VietTESOL International Convention 2026

REPOSITIONING ENGLISH:
FROM FOREIGN TO SECOND LANGUAGE
Danang, 27 - 29 August 2026

VIETNAM ASSOCIATION OF ENGLISH LANGUAGE TEACHING AND RESEARCH
THE UNIVERSITY OF DANANG UNIVERSITY OF FOREIGN LANGUAGE STUDIES
MINISTRY OF EDUCATION AND TRAINING NATIONAL FOREIGN LANGUAGES PROJECT

[#1831] A COMPARATIVE STUDY OF IELTS WRITING TASK 2 EVALUATION BETWEEN CHATGPT AND CERTIFIED EXAMINERS

VU THI HONG MAI
Faculty of English, Hanoi Pedagogical University 2, Vietnam
hongmai4112004@gmail.com

TRAN THI NGAN, MA.
Faculty of English, Hanoi Pedagogical University 2, Vietnam
tranthingan@hpu2.edu.vn

01 INTRODUCTION

Generative AI is increasingly used in writing assessment, yet accessibility does not establish scoring validity. The central issue is whether ChatGPT produces stable criterion-level scores that align closely enough with certified IELTS examiner judgment. This matters as AI-mediated feedback becomes more common in Vietnamese EFL and emerging ESL-oriented classrooms.

RESEARCH GAP
Few studies combine real Vietnamese IELTS candidates, criterion-level official scores, and repeated AI ratings.

02 RESEARCH QUESTION

To what extent are ChatGPT's Task 2 scores across the four analytic criteria consistent with those assigned by certified IELTS examiners?

OPERATIONALIZED AS
repeatability • association • agreement • directional bias

03 THEORETICAL FRAMEWORK

The study combines the criterion-referenced IELTS analytic rubric with a reliability lens that separates repeated-score stability from human-AI agreement.

TR	CC	LR	GRA
Task Response	Coherence & Cohesion	Lexical Resource	Grammar Range & Accuracy

- **Consistency:** Are repeated ChatGPT ratings stable?
- **Association:** Do AI and human raters rank scripts similarly?
- **Agreement:** Do they assign the same band scores?
- **Bias:** Does either rater systematically score higher or lower?

04 METHODOLOGY

10	3	4	GPT-5
candidate scripts	ChatGPT runs	IELTS criteria	2025 free version

10 IELTS Writing Task 2 retake scripts
40 minutes • no external help • anonymized

HUMAN BENCHMARK
Official examiner scores
TR • CC • LR • GRA baseline

AI CONDITION
GPT-5 (2025) × 3 runs
Fixed rubric-aligned prompt

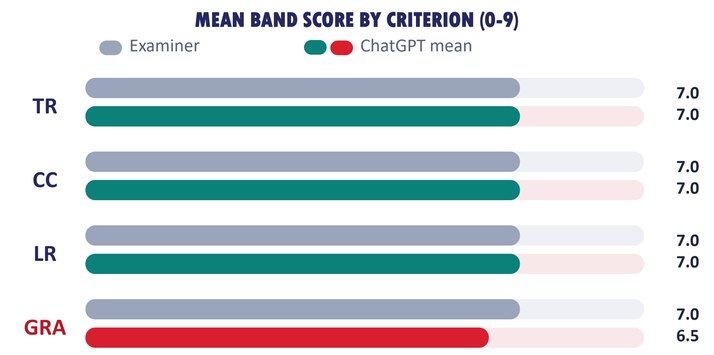
Criterion-by-criterion comparison
TR | CC | LR | GRA

STATISTICAL ANALYSIS
Descriptive statistics • Pearson correlation
Intraclass correlation coefficients • Mean differences

05 HUMAN-AI AGREEMENT BY CRITERION

06 KEY FINDINGS

CONSISTENT, BUT NOT INTERCHANGEABLE



Across three AI runs, SD remained below 1.0:
TR .95 • CC .74 • LR .84 • GRA .83

Only GRA showed a mean bias:
ChatGPT – examiner = -0.5 band

07 IMPLICATIONS FOR RESPONSIBLE USE

FORMATIVE USE IS PROMISING
- Accessible supplementary scoring and immediate feedback.
- Useful for self-study or as a teacher's second-pass review.
- Strongest use case: low-stakes, iterative writing practice.

HIGH-STAKES USE STILL REQUIRES HUMANS
- Retain certified examiner judgment and moderation.
- Audit criterion-level bias, especially the -0.5 GRA shift.
- Do not treat stable outputs as score equivalence.

FOR ESL-ORIENTED CLASSROOMS
- Build teacher and learner AI-assessment literacy.
- Standardize prompts; record model/version and scoring conditions.
- Validate across scripts, proficiency bands, tasks, and models.
- Position AI as decision support, never an autonomous examiner.

08 LIMITATIONS AND NEXT STEPS

- Limitations: n=10, controlled retake scripts, one fixed prompt, free-model updates, and no examiner comments.
- Next: larger/diverse corpora, multiple raters/models, prompt variation, and criterion-level calibration.

CONCLUSION

ChatGPT can support but should not replace certified examiner judgment.

Stable repeated scores and high ICCs are encouraging; partial band agreement and GRA under-scoring still require human oversight.

REFERENCES AND STUDY RECORD

Ahn, J., Lee, J., & Son, M. (2024). ChatGPT in ELT. ELT Journal, 78(3), 345-355.
Chapelle, C. A., & Voss, E. (2016). Technology and language assessment. LLT, 20(2), 116-128.
Stemler, S. E. (2004). Approaches to interrater reliability. PARE, 9(1).
Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. Pedagogies, 3(1), 22-36.

POSTER INFORMATION

ID: #1831 | Strand 1: EFL to ESL Transition
29 August 2026 | 08:45 | Poster session: 45 min
Administrative Area – Campus Square (UD-UFLS)
convention.viettesol.org.vn/event/11/contributions/3453/