International Journal of Pedagogical Language, Literature, and Cultural Studies (i-Plural)
Volume 03 No 03
http://dx.doi.org/10.63011/ip.v3i3.88

"Jeder hat seinen eigenen Geschmack": Comparative Analysis of AI and Human Raters in German Proverb Oral Performance Assessment

Sudarmaji[1]; Iman Santoso[2*]; Retna Endah Sri Mulyati[3]; Aditya Rikfanto[4]

[1] Universitas Negeri Yogyakarta, Indonesia, sudarmaji@uny.ac.id
[2] Universitas Negeri Yogyakarta, Indonesia, iman_santoso@uny.ac.id
[3] Universitas Negeri Yogyakarta, Indonesia, retna_endah@uny.ac.id
[4] Universitas Negeri Yogyakarta, Indonesia, adityarikfanto@uny.ac.id

*Corresponding author:
E-mail:
iman_santoso@uny.ac.id

Abstract
This study aims to analyze scoring variations and test the statistical significance of differences between AI-based evaluation (Large Language Model) and human raters in assessing the oral performance of German proverbs. Using a quantitative, within-subjects design, the study involved 30 students who were evaluated on four parameters: grammar, pronunciation, fluency, and the use of proverbs. The results of a paired-samples t-test reveal highly significant differences ($p < 0.001$) across all assessment aspects. The AI consistently demonstrates a leniency bias, assigning higher absolute scores than human raters. The largest discrepancy is in the use of proverbs, with a mean difference of 3.73 points. These findings indicate that AI tends to operate at the level of structured linguistic surface features, yet remains limited in capturing cultural nuances, implicit meanings, and pragmatic contextual appropriateness. These are areas that constitute the core strength of human cognitive sensitivity. This study recommends implementing a hybrid app using language to evaluate, in which AI assesses structural-mechanistic aspects, while interpretive sociolinguistic dimensions remain under human raters' control.

Keywords: *Artificial Intelligence, language assessment, human rater, oral performance, German proverbs*

Introduction

The application of artificial intelligence in the global educational landscape has triggered a fundamental transformation in foreign language assessment practices. The integration of technologies based on Natural Language Processing (NLP) and Large Language Models (LLMs) enables systems to serve as evaluative entities that automatically, systematically, and at scale assess linguistic performance. In the context of language assessment, AI operates as an automated rater that offers high consistency, particularly in evaluating structural aspects such as grammar and syntax.

However, oral performance is a complex and multidimensional construct, encompassing linguistic accuracy, phonological aspects, fluency, and pragmatic competence. In German, this complexity is further intensified by the presence of idiomatic and paremiological expressions that carry implicit and context-dependent meanings. Proverbs, for instance,

How to cite:
Sudarmaji, Santoso, I., Mulyati, R. E. S., & Rikfanto, A. (2026). *"Jeder hat seinen eigenen Geschmack":* Comparative Analysis of AI and Human Raters in German Proverb Oral Performance Assessment. *International Journal of Pedagogical Language, Literature, and Cultural Studies*. Nexus Publishing. ISSN: 3047-2202. Pages 115-122. doi: 10.63011/ip.v3i3.88

function not only as linguistic elements but also as pedagogical stimuli that can foster creative language production by developing stories or dialogues (Aliyeva, 2025). Such activities require the simultaneous integration of language proficiency and cultural understanding.

Despite the advanced capabilities of contemporary AI models such as Gemini in formal language analysis, their ability to capture cultural nuances, non-literal meanings, and pragmatic context remains limited. This limitation becomes particularly significant when assessment involves the interpretation of implicit meaning, which has traditionally been carried out more effectively by human raters through cognitive sensitivity and contextual experience. Therefore, there is a need to examine the extent to which AI can approximate the quality of human evaluation in the domain of complex oral performance.

On the other hand, previous research has generally focused on written-language performance or on structured linguistic aspects, thereby failing to fully capture the authentic characteristics of oral performance (Hartwell & Aull, 2023). When speaking is analyzed, the approaches often rely on transcription-based data (Margetson et al., 2023). This approach overlooks prosodic dimensions such as intonation, stress, and speech rhythm, which are essential components in the assessment of pronunciation and fluency. Consequently, the construct validity of measures of speaking ability is limited.

In a previous study, Li et al. (2026) mapped general differences between AI and human scoring. However, there remains a scarcity of research specifically examining how AI evaluates the use of proverbs as non-compositional units in the oral performance of German language (DaF) students.

Based on this research gap, the present study aims to identify score differences between AI and human raters, analyze variations in assessment across aspects of oral performance, test the statistical significance of the observed differences, and evaluate the level of alignment between the two evaluators in the context of German oral performance.

Unlike previous studies that mainly compared AI and human scoring in general language assessment, this study specifically focuses on German proverb oral performance, which combines linguistic accuracy with pragmatic and cultural competence. Therefore, the study contributes empirical evidence regarding the strengths and limitations of AI in evaluating culturally contextualized oral language performance.

Automated Scoring by Artificial Intelligence

Automated scoring in language learning meets the need for efficient, scalable evaluation (Ercikan & McCaffrey, 2022). AI-based systems use machine learning to recognize linguistic patterns and generate consistent scores. The chief advantages are scoring stability and the ability to handle large data volumes without cognitive fatigue.

However, the complexity of assessing oral performance cannot be reduced to measurable linguistic patterns alone. While AI systems excel at evaluating formal linguistic features, they are limited in capturing important aspects such as pragmatic nuances, contextual meaning, sarcasm, and implicit intentions that depend heavily on cultural and situational context (Mahowald et al., 2023). These systems are often termed "stochastic parrots," reflecting their tendency to generate predictions based on statistical patterns rather than truly understanding the underlying intentions or culturally embedded meanings (Bender et al., 2021).

The Nature of Proverbs and the Pragmatic Dimension

Proverbs, or Sprichwörter in German, represent cultural wisdom using metaphor and specific communicative functions. Mieder (2004) notes that proverbs are concise, widely recognized statements that convey general truths in fixed forms. In the context of oral performance, the use of proverbs introduces several layers of complexity:

1. Proverbs are understood idiomatically; their overall meaning cannot be derived from the literal meanings of individual words (Cram, 2015).
2. Proverbs are highly context-dependent, serving as strategies to reinforce arguments, give advice, or deliver subtle criticism (Charteris-Black, 1995).
3. Choosing proverbs appropriately shows high sociolinguistic competence, signaling when a traditional expression fits a situation. Yanga (2017) explains that proverbs resolve conflicts or provide advice without appearing personal or didactic. Proverbs convey collective cultural values, letting speakers invoke society's authority in arguments.

Assessment of Oral Competence

The assessment of oral competence aims to measure an individual's ability to use language effectively in spoken communication. According to the Common European Framework of Reference for Languages (CEFR), this competence encompasses several sub-dimensions:

1. Linguistic Dimension: This dimension includes grammatical accuracy (grammar), lexical range, and pronunciation accuracy.
2. Pragmatic Dimension: Fluency and the ability to organize messages coherently are central to this dimension.
3. Sociolinguistic Dimension: This dimension relates to an individual's ability to adapt language use in accordance with social and cultural norms.

AI-based oral performance assessments often miss prosodic features like intonation, stress, and rhythm when relying solely on transcribed text, evaluating only what is said rather than how it is expressed.

Comparison of Assessment Capacity: AI vs. Human Rater

In a speaking assessment, human raters have a significant advantage in interpretation. They are capable of interpreting implicit meaning and evaluating whether a proverb is used "naturally" within a conversation. Human evaluators rely on linguistic intuition to assess non-linear features that are difficult to quantify algorithmically.

In contrast, AI offers a high degree of objectivity in parameters that are acoustically and structurally measurable. However, German proverbs, which often exhibit archaic structures or complex metaphors, may lead to misinterpretation by AI systems, particularly when training data do not sufficiently capture the breadth of pragmatic variation. Therefore, comparing the consistency of AI with the interpretative depth of human raters becomes crucial in determining the validity of oral performance assessment.

Method

This study employs a quantitative, within-subject, comparative design to enable direct comparisons of two assessment sources within the same subjects. The participants consisted of 30 students who were asked to perform an oral task in the form of a monologue on the topic of food (*Lieblingsessen*).

The assessment of German oral performance was conducted by two evaluators: an AI system based on a Large Language Model (LLM), represented by Gemini©, and a doctoral-qualified German language instructor. Both evaluators applied the same assessment rubric to ensure consistency in the evaluation criteria, allowing any differences in scores to be attributed to the characteristics of the evaluators rather than to the assessment instrument. The rubric consisted of four main parameters: grammar, pronunciation, fluency, and the use of German proverbs. To ensure consistency across all evaluations, the AI system was provided with a standardized prompt containing the assessment rubric and detailed scoring criteria. In the grammar component, the assessment focused on the accuracy of sentence structure and the syntactic complexity demonstrated by the students. In the pronunciation component, evaluation emphasized clarity of articulation, intonation, and the accuracy of stress, all of which influence overall comprehensibility. In the fluency component, the assessment assessed students' ability to convey ideas smoothly, without unnatural pauses or excessive repetition. In the proverbs (idiomatic expression) component, the evaluation targeted students' sociolinguistic ability to appropriately integrate idiomatic or proverbial expressions to enrich communicative nuance. Each student's performance was transcribed and subsequently evaluated by both raters using the same scoring scale. Each aspect contributed 25% to the total score of 100.

Both evaluative entities employed identical scoring scales to ensure data comparability. Data analysis was conducted through several systematic stages. First, descriptive statistics were used to identify general patterns in score distribution. Second, an inferential statistical analysis, a paired-samples t-test, was applied to test the significance of score differences (Rietveld & van Hout, 2017). Finally, an analysis of the direction of the relationship between scores was conducted to determine the level of convergence or divergence between AI-based and human assessments.

Results

General Description of the Data and Significance Testing of Differences

The data analysis demonstrates a consistent pattern: AI assigns higher scores than human raters across all subjects and assessment aspects. This observation aligns with Lundgren (2024), who reports that AI systematically provides higher scores in areas such as fluency and coherence. AI tends to give the "benefit of the doubt" to complex sentence structures containing minor errors, while human raters apply more stringent evaluative standards.

AI scores for the grammar component range from 22 to 24, whereas human rater scores range from 18 to 20. Similar patterns are observed in pronunciation, fluency, and proverb usage, with mean differences of 1 to 3 points. The consistency of these discrepancies suggests a systematic tendency in AI-based assessment that is unlikely to result from random variation.

A paired-sample t-test indicates that the differences between AI and human rater scores are statistically significant across all assessment aspects. The stable score differences across

subjects produce high t-values, resulting in the rejection of the null hypothesis of no difference between the two evaluators.

This statistical significance suggests that the differences between the two raters are structural and consistent, rather than attributable to individual fluctuations or measurement error.

Table 1. Results of Paired Sample t-test Analysis

Assessment Aspect	Mean (AI)	Mean (Human)	Mean Difference	p-value	Judgement
Grammar	22.67	19.67	+3.00	< 0,001	Significant
Pronunciation	21.40	18.40	+3.00	< 0,001	Significant
Fluency	21.80	20.80	+1.00	< 0,001	Significant
Proverbs	21.10	17.37	+3.73	< 0,001	Significant

The results of the statistical tests show that in all categories (Grammar, Pronunciation, Fluency, and Proverbs), the p-values are far below the significance threshold of 0.05 ($p < 0.001$). This indicates that the differences between scores assigned by AI and human raters are not due to chance, but rather reflect statistically highly significant differences in evaluation.

From a statistical perspective, the AI demonstrates a tendency toward over-scoring or a leniency bias compared to human raters. Notably, in the Grammar and Pronunciation aspects, the AI consistently assigns scores that are exactly 3 points higher than those of human raters for every student. Similarly, in Fluency, the AI consistently provides scores that are 1 point higher. This suggests that the AI operates with an internal scoring standard that is linearly shifted relative to human evaluators.

The Proverbs (idiomatic expression) category shows the highest mean difference, namely 3.73 points. Unlike other aspects where the differences are fixed, this component exhibits variability ranging from 3 to 4 points. Statistically, this reinforces the argument that AI applies a different evaluative perspective or "preference" when assessing metaphorical and idiomatic language use compared to human raters.

Although there is a degree of correlation between AI and human raters in ranking student performance, the absolute scores generated by AI cannot be equated with those of human raters. AI tends to evaluate oral performance using comparatively more lenient criteria.

Discussion

Correlation and Consistency of Assessment

The correlation analysis indicates a positive relationship between AI and human rater scores. Students who receive high scores from AI also tend to receive high scores from human raters, and vice versa. This suggests that both evaluators demonstrate alignment in identifying the relative performance levels of students. Statistically, the correlation coefficient is 0.688, indicating a significant relationship between AI scores and human rater scores (see Appendix C). However, this alignment does not imply absolute equivalence. AI consistently assigns higher scores, meaning that although the evaluative patterns are similar, differences exist in the absolute scoring outcomes.

Analysis of Systematic Bias

The consistent differences between AI and human raters indicate the presence of systematic bias. AI demonstrates a tendency toward more permissive scoring, whereas human raters tend to be stricter in evaluating student performance. No cases were found in which human rater scores exceeded AI scores, further reinforcing the indication of a unidirectional bias.

This difference is related to the fundamental characteristics of both evaluation systems. AI operates based on the detection of explicit linguistic patterns, whereas human raters integrate more complex implicit judgments. Human evaluators possess a "mental model" of good language performance, developed through extensive experience across different proficiency levels, allowing them to make rapid judgments without always explicitly referring to rubrics.

Aspect-by-Aspect Analysis

In the grammar aspect, AI shows consistent superiority due to its ability to recognize rule-based linguistic structures. This finding aligns with Ranalli, Link & Chukharev-Hudilainen (2017), who found that AI demonstrates consistent advantages in detecting rule-based errors such as subject–verb agreement and article usage, compared to human raters who may be less consistent in identifying similar errors in longer texts. The significant differences observed reflect the algorithmic efficiency in processing formal linguistic structures.

In the pronunciation aspect, the differences remain statistically significant; however, their validity must be interpreted in the context of data limitations. Since the assessment is based on transcription, the phonological dimension is not fully represented, thus limiting the interpretation to textual representation only.

In the fluency aspect, the differences are relatively smaller, although still significant. This suggests that AI is approaching human evaluation in identifying fluency based on textual structure, but has not yet fully captured the dynamics of actual speech production.

The most pronounced differences appear in the use of proverbs. Human raters assign lower and more variable scores, reflecting the complexity of interpreting idiomatic meaning. AI tends to fail in capturing contextual appropriateness, resulting in higher scores without fully considering pragmatic nuance. In contrast, human raters evaluate proverbs as lexical units that carry cultural and expressive value, as illustrated in expressions such as: *"Die beste Zutat in der Küche ist Liebe!"*

The analysis shows that the relationship between AI and human raters is convergent in pattern but divergent in absolute values. AI is able to replicate the general tendencies of human assessment in determining learner proficiency rankings, but not in establishing equivalent evaluation standards.

This difference reflects a duality between mechanistic and interpretative approaches in language assessment. AI operates at a structured surface linguistic level, whereas human raters work at a deeper, context-dependent level of meaning.

Conclusion

This study concludes that AI and human raters evaluate German oral performance differently, particularly when assessment involves culturally embedded language such as proverbs. Although AI demonstrates consistent scoring patterns and efficiently evaluates

structural linguistic features, it remains less capable of interpreting pragmatic meaning and sociocultural appropriateness. These findings contribute to the growing body of research on AI-assisted language assessment by demonstrating that AI is better positioned as a complementary assessment tool rather than a replacement for human evaluators in complex oral performance assessment.

The positive correlation between AI and human ratings indicates that AI can reliably identify general proficiency patterns, despite systematic differences in absolute scores. Therefore, AI has considerable potential to support formative assessment and reduce teachers' workload, while human raters remain essential for evaluating contextual meaning, cultural appropriateness, and pragmatic language use. Future research should investigate advanced prompting strategies, domain-specific AI models, or hybrid human–AI assessment approaches to improve the evaluation of culturally nuanced oral

Recommendation

As a recommendation for future development, the integration of AI in language assessment should be viewed as a collaborative system between complementary agents. A hybrid approach can be applied by optimizing AI for instant evaluation of structurally measurable aspects, while human raters retain full control over dimensions requiring sociolinguistic interpretation, such as idiomatic competence and communicative appropriateness. Furthermore, to improve construct validity, future AI systems should be developed to become more sensitive to prosodic features such as intonation, stress, and speech rhythm, ensuring that assessment is not limited to transcribed text alone. AI models should also be trained on datasets that include a wider range of pragmatic and metaphorical usage to reduce systematic bias and move closer to human interpretative depth.

Ultimately, the difference in evaluation between AI and humans shows that each entity has its own lens in perceiving language proficiency. AI tends to value structural regularity and pattern consistency, while human raters appreciate the "spirit" and underlying meaning of communication. As the German proverb states, *"Jeder hat seinen eigenen Geschmack",* meaning that everyone or every system has its own preferences. In evaluation contexts, "preference" refers to differing assessment paradigms. The combination of AI's algorithmic precision and human emotional intuition may therefore enrich the overall spectrum of language assessment.

Acknowledgements

We would like to express our gratitude to the Faculty of Languages, Arts, and Culture at Yogyakarta State University for providing the authors with the facilities to conduct this joint research. In addition, we would also like to thank the organizing committee of the Iconels 2026 international seminar at Malang State University for giving the authors the opportunity to present the results of this research at the seminar.

References

Aliyeva, E. (2025). The Role of Teaching Proverbs and Sayings in Enhancing Students' Speaking Skills. Acta Globalis Humanitatis et Linguarum, 2(1),54-61. https://doi.org/10.69760/aghel.02500107

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623). https://doi.org/10.1145/3442188.3445922

Charteris-Black, J. (1995). Proverbs in communication. Journal of Multilingual & Multicultural Development, 16(4), 259-268. https://www.google.com/search?q=https://doi.org/10.1080/01434632.1995.9994609

Cram, D. (2015). The linguistic status of the proverb. In Wise Words (RLE Folklore) (pp. 73-97). Routledge. https://www.google.com/search?q=https://doi.org/10.4324/9781315662138-11

Ercikan, K., & McCaffrey, D. F. (2022). Optimizing implementation of artificial-intelligence-based automated scoring: An evidence centered design approach for designing assessments for AI-based scoring. Journal of Educational Measurement, 59(3), 272-287. https://doi.org/10.1111/jedm.12332

Hartwell, K., & Aull, L. (2023). Editorial Introduction–AI, corpora, and future directions for writing assessment. Assessing Writing, 57, 100769. https://doi.org/10.1016/j.asw.2023.100769

Li, B., Qunhan, X., & Mao, C. (2026). Differences between human and AI scoring: A meta-analysis of English language assessments. Scientific Reports. https://doi.org/10.1038/s41598-026-48053-w

Lundgren, M. (2024). Large language models in student assessment: Comparing ChatGPT and human graders. arXiv preprint arXiv:2406.16510. https://doi.org/10.48550/arXiv.2406.16510

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. Trends in cognitive sciences, 28(6), 517-540. https://www.sciencedirect.com/science/article/pii/S1364661324000275

Margetson, K., McLeod, S., Verdon, S., & Tran, V. H. (2023). Transcribing multilingual children's and adults' speech. Clinical Linguistics & Phonetics, 37(4-6), 415-435. https://doi.org/10.1080/02699206.2022.2051073

Mieder, W. (2004). Proverbs: A Handbook. Greenwood Press.

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of Grammarly. CALICO Journal, 34(2), 149-181. https://doi.org/10.1080/01443410.2015.1136407

Rietveld, T., & van Hout, R. (2017). The paired t test and beyond: Recommendations for testing the central tendencies of two paired samples in research on speech, language and hearing pathology. Journal of communication disorders, 69, 44-57.

Yanga, T. (2017). Inside the proverbs: A sociolinguistic approach. In African Languages/Langues Africaines (pp. 130-157). Routledge.