Jurnal Evaluasi Pendidikan
Volume 14, Nomor 1, 2023

SPEAKING ENGLISH PERFORMANCE ASSESSMENT WITH THE FACET RASCH MEASUREMENT MODEL

Wahyu Hidayat
Universitas Islam Negeri (UIN) Sultan Maulana Hasanuddin Banten, Indonesia

M. Noor Anzali
Universitas Islam Negeri (UIN) Sultan Maulana Hasanuddin Banten, Indonesia

Muhammad Turmudi
Universitas Islam Negeri (UIN) Sultan Maulana Hasanuddin Banten, Indonesia

Alamat Korespondensi
Wahyu.hidayat@uinbanten.ac.id

ABSTRACT
This study aims to assess students' English-speaking abilities based on peer assessment. This study is a quantitative study involving 10 students. Data was collected using tests and student speaking assessment rubrics with score criteria from 1 to 5. Speaking assessment criteria are pronunciation, grammar, vocabulary, fluency and understanding. Data were analyzed using Many Faceted Rasch Measurement (MFRM). The Facets Rasch Measurement model is able to see the interaction between respondents and items at once. The research results show that the item index for criteria/quality (6.39), speaker (0.51), and rater (5.32) as well as the standard deviation value clearly shows a good distribution of item difficulty. Criterion reliability is 0.98 for raters is 0.21, for raters is 0.97.
Keywords: *Speaking Performance Assessment, and Facet Rasch Measurement Mode.l*

ABSTRAK
Penelitian ini bertujuan untuk menilai kemampuan berbicara bahasa Inggris siswa berdasarkan penilaian teman sejawat. Penelitian ini merupakan penelitian kuantitatif yang melibatkan 10 orang siswa. Data dikumpulkan dengan menggunakan tes dan rubrik penilaian berbicara siswa dengan kriteria skor 1 sampai 5. Kriteria penilaian berbicara adalah pengucapan, tata bahasa, kosa kata, kefasihan dan pemahaman. Data dianalisis menggunakan Many Faceted Rasch Measurement (MFRM). Model Facets Rasch Measurement mampu melihat interaksi antara responden dan item sekaligus. Hasil penelitian menunjukkan bahwa indeks butir soal kriteria/ kualitas (6,39), pembicara (0,51), dan penilai (5,32) serta nilai simpangan baku jelas menunjukkan sebaran kesukaran butir yang baik. Keandalan kriteria adalah 0,98 untuk penilai adalah 0,21, untuk penilai adalah 0,97.
Kata Kunci: *Penilaian Kinerja Berbicara, dan Model Pengukuran Facet Rasch*

1. Introduction

Language learning includes speaking as a communicative skill and other important aspects, such as pronunciation, intonation, grammar, vocabulary, and so on. To ensure that students can communicate in the target language, these things must be taught during the process of learning any language. One of the skills most valued by students is speaking. It's an important part of everyday interactions, and a person's first impression is often based on their ability to speak fluently and provide information thoroughly. Teachers should prepare students as best as possible to speak English in real life situations (Byrne, 1986).

In language learning the main goal is mastery of language skills. Language skills refer to skills in using language in communication. With language skills, someone can express their thoughts and feelings to other people. This is the main goal of language learning as a form of communication. In linguistic studies, language skills are concrete and refer to the actual use of language, in spoken form that can be heard or in written form that can be read (Leong & Ahmadi, 2017; Sanjaya & Hidayat, 2021).

Mastery of these skills is an important aspect that determines the success of the second or foreign language teaching and learning process (Brown, 2000; Nunan, 1991)and also characterizes speaking proficiency as a sign of a successful level of language proficiency. When someone speaks, listeners will give specific responses to personality and attitudes (Louma, 2004).

The majority of Indonesian students still believe that English is difficult to learn. This phenomenon shows that students face difficulties in learning English (Pratolo, 2017). Even though they are over 17 years old and have studied English for more than six years, most Indonesian students cannot speak English (Fahmi et al., 2020). Students at universities are even in their third or fourth semester of class having difficulty speaking English. This shows that the teaching and learning process in Indonesia faces serious problems (Fahmi et al., 2020). This phenomenon also occurs in several other countries in Asia where university students cannot speak English well (Ramdani & Rahma, 2018). Thus, assessment is an alternative to monitoring and assistance.

Assessment is an important method for identifying differences and communicating between what the student is teaching and what the evaluator is teaching

about a particular topic or subject (Mulianah & Hidayat, 2021; Sanjaya & Hidayat, 2022). Effective assessment planning must identify the main problems students face. These assessments should help students feel more comfortable in learning, acknowledge their weaknesses, help them express their confusion and increase their motivation to learn (Hidayat, Lawahid, et al., 2021).

2. Research Methods

This research uses a quantitative descriptive approach to collect and analyze numerical data. Descriptive research is a research method that aims to provide a systematic and careful description of the facts and characteristics of a particular population with the aim of solving actual problems and collecting data or information to then compile, describe and analyze (Arikunto, 2002; Hidayat, Musab, et al., 2021). This research involved 10 students to assess the speaking abilities of 10 students. Students' speaking abilities are measured using tests and assessed using a performance rubric. Then the data was analyzed using Many Faceted Rasch Measurement (MFRM). In the Facets Rasch Measurement model, we can see the interaction between respondents and items at once (Aryadoust et al., 2021). In the Rasch model, the value is seen based on the logit value, which shows the probability of an item being selected in a group of respondents (Aryadoust et al., 2019; Maryati1 et al., 2019).

3. Results and Discussion

3.1. Rubric Items for Assessment of Speaking Performance Quality

Table 1 presents the MFRM data analysis, providing summary statistics of the reliability and discount indices of the items and raters of the MFRM analysis results. Item and rater reliability was considered excellent for the measurement. Items with high reliability indicate that the items as a whole define the latent variable well. This shows that the seven items are reliable and can be applied to various groups of respondents. However, the item index presents the item difficulty level.

In this study, a good distribution of item difficulty was demonstrated by the item separation indices of criterion/quality (6.39), speaker (0.51), and rater (5.32), along with clear standard deviation values. While the separation index for raters shows how well this rubric can assess "person's abilities" in terms of speaking performance assessment, which is latent, indicating that this rubric assessment instrument is suitable and reliable for identifying speaking performance assessment.

Table 1. Reliability and Separation of MFRM

	Reliability	Separation
Criteria/Quality	0.98	6.39
Speakers	0.21	0.51
Rater	0.97	5.32

Table 2 shows that the questions and test takers are of good quality (Wind & Engelhard, 2016). The above information indicates a high vacancy rate; This level of separation indicates that only one group of speakers has speaking skills. The results show that the criteria/quality show High/Very Good reliability (0.91–0.94), while speaker reliability is low (less than 0.67) and rater reliability is very good (more than 0.94).

3.2. Speaking Performance Appraisal Analysis

Below is information about the results of the assessors involved in the study, the metrics used to assess students, the standards used to assess, how well the rubric levels performed, and how the achievement levels of fellow assessors impacted the process.

This part of the study also included student responses to open-ended questions from peer assessment. The logistic map includes student scores for the peer assessment process, assessment criteria, and raters' levels of rigor and generosity.

Figure 1 below shows the analytical results of the analysis; The first column shows the scale size of the logistics map, with the scale level being between -2 and +2. The results of the speaker's analysis are displayed in the second column. In the third column, the logit map displays the students' assessment criteria, and the second column shows the distribution of students' scores based on their performance in the assessment, distributed from top to bottom, from students with the highest scores to students with the lowest scores. On the other hand, column four shows the division of the raters, and the last column shows the division of the degree of assessment, which has a score between 1 and 5. We can see the elements visually in the same table thanks to the logistic map that has all this data.

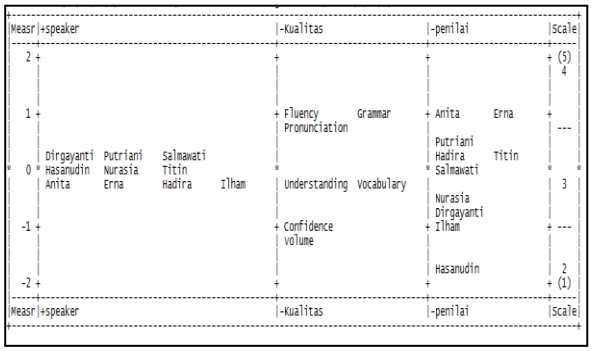

Figure 1. Map of Rater, Item, and Speaker Variables

The results showed that the speaker with the lowest score was Ilham and the highest was Dirgayanti. After that, the quality or criteria can be seen. Fluency, grammar, and pronunciation are the highest levels of difficulty. Comprehension and vocabulary are at an intermediate or moderate level. Confidence and volume are also easy. There were two assessors—Anita and Erna—who gave high marks, but Hasanuddin was the stingier assessor.

3.3. Quality of Speaking Item

Figure 2 below describes the indicators of student speaking quality or performance in detail.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Kualitas
223	90	2.48	2.49	.97	.13	.93	-.4	.93	-.4	1.09	.57	.57	4 Fluency
228	90	2.53	2.55	.88	.13	.86	-1.0	.85	-1.1	1.18	.61	.57	2 Grammar
239	90	2.66	2.68	.69	.13	.98	.0	.99	.0	1.03	.77	.57	1 Pronunciation
286	90	3.18	3.21	-.15	.14	.96	-.2	.97	-.1	1.07	.66	.55	3 Vocabulary
286	90	3.18	3.21	-.15	.14	.79	-1.4	.83	-1.1	1.17	.45	.55	5 Understanding
325	90	3.61	3.64	-.91	.14	1.02	.1	1.01	.1	.97	.44	.53	6 Confidence
344	90	3.82	3.86	-1.31	.15	1.31	1.9	1.27	1.7	.64	.28	.51	7 volume
275.9	90.0	3.07	3.09	.00	.14	.98	-.2	.98	-.2		.54		Mean (Count: 7)
44.3	.0	.49	.50	.83	.01	.15	1.0	.13	.9		.15		S.D. (Population)
47.8	.0	.53	.54	.89	.01	.16	1.1	.14	1.0		.16		S.D. (Sample)

Model, Populn: RMSE .14 Adj (True) S.D. .81 Separation 5.90 Strata 8.20 Reliability .97
Model, Sample: RMSE .14 Adj (True) S.D. .88 Separation 6.39 Strata 8.85 Reliability .98
Model, Fixed (all same) chi-square: 241.6 d.f.: 6 significance (probability): .00
Model, Random (normal) chi-square: 5.9 d.f.: 5 significance (probability): .32

Figure 2. Quality of Speaking Item

The speaking fluency indicator is the lowest or most difficult criterion to master, as shown in Figure 2. Apart from fluency, grammar and pronunciation, the confidence and volume indicators are the easiest for students to master. This is displayed in the measure column with positive values for fluency, grammar, and pronunciation, and negative values for confidence and volume. If the performance shows a positive number, then the indicator is difficult or difficult for the student to master, but if the performance shows a negative number, then the indicator is easy for the student to master(Weigle, 1998).

Rater of Speaking

Figure 3 below shows the Rater's results on speaking performance.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu penilai
152	63	2.41	2.40	1.06	.16	.86	-.8	.86	-.8	1.18	.73	.57	7 Erna
152	63	2.41	2.40	1.06	.16	.84	-.9	.84	-.9	1.24	.82	.57	8 Anita
169	63	2.68	2.71	.59	.16	1.05	.3	1.04	.3	.94	.51	.57	5 Putriani
178	63	2.83	2.85	.37	.16	.90	-.5	.92	-.4	1.09	.66	.57	10 Titin
185	63	2.94	2.95	.22	.16	.99	.0	1.00	.0	.98	.68	.56	4 Hadira
189	63	3.00	3.04	.08	.16	1.15	.9	1.20	1.1	.76	-.01	.56	9 Salmawati
211	63	3.35	3.40	-.52	.17	1.02	.1	.98	.0	1.01	.55	.54	3 Nurasia
217	63	3.44	3.50	-.69	.17	.85	-.8	.85	-.8	1.14	.52	.53	1 Dirgayanti
227	63	3.60	3.61	-.89	.17	1.22	1.2	1.23	1.2	.74	.35	.53	6 Ilham
251	63	3.98	4.02	-1.69	.18	.83	-.9	.85	-.8	1.15	.46	.50	2 Hasanudin
193.1	63.0	3.07	3.09	-.04	.16	.97	-.2	.98	-.1		.53		Mean (Count: 10)
31.1	.0	.49	.51	.85	.01	.13	.8	.14	.8		.22		S.D. (Population)
32.8	.0	.52	.54	.89	.01	.14	.8	.14	.8		.23		S.D. (Sample)

Model, Populn: RMSE .17 Adj (True) S.D. .83 Separation 5.03 Strata 7.05 Reliability .96
Model, Sample: RMSE .17 Adj (True) S.D. .88 Separation 5.32 Strata 7.42 Reliability .97
Model, Fixed (all same) chi-square: 247.0 d.f.: 9 significance (probability): .00
Model, Random (normal) chi-square: 8.7 d.f.: 8 significance (probability): .37

Figure 3. Raters of Speaking

The order of Raters (Erna, Anitah, Putriani, Titin, Haida, and Salmawati) who gave easy marks or high scores is shown on Figure 3 above. Nurasia, Dirga, Ilham, and Hasanuddin then gave rather low scores. The Nu Assessor table shows the numbers, and the Measure table shows the raters who gave high marks. Raters with low scores tend to give low scores.

The process of teaching, learning and evaluation in education is very complex, so it is very important for teachers to be able to differentiate between various elements of evaluation. Educational researchers, especially in the field of language education, pay great attention to the ability to identify each element in evaluation items to assess and improve the language quality of educators and students in the future.

Researchers reviewed the results of their research based on previous findings regarding the results of the analysis of assessors who assessed speaking performance using the facet Rasch model. The reliability value of the criteria/item quality is 0.98, indicating the consistency of the quality criteria is very good, while the reliability value of the resource person is 0.21, indicating the consistency of the answers from the resource person is 0.97, this shows that the reliability of this assessor is very good (Fisher, 2007). based on reliability criteria, the item person reliability and reliability values are (1) <0.67: Weak, (2) 0.67 - 0.80: Fair, (3) 0.81-0.90: good, (4) 0.91-0.94: Very Good, (5)> 0.94: Very Good.

For the criteria, fluency, grammar, and pronunciation indicate difficulties for students to learn and master. The results showed that vocabulary and comprehension were medium level items, which were not difficult and easy to learn and master. Items that are easy to master are volume and confidence, which only require effort and don't require much thought.

4. Conclusion

From the analysis that has been carried out, it can be concluded that the criteria for items, assessors and

sources vary in level. This is evidenced by the fact that, because each speaker has the right to rate other speakers, this research is a peer-reviewed study with sources given high marks but other sources given low marks.

5. References

Arikunto, S. (2002). *Prosedur penelitian*. Rineka Cipta.

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), 6–40. https://doi.org/10.1177/0265532220927487

Aryadoust, V., Tan, H. A. H., & Ng, L. Y. (2019). Scientometric review of Rasch measurement: The rise and progress of a specialty. *Frontiers in Psychology*, *10*(2197). https://doi.org/10.3389/fpsyg.2019.02197

Brown, H. (2000). *Principles of language learning and teaching*. Prentice Hall.

Byrne, D. (1986). *Teaching Oral English: Longman Handbooks for English Teacher*. Longman Group.

Fahmi, Pratolo, B. W., & Zahruni, N. A. (2020). Dynamic assessment effect on speaking performance of Indonesian EFL learners. *International Journal of Evaluation and Research in Education (IJERE)*, *9*(3), 778–790. https://doi.org/10.11591/ijere.v9i3.20466

Fisher, W. P. J. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*(1), 195–201.

Hidayat, W., Lawahid, N. A., & Mujahidah. (2021). Problems and Constraints of Authentic Assessment among Children ' s Early Education Teachers. *Asia-Pacific Journal of Research in Early Childhood Education*, *15*(2), 87–109. https://doi.org/dx.doi.org/10.17206/apjrece.2021.15.2.87

Hidayat, W., Musab, M., Lawahid, N. A., & Mujahidah, M. (2021). Developing the flipped learning instrument in an ESL context: The experts' perspective. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *25*(1), 35–48. https://doi.org/10.21831/pep.v25i1.38060

Leong, L.-M., & Ahmadi, S. M. (2017). An Analysis of Factors Influencing Learners' English Speaking Skill. *International Journal of Research in English Education*, *2*(1), 34–41.

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

Maryati1, Prasetyo, Z. K., Wilujeng, I., & Sumintono, B. (2019). Measuring teachers' pedagogical content knowledge Using many-facet Rasch model. *Cakrawala Pendidikan, 38*(3), 1–14. https://doi.org/10.21831/cp.v38i2.26098

Mulianah, S., & Hidayat, W. (2021). Relationship between Teacher Communication Patterns and. *Jurnal Tarbiyatuna*, *12*(2), 146–155. https://doi.org/doi.org/10.31603/tarbiyatuna.v12i2.4388

Nunan, D. (1991). *Second language teaching*. McGraw Hill.

Pratolo, B. W. (2017). *"Exploring Indonesian Learners' Beliefs about Language Learning Strategies through Reflection", Figshare* [Monash University, Clayton, Australia]. http://www.adb.org/sites/default/files/publication/159308/adbi-financial-inclusion-asia.pdf

Ramdani, J. M., & Rahma. (2018). No Title. *Indonesian Journal of Applied Linguistics*, *8*(2), 388–401.

Sanjaya, B., & Hidayat, W. (2021). Evaluasi Keterampilan Berbicara Bahasa Arab Siswa Madrasah Aliyah Di Provinsi Jambi. *Journal of Arabic Studies*, *6*(2), 220–235. http://dx.doi.org/10.24865/ajas.v6i2.384

Sanjaya, B., & Hidayat, W. (2022). Student speaking skill assessment: Techniques and results. *International Journal of Evaluation and Research in Education*, *11*(4), 1741–1748. https://doi.org/10.11591/ijere.v11i4.22782

Weigle, S. (1998). Using FACETS to model rater training effects. *LanguageTesting*, *15*(263–287).

Wind, S. A., & Engelhard, G. J. (2016). Exploring Rating Quality in Rater-Mediated Assessments Using Mokken Scale Analysis. *Educational and Psychological Measurement*, *76*(4), 685–706. https://doi.org/10.1177/0013164415604704