DYNAMIC SPEAKING TEST

Reliability in ratings: A comparative study of human vs. AI evaluators in a speaking assessment

R. Sean McDonald

Executive summary

The Dynamic Speaking Test (DST) is a standardized assessment designed to measure language competence according to the Common European Framework of Reference for Languages (CEFR). Utilizing a task-based, communicative approach, DST evaluates test takers' speaking abilities across all CEFR levels, accommodating both beginners and advanced speakers (C1-C2 levels). The test comprises six tasks, where test takers record their responses, which are then evaluated by an AI system based on five criteria: Pronunciation, Fluency, Grammar, Vocabulary and Task Achievement.

Reliability and validation of DST:

- High inter-rater reliability: The Krippendorff's Alpha values for individual questions ranged from .76 to .87. This indicates robust reliability in the AI ratings across different questions, ensuring consistent and dependable assessments.

- Correlations: Significant correlations between machine and human ratings ($p < .01$) confirm that the AI's evaluations are reliable and comparable to human judgement, reinforcing the AI's capability to provide accurate and consistent evaluations.

- Rater characteristics analysis: The multi-facet Rasch analysis (MFRM) positioned the AI within the expected range of rater stringency. This confirms that the machine marking functions within acceptable reliability parameters, aligning closely with human evaluators in terms of rating stringency and consistency.

The integration of AI in DST enhances the reliability and validity of language assessments, providing standardized and objective evaluations across different proficiency levels. AI has been proven to be a reliable and valid tool for language proficiency assessment, offering consistency and efficiency comparable to human raters. Its ability to provide detailed performance insights and handle large-scale evaluations efficiently underscores its potential as a valuable asset in educational settings.

1. General information

A data set with ratings for speaking performance from 34 participants was provided. Six questions were evaluated by six human raters (Human 1 - 6; H1 - H6) and one evaluation run by an AI. The global rating (SCO) is also available.

The assessments were provided in CEFR levels, which were coded as follows for descriptive analysis purposes:

- A0 = 0
- A1 / A1- = 1
- A1+ = 1.5
- A2 / A2- = 2
- A2+ = 2.5
- B1 / B1- = 3
- B1+ = 3.5
- B2 / B2- = 4
- B2+ = 4.5
- C1 / C1- = 5
- C1+ = 5.5
- C2 = 6

For the multi-facet Rasch analysis (MFRM), the data was coded as follows in the interests of a clearer presentation:

- A1 = 1
- A1+ = 2
- A2 = 3
- A2+ = 4
- B1 = 5
- B1+ = 6
- B2 = 7
- B2+ = 8
- C1 = 9
- C1+ = 10
- C2 = 11

The informative value/generalizability of this study is limited due to the small number of test subjects (N = 34).

The following programs were used for the analyses: SPSS 29.0, jamovi 2.4, Facets 4.1.7.

2. Descriptive statistics

First, the average values of the six ratings of AI and H1-H6 were calculated, resulting in the following picture:

Table 1: Descriptive statistics

	N	Minimum	Maximum	Mean	Std. Error	Std. Deviation
SCO	34	0	5.5	2.912	0.252	1.4692
AI_Mean	34	0.6	5.4	3.044	0.2067	1.2052
H1_Mean	34	1	4.8	3.201	0.1647	0.9603
H2_Mean	34	0.8	4.8	2.966	0.1901	1.1084
H3_Mean	34	1.2	4.8	2.956	0.1754	1.0229
H4_Mean	34	1	4.9	3.245	0.1702	0.9926
H5_Mean	34	1	4.8	2.98	0.1913	1.1152
H6_Mean	34	0.7	4.8	2.953	0.1777	1.0362

We found a high degree of agreement between human and AI marking, with the scores demonstrating consistency among all raters. Even though there was some variation, the human and AI markers are very close and create a visible 'fish swarm', illustrating that the group is marking similarly overall.

As the data is ordinally scaled, a Spearman correlation matrix was calculated for the average ratings:

Table 2: Correlation coefficients

	AI Global	AI_Mean	H1_Mean	H2_Mean	H3_Mean	H4_Mean	H5_Mean	H6_Mean
AI Global	1	0.986	0.916	0.927	0.944	0.934	0.942	0.943
AI_Mean	0.986	1	0.899	0.916	0.937	0.938	0.942	0.941
H1_Mean	0.916	0.899	1	0.934	0.937	0.901	0.913	0.899
H2_Mean	0.927	0.916	0.934	1	0.939	0.937	0.926	0.933
H3_Mean	0.944	0.937	0.937	0.939	1	0.95	0.955	0.933
H4_Mean	0.934	0.938	0.901	0.937	0.95	1	0.94	0.921
H5_Mean	0.942	0.942	0.913	0.926	0.955	0.94	1	0.948
H6_Mean	0.943	0.941	0.899	0.933	0.933	0.921	0.948	1

Table 3: Significance values

	AI Global	AI_Mean	H1_Mean	H2_Mean	H3_Mean	H4_Mean	H5_Mean	H6_Mean
AI Global	-	< .001	< .001	< .001	< .001	< .001	< .001	< .001
AI_Mean	< .001	-	< .001	< .001	< .001	< .001	< .001	< .001
H1_Mean	< .001	< .001	-	< .001	< .001	< .001	< .001	< .001
H2_Mean	< .001	< .001	< .001	-	< .001	< .001	< .001	< .001
H3_Mean	< .001	< .001	< .001	< .001	-	< .001	< .001	< .001
H4_Mean	< .001	< .001	< .001	< .001	< .001	-	< .001	< .001
H5_Mean	< .001	< .001	< .001	< .001	< .001	< .001	-	< .001
H6_Mean	< .001	< .001	< .001	< .001	< .001	< .001	< .001	-

All mean values correlate significantly at the $p < .01$ level.

Krippendorf's alpha for ordinal scaled data is $\alpha = .92$ for the eight average ratings of the 34 participants.

Values > .8 indicate a high inter-rater reliability.

In the following, Krippendorff's alpha is reported separately for each question Q1 - Q6, initially without consideration of global score and thus for seven raters (AI, H1 - H6):

- Q1: $\alpha = .77$
- Q2: $\alpha = .77$
- Q3. $\alpha = .79$
- Q4: $\alpha = .77$
- Q5: $\alpha = .84$
- Q6: $\alpha = .87$

If the global rating is simulated as the eighth rating for the respective questions, the following alpha values result:

- Q1: $\alpha = .79$
- Q2: $\alpha = .76$
- Q3. $\alpha = .79$
- Q4: $\alpha = .79$
- Q5: $\alpha = .84$
- Q6: $\alpha = .86$

Q5 and Q6 show the strongest inter-rater reliability with $\alpha > .8$.

3. Cross-reference tables

Cross-reference tables (or cross-tabulations) are used to analyze the relationship between two or more variables. In this context, they were used to compare the ratings given by AI and human markers for the same set of items. The cross-reference tables presented for AI vs. H1, H2, etc., illustrate the distribution of ratings given by AI compared to each human marker. They help in visualizing how often the AI and human markers agree or disagree on their ratings. Analysis of any discrepancies between AI and human ratings can be used to improve the AI model.

As these tables demonstrate, the AI consistently gives similar ratings to those of human markers, indicating reliability. Furthermore, these tables demonstrate that there are no systematic differences or biases between AI and human ratings. However, it can also be seen that the ratings are not exact matches. This is due to the following factors:

- Expected variability: Some degree of variability between ratings is expected and normal.
- Assessing AI robustness: If the AI ratings are too similar to human ratings in every case, it might indicate 'overfitting'. Some level of independent variation is desirable to ensure that the AI system is robust and not simply mimicking human ratings without its own analytical process.
- Human-AI complementarity: The differences between AI and human ratings can also be seen as complementary. In some cases, AI might detect nuances that human markers miss, or vice versa.

As an example of cross-tabulation, below are the tables for the matrix A1 vs Human markers 1-6. In this example, we focus on the data provided by Q1, which serves as a model.

Table 4: AI_Q1 vs H1_Q1 Cross Reference

AI_Q1	A1	A2	A1+	B1	B1+	B2	C1	Total
0	0	0	0	1	0	0	0	1
A2+	0	3	0	0	0	0	0	3
B1	1	5	0	2	0	0	0	8
B1+	0	3	0	6	0	1	1	11
B2	0	0	0	2	0	3	1	6
C1	0	0	0	0	0	5	0	5
Total	1	11	0	11	0	9	2	34

Comment: AI has rated a Q1 utterance with 0, which was rated B1 by H1. H1 rated 2 utterances with C1, AI rated 7 utterances with C1. H1 and AI agreed on 2 utterances, H1 rated the 5 other utterances that AI rated C1 as B2.

Table 5: AI_Q1 vs H2_Q1 Cross Reference

AI_Q1	A1	A2	A2+	B1	B1+	B2	C1	Total
A1	0	1	0	0	0	0	0	1
A2	0	2	1	0	0	0	0	3
B1	2	2	3	1	0	0	0	8
B1+	0	3	0	4	2	1	0	10
B2	0	1	0	0	0	4	0	5
C1	0	0	0	0	0	3	4	7
Total	2	9	4	5	2	8	4	34

Table 6: AI_Q1 vs H3_Q1 Cross Reference

AI_Q1	A1	A2	A2+	B1	B1+	B2	C1	Total
A1	0	1	0	0	0	0	0	1
A2	1	2	0	0	0	0	0	3
B1	1	5	0	2	0	0	0	8
B1+	0	3	0	5	1	1	0	10
B2	0	1	0	1	0	3	0	5
C1	0	0	0	0	0	6	1	7
Total	2	12	0	8	1	10	1	34

Table 7: AI_Q1 vs H4_Q1 Cross Reference

AI_Q1	A2	A2+	B1	B1+	B2	B2+	C1	Total
A1	1	0	0	0	0	0	0	1
A2	2	1	0	0	0	0	0	3
B1	4	3	1	0	0	0	0	8
B1+	2	1	4	2	1	0	0	10
B2	0	1	0	1	3	0	0	5
C1	0	0	0	0	4	2	1	7
Total	9	6	5	3	8	2	1	34

Table 8: AI_Q1 vs H5_Q1 Cross Reference

AI_Q1	A2	A2+	B1	B1+	B2	B2+	C1	Total
A1	1	0	0	0	0	0	0	1
A2	2	1	0	0	0	0	0	3
B1	4	3	1	0	0	0	0	8
B1+	2	1	4	2	1	0	0	10
B2	0	1	0	1	3	0	0	5
C1	0	0	0	0	4	2	1	7
Total	9	6	5	3	8	2	1	34

Table 9: AI_Q1 vs H6_Q1 Cross Reference

AI_Q1	A1+	A2	B1	B1+	B2	B2+	C1	Total
A1	0	1	0	0	0	0	0	1
A2	1	2	0	0	0	0	0	3
B1	2	3	3	0	0	0	0	8
B1+	0	3	4	2	1	0	0	10
B2	0	1	1	0	3	0	0	5
C1	0	0	0	0	4	2	1	7
Total	3	10	8	2	8	2	1	34

The cross-reference tables provide detailed comparisons between AI and human ratings for Question 1 (Q1). These tables reveal several key insights:

- Discrepancy in C1 ratings:
  - The AI rated seven responses at the C1 level for Q1.
  - The maximum number of C1 ratings for Q1 by a human rater was four.

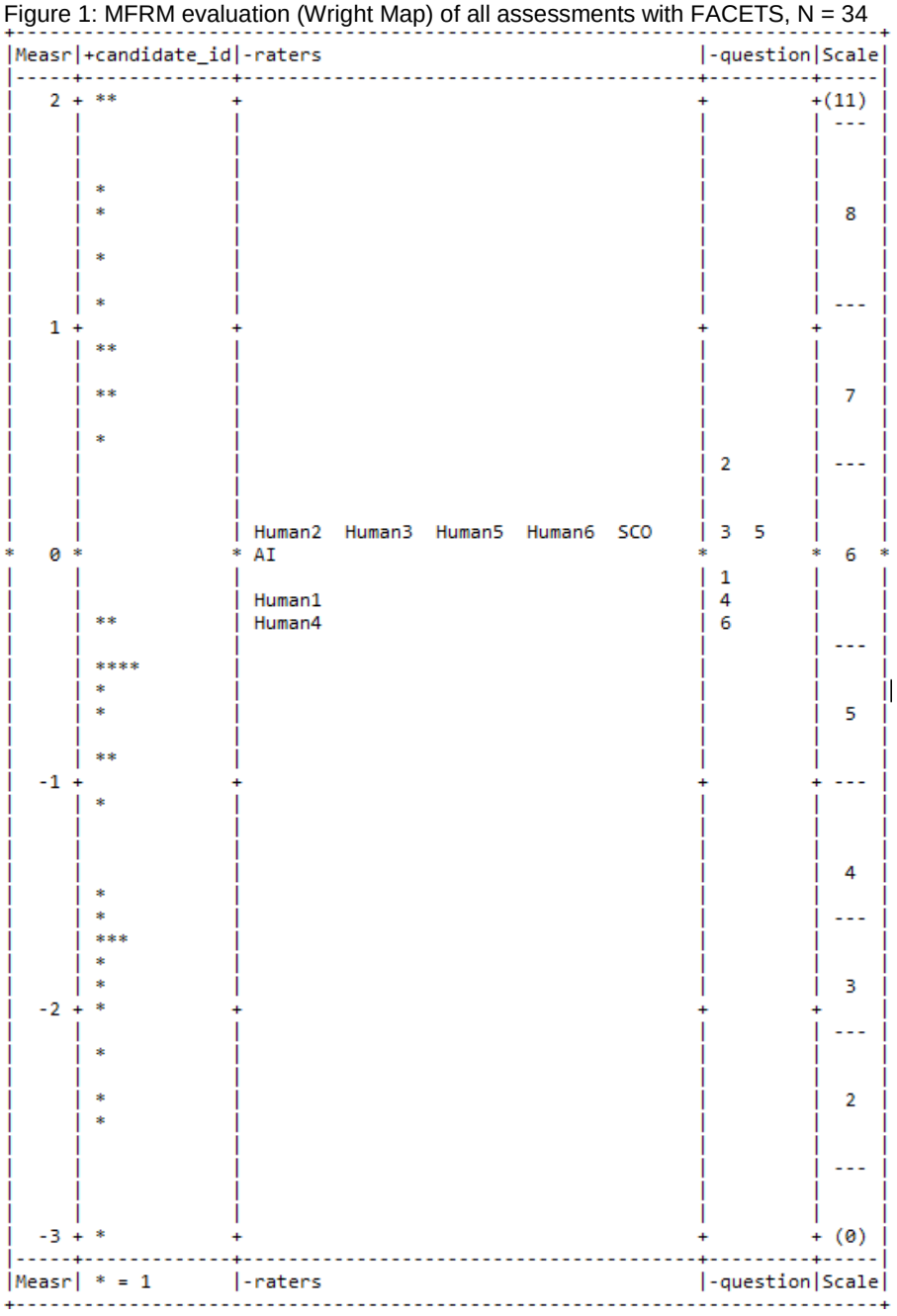
- Consistency in high ratings:
  - Whenever human raters assigned a C1 rating, it always corresponded to the cases where the AI also rated the responses as C1.
  - This indicates a consistent agreement between AI and human raters when identifying the highest level of proficiency (C1).

- Minimum ratings by human raters:
  - All responses that the AI rated as C1 were rated at least B2 or higher by the human raters.
  - This shows a baseline level of higher proficiency judgement, as human raters did not assign ratings below B2 for responses that the AI considered C1.

4. Further analysis

In order to better interpret the available data, a multi-facet Rasch analysis (MFRM) was carried out with Facets 4.1.7 with the facets Candidate (1-34), Rater (AI, H1- H6) and Question (Q1-Q6). The global assessment of AI was treated as if it were also the respective assessment for Q1-Q6. The following picture emerged:

Figure 1: MFRM evaluation (Wright Map) of all assessments with FACETS, N = 34



More competent candidates are displayed higher up in the *candidate_id* area. The raters are positioned on the scale according to their rigour. The rater facet is negatively oriented, as indicated by the "-" in front of the facet name (*-raters*), i.e. more stringent raters appear higher up in the column (Human2; 3; 5; 6; SCO; AI) and less stringent (or more lenient) raters lower down (Human 1 and 4). This result also reflects the values from the descriptive statistics, in which H1 and H4 gave the mildest rating with an average rating of approximately 3.2.

Tasks 1 - 6 (Question section) are clustered closely together with similar difficulty parameters, Q2 is more demanding than Q6.

The characteristics of the raters are shown in detail in Fig. 2. The ‘strictest’ rater with Measure = .12 is SCO (note the restrictions due to the transfer of the global rating to Q1-Q6), the mildest rater is Human4. The rater severity (Measure) should be in the interval of -1 / 1 logits; all raters were in this range.

Infit MnSq and Outfit MnSq have an ideal value of 1, values in an interval of .5 / 1.5 are permissible. With the exception of rater H4, all fulfil this criterion. Values above 1 (AI, H1, H4) indicate that there is more variance in the ratings than expected. Values below 1 (SCO, H3, H6, H2, H5) indicate ‘overfit’, i.e. less variance than expected. The standard error of the Model S.E. assessment measure is identical for all raters.

Figure 2: Characteristics of the raters (MFRM)

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	N raters
1002	204	4.91	5.09	.12	.06	.59	-4.9	.57	-5.2	1.32	.94	.86	29.9	28.4	8 SCO
1005	204	4.93	5.10	.11	.06	.65	-4.0	.66	-3.9	1.29	.89	.86	43.5	28.5	4 Human3
1006	204	4.93	5.11	.11	.06	.72	-3.2	.72	-3.1	1.19	.88	.86	45.8	28.5	7 Human6
1010	204	4.95	5.13	.10	.06	.79	-2.2	.80	-2.1	1.13	.89	.86	44.3	28.5	3 Human2
1019	204	5.00	5.17	.07	.06	.62	-4.4	.63	-4.4	1.06	.91	.86	43.8	28.6	6 Human5
1054	204	5.17	5.33	-.04	.06	1.77	6.5	1.77	6.4	.14	.81	.86	26.2	28.7	1 AI
1105	204	5.42	5.56	-.20	.06	1.24	2.2	1.23	2.2	1.20	.80	.86	42.7	28.3	2 Human1
1127	204	5.52	5.66	-.28	.06	1.54	4.7	1.50	4.4	.77	.78	.86	38.1	27.9	5 Human4
1041.0	204.0	5.10	5.27	.00	.06	.99	-.7	.98	-.7		.86				Mean (Count: 8)
46.3	.0	.23	.21	.15	.00	.43	4.2	.43	4.2		.05				S.D. (Population)
49.5	.0	.24	.22	.16	.00	.46	4.5	.46	4.5		.06				S.D. (Sample)

Model, Populn: RMSE .06 Adj (True) S.D. .14 Separation 2.43 Strata 3.57 Reliability (not inter-rater) .85
Model, Sample: RMSE .06 Adj (True) S.D. .15 Separation 2.62 Strata 3.83 Reliability (not inter-rater) .87
Model, Fixed (all same) chi-squared: 54.3 d.f.: 7 significance (probability): ,00
Model, Random (normal) chi-squared: 6.2 d.f.: 6 significance (probability): ,40
Inter-Rater agreement opportunities: 5712 Exact agreements: 2244 = 39.3% Expected: 1623.7 = 28.4%

5. Argument for using artificial intelligence (AI) for testing English

Overall, the automatic ratings from AI/SCO are very similar to the assessment of four of the six human raters, with two raters deviating from this and giving milder ratings. Nevertheless, the human raters appear to award very good ratings (C1) less frequently, for example (see descriptive statistics).

Chief advantages of AI

1. Significant correlation with human ratings: The ratings provided by AI showed significant correlations with those of the human raters, indicating that the AI scoring aligns well with human judgement while also offering consistent automated scoring.
2. High inter-rater reliability: The high Krippendorff's Alpha value (α = .92) for the average ratings across all raters, including the AI, demonstrates that the AI evaluations are reliable and consistent with human raters. Reliability values greater than .8 are considered high, reinforcing the trustworthiness of machine marking.
3. Efficiency, scalability and consistency: Unlike human raters, the AI can evaluate numerous responses quickly and consistently without fatigue, bias, or variation in judgement over time. This makes it an ideal tool for large-scale testing environments.
4. Detailed analyses: The AI marking provides granular insights into each question and candidate's performance, similar to the multi-facet Rasch analysis used in the report. This can help in diagnosing specific strengths and weaknesses in a learner's language proficiency, which is beneficial for targeted learning and teaching interventions.

6. Reliability and validity analysis

Reliability:

- High inter-rater reliability: The Krippendorff's Alpha values for individual questions, with and without the global rating (SCO), were consistently high (ranging from .76 to .87), indicating robust reliability in AI ratings across different questions.
- Correlations: Significant correlations between machine and human ratings ($p < .01$) confirm that the AI evaluations are reliable and comparable to human judgement.
- Rater characteristics analysis: The multi-facet Rasch analysis positioned the AI within the expected range of rater stringency, confirming that it functions within acceptable reliability parameters.

Validity:

- Alignment with human raters: The high correlation and agreement with human raters validate that the AI accurately reflects human evaluative standards, ensuring that its scores are meaningful and representative of a candidate's actual language proficiency.
- Descriptive statistics: The consistency in the AI maximum ratings compared to human ratings validates its capability to recognize high proficiency accurately, adding to its criterion validity.

7. Conclusion

Using artificial intelligence machine marking for testing English presents several advantages, including high reliability, consistency with human raters, detailed performance insights and the ability to handle large-scale assessments efficiently. The data from the report underscores the reliability and validity of the AI, making it a valuable tool for accurate, scalable and efficient language proficiency testing.

Bibliography

Eckes, T. (2020). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In Aryadoust, V. & Raquel, M. (Eds.) *Quantitative data analysis for language assessment: Volume II: Advanced methods* (pp.153-176). Abingdon: Routledge.

McKay, T.H., & Plonsky, L. (2021). Reliability analyses: Estimating error. In Winke, P & Brunfaut, T. (Eds.) *The Routledge handbook of second language acquisition and language testing* (pp.428-482). New York: Routledge.

Zapf, A., Castell, S., Morawietz, L. & Karch, A. (2016). Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology* 16, 93.

DYNAMIC SPEAKING TEST

Developing the AI algorithm

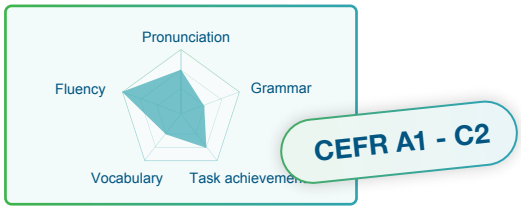
Test development and the use of natural language processing

The Dynamic Speaking Test is made fully automatic through extensive use of natural processing techniques. The accuracy of the scoring was achieved through painstaking data gathering, manual rating, machine learning modeling and user testing over an 18-month period.

Five thousand user audio samples were transcribed using industry-leading speech recognition and then each was manually graded by three qualified examiners on a variety of parameters including pronunciation, vocabulary, grammar, coherence and relevance. If two raters had more than one point difference in rating then the third rater was asked to arbitrate.

Once the data was graded, the responses and the validated scores were used to train the AI system. Algorithms evaluated thousands of English syntactic and semantic rules to determine which rules mattered the most for assessing pronunciation, fluency, vocabulary, grammar, cohesion and relevance. Deep learning machine learning models were then built on the filtered set of rules to accurately predict scores for any arbitrary audio sample. Note that test re-test reliability of our models is found to be 0.82.

The original validation was based on IELTS scores and the system is capable of rating students within 1 point of qualified IELTS examiners. In the development of the Dynamic Speaking Test, the scoring has now been calibrated and validated against CEFR levels.



Here are the key statistics observed with regard to human inter-rater agreement based on the context of IELTS scores:

% of items on which raters gave exactly the same grade = 21.8%

% of items on which raters were within 0.5 IELTS points = 72.7%

% of items on which raters were within 1 IELTS points = 98.2%

Cohen's kappa = 0.794

Pearson's correlation between raters = 0.883

RMSE = 0.674