Hallo Language Test (HLT)

Human–AI Score Correspondence

A Preliminary Validation Study of AI Scoring for the Hallo Language Test

Abstract

This white paper examines the correspondence between speaking proficiency scores assigned by Hallo's automated scoring system and independent, trained third-party human evaluators. In 234 paired observations, the correlation between AI scores and the mean of available human ratings was strong (Pearson r = 0.89, 95% confidence interval [0.86, 0.91]). As a human-only reference, the mean correlation across the human–human evaluator pairings was reported as r = 0.85 (N = 210). Both findings indicate strong positive linear relationships within the analyzed samples.

These findings provide preliminary evidence supporting the use of automated scoring in the Hallo Language Test, while further research is needed to evaluate additional aspects of scoring accuracy, reliability, and overall test validity.

1. Introduction

The Hallo Language Test is a computer-adaptive test (CAT) intended to evaluate proficiency across reading, listening, speaking, and writing. Its scoring framework reports numerical proficiency scores alongside interpretations associated with the Common European Framework of Reference for Languages (CEFR). As described in the Hallo Language Test Technical Manual (Hallo, 2026a), the assessment uses automated scoring and reports diagnostic information across language skills.

For automated language assessment, a central validation question is whether higher scores awarded by human evaluators tend to coincide with higher scores from the automated system. Pearson's product-moment correlation coefficient provides an interpretable measure of this linear relationship. The present white paper reports this analysis as a focused piece of preliminary scoring evidence; it does not attempt to establish the complete validity argument for all intended test uses.

2. Study design

2.1. Ratings and analytic sample

The underlying validation study assembled ratings from independent, trained third-party human evaluators who were not affiliated with Hallo, alongside scores generated by Hallo's AI scoring system. Human evaluators supplied numerical proficiency scores on the same 0-10 scale as the automated assessment. The present analysis compares each AI numerical score with the mean of human numerical scores available for that response. Where not all three human ratings were present, the available human ratings contributed to the mean.

This analysis focuses on speaking responses and evaluates correspondence between human-assigned proficiency scores and Hallo's automated speaking scores. The analysis includes 234 responses with both a valid AI score and a calculable human mean, after applying the study's response-quality criteria. Cases requiring review due to issues such as unavailable or compromised audio and technical problems were not included. This is an analytic sample, and its characteristics should not be assumed to represent all Hallo test takers.

The sample responses were collected in at-home testing environments, where some recordings contained background noise or variations in audio clarity. Responses were included as long as the speech remained

comprehensible, and background noise alone was not grounds for exclusion. Although the independent effects of recording conditions on scoring outcomes were not evaluated in this study, the inclusion of these responses reflects the variability encountered in real-world testing environments.

2.2. Statistical analysis

Pearson's r was used to quantify the strength and direction of linear association between the AI numerical score and the corresponding human mean. The human-AI result is presented alongside a human-human benchmark reported as the mean correlation across the human evaluator pairings. To establish a human inter-rater benchmark, Pearson correlation coefficients were calculated for each of the possible pairwise comparisons among the human evaluators. The analysis was restricted to 210 assessment responses independently scored by all evaluators, ensuring a consistent sample across comparisons. The coefficients were averaged arithmetically, yielding a mean human–human correlation of r = 0.85.

This benchmark provides a reference for interpreting the correlation between the automated scoring system and human evaluations (r = 0.89). The human–AI correlation was slightly higher than the mean human–human correlation, suggesting that the automated system demonstrated a comparable level of linear score correspondence within the evaluated samples.

The correlation analysis is an association analysis: r = 1 denotes a perfect positive linear relationship, whereas r = 0 indicates no linear relationship. Importantly, a strong correlation can occur when one scoring method systematically produces higher or lower scores than the other. Accordingly, correlation alone is not a direct measure of exact agreement.

Hallo also uses Many Facet Rasch Measurement (MFRM) analyses to monitor scoring consistency and evaluator severity. These analyses form part of Hallo's broader validation and quality assurance framework.

3. Results

3.1. Study results

The automated scores and human-mean scores showed a strong positive linear relationship: Pearson r = 0.89 (95% CI [0.86, 0.91], p <0.001), based on 234 paired observations. The human-human benchmark was also strong (mean r = 0.85; reported N = 210). These coefficients are descriptive comparisons, not evidence that the difference of 0.04 is statistically significant. A combined p-value for the mean human–human correlation was not calculated; however, each individual evaluator pairing yielded p < 0.001.

Table 1. Pearson correlation for the quality-screened human-AI comparison and the reported human-human benchmark. The human-human estimate is an average across multiple evaluator pairings.

Comparison	N	Pearson r	p-value
Human mean vs. AI	234	0.89	<0.001
Human–human mean	210	0.85	N/A*

Note: The human–human correlation represents the arithmetic mean of the pairwise correlations among human evaluators. Each individual pairing had p < 0.001; a p-value for the mean correlation was not calculated.

Both reported correlations indicate that higher scores from one scoring source generally correspond to higher scores from the other. The human-AI coefficient (0.89) is numerically above the human-human mean (0.85), but this difference should not be interpreted as a demonstrated superiority of automated scoring.

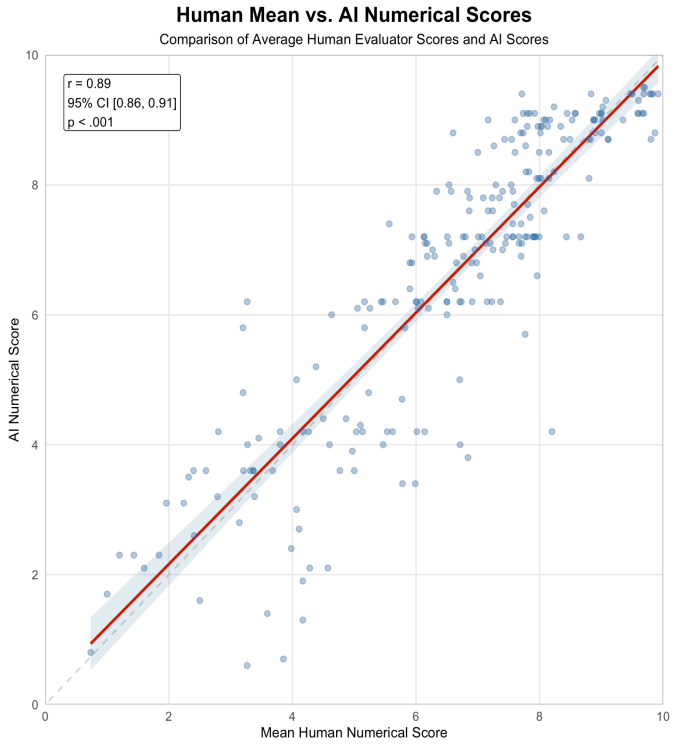


Figure 1. Relationship between mean human evaluator scores and Hallo AI numerical scores (N = 234). The dotted gray line represents perfect agreement, the blue line represents the fitted linear regression, and the shaded region represents the 95% confidence interval for the fitted regression mean.

3.2. External Benchmark: Duolingo English Test

To contextualize the observed human–AI score correspondence, this study provides a descriptive comparison with published findings from the Duolingo English Test (DET), an established English language proficiency assessment used for university admissions. The DET Technical Manual reports human–machine Pearson correlations of 0.84 for speaking and 0.85 for writing, with corresponding human–human correlations of 0.85 and 0.87 (Naismith et al., 2025, Table 4, p. 25).

Table 2. Descriptive comparison of published human–AI and human–human Pearson correlations across independently conducted studies.

Assessment / Skill	Human–AI (r)	Human–Human (r)
Hallo (human–AI validation sample)	0.89	0.85
DET Speaking	0.84	0.85
DET Writing	0.85	0.87

Note: Hallo's analysis included 234 human–AI paired responses and 210 responses for the human–human benchmark. DET evaluated 1,922 speaking and 2,460 writing test sessions. Results are from independently conducted studies with different samples and methodologies and are not directly comparable.

Hallo demonstrated a strong human–AI score correlation (r = 0.89), compared with published DET correlations of 0.84 for speaking and 0.85 for writing. These findings provide a useful external reference for

interpreting Hallo's results. However, differences in assessment design, samples, and scoring methodologies mean that the correlations should not be interpreted as evidence of superior scoring accuracy.

4. Discussion

The observed correlations provide encouraging evidence that Hallo's automated numerical scores track human judgments of relative language proficiency. The human–human benchmark offers a useful context for interpreting that association: variation is also present between independent human evaluators. These findings support describing the automated results as strongly associated with human scoring patterns, not as identical to any specific human scores.

These results are not, on their own, evidence that AI and human scores are interchangeable for individual examinees. Pearson correlation is unaffected by some systematic score shifts and does not establish how frequently the two scoring methods assign identical values or the same CEFR category. The analysis also does not establish that either human or AI scores recover a test taker's true proficiency without error. Speaking proficiency involves multiple dimensions of performance, making consistent application of standardized scoring rubrics essential.

The internal validation program also considers other statistical approaches, including categorical agreement, absolute numerical agreement, and evaluator-severity modeling. Those analyses provide complementary perspectives, but their results are not reported in this focused white paper. This narrower presentation should therefore be read as evidence of score correspondence rather than a comprehensive account of all measurement properties.

Pearson correlation was selected as the primary measure of numerical score correspondence in this study because it captures the consistency of scoring patterns across the continuous 0-10 scale. While CEFR proficiency bands (A1-C2) provide meaningful categorical interpretations, numerical scores allow for a more granular evaluation of the relationship between human and automated scoring. Correlation measures the strength of this relationship rather than exact scoring agreement. Hallo's CEFR banding is informed by ongoing validation efforts, including concordance research involving established standardized assessments such as TOEFL and IELTS.

5. Limitations and Further Validation

Pearson's correlation measures the strength of a linear relationship between scores but does not directly measure agreement or scoring accuracy. Two raters may demonstrate a strong correlation while exhibiting systematic differences in scoring severity. Additionally, correlation coefficients can be influenced by the distribution and range of observed scores, unusually discrepant observations, and nonlinear relationships. Consequently, these results should be interpreted alongside the complementary agreement and reliability analyses presented in the technical manual.

The human comparison score was calculated as the mean of available human ratings, meaning that the number of contributing evaluators may differ across responses. This variation may influence the stability of the human composite and its correlation with AI scores. Additionally, the human–human mean correlation represents a summary of pairwise relationships rather than a single uniformly rated comparison. As such, direct statistical comparisons between the human–human and human-AI correlations require consideration of their different calculation methods and associated uncertainty.

These findings should be interpreted within the broader validity framework established in the Hallo Language Test Technical Manual, which addresses the assessment's design, sampling, scoring procedures, and additional reliability and validity evidence. Pearson's correlation provides evidence of consistency in relative scoring patterns but should not be interpreted independently as evidence of scoring equivalence or overall assessment validity.

Further research on concordance with TOEFL, IELTS, and DET, as well as other validity and reliability research will follow, and may be added to the technical manual as it becomes available.

6. Conclusion

In this quality-screened validation sample, Hallo's automated numerical proficiency scores were strongly correlated with the mean of available human ratings (r = 0.89, 95% CI [0.86, 0.91]; N = 234). The reported average human–human correlation was similarly strong (mean r = 0.85; reported N = 210). Together, these descriptive findings support the conclusion that automated scores follow human-assigned score patterns in the analyzed sample.

References

- Hallo. (2026a, May). Hallo Language Test: Technical Manual. Internal technical documentation.

- Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment—Companion volume. Council of Europe.

- Naismith, B., Cardwell, R., LaFlair, G. T., Nydick, S., & Kostromitina, M. (2025). Duolingo English Test: Technical Manual. Duolingo Research Report. https://englishtest.duolingo.com/research