AI, Instructor, Observer, and Peer Feedback in Online Microteaching: A Multi-Rater Comparison Study

Veronika Vrábelová, Zuzana Sándorová, Zdena Kráľová, and Elena Kováčiková
Department of English Language and Culture, Constantine the Philosopher University in Nitra, Nitra, Slovakia

veronika.vrabelova@ukf.sk
zsandorova@ukf.sk
zkralova@ukf.sk
ekovacikova@ukf.sk

Abstract: Feedback is central to the effectiveness of microteaching as a teacher training method, yet how different feedback sources compare when evaluating the same performance is an underdeveloped area. In online environments, peer feedback, instructor assessment, external observation, and AI-generated evaluation can all operate alongside one another. Understanding which sources produce meaningful and consistent judgements has practical consequences for teacher educators. This paper presents a case study conducted at Constantine the Philosopher University in Nitra, Slovakia, within a master's-level course on the Didactics of Distance Online Education. Students of the course each delivered an online microteaching session via Google Meet, incorporating at least one AI tool connected to their lesson objective. All sessions were recorded and re-analysed by an external observer. Five rater sources evaluated each microteaching against a shared four-criterion rubric: course peers, the course instructor, the external observer, and two AI models applied under three distinct evaluation conditions: isolated, sequential, and simultaneous. The study investigates how these rater sources compare in their assessments of the same teaching events and whether AI-generated feedback converges with trained human judgement depending on evaluation structure. Findings reveal differences between rater sources. Peer feedback shows a pattern of inflation and an inverse relationship with expert judgement, raising questions about its reliability as a standalone formative instrument. AI evaluation is highly sensitive to evaluation conditions. Isolated assessment produces variable and unreliable results. Sequential evaluation, in which all presenters are assessed in a single continuous conversation, converges closely with the judgements of trained human raters. This convergence was confirmed across both AI models tested. These findings have implications for feedback design in online teacher education. Rater source is as consequential as rubric design. Peer feedback requires pedagogical framing and should not function as a standalone assessment instrument. AI can serve as a supplementary tool when deployment conditions are understood in advance. Pedagogical expert judgement remains the validity benchmark against which any automated or peer-based system must be measured. The study contributes empirical data to a field where AI tools are being adopted faster than the research can evaluate them.

Keywords: Online Microteaching, AI Assessment, Peer Feedback, Multi-Rater Comparison, EFL Teacher Education

1. Introduction

Large Language Models (LLM) propose the possibility of being used as supplementary assessors of student performance. Instructors managing large cohorts can now turn to models such as ChatGPT (OpenAI, 2026) or Gemini (Google, 2025) to help with generating evaluations and reducing the time pressure that comes at scale. Whether these tools can reliably fulfil this role still depends on factors that are not fully explored.

This study investigates that question in pre-service EFL teacher education. Twenty-two online microteaching sessions were evaluated by nine raters: an Instructor, an Observer, Peers, and two AI models. All of them were assessing identical performances against a shared rubric. Three AI evaluation conditions were tested: Isolated, Sequential, and Simultaneous. Peer feedback is examined alongside AI evaluation, as it raises parallel reliability concern and research shows peer scores tend toward leniency due to familiarity with other students (Falchikov and Goldfinch, 2000; Tomal, 2010).

The overreaching aim is to determine whether AI models can serve as dependable supplementary raters and to identify the conditions under which their evaluations most closely align with expert human judgement. All raters used a four-criterion rubric on a five-point Likert scale covering: Overall Effectiveness, Clarity of Instruction, Engagement, and Use of Technology. Three research questions guide this study:

> (RQ1): How do different rater sources compare in their assessments of the same microteaching when evaluated against a shared rubric?
>
> (RQ2): To what extent does peer feedback converge with expert judgement, and what patterns characterise peer rating behaviour?

(RQ3): How does evaluation condition (Isolated, Sequential, Simultaneous) affect the consistency and accuracy of AI-generated feedback relative to human raters?

2. Literature Review

2.1 Background

Microteaching was formalised by Allen and Ryan (1969) as a teacher training method that reduces classroom complexity into a short teaching episode delivered to a small group, followed by structured observation and feedback. The main principle is that isolating a specific teaching skill in a controlled environment allows for targeted development before student teachers enter real-life classrooms. The method has since been adapted widely across teacher education programmes internationally, including online microteaching where videoconferencing platforms replace traditional classroom settings (Tomal, 2010).

Feedback is the core instrument that makes microteaching functional and an important learning experience. Nicol and Macfarlane-Dick (2006) argue that effective feedback must be specific and credible. This is relevant to peer feedback specifically as students cannot provide meaningful and evaluative feedback without being prepared to apply assessment criteria first. In microteaching courses, feedback typically comes from the instructor and from peers but with AI come new possibilities, though Zhang et al. (2026) found that while AI-generated feedback enhanced pre-service teachers' self-efficacy comparably to teacher feedback, it was rated significantly lower in credibility and usefulness.

2.2 Current Position in Research

The reliability of peer feedback in higher education has been thoroughly researched. In a meta-analysis of peer assessment studies, Falchikov and Goldfinch (2000) found that peer scores tend to be inflated compared to instructor scores, with the gap widening as personal familiarity increases. Tomal (2010) documented the same pattern specifically for microteaching contexts, where pre-service teachers tend to assign socially affirming scores. These tendencies appear across contexts and a question how they compare against AI raters gains relevance. Usher (2025) found that peer and instructor grades aligned with each other, while AI assigned higher grades. Usher and Faraon (2025) refined this picture further, showing that the relationship is dependent on quality. In the study, ChatGPT's alignment with instructors improved as project quality increased, while peer-instructor alignment was strongest for lower-quality work. Students themselves perceived ChatGPT as more generous, neutral, consistent, and less attentive. This positions leniency as an aspect that shifts based on what is being evaluated and by whom.

Understanding why AI raters behave as they do requires a closer look beyond the actual score. Seo et al. (2025) identified systematic biases in LLM evaluation, including position bias, knowledge bias, and format bias. Each of these three biases are capable of distorting the scores regardless of the student's performance. The same study found that five-point Likert scale produced inconsistent AI output and models gravitated towards extremes scores. Yavuz et al. (2025) reported strong Intraclass Correlation Coefficient scores for ChatGPT and Bard against experienced EFL teachers but noted that reliability weakened on non-objective criteria such as style and expression, suggesting that AI consistency is criterion dependent. Usher (2025) also found that AI provided more detailed and structured feedback than peers, although sometimes with irrelevant or contradictory content. All these findings suggest that AI evaluation has a unique behaviour that is not strictly lenient or reliable, it is shaped by what the rubric asks it to score and how that judgement is structured.

Evaluation conditions became crucial in Zhuo et al.'s study (2024), where they mention their ProSA framework. They found that prompt sensitivity is caused especially in subjective evaluation contexts, which helps to explain why the same rubric can produce different results depending on how and when it is presented in the model. Further layer was added by Song et al. (2025), that prior conversational history functions as in-context learning, meaning that each additional student a model evaluates in sequence changes the context against which the next is judged. They also identified model-specific positional bias. According to their research, GPT-4o favoured responses presented earlier in a sequence, while Gemini favoured those presented later.

AI evaluation is not stabilised and produces inconsistent results. Its outputs are shaped by model architecture, evaluation structure, prompt design, and conversational history. All of these factors interact with each another in ways that we are slowly beginning to understand. This study addresses the question of how these factors play out in real educational assessment context, alongside peer, instructor, and observer judgement.

3. Methodology

3.1 Participants and Setting

The study was conducted at Constantine the Philosopher University in Nitra within the master student level course *Didactics of Distance Online Education* in the EFL teacher education programme. Twenty-two students participated during the winter semester of 2024/2025 academic year. From the cohort, 19 students were fourth year, and 3 were fifth year students. As part of the course, each had to deliver a 15-minute microteaching via Google Meet with the instructions to target their classmates, as C1 students, and to incorporate at least one AI tool pedagogically connected to their lesson objective. At the beginning of the course, all participants were informed about the nature of the study and signed informed consents. All sessions were recorded with OBS software and later analysed for detailed transcript construction. Throughout the whole study, all information of the participants was anonymised, and we refer to them simply as Presenter Number (Pn).

3.2 The Evaluation Rubric

All raters used the same four-criterion rubric, presented in full in the Figure 1. The rubric covers Overall Effectiveness (pedagogical coherence, achievement of learning objective, contextual appropriateness, and time management), Clarity of Instruction (precision, accessibility, and Multimodal delivery), Engagement (active participation and learner involvement), and Use of Technology (integration, functionality, and pedagogical value of AI and digital tools used). Each criterion is scored on a 1-5 Likert Scale with detailed descriptions at every level.

Category	Criteria Focus	1.	2.	3.	4.	5.
Overall Effectiveness	Achievement of learning objective Pedagogical coherence Contextual appropriateness Time management	Learning objectives are absent or unattainable within the time range Activities are incoherent, disconnected Topic is inappropriate for C1 academic context	Learning objective is visible but unmet Time management is poor Activities have minimal logical connection to the objective Topic is of limited sustainability for C1 academic learners	Learning objectives are evident and partially achieved Time is managed with some inconsistencies Most activities relate to the objective with rationale provided Topic is generally appropriate	Learning objectives are clearly visible and sustainably achieved Time is well-managed Activities are purposefully connected with clear rationale Topic is appropriate for C1 academic learners Comprehension is monitored	Learning objectives are explicitly stated Objectives can be fully achieved Time is managed All activities are purposefully organised Topic is suited for C1 academic level Learner comprehension is monitored
Clarity of Instruction	Precision Accessibility Multimodal delivery	Instructions are absent Incomprehensible Verbal without visual support Repeated clarification is needed	Instructions are partially clear Visual support is inconsistent or absent Unnecessary complexity Requests for clarification	Instructions are generally clear Instructions are occasionally supported both verbally and visually Most participants can proceed without further clarification	Instructions are clear Instructions are verbal and visible on the screen Students can engage with tasks with little need for clarification	Instructions are precise Instructions are delivered verbally and visible on the screen No clarification is needed Students can engage with activity immediately
Engagement	Active participation Learner involvement	Participants remain passive No facilitation strategies Activities require no meaningful participation Learner engagement is not monitored	Engagement is limited to minority Facilitation attempts are minima Activities allow passive participation Monitoring is absent or superficial	Most participants engage adequately The presenter employs basic facilitation strategies Activities encourage some active participation Monitoring occurs periodically	Most participants can engage consistently The presenter uses facilitation strategies Activities require meaningful participation Learner engagement is monitored	All students can engage throughout Presenter continuously facilitates involvement All activities create meaningful engagement Participation is monitored systematically
Use of Technology	Integration Functionality Pedagogical value	Technology is absent Technology disrupts the learning process Tools have not been tested Tools do not add value to activities	Technology is employed but bears little connection to the learning objective Technical difficulties No advantage over traditional instruction	Technology supports learning objectives in a basic manner Tools function without major disruption Some advantage over traditional instruction	Technology is well-integrated Technology meaningfully supports the learning objective Tools function reliably Prior test of tools was done Digital use offers advantage over traditional instruction	Technology is seamlessly integrated Technology enriches the learning objective All tools have been pre-tested and function correctly Technology can enable learning experience that is not possible in traditional instruction

Figure 1: The Evaluation Rubric

3.3 Human Raters and Peer Data Collection

Two raters scored all 22 performances, course instructor and an external observer. Both re-analysed all recordings, transcripts, and scores accessed after the course ended and again a year later for the purpose of this study. The external observer is also one of the authors of this study and to manage this positionality, the re-analysis was introduced from video recordings to ensure that scores were not influenced by memory or context of the course itself.

The peer feedback form conducted via Google Forms used the same criteria and was distributed to all participants except the presenter immediately after they finished their microteaching. Session attendance naturally varied between 19 and 22 participants.

3.4 AI Evaluation Conditions

Two AI models were used to collect data in the form of identical feedback, ChatGPT GPT-5.3 Instant (OpenAI) and Gemini 3 Flash (Google). Both were accessed through paid subscriptions purchased for this study, using the standard conversational type as the most employed among practitioners. Deep-thinking or reasoning mode were strategically not employed for the same reason. The evaluation procedure was identical for both models but varied based on conditions. Three evaluation conditions were applied:

- **Isolated condition:** each transcript with prompt was submitted in a completely new chat session, with no prior context. This was repeated three times per presenter to assess within-condition consistency. Evaluation rubric was included each time with the transcript.
- **Sequential condition:** all 22 participants were submitted one by one within the same continuous conversation which allowed the model to retain all prior evaluations as context. This was repeated as three separate full sessions. Evaluation rubric and the prompt were pasted only once at the beginning of the chat.
- **Simultaneous condition:** all 22 transcripts and evaluation rubric were submitted together in a single prompt. This was also repeated three times.

In all three types of conditions identical prompt was used, except Simultaneous condition, which required to change it from singular to plural form as a technical necessity (*"evaluate student – evaluate students")*. An initial test without these instructions showed that the model returned feedback for only the first student. All AI scores reported in the findings represent means across the three runs per condition.

3.5 Statistical Analysis

Intraclass Correlation Coefficients (ICC) for absolute agreement (A,1) assessed reliability between instructor and observer under a two-way mixed effects model. ICC was selected because both raters used a rubric with numeric scale from 1 to 5. Spearman correlations were computed against all sources using means across all four criteria. A Friedman test compared all nine-rater source and Wilcoxon signed-rank post-hoc comparisons applied to the six AI condition comparisons of interest. All AI scores are means across three runs per conditions. Each statistical method was strategically chosen to answer following questions: What were the absolute scores given? Who was systematically too strict or lenient? What is the overall agreement level? Are the drops and jumps between specific AI conditions real? Does the AI's internal student ranking mimics Instructor and Observer? Statistical analyses were conducted using IBM SPSS Statistics (Version 32).

This study is exploratory and it does not seek to generalise findings across institutions or AI systems. Its main purpose is to surface patterns and generate hypotheses that future and larger-scale research can test in more detail.

4. Findings

4.1 Human Benchmark Reliability

The Instructor and Observer agreed well enough to be taken as reliable benchmark. Their ICC(A,1) was .858, 95% CI [.791, .904], $p < .001$, placing it in the good range (Koo and Li, 2016). They gave identical scores on 68.2% of all cells and differed at maximum only by one point on 94.3%. Their combined mean was 3.19 and it also serves as a benchmark for the statistical analysis. Instructor and Observer used the Likert scale of 1-5 almost fully, with their means minimum scaling from 1.25 to maximum 5.00.

4.1.1 (RQ1) Rater Source Comparison

A Friedman test across all nine rater sources confirmed statistically significant differences, chi-squared (8)=68,983, p<.001, Kendall's W =.392. This reflects moderate agreement across all nine rater sources on the ordering of presenters. It is worth mentioning that it is still closer to the lower end of the scale which can mean that some sources were misaligned with the rest. This can be clearly seen with GPT Isolated. The differences between raters are evident and not random and the data in Table 1 supports this claim. The combination and further division of Instructor and Observer mean created Benchmark for this study.

Table 1: Descriptive statistics, Friedman mean ranks, Spearman r against Instructor and Observer (N=22)

Rater Source	Mean	Standard Deviation	Minimal mean	Maximal mean	Friedman rank	Spearman r vs. Benchmark ((Inst.+Obs.)/2)	p
Instructor	3.16	1.111	1.25	5.00	4.34	-	-
Observer	3.22	1.073	1.50	5.00	4.73	-	-
Peer Feedback	4.27	0.495	2.68	4.70	7.82	-.300	.175 (ns)
GPT Isolated	2.68	0.591	1.42	3.58	1.93	.774***	<.001
GPT Sequential	3.23	0.700	1.92	4.50	5.23	.797***	<.001
GPT Simultaneous	3.10	0.704	1.83	4.25	4.18	.840***	<.001
Gemini Isolated	3.22	0.782	1.67	4.25	4.86	.745***	<.001
Gemini Sequential	3.20	0.747	1.83	4.50	4.82	.687***	<.001
Gemini Simultaneous	3.47	0.801	2.00	4.75	7.09	.761***	<.001

4.2 (RQ2) Peer Feedback

The peer feedback showed interesting divergence from other rater sources. Visible from Table 1, their average mean sits one point higher in comparison to both Instructor and Observer, as well as all 6 AI raters. Their standard deviation of 0.495 is the smallest, which shows their uniform rating style. Interestingly enough, even though peers rated their classmates highly, they never reached complete agreement to give someone mean of 5.00. The Spearman correlation did not reach significance against the Benchmark (r = -.300, p= .175), and it correlated negatively even with all AI raters as well. Two presenters are worth mentioning, specifically P21 and P22. Both are fifth-year students. Their peer mean scores were 3.18 and 2.68, the two lowest in the entire group against average a mean of 5.00 and 4.13 from the Benchmark. Qualitative open-ended feedback, although not included for the statistical purpose of this study, clarifies possible reasons. Peers had following suggestions for improvement for both presenters (P21 and P22): "*Given that it's a micro teaching presentation I feel like it should have been some kind of presentation of some knowledge on the topic.*" Another student stated:" *Longer introduction, not just the activity*," or "*Honestly just re-do it all. It was boring, not engaging, and again he could have checked if those who were using tablets, phones, etc. could upload document into Google Doc.*" Third fifth-year student who followed more formulaic microteaching session similar to the rest of the cohort scored significantly higher (Peer mean = 4.25; Benchmark=4.38).

4.3 (RQ3) AI Evaluation Conditions

Our claim states that AI conditions are important for final evaluation, which can be seen in Table 2. For each condition and each run, the same prompt, evaluation rubric, and transcripts were used, and yet it produced different scores. GPT general mean across three conditions varied from 2.68 to 3.23 to 3.10, Gemini produced more similar results, but means were still not identical.

GPT Isolated produced the lowest mean rank in the Friedman analysis at 1.93. The lowest mean was 1.42 and the highest 3.58. A student who received full score (5/5) from the Benchmark was able to receive only 3.58 in the GPT Isolated condition. Data implies that GPT Isolated treated each presenter's transcript strictly against the evaluation rubric.

GPT Sequential behaved differently. The average mean of 3.23 is very close to the average Benchmark mean of 3.19, and its Spearman correlation against the Benchmark was r = .797, p<.001. In this case, the maximum mean was 4.50, GPT Sequential was able to recognise strong presenters and score them adequately. The Wilcoxon test confirmed that sequential scored 20 out of 22 students higher than isolated (Z= -3.916, p< .001) and Simultaneous scored higher 18 out of 22 students than Isolated (Z= -3.567, p< .001). Comparison between GPT Sequential and Simultaneous showed that Sequential scored 14 out of 22 students higher (Z= -2.302, p< .001), which is the weakest of the three GPT comparisons. Ranking emerges: GPT Isolated < GPT Simultaneous < GPT Sequential.

Gemini conditions produced different results in comparison to GPT. Gemini Isolated and Gemini Sequential were quite well aligned with the Benchmark. Gemini Isolated produced a mean of 3.22, a distance of only +0.03 from the Benchmark of 3.19, and Sequential produced a mean of 3.20, a distance of +0.01. The comparison between them produced a null result (Z= -0.599, p= .549). With only two ties, scores were moving up and down for the same number of students. A possible explanation is that the two conditions approached individual criteria differently, but this claim would require a further analysis to test each criterion separately, and therefore, it remains as a limitation of this study.

Gemini Simultaneous produced repeating inflation as it scored higher than Isolated for 17 out of 22 students (Z= -3.101, p= .002) and higher than Sequential for 21 out of 22 students (Z= -4.020, p< .001). Its mean and Friedman mean rank place it well above the Observer and Instructor scores.

From these findings, GPT Isolated and Gemini Simultaneous seem to be at opposite ends of the spectrum. GPT Isolated is too strict, and Gemini Simultaneous more lenient. Table 2 presents Wilcoxon results, with each condition presented in a shortened version (ChatGPT Isolated = GPT Iso, etc.). Z^a is based on negative ranks, and Z^b is based on positive ranks.

Table 2: Wilcoxon signed-rank results for AI condition comparison

Comparison	Negative Ranks	Positive Ranks	Ties	Z	p
GPTSeq>GPTIso	1	20	1	-3.916[a]	< .001
GPTSim>GPTIso	1	18	3	-3.567[a]	< .001
GPTSeq>GPTSim	14	5	3	-2.302[b]	.021
GemSeq vs GemIso	10	10	2	-0.599[b]	.549
GemSim>GemIso	3	17	2	-3.101[a]	.002
GemSim>GemSeq	0	21	1	-4.020[a]	< .001

Table 3 shows all runs for each condition with the standard deviation and distance from the Benchmark. The standard deviations report on how much the average mean varied between runs. GPT Isolated was the most stable yet had the biggest distance from the Benchmark (-0.51). GPT Sequential, Gemini Sequential, and Gemini Isolated were very close to the Benchmark.

Table 3: AI scores across three runs per condition

Condition	Run 1	Run 2	Run 3	Mean	SD	Distance from the Benchmark (3.19)
ChatGPT Isolated	2.68	2.69	2.66	2.68	.017	-0.51
ChatGPT Sequential	3.19	3.19	3.30	3.23	.059	+0.04
ChatGPT Simultaneous	3.05	3.06	3.19	3.10	.082	-0.09
Gemini Isolated	3.20	3.26	3.20	3.22	.035	+0.03
Gemini Sequential	3.20	3.12	3.26	3.20	.070	+0.01
Gemini Simultaneous	3.42	3.51	3.48	3.47	.046	+0.28

5. Discussion

The main finding of this study is that rater source matters significantly. Although at first, this may not sound like a new finding, after close examination of all rater sources against each other, especially all AI raters, we can see how vast the differences are. All nine raters used the same four-criterion rubric and evaluated the same 22 microteaching transcripts, but the score differences were systematic. Kendall's W of .392 indicates that the nine sources shared a moderate common signal about which students performed better and which worse. This means that the evaluation system as a whole was moderately coherent, and that two sources, peer feedback, and GPT Isolated in particular, were pulling in directions that weakened the overall agreement. Understanding why each source behaved as it did is as important as observing that they differed.

5.1 Peer Feedback

Peer feedback showed almost no variance across students, with a mean of 4.27 (Benchmark mean=3.19) and a standard deviation of only .495. Peers gave nearly identical high scores to almost everyone regardless of the actual performance. The P21 and P22 cases serve as evidence. Both were fifth-year students who received very high scores from the Instructor and Observer but the two lowest peer scores in the cohort. This is consistent with Falchikov and Goldfinch (2000), who found that peer scores inflate in comparison to instructor scores and that personal familiarity between raters and presenters widens the gap. The negative Spearman correlation between peers and the human Benchmark, which extended across all six AI conditions too, suggests peers were scoring highly and in a direction that diverges from all other rater sources. Since these presenters are pre-service EFL teachers who will assess their students, it is important to note that genuine understanding of performance criteria requires structured development (Nicol and Macfarlane-Dick, 2006). Evaluation exercises and post-observation reflection where peer scores are compared directly against instructor's ratings are the kinds of activities that help to build this skill.

5.2 AI Rater Findings

The AI findings reveal that evaluation results produce different scores depending on the condition. Without cohort context, models apply the rubric as an absolute standard with no reference to how the group performed relative to each other. GPT Isolated shows this clearly, as it produced consistent deflation and a ceiling that never reached what the Instructor gave. The mechanism for why Sequential condition corrects this is that processing transcripts one by one within the same conversation provides the model with a frame of reference. It learns what the specific cohort looks like before scoring. This is similar to Song et al. (2025), who found that prior conversational history functions as in-context learning for GPT models specifically.

Gemini behaved differently, which asks for further explanation. Gemini Isolated was already close to the human Benchmark without any cohort context, and the comparison between Gemini Sequential and Isolated produced a null result ($p = .549$). This suggests that Gemini calibrates differently in Isolation, possibly because it can be more sensitive to the evaluation rubric and prompts than GPT is. The asymmetry in Simultaneous condition is also quite revealing. GPT Simultaneous improved on GPT Isolated, which suggests that having all transcripts present in one prompt provided some comparative context. Gemini Simultaneous inflated scores for 21 out of 22 presenters above Sequential levels, which produced the strongest Wilcoxon result. A possible hypothesis is that Gemini's higher context sensitivity makes it vulnerable to the structure of a simultaneous prompt. When all 22 transcripts were approached at once, it may have read the amount of content as a signal that this is a high-performing group and scored according to that. This makes its Isolated condition more calibrated but Simultaneous more vulnerable. Although it remains as a hypothesis and requires further evidence.

GPT Simultaneous produced the highest Spearman correlation out of all AI rates across conditions ($r = .840$), which indicated strong rank-order agreement with the Benchmark. It reflects a correlation but does not prove the accuracy of absolute scores. GPT Simultaneous remained below the Benchmark in mean score and was outranked by GPT Sequential in direct Wilcoxon comparison, making Sequential more suitable. This is worth mentioning as it poses an interesting addition to our findings.

For future study, each criterion should be analysed separately. The current analysis worked at the presenter level, averaging the mean across all four criteria. Conditions may have differed not only in overall score but also in how much importance they assigned to specific criteria, which could explain some of the individual-level variation between runs. Knowing which criterion is most sensitive could help instructors use these tools more precisely.

5.3 Limitations

This study has several limitations. It is a single case study at one institution with 22 presenters, which limits generalisability. AI models evaluated written transcripts, while human raters had access to both video recordings and transcripts. The three AI conditions were not structurally identical, as the Simultaneous required a slightly different framing, changing the noun for "*presenter*" from singular to "*presenters*" in plural. The study tested two specific model versions at a specific timeframe, and since both models are updated continuously, the same study may produce different results in future.

6. Conclusion

This study compared nine rater sources evaluating the same 22 microteaching sessions with an identical rubric. Rater sources had a statistically significant effect on scores. Peer feedback inflated scores and could not differentiate between stronger and weaker performances. It correlated negatively with Instructor and Observer Benchmark but also with all AI conditions. Peers were ranking the cohort in the opposite direction from every trained human or AI raters. Sequential AI evaluation aligned with the human Benchmark for both ChatGPT and Gemini. GPT Isolated underperformed, while Gemini Isolated was already similar to the Benchmark. Gemini Simultaneous produced inflation, while GPT Simultaneous performed moderately.

The practical recommendation is that Sequential evaluation for both models proved to be the safest. It is the only condition that produced results close to the human Benchmark for both models. It avoided the inflation of Gemini Simultaneous and the strict scoring of GPT Isolated. Without understanding of how evaluation conditions affect AI output, using AI feedback risks adding to instructors' workload. It may also create an unfair environment for students who are assessed differently depending on how the tool was set up and what condition was used. The trained pedagogical professionals should remain the primary evaluators of student performance, especially where feedback needs to be given to complex performance. AI models in their current form are best used sparingly as a supplementary source that can support their judgement or help manage large cohorts. Until AI models are calibrated specifically for pedagogical evaluation and tested to perform reliably, instructors should treat their output as one of the possibilities but not as a final decision. Given the exploratory nature of this study, all recommendations should be treated as working hypotheses.

Ethics Declaration

Signed consent forms were collected at the beginning of the course, and all participants were informed about the nature of the study. All data were anonymised and no personal information was necessary for conducting this research.

AI Declaration

AI models were used as evaluators of microteaching transcripts and formed central component of the research. AI assistance was used during the writing process to check for grammatical errors.

Acknowledgments

The current study is part of the KEGA 027UKF-4/2025 project titled *Didactics of Online and Distance English as a Foreign Language Teaching - Textbook and Multimedia Support Design*. It also forms part of the UGA project V/16/2026 titled *Využívanie umelej inteligencie v dištančnom vyučovaní anglického jazyka.*

References

Allen, D.W. and Ryan, K. (1969) *Microteaching*. Reading, MA: Addison-Wesley.

Falchikov, N. and Goldfinch, J. (2000) "Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks", *Review of Educational Research*, Vol 70, No. 3, pp. 287-322.

Google (2025) *Gemini 3 Flash,* Large language model, https://gemini.google.com/app [Accessed: 28 March 2026]

Koo, T.K. and Li, M.Y. (2016) "A guideline of selecting and reporting intraclass correlation coefficients for reliability research", *Journal of Chiropractic Medicine*, Vol 15, No. 2, pp. 155-163, https://doi.org/10.1016/j.jcm.2016.02.012.

Nicol, D.J. and Macfarlane-Dick, D. (2006) "Formative assessment and self-regulated learning: a model and seven principles of good feedback practice", *Studies in Higher Education*, Vol 31, No. 2, pp. 199-218, https://doi.org/10.1080/03075070600572090.

OpenAI (2026) *ChatGPT GPT-5.3 Instant*, Large language model, https://chatgpt.com/ [Accessed: 28 March 2026]

Seo, H., Hwang, T., Jung, J. et al. (2025) "Large language models as evaluators in education: verification of feedback consistency and accuracy", *Applied Sciences*, Vol 15, No. 2, p. 671, https://doi.org/10.3390/app15020671.

Song, M., Zheng, M. and Luo, X. (2025) "Can many-shot in-context learning help LLMs as evaluators? A preliminary empirical study", in: *Proceedings of the 31st International Conference on Computational Linguistics.* Abu Dhabi: Association for Computational Linguistics, pp. 8232-8241, https://aclanthology.org/2025.coling-main.548/.

Tomal, D.R. (2010) *Action research for educators.* 2nd edn, Rowman & Littlefield, Lanham, MD.

Usher, M. (2025) "Generative AI vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education", *Assessment & Evaluation in Higher Education*, Vol 50, No. 6, pp. 912-927 https://doi.org/10.1080/02602938.2025.2487495.

Usher, M. and Faraon, M. (2025) "Who grades best? Comparing ChatGPT, peer, and instructor evaluations across varying levels of student project quality", *Assessment & Evaluation in Higher Education,* pp. 1-20, https://doi.org/10.1080/02602938.2025.2588682.

Yavuz, F., Çelik, Ö. and Yavaş Çelik, G. (2025) "Utilizing large language models for EFL essay grading: an examination of reliability and validity in rubric-based assessments", *British Journal of Educational Technology*, Vol 56, pp. 150-166, https://doi.org/10.1111/bjet.13494.

Zhang, C., Hu, M., Wu, W., Chen, Y., Wang, K., Gao, T. and Fu, Q. (2026) "AI versus teacher feedback in developing pre-service teachers' teaching self-efficacy: a quasi-experimental study in simulated teaching", *Teaching and Teacher Education*, Vol 172, article 105355, https://doi.org/10.1016/j.tate.2025.105355.

Zhuo, J., Zhang, S., Fang, X., Duan, H., Lin, D. and Chen, K. (2024) "ProSA: assessing and understanding the prompt sensitivity of LLMs", in: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Miami, FL, pp. 1950–1976, https://aclanthology.org/2024.findings-emnlp.108/.