with some inaccuracy and/or may not be used at completely appropriate places in the discussion, though meaning clear. One question is used with no surface errors **and** no issues with placement in the discussion.
4 – 5	One or Two questions used, but errors/usage significantly interferes with meaning. Noticeable errors are present in question forms and placement in the discussion, leading to misunderstanding/marked interaction.
0 - 3	No questions used/Questions used are highly inaccurate/incomprehensible and placement is extremely clumsy.
DISCUSSION SKILLS	
19 – 20	All 4 skills used highly effectively. Language relating to the skills is highly accurate. Skills are used at completely appropriate places in the discourse and help develop the discussion in a highly effective way.
15 – 18	3 - 4 skills used. Language relating to skills is accurate. Skills are used at appropriate places in the discourse and help develop the discussion.
12 – 14	2 skills are used with reasonable accuracy, or alternative but generally appropriate language is used, Skills are used at generally appropriate places in the discussion, though there may be some issues with timing/relevance to previous utterance. Contributions mostly help develop the discussion.
7 – 11	Only one skill used (even if appropriately). Two skills used very inappropriately & use interferes with meaning. Noticeable errors are present in all skill-related language, which may obscure meaning. Placement in the discourse is marked and usage is fairly clumsy. May lack relevance to previous utterance.
0 – 6	0 /1 skill(s) used. Skill-related language has significant form errors, making meaning unclear. Skills used at completely inappropriate points in the discussion and usage causes some confusion.
DELIVERY AND EFFECTIVENESS	
19 – 20	Very wide range of unit language used. All language is used highly accurately with appropriate stress & pronunciation. (9 – 10) All speech is delivered fluently and smoothly with minimal hesitation. (5) Highly active throughout discussion. Contributions considerably develop the discussion (5)
15 – 18	Good range of unit language used. Language use, stress and pronunciation is generally accurate and comprehensible. (7 – 8) Speech is delivered fairly fluently with some hesitation. (4) Active throughout the discussion & contributions generally develop the discussion. (4)
12 – 14	Some unit language used. Language use, stress and pronunciation is mostly accurate, but may require some listener effort to comprehend. (6) Flow of speech is generally maintained but with noticeable hesitation and/or repetition. (3) Sufficiently active throughout the discussion. Contributions maintain the discussion (3)
7 – 11	Few unit language items used. There are some issues with use, stress and pronunciation, meaning may be unclear at times. (4 – 5) Flow of speech is sometimes not maintained and there is significant hesitation and/repetition. (2) Relatively inactive in the discussion, appears reticent to speak/overly dominant in the discussion (2)
0 - 6	No/almost no unit language used. Language use, stress and pronunciation is mostly inappropriate and causes considerable strain for the listener. (0 – 3) Participation is extremely limited and participation minimal/extremely dominant throughout the discussion (1) Delivery is extremely halting/follow of speech generally not maintained beyond one clause/phrase. (1)

Appendix C

Revised rubric

	5	4.5	4	3.5	3	2	1
Interaction	Participants cooperate to build on and develop each other's ideas very effectively. There is little or no inter-turn pausing. Interaction is highly effective and features a wide range of interactional strategies.	Between 4 & 5.	Participants generally cooperate to build on each other's ideas, although some topics may not be adequately developed. There is some minor inter-turn pausing. Interaction is mostly effective and features a range of interactional strategies.	Between 3 & 4.	Participants attempt to cooperate and build on each other's ideas, although topic development may be limited. There is some, short inter-turn pausing. Interaction is generally acceptable using a limited range of interactional strategies.	There is some attempt to cooperate and develop ideas, but contributions are often not sufficiently related to the previous turn. There is some noticeable, inter-turn pausing. Interaction is faulty and often breaks down.	Participants generally do not cooperate. Speakers appear to produce pre-planned language with little or no topic development or mutual interaction. There is significant mid-turn pausing. Interaction is limited/non-existent
Fluency	Speaks fairly fluently throughout although occasional hesitation, repetition or filler use may be present. Can generally produce complex speech with good fluency.	Between 4 & 5.	Speaks with a mixed degree of fluency with hesitation, repetition or filler use. Complex speech causes dysfluency, but simple language is generally fluent.	Between 3 & 4.	Can generally maintain flow of speech albeit slowly with some use of pausing, hesitation and/or repetition. Even simple speech can be quite slow.	There is noticeable hesitation. Speaker cannot form even simple utterances smoothly.	Speech is highly disfluent characterised by short utterances and significant pauses. Speech is typically simple utterances.
Language use	Language use has a degree of sophistication/complexity and is generally accurate. Meaning is generally clear.	Between 4 & 5.	Complex/sophisticated language is attempted but is not always successful. Minor errors are fairly frequent and can sometimes interfere with meaning, especially with more complex language.	Between 3 & 4.	Language is mostly simple and complex language is rarely attempted. Errors are frequent and there are a number of errors that affect comprehensibility.	Language is mostly simple but complex forms are attempted occasionally. There are some errors that significantly affect meaning.	The language is very simple and is largely memorized utterances /single phrases. There are frequent errors, even in basic language. Errors affect the meaning, such that communication is very difficult.
Relevance and content	Contributions are detailed and relevant demonstrate original thinking to extend and explore the topic.	Between 4 & 5.	Contributions are relevant, detailed and varied with some originality and depth.	Between 3 & 4.	Contributions are generally relevant, although tend to be simplistic.	Contributions are minimally relevant/overly simplistic.	Contributions are not relevant to the topic or prior turn.