Computers and Education Open 10 (2026) 100341

Contents lists available at ScienceDirect

Computers and Education Open

journal homepage: www.sciencedirect.com/journal/computers-and-education-open

Comparing GPT and human raters in essay assessment: Variability, bias, and the potential of LLM-based scoring

Hsiang-Ning Wu [a], Man-Ni Chu [a], Jia-Lien Hsu [b,*]

[a] Graduate Institute of Cross-Cultural Studies, Fu Jen Catholic University, Taiwan
[b] Department of Computer Science and Information Engineering, Fu Jen Catholic University, Taiwan

ARTICLE INFO

Dataset link: https://drive.google.com/drive/folders/1fRVUr8S-veXXeUkoQDWnpUjOnI163hyM?usp=share_link

Keywords:
English writing assessment
Rater variability
Automated essay scoring (AES)
Natural language processing (NLP)
GPT
Many-Facet Rasch Measurement (MFRM)

ABSTRACT

Ensuring reliable essay scoring is challenging when rater severity and bias vary by linguistic background. This study investigates whether a GPT-based Automated Essay Scoring (AES) system can complement human raters in second-language (L2) writing assessment. Using a Many-Facet Rasch Measurement (MFRM) model, this study analyzed ratings from 20 human raters—10 native English-speaking (NES) and 10 non-native English-speaking (NNES)—and GPT across four analytic criteria (*Content*, *Organization*, *Grammar*, and *Vocabulary*) on 181 university L2 essays. The results indicate that raters demonstrated variability and potential rater bias; GPT exhibited high internal consistency but range compression at the upper end compared to humans. These findings offer insights into how GPT can supplement or complement traditional rater-based methods, potentially alleviating the time-consuming and subjective aspects of human scoring. Nonetheless, concerns remain regarding its sensitivity to subjective writing features such as creativity, tone, and nuanced lexical use. This study contributes to the growing understanding of large language models in educational contexts and highlights the need for further refinement and validation of AES systems.

Structured practitioner notes

What is already known about this topic

- Human raters often demonstrate variability in rating severity and consistency, influenced by factors such as linguistic background and personal bias.
- Automated Essay Scoring (AES) systems, particularly those leveraging Large Language Models like GPT, have shown promise in delivering consistent and objective assessments.
- There is ongoing debate over whether LLMs can accurately capture the nuanced, context-specific judgments of human raters, especially when evaluating subjective criteria like content development and organization.

What this paper adds

- This study compares human rater variability with the consistency of a GPT-based AES system, examining differences in rating severity, bias, and consistency across various criteria.
- NES raters typically apply stricter evaluations, whereas NNES raters exhibit broader severity ranges. In contrast, GPT shows a more neutral and stable rating pattern, albeit with a tendency toward stricter content evaluation.
- Findings illuminate how the GPT-based AES system aligns with or diverges from human raters, offering insight into practical strategies for integrating LLM-driven scoring tools into educational assessment.

Implications for practice and/or policy

- The findings suggest that a GPT-based AES system can serve as a valuable tool for ensuring consistency in writing assessments, particularly in large-scale educational settings where human variability can pose challenges.
- The stricter stance on content observed in GPT underscores the need for further fine-tuning and prompt engineering to better capture nuanced human judgments.
- Integration of hybrid models that combine the strengths of human and automated scoring could be explored for more balanced and comprehensive assessments.

* Correspondence to: Department of Computer Science and Information Engineering, Fu Jen Catholic University, SF 618, No. 510, Zhong-Zheng Rd., Xinzhuang Dist., New Taipei City 242062, Taiwan.
E-mail addresses: shuake0225@as.edu.tw (H.-N. Wu), 081733@mail.fju.edu.tw (M.-N. Chu), alien@csie.fju.edu.tw (J.-L. Hsu).

https://doi.org/10.1016/j.caeo.2026.100341
Received 11 June 2025; Received in revised form 15 February 2026; Accepted 17 February 2026
Available online 28 February 2026

1. Introduction

Ensuring consistency and fairness in essay rating has long been a central concern in language assessment. Traditionally, English essays are evaluated by human raters, yet manual scoring is time consuming and susceptible to rater variability stemming from factors such as the linguistic and cultural background of the rater or personal constructs of writing quality, raising concerns about reliability and fairness [1–5].

A well-established approach to examining rater behavior is Many-Facet Rasch Measurement (MFRM) [6], which jointly estimates rater severity, bias interactions, and other facets (e.g., criteria). Prior MFRM work has documented systematic rater effects associated with task type and rater background [3,7] as well as criterion-specific biases and proficiency-related differentials [5,8–10].

Against this backdrop, automated essay scoring (AES) emerged to improve efficiency and reduce subjectivity [11–13]. Early systems, such as PEG, e-rater and IntelliMetric, relied on handcrafted linguistic features coupled with regression or traditional machine-learning models. Although these approaches achieved moderate–high correlations with human scores, they were limited in semantic understanding and in handling off-topic or content-driven criteria [12]. Hybrid architectures that fuse neural contextual representations with linguistically motivated features suggest a transitional path from feature-based AES to LLM-enhanced scoring, often yielding higher alignment with human judgments [14].

With the advent of large language models (LLMs), AES has developed rapidly. A growing body of work shows that LLMs can approximate human-like evaluations and correlate strongly with human ratings, while still exhibiting residual severity or criterion-specific biases that require calibration [11,15–17]. Studies also document high overall accuracy with comparatively lower specificity for higher-order constructs, such as task response and coherence, relative to lexico-grammar [18]. Complementing these findings, classroom-oriented studies report strong human–AI correlations for GPT-based AWE with slightly lower MFRM fit, underscoring both feasibility and the need for routine calibration in instructional contexts [19].

Findings from multilingual and EFL settings are mixed. Evidence includes lower GPT–human agreement for some L1 groups [20], persistent cross-L1 disparities even after fine-tuning [21], and language-/task-specific moderation of alignment [22]. Large-scale analyzes also document demographic differentials such as race or ethnicity raising ethical concerns for deployment [23]. Moreover, comparative evaluations show that model–human alignment can fluctuate with task design, model version, and interface [24,25], even as some contemporary models exhibit reduced rater effects under MFRM [26]. In parallel, psychometric perspectives on LLM raters (e.g., GPT, Claude) suggest near-human reliability while flagging concerns about bias, central tendency, and transparency [27]. Hybrid architectures that combine neural representations with linguistic features have also reported improved alignment with human judgments [14].

Taken together, the literature portrays LLM-based AES as promising yet imperfect. While overall reliability can approach human levels in rubric-based settings, less is known about how LLM raters behave across analytic criteira when situated alongside human raters within a shared measurement framework. In particular, few studies have simultaneously examined (a) criterion-level alignment, (b) relative severity and variance dispersion, and (c) rater–criterion bias interactions between LLMs and raters of different linguistic backgrounds using Many-Facet Rasch Measurement.

To address these gaps, the present study benchmarks 21 raters, including NES, NNES, and a GPT-based AES system, across four analytic criteria (*Content*, *Organization*, *Grammar*, *Vocabulary*) within an MFRM framework. By placing GPT and human raters on the same logit scale, this design enables direct comparison of severity, bias patterns, and construct-level alignment. Rather than inferring validity from aggregate correlations alone, the study examine whether GPT reproduces, mitigates, or diverges from typical human rater variability at the criterion level.

Through this rater-centric and construct-sensitive approach, the study contributes empirical evidence on the dimensional validity, fairness, and operational positioning of GPT-based AES in EFL writing assessment.

Building upon these objectives, the study addresses the following research questions:

1. To what extent do GPT-based AES scores align with human ratings across the four analytic criteria (*Content*, *Organization*, *Grammar*, *Vocabulary*)?
2. Does the degree of alignment differ by analytic criteria?
3. To what extent does GPT-based AES exhibit bias or rater effects similar to—or distinct from—those of human raters?

Collectively, these questions target the properties of GPT-based AES, laying the groundwork for reliable and equitable deployment in L2 writing assessment.

2. Methods

This study compared human raters and a GPT-based AES system in their evaluation of student essays, highlighting the potential advantages and limitations of machine-based scoring. Within the group of human raters, participants were further divided into NES and NNES to explore variability in rating severity and consistency. The intention was to examine how linguistic background could reflect potential variability among human raters, and whether GPT-based evaluations might offer more consistent scoring across diverse essay types. Therefore, a Many-Facet Rasch Measurement (MFRM) model was employed to analyze severity, consistency, and bias within these two broad rater categories (Human vs. GPT).

The measurement model for assessing writing ability in this study is based on the framework proposed by Engelhard [28], as represented in Fig. 1. In this model, rater severity, task difficulty, and criterion difficulty are considered intervening variables. Task difficulty specifically refers to the challenges associated with different tasks, aligning with the study's focus on analyzing the evaluation and rating of a diverse range of tasks. Criterion difficulty refers to the inherent difficulty of each rating dimension used to assess writing ability. Specifically, criterion difficulty captures the extent to which a particular evaluative criterion—such as *Content*, *Organization*, *Grammar*, or *Vocabulary*—poses a challenge for raters to apply consistently.

2.1. Participants

The study involved 20 human raters and an GPT-based AES system. Among the human raters, there were 10 NES and 10 NNES. All human raters had prior experience in language assessment, and each was studying for a master's degree in Teaching English as a Foreign Language (TEFL) in the UK. The NES raters (5 males and 5 females) were from the UK, while the NNES raters (4 males and 6 females) were from Taiwan. All the NNES raters possessed high proficiency in English, evidenced by their achievement of at least a CEFR C1 level and an IELTS score of 8.0 or above. In addition, these raters had 1–3 years of previous classroom or testing experiences involving English writing assessment, which provided familiarity with rating procedures.

A GPT-4 model served as the automated rater, providing a computational perspective on the essays. The GPT-4 version accessed in July 2024 was utilized. Under these conditions, the GPT-4 model functioned as the core of the GPT-based AES group for comparison with human raters.

Each rater was assigned a unique identification code: R1–R10 for NES raters, R11–R20 for NNES, and GPT for the GPT-based AES system. These labels are used consistently throughout the manuscript, and in the Results section, the rater positions are clearly indicated according to these IDs.

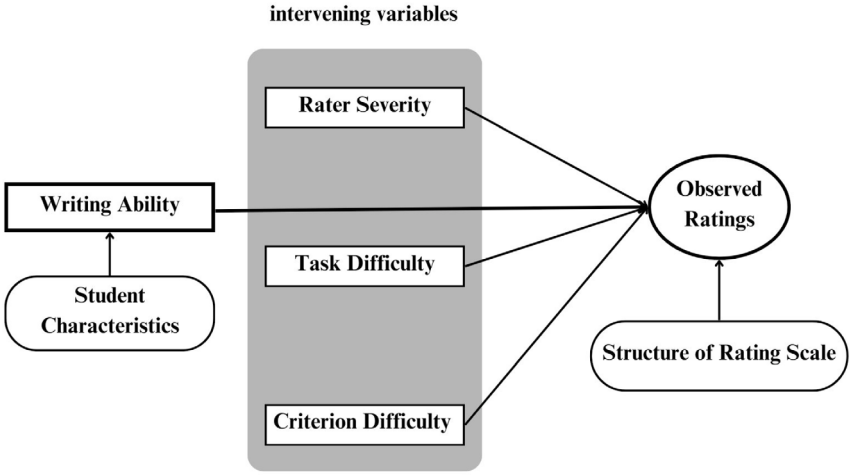

Fig. 1. Measurement Model for Assessing Writing Ability.
Source: Adapted from Engelhard, 1992.

Table 1
Writing tasks.

Examinees	Task
Group 1	Summarizing Silvia Percu's "Designing for an Ageing Society Products and Services"
Group 2	Translating of Chinese passages focused on topics such as economics, social issues, and cultural reflections into English.
Group 3	Do you agree or disagree with the following statement? Most advertisements make products seem much better than they really are. Use reasons and examples to support your answers.

2.2. Essays

A total of 181 essays were collected from three distinct groups of university-level EFL learners in Taiwan, each of whom completed a different writing task. The writing tasks are summarized in Table 1, and each task was produced by a distinct group of EFL learners.

Group 1 was asked to read and summarize the article "Designing for an Ageing Society: Products and Services" by Silvia Percu.

Group 2 was assigned to translate a series of Chinese passages into English. The source materials focused on topics such as economics, social issues, and cultural reflections. Each text yielded between 7 and 9 student translations, resulting in a balanced collection of 81 translated essays.

Students in Group 3 responded to the following prompt: "Do you agree or disagree with the following statement? Most advertisements make products seem much better than they really are. Use reasons and examples to support your answers".

Each essay was assessed across four criteria: *Content*, *Organization*, *Grammar*, and *Vocabulary* . Scores were assigned on a scale from 0 to 5 for each criterion, with 0 indicating the lowest level of performance and 5 the highest. The rating rubric is shown in Appendix. The use of diverse tasks and multiple rating criteria was intended to provide a comprehensive evaluation of each rater's performance, while also highlighting how each rater group (Human vs. GPT) might handle different textual challenges.

2.3. Procedure

Twenty human raters (10 NES and 10 NNES) and the GPT-based AES system independently evaluated 181 anonymized essays using a standardized rubric (Appendix A1 Table A.1). Each essay was assessed on a 6-level scale (0–5) in four criteria: *Content*, *Organization*, *Grammar*, and *Vocabulary*, with 0 indicating the lowest level of performance and 5 the highest. The essays were coded to ensure that raters were unaware of the writers' identities or the specific task instructions used by the writers.

Importantly, each essay was independently rated as a standalone text. For tasks 1 and 2, raters were not provided with the original source texts, such as the article for summary or the Mandarin passage for translation. Raters were instructed to evaluate each essay solely based on its quality and coherence.

2.3.1. Calibration and training for human raters

Before the main evaluation process, each human rater participated in a calibration session. A pool of 300 pre-labeled essays was available, ranging from CEFR A2 to C2 levels. From this pool, one essay was selected for every ten essays at each level, resulting in six essays per level and a total of 30 sample essays. These labeled samples were intended to help raters align their scoring practices with the study's proficiency standards. Each human rater scored these 30 sample essays using the provided rubric and submitted total scores. If a rater's total scores deviated by more than one proficiency level from the labeled level, that rater would have been removed from the study. Although this calibration process might reduce inherent variability among raters, it was deemed necessary to ensure a baseline level of rating consistency. In this study, all 20 raters met the calibration criteria and proceeded with the main evaluation.

2.3.2. Calibration for GPT-4

The GPT-based AES system also underwent a brief calibration session to validate the prompt. Specifically, the system rated the same set of 30 sample essays used for the human raters, following the same 0–5 scale across the four criteria. Its assigned levels were then compared to the established proficiency labels. If the GPT-4 ratings had deviated by more than one proficiency level, the prompt would have been refined and the process repeated. However, the system successfully remained within the acceptable range on the first attempt.

2.3.3. GPT prompting

We prompted GPT-4 (temperature = 0, top_p = 1.0, max_tokens = 150, n = 1) with a fixed system message and rubric-aligned instructions to return analytic scores (0–5) for *Content*, *Organization*, *Grammar*, and

Vocabulary without giving feedback. A short excerpt is shown below; the full prompt of the exact interface is provided in Table A.2.

Excerpt: "I would like you to mark essays... Use a scale from 0 to 5... Output the scores as follows: Content, Organization, Grammar, Vocabulary".

2.3.4. Main rating process

After calibration, all 181 essays were evaluated within three days. The human raters used the rubric descriptions to assign scores manually, while the GPT-based AES system was prompted to interpret and apply the same rubric. Table A.2 shows the scoring prompt used for GPT-4. For each essay, the rubric was input in text format, and GPT-4 returned an output with four individual scores. These scores were exported to a CSV file and subsequently merged with the human raters' results, producing 21 sets of scores (20 human raters and 1 GPT rater) for each essay.

The rubric, detailed in Appendix A1, provided descriptions for each rating level within the four criteria. To enable a direct comparison of the ratings, the GPT-based AES system applied this same rubric under the specified prompt. The implementation utilized the GPT-4 model as of July 27, 2024, processing each of the 181 essays individually in a for-loop. The system was prompted, additional functions were defined to process the scores obtained, and the final outputs were then saved in CSV format. This procedure generated a dataset of 21 rating sets per essay, allowing for a comprehensive analysis of rating severity, consistency, and bias across human raters and the GPT-based AES system. It also enabled the examination of potential intra-group differences among human raters (NES and NNES), illustrating whether GPT might provide more stable scores than humans with different linguistic backgrounds.

2.4. Data analysis

To analyze the rating behaviors of both human raters and the GPT-based AES system, we applied MFRM using FACETS version 3.86.0. The analysis focused on three key facets: (1) raters (categorized overall as human and GPT, with human raters further sub-categorized as NES or NNES), (2) tasks and (3) criteria (*Content*, *Organization*, *Grammar*, *Vocabulary*).

The MFRM model can be represented mathematically as follows:

$$\log\left(\frac{P_{nmijk}}{P_{nmijk-1}}\right) = B_n - A_m - D_i - C_j - F_k \quad (1)$$

where: B_n = ability of examinee n, A_m = difficulty of task m, D_i = difficulty of criteria i, C_j = severity of rater J, F_k = difficulty of category k relative to category k − 1, P_{nmijk} = probability of rating of k under these circumstances, $P_{nmijk-1}$ = probability of rating of k − 1

Raw scores from the 21 raters were structured and categorized by rater group, task, and rating criteria. The MFRM model was then specified to estimate the severity of each rater, the difficulty of each task and criterion, and to explore potential rater bias across criteria. Whether certain raters were consistently more severe or lenient was examined through bias analyses.

The MFRM analysis provided valuable insights into patterns of severity and leniency among raters, revealing potential biases and inconsistencies in the assessment process.

In addition to the main MFRM parameters, the model output includes two fit indices—Infit and Outfit Mean Square Error (MSE)—that serve as indicators of rater fit. These indices quantify the extent to which the observed scores deviate from the model's expectations. Specifically, the Infit MSE is calculated as:

$$\text{Infit MSE} = \frac{\sum_i w_i (O_i - E_i)^2}{\sum_i w_i},$$

and the Outfit MSE is given by:

$$\text{Outfit MSE} = \frac{\sum_i (O_i - E_i)^2}{N},$$

where O_i represents the observed score, E_i is the expected score as predicted by the MFRM, w_i is the weight associated with each observation, and N is the total number of observations. These indices are automatically computed by FACETS. Values within an acceptable range (typically 0.6 to 1.4) suggest that the rater's assessments are consistent with the model expectations; values outside this range may indicate problematic rating behaviors.

To assess potential biases in the rating process, an analysis of rater-criterion interactions was conducted. In this study, "bias" is defined as the systematic deviation of a rater's observed scores from the expected scores predicted by the MFRM. FACETS computes bias measurement values for each rater by comparing the observed scores with the expected scores derived from the MFRM model. These bias values are provided for each essay and are then summarized by rating criterion to yield an overall bias measurement for each rater in a given context. According to the methodology proposed by McNamara [29], bias measurements within the range of −0.5 to 0.5 and t-values between −2.0 and 2.0 are considered acceptable, indicating no significant bias. Bias values and their corresponding t-values outside these ranges suggest significant rating bias, reflecting either leniency or severity in the rater's evaluations.

3. Results

Using MFRM, this study examined rater severity, rater × criterion interactions, and GPT–human alignment across four analytic criteria for NES, NNES, and a GPT-based AES system. Fig. 2 provides a variable map locating raters and criteria on a common logit scale.[1]

Among the 21 raters, Rater 10 (NES) and Rater 18 (NNES) appeared at the severe end, whereas Rater 19 (NNES) was notably lenient. GPT was positioned in the lower–mid range, indicating moderate severity relative to the human raters. Appendix materials (Table A.4) report estimated criterion difficulties; because our focus is rater behavior, we summarize those facets briefly and concentrate on rater-level findings below.

3.1. Rater severity and model fit

On the logit scale, rater severity ranged from −0.70 (Rater 19, NNES) to 4.40 (Rater 10, NES). Most raters showed acceptable fit (infit and outfit mean squares ≈ 0.60–1.40); several NNES raters underfit (e.g., Rater 18: infit = 2.05, outfit = 2.00), whereas GPT exhibited near-ideal fit (infit = 0.89, outfit = 0.93). Full rater-by-rater estimates and fit statistics are provided in Appendix Table A.3.

Summarizing by group, Table 2 indicated a consistent ordering of severity. NES raters were the most severe, NNES next, and GPT the least severe. The NES–NNES gap was sizeable (roughly 0.7 logits), and GPT was notably less severe than both human groups (approximately −1.2 logits relative to NES and −0.4–−0.5 relative to NNES). As a robustness check, we contrasted *All Humans* (weighted across 20 raters) with GPT; results are provided in the Appendix (Tables A.5–A.6).

3.2. Rater × criterion bias/interaction

Table 3 shows distinct tendencies among NES raters, NNES raters, and GPT across criteria. We identified rater-by-criterion bias/interaction using two-tailed t-tests from the MFRM output, flagging cells when both $|t| > 2.0$ and $|\text{bias}| > 0.5$. Under this rule, 16 rater–criterion effects were flagged. *Content* yielded the most flags, with *Organization* second; *Grammar* and *Vocabulary* were sparse. At the group level, NES raters more often showed stricter tendencies on *Content*, whereas NNES raters more often showed stricter tendencies on *Organization*. GPT showed a stricter tendency on *Content* only. These patterns are summarized by counts in Table 4 and visualized in Fig. 3.

[1] FACETS Version 3.86.

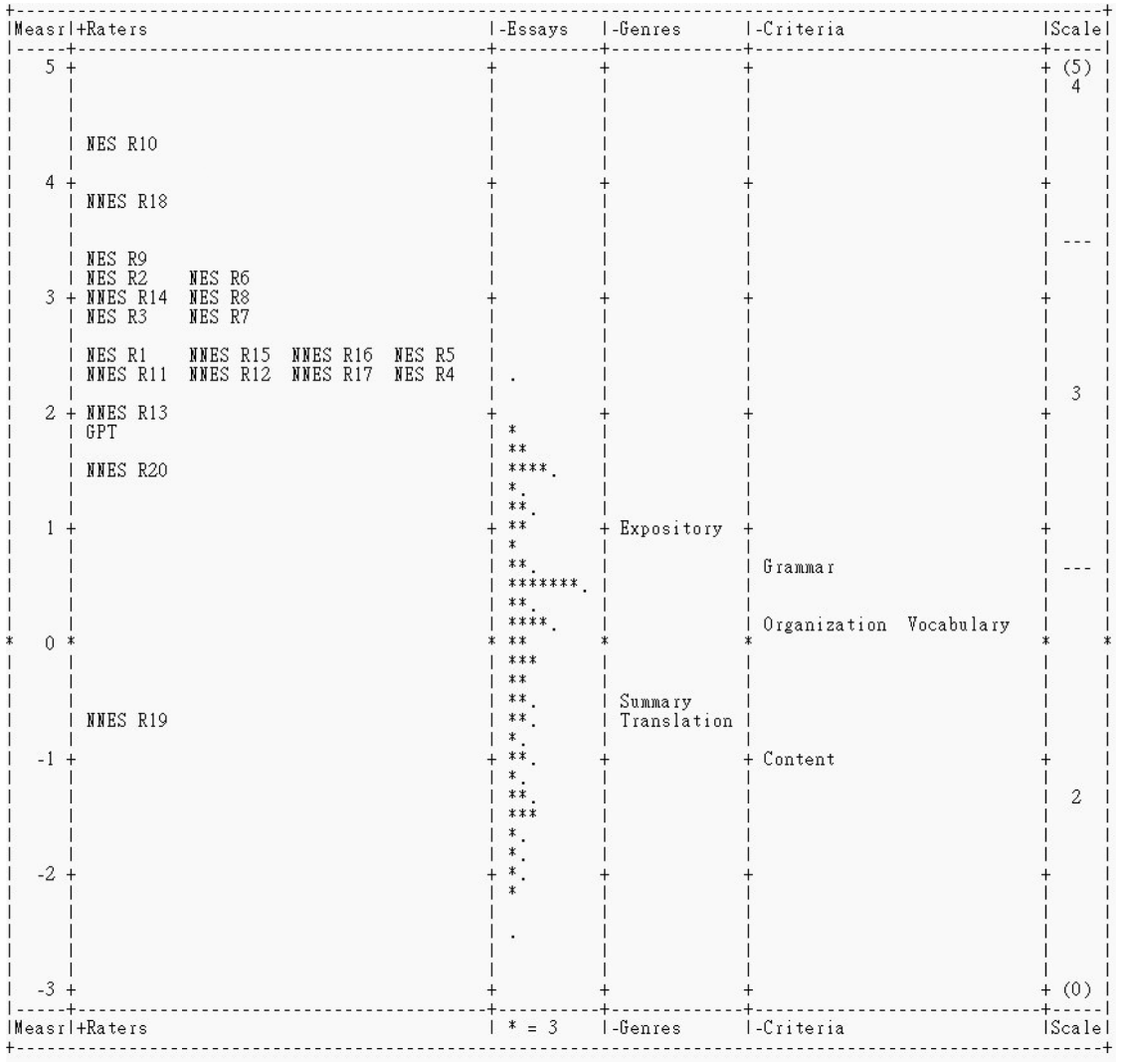

Fig. 2. Variable map of raters and criteria (FACETS v3.86; Linacre, 2004).

Table 2
Group-level rater severity: within-group structure and between-group contrasts.

Panel A. Within-group summaries							
Group	Mean	SE($\bar{M}_g$)	95% CI L	95% CI U	SD (within)	Separation	Reliability
NES[a]	2.994	0.019	2.957	3.031	0.558	9.31	0.989
NNES[b]	2.264	0.020	2.225	2.303	1.109	18.48	0.997
GPT	1.820	0.070	1.683	1.957	–	–	–
Panel B. Between-group contrasts (Δ in logits)							
Contrast	Δ	SE(Δ)	t	95% CI L	95% CI U	p	
NES–NNES	0.730	0.027	26.823	0.676	0.783	< 0.001	
GPT–NES	−1.174	0.073	−16.187	−1.316	−1.032	< 0.001	
GPT–NNES	−0.444	0.073	−6.114	−0.587	−0.302	9.74×10^{-10}	

[a] NES = Rater 1–10.

[b] NNES = Rater 11–20.

Notes. Panel A means and $SE(\bar{M}_g)$ are inverse-variance weighted; 95% CIs are Wald: $\bar{M}_g \pm 1.96\, SE(\bar{M}_g)$. "SD (within)" is the unweighted SD of rater logits within group. Separation ≈ SD/RMSE, Reliability ≈ $\text{Sep}^2/(1+\text{Sep}^2)$ with RMSE = .06. GPT is a singleton (N=1), so SD/Separation/Reliability are not applicable. Panel B reports pairwise contrasts on the common Rasch logit scale.

A BH-controlled analysis ($q < .05$) yielded a similar pattern. Significant cells concentrated on *Organization*, with a secondary cluster on *Content*; GPT showed a stricter tendency on *Content*. Full p and BH-adjusted q values appear in Appendix Table A.7.

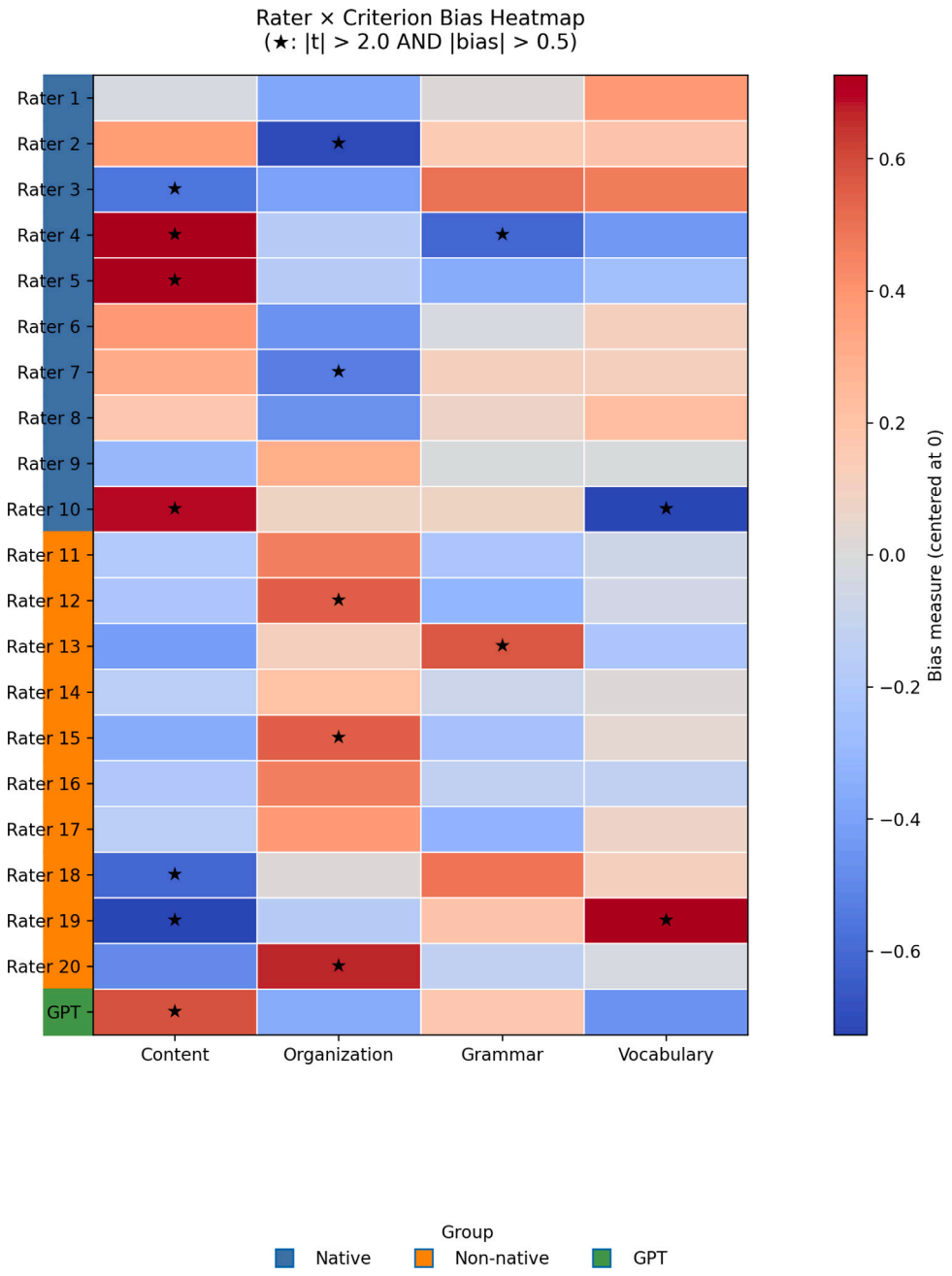

Fig. 3. Rater × criterion bias heatmap (pre-BH, two-tailed $|t| > 2$ and $|\text{bias}| > 0.5$). Color encodes bias magnitude (red = stricter, blue = lenient); significant cells are with star sign. After BH correction ($q < .05$), the flagged set was unchanged.

Overall, *Content* produced the most flagged interactions, with *Organization* second; effects on *Grammar* and *Vocabulary* were sparse. At the same time, *Organization* exhibited a clear group split (NNES more often stricter; NES more often lenient).

3.3. GPT–human associations by criterion

We compared GPT-assigned scores with the mean of 20 human raters across the four analytic criteria using Pearson correlation and ICC(A,1).[2] As summarized in Table 5, the associations were moderate overall (Pearson r = 0.63–0.73; ICC(A,1) = 0.58–0.63), indicating that GPT's analytic scoring patterns broadly aligned with the aggregated human judgments. Among the four criteria, the model showed the strongest linear alignment with human raters on *Vocabulary* (r = 0.73, ICC = 0.60) and *Organization* (r = 0.71, ICC = 0.63). *Content* yielded the smallest mean error (MAE = 0.42, RMSE = 0.55) and the highest absolute-agreement consistency (ICC = 0.63), suggesting that GPT's content scoring was the most stable relative to human averages. In contrast, *Grammar* exhibited the lowest association (r = 0.63, ICC = 0.58), implying greater divergence in sensitivity to linguistic accuracy.

Fig. 4 visualizes these associations via scatter plots of GPT versus human mean scores for each criterion. The diagonal line ($y = x$) represents perfect agreement, while the orange line highlights systematic deviations. Across all panels, GPT slightly compress the human score range, reflecting a narrower scoring variance. In *Organization*, *Grammar*, and *Vocabulary*, GPT's regression line fell below the identity

[2] ICC(A,1) denotes a two-way mixed-effects model, absolute agreement, single measurement.

Table 3
Bias magnitudes for rater–criterion interactions.

Rater	Content	Organization	Grammar	Vocabulary
1				
2		−.71		
3	−.56			
4	1.12		−.61	
5	.75			
6				
7		−.53		
8				
9				
10	.69			−.74
11				
12		.55		
13			.57	
14				
15		.55		
16				
17				
18	−.61			
19	−.76			.73
20		.67		
GPT	.58			

Note: Blank cells indicate no significant bias under the joint thresholds ($|t| > 2.0$ and $|\text{bias}| > 0.5$).

Table 4
Bias frequency by criterion and rater group.

Rater Group	Content S/L	Organization S/L	Grammar S/L	Vocabulary S/L
NES	3/1	0/2	0/1	0/1
NNES	0/2	3/0	1/0	1/0
GPT	1/0	0/0	0/0	0/0

Note: S/L indicates the frequency of significant severity (S) and significant leniency (L) in bias scores, as identified by the t-test results.

Table 5
Comparison of GPT and human mean scores across four writing criteria.

Criterion	n	r	MAE	RMSE	ICC(A,1)
Content	181	0.65	0.42	0.55	0.63
Organization	181	0.71	0.58	0.68	0.63
Grammar	181	0.63	0.51	0.61	0.58
Vocabulary	181	0.73	0.56	0.64	0.60

Table 6
ICC(A,1) by criterion: human-only and GPT–human comparisons.

Criterion	Human-only	GPT vs All	GPT vs NES	GPT vs NNES
Content	0.380	0.382	0.409	0.472
Organization	0.556	0.549	0.676	0.456
Grammar	0.393	0.393	0.425	0.402
Vocabulary	0.489	0.485	0.537	0.472

Human-only uses Rater 1–20; GPT vs All uses Rater 1–20 plus GPT; GPT vs NES uses Rater 1–10 plus GPT; GPT vs NNES uses Rater 11–20 plus GPT.

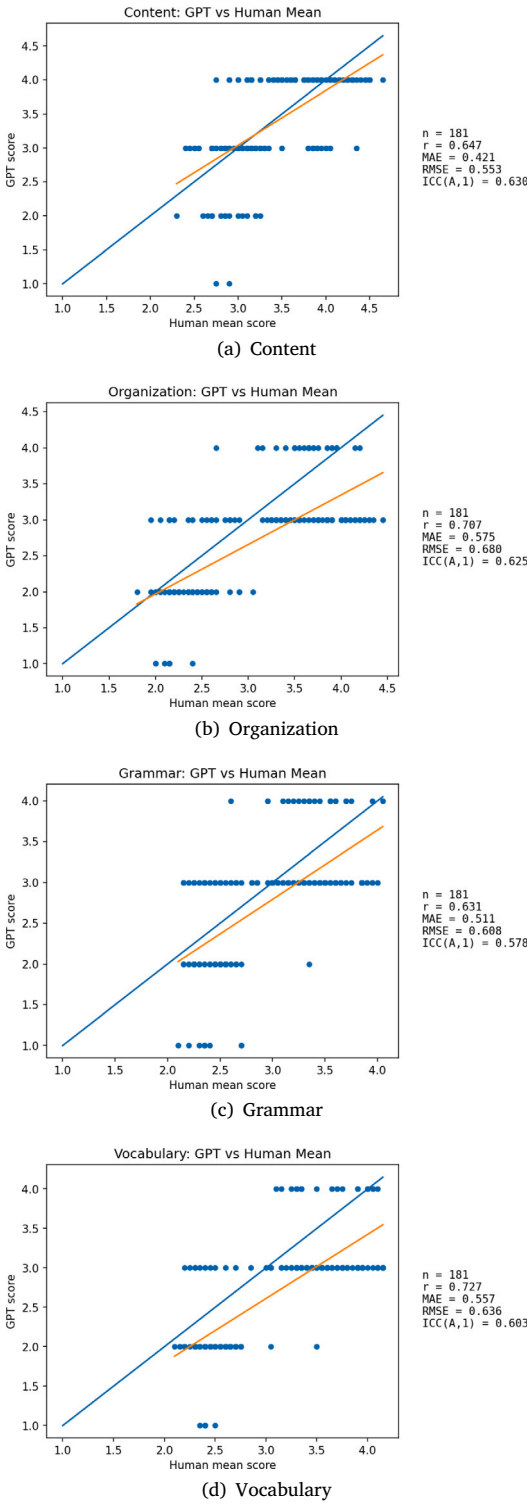

Fig. 4. GPT vs. human mean scores by criterion.

line, indicating a general tendency toward stricter scoring relative to human raters, especially in the higher score range, where the deviation was largest. This downward bias, together with the visibly narrower score spread, suggests more consistent yet range-compressed ratings, implying reduced discriminative sensitivity for very high or very low quality essays.

Rater-wise GPT–human correlations (Appendix Table A.8) broadly mirrored this ordering: most raters aligned better with GPT on *Organization/Vocabulary* than on *Grammar*, with a small number of low-alignment cases (e.g., Rater 19). Comparative ICCs (Table 6) indicated that adding GPT as the 21st rater left human-only reliability essentially unchanged (human-only: 0.38–0.56; humans+GPT: 0.38–0.55), suggesting broad alignment with the collective human standard and somewhat stronger agreement with NES than with NNES (notably on *Organization*).

4. Discussion

This study compared a GPT-based AES system with NES and NNES raters on four analytic criteria, *Content*, *Organization*, *Grammar*, and

Vocabulary, to address: (1) relative severity among GPT, NES, and NNES; (2) alignment of criterion level between GPT and human ratings; and (3) whether GPT exhibits rater–criterion biases comparable to or distinct from those of human raters.

Overall, the GPT-based AES demonstrated consistent performance across criteria, consistent with reports that LLMs can approach near-human-level reliability under analytic, rubric-based conditions [26,27] and thereby offer objective support for L2 writing assessment [16,31]. However, alignment was construct-dependent rather than uniform.

The model showed the strongest linear alignment with human raters on *Vocabulary* and *Organization*, indicating that GPT captured relative quality rankings in lexical and structural dimensions with human-like consistency. Meanwhile, GPT's *Content* scores yielded the smallest mean error and the highest absolute-agreement reliability, suggesting highly stable content evaluations that closely approximate human scoring magnitudes. By contrast, *Grammar* yielded the weakest correspondence, echoing observations that LLMs may diverge from human raters in grammatical sensitivity and error weighting [11,15,18]. This criterion-dependent variability is consistent with findings that LLMs emulate human-like scoring tendencies yet remain sensitive to the construct being judged and to interface or task design factors [24,25].

Across four criteria, GPT's mean scores were generally lower than human averages, with the largest discrepancies in *Organization*, *Grammar*, and *Vocabulary*, particularly at the upper end of the score range. This downward bias implies that GPT applied slightly stricter scoring standards than human raters in these criteria. In addition, GPT exhibited a narrower score variance than human raters, indicating higher internal consistency but reduced discriminative range. Practically, while GPT provides stable and reliable evaluations, it may be less sensitive in differentiating essays of exceptionally high or low quality, especially in structural and linguistic dimensions.

GPT's stricter stance on *Content* also parallels reports that, while LLMs may appear lenient on high-quality texts, they can be less forgiving of departures from expected norms, particularly in highly creative or culturally situated writing [16,17,32]. In other words, GPT may mitigate some facets of human subjectivity yet lack deeper critical or cultural sensitivity in certain contexts. Thus, while GPT's consistency can counterbalance the variability of human raters, further calibration is needed to expand its effective scoring range and to improve detection of fine-grained issues.

Human raters, by contrast, displayed greater variability in severity and criterion-specific tendencies. NES raters were generally stricter overall, whereas NNES raters showed relative leniency on *Content* but stricter judgments on *Organization* and *Grammar*. These findings align with established evidence of rater variability in EFL writing assessment [3,5,8–10]. Notably, while our data indicate higher overall severity among NES raters, Kim and Gennaro [7] reported stricter NNES ratings. Differences in task types and rating contexts (compare and contrast essays in 2012 and the multi-task design here), as well as interactions with rater background, likely mediate these effects [12,33,34]. The observed dispersion also accords with MFRM applications showing that some LLMs can exhibit lower rater effects than individual human raters, though such advantages are model- and context-dependent [26].

Notably, GPT's stronger alignment on *Organization* and *Vocabulary* but weaker alignment on *Grammar* maps to findings that LLMs may provide uneven diagnostic granularity across rubric dimensions. Prior work shows that GPT-4 can accurately identify issues, yet feedback specificity tends to be richer for grammatical/lexical phenomena and more generic for task fulfillment and coherence, with occasional misclassification in language-form judgments [18]. These asymmetries caution against inferring analytic-level quality from a single global agreement figure and highlight the need to calibrate dimension weights and decision thresholds when deploying LLM-assisted rating.

While GPT showed a content-strict but otherwise unbiased pattern, this selective tendency echoes broader evidence that LLM-based raters, though generally consistent overall, may vary subtly across language or learner groups. For instance, GPT's scoring has been shown to be reliable in multilingual settings such as Turkish L2 writing [22], yet other studies report lower agreement for certain L1s groups [20] and cross-L1 disparities even after fine-tuning [21]. High global correlations therefore do not guarantee fairness for all subgroups, highlighting the need for systematic bias auditing by dimension and learner background.

The weakest GPT–human alignment on *Grammar* is notable given the presumed objectivity of grammatical accuracy. One explanation is that GPT and human raters may operationalize "grammar" differently. Human raters sometimes discount minor errors when overall fluency or content quality is strong, whereas GPT may prioritize surface correctness and underweight syntactic variety. Prior studies document variable or misclassified judgments on grammatical dimensions [18,24], over- or under-sensitivity to formal features under rubric-based scoring [27], and reliance on distributional regularities rather than rule-governed grammaticality, especially for non-native writing [16,17]. In addition, prompt instructions may not have sufficiently constrained GPT's grammatical criteria, producing inconsistent weighting of errors versus complexity. Even seemingly objective criteria can involve latent subjectivity, reinforcing the need for targeted calibration in LLM-based analytic scoring.

Building upon these findings, this study offers several significant strengths and contributions to the literature on LLM-assisted analytic scoring. First, this study provides a criterion-level, facet-sensitive comparison across four analytic criteria. By disentangling alignment patterns across Content, Organization, Grammar, and Vocabulary, the findings demonstrate that GPT does not uniformly approximate human raters. Instead, its alignment is construct-dependent. It is strongest in Vocabulary and Organization, most stable in absolute agreement for Content, and weakest in Grammar. This criterion-specific mapping advances the field beyond global reliability claims and offers a more nuanced validity argument for LLM-based AES. Second, the study identifies GPT's systematic severity pattern and reduced score dispersion relative to human raters. While previous studies often emphasize near-human reliability, fewer have examined variance compression and its operational implications. The narrower variance observed here suggests that GPT functions as a consistency stabilizer, potentially reducing rater bias, yet may under-differentiate performances at the score extremes. This trade-off between internal consistency and discriminative range has direct implications for high-stakes and fine-grained assessment contexts. Third, by simultaneously analyzing NES and NNES rater behavior alongside GPT, this study situates LLM-based scoring within a human rater variability framework. The findings reveal that GPT exhibits a content-strict but otherwise balanced pattern compared with background-related severity shifts observed among human raters. This comparative positioning clarifies that GPT does not simply replicate human bias but displays its own construct-sensitive tendencies. Such evidence advances current debates on fairness and human–AI integration by demonstrating that LLM-based scoring systems exhibit construct-sensitive patterns distinct from, rather than merely replicating, human rater biases. Taken together, these contributions advance the evaluation of GPT-based AES within a construct validity framework, demonstrating that LLM-based scoring systems must be examined criteria by criteria rather than through aggregate reliability indices alone.

While the findings provide important insights into GPT-based analytic scoring, several limitations should be acknowledged. First, the study is based on a specific task design and rubric structure. Although the four analytic criteria represent common constructs in L2 writing assessment, the results may not generalize to other genres, proficiency levels, or holistic scoring frameworks. Different prompt designs, task types, or rubric weightings could meaningfully alter alignment patterns. Second, the GPT scoring procedure relied on a fixed prompt configuration. As prior research suggests, prompt can substantially influence LLM outputs. The present findings therefore reflect the performance of GPT under a particular prompting strategy rather than an intrinsic

property of the model. Alternative prompting, iterative refinement, or fine-tuning procedures may yield different severity and variance patterns. Third, the study focused on agreement and severity indices but did not directly examine fairness across learner subgroups such as by L1 background, proficiency band, or demographic characteristics. Although overall alignment was strong, subgroup-level bias or differential functioning cannot be ruled out. Future studies should incorporate systematic bias auditing and differential item functioning analyses to further validate fairness claims. Fourth, while the comparison included NES and NNES raters, the rater sample size and contextual background were bounded by the specific assessment setting. Broader rater pools or cross-background comparisons would provide stronger generalizability regarding human–AI severity contrasts. Finally, the study evaluated score-level alignment rather than the interpretability of GPT's reasoning processes. Because LLM scoring decisions are generated through probabilistic inference rather than transparent rule-based evaluation, questions remain regarding construct representation and cognitive validity. Further research should examine explanation consistency, decision traceability, and the pedagogical interpretability of LLM-based feedback. These limitations underscore that GPT-based AES should not be evaluated solely through aggregate reliability indices but within a broader construct validity framework encompassing fairness, interpretability, and contextual sensitivity. Within such a framework, GPT-based AES functions most effectively as a calibrated and monitored component of hybrid human–AI assessment systems rather than as a standalone replacement for human judgment.

5. Conclusion

The study contribute to the growing literature on AI-assisted writing assessment by demonstrating that GPT-based scoring must be examined at the criterion level rather than inferred from overall agreement statistics. By situating GPT alongside NES and NNES raters within a shared severity framework, this study clarifies that LLM-based raters exhibit construct-sensitive patterns distinct from human variability. This perspective reframes GPT-based AES not as a wholesale substitute for human judgment, but as a calibrated scoring agent whose strengths and limitations vary across analytic criteria. Pedagogically and operationally, GPT-based AES can function as a consistency stabilizer, helping to dampen rater variability in large-scale or classroom-based assessments. However, its content-strict profile, reduced score dispersion, and weaker grammatical alignment require ongoing calibration, fairness auditing, and interpretive transparency to maintain construct validity. Several limitations suggest directions for future research. This study used a single model version in a zero-shot, rubric-prompted configuration; future work should explore alternative prompting strategies, fine-tuning, and cross-model comparisons to evaluate robustness. Cross-task and cross-population generalizability must also be examined to assess fairness across genres, proficiency levels, and learner backgrounds. In sum, GPT-based AES provides stable and analytically structured scoring that can mitigate certain forms of human rater variability. Yet, its criterion-specific deviations and potential subgroup sensitivities reinforce that LLM-based scoring systems require principled validation, monitoring, and human oversight. Within such a construct validity framework, GPT can serve as a credible and auditable partner in L2 writing assessment.

CRediT authorship contribution statement

Hsiang-Ning Wu: Writing – original draft, Methodology, Conceptualization. **Man-Ni Chu:** Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Jia-Lien Hsu:** Writing – original draft, Methodology, Formal analysis, Conceptualization.

Funding

The research leading to these results received funding from the National Science and Technology Council, Taiwan under Grant Agreement NSTC-113-2221-E-030-013, NSTC-114-2410-H-030-018, NSTC-114-2221-E-030-007.

Ethics statement

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Fu Jen Catholic University (protocol code FJU-IRB NO: C111010, 2022/11/30.)

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to express our gratitude to the students who participated in the essay writing, whose contributions were invaluable to this research. We also acknowledge the assistance provided by the 20 raters from the University of Leeds for their generous support and dedication in evaluating the essays. Your time and effort have been crucial in making this study possible.

Appendix

A.1. Rating rubric

The rating rubric for the writing test is shown below (see Table A.1).

A.2. Full GPT-based AES prompt

See Table A.2.

A.3. Descriptive statistics for rater severity

See Table A.3.

A.4. Descriptive statistics for criteria difficulty

Table A.4 provides insights into the difficulty associated with the four criteria logit measures. *Grammar* was the most challenging criterion, followed by *Organization*, *Vocabulary*, and *Content*.

A.5. Group-level rater severity between all humans and GPT

See Tables A.5 and A.6.

A.6. Rater × criterion bias/interaction with BH correction

See Table A.7.

A.7. Pearson correlations between GPT and individual human raters

See Table A.8.

Table A.1
Rating rubric.

Criteria	Level	Description
Content	5 Excellent	The essay fulfills the writing task and addresses the topic richly and fully.
	4 Strong	The essay fulfills the writing task and addresses the topic appropriately.
	3 Good	The essay generally fulfills the writing task and addresses the topic adequately.
	2 Acceptable but Limited	The essay only partially fulfills the writing task and addresses part of the topic.
	1 Weak	The essay shows a vague or incomplete understanding of the writing task and topic.
	0 Seriously Flawed	The essay suggests a lack of understanding of the writing task and topic.
Organization	5 Excellent	Coherence between paragraphs, sentences, and ideas is successfully achieved through a variety of methods.
	4 Strong	Coherence is adequate and achieved through the use of transitional words and phrases.
	3 Good	The essay is generally well organized; there may be a limited lack of coherence and difficulty with para- graphing.
	2 Acceptable but Limited	Paragraphs often lack coherence; sentences are not well connected; paragraphs are not appropriately connected to each other.
	1 Weak	There is little coherence within and across paragraphs; sentences are strung together somewhat haphazardly; there is little or no clear attempt to connect paragraphs.
	0 Seriously Flawed	There is no attempt to divide the essay into conceptual paragraphs, or the paragraphs are unrelated.
Grammar	5 Excellent	Except for rare minor errors, the writer's use of grammar is precise and does not obscure meaning.
	4 Strong	Effective and appropriate use of grammar in general; minor errors in articles, agreements, verb forms, and no incomplete sentences; meaning is never obscured.
	3 Good	Competent control of grammar in general; there may be errors in article use and verb agreement and several errors in verb form; Errors affect clarity occasionally, but generally do not obscure meaning.
	2 Acceptable but Limited	Limited competence in use of gram- mar; occasional major errors or frequent minor errors in grammar occasionally hinder comprehension.
	1 Weak	Frequent errors in verb formation, articles, and incomplete sentences; sentence structures make comprehension difficult.
	0 Seriously Flawed	Numerous problems in all areas of grammar; sentence construction is so poor that sentences are often incomprehensible.
Vocabulary	5 Excellent	There is a wide range of appropriately used vocabulary; there are few problems with word choice.
	4 Strong	Vocabulary is broad and is generally used appropriately.
	3 Good	Vocabulary is generally adequate but may sometimes be inappropriately used.
	2 Acceptable but Limited	Vocabulary shows some flexibility but may be inaccurate or repetitive in places.
	1 Weak	Vocabulary is quite basic and word choice is inaccurate. The writer may rely on repeating words and expressions from the prompt.
	0 Seriously Flawed	The essay suggests a lack of understanding of the writing task and topic.

Table A.2
Scoring prompt.

I would like you to mark essays written by EFL learners based on the following rubric. Evaluate and provide scores in the categories of Content, Organization, Grammar, and Vocabulary. Use a scale from 0 to 5, where 0 indicates poor performance and 5 indicates excellent performance.

[Rubric in text format inserted here]

Essay:
[Each of the 181 essays inserted here through a for loop]

Output the scores as follows:
Content:?
Organization:?
Grammar:?
Vocabulary:?

Model: GPT-4 (run date: July 27, 2024); temperature = 0; top_p = 1.0; max_tokens = 150; n = 1.

Table A.3
Rater severity: descriptive statistics.

	Rater	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd
19	NNES	2.16	2.14	−.70	.07	1.94	9.0	2.01	9.0
20	NNES	2.79	2.78	1.45	.07	1.41	6.6	1.45	7.0
21	GPT	2.90	2.91	1.82	.07	.89	−2.2	.93	−1.2
13	NNES	2.97	2.98	2.02	.06	.70	−6.4	.69	−6.5
4	NES	3.08	3.10	2.36	.06	.94	−1.1	.93	−1.4
12	NNES	3.08	3.10	2.36	.06	.78	−4.7	.79	−4.4
11	NNES	3.09	3.11	2.39	.06	.83	−3.4	.82	−3.6
17	NNES	3.10	3.11	2.41	.06	.70	−6.4	.70	−6.4
5	NES	3.12	3.14	2.47	.06	.82	−3.6	.81	−3.9
16	NNES	3.13	3.14	2.49	.06	.71	−6.2	.69	−6.7
15	NNES	3.13	3.14	2.49	.06	.85	−3.1	.83	−3.6
1	NES	3.13	3.15	2.51	.06	.80	−4.1	.85	−3.1
7	NES	3.24	3.26	2.83	.06	.92	−1.5	.95	−1.0
3	NES	3.24	3.26	2.83	.06	.86	−2.9	.88	−2.4
8	NES	3.29	3.31	2.97	.06	.88	−2.3	.90	−1.9
14	NNES	3.29	3.32	2.97	.06	.83	−3.4	.82	−3.8
6	NES	3.33	3.36	3.09	.06	.80	−4.1	.82	−3.9
2	NES	3.33	3.36	3.09	.06	.97	−.6	.98	−.3
9	NES	3.44	3.47	3.39	.06	.96	−.6	.97	−.6
18	NNES	3.56	3.61	3.76	.06	2.05	9.0	2.00	9.0
10	NES	3.78	3.85	4.40	.06	1.09	1.7	1.07	1.3
M	(Count: 21)	3.15	3.17	2.54	.06	.99	−1.5	.99	−1.4
S.D.	(Population)	.31	.33	.96	.00	.36	4.4	.36	4.5
S.D.	(Sample)	.32	.33	.98	.00	.37	4.6	.37	4.6

RMSE .06 Adj S.D. .96 Separation 14.92 Reliability 1.00.
Fixed (all same) chi-squared: 4294.8 d.f.: 20 significance: .00.

Table A.4
Descriptive statistics for criteria difficulty.

	Criterion	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd
1	Content	3.48	3.52	−.97	.03	1.04	1.7	1.05	2.1
4	Vocabulary	3.10	3.11	.13	.03	.86	−6.8	.88	−5.6
2	Organization	3.10	3.11	.16	.03	1.03	1.2	1.04	1.5
3	Grammar	2.92	2.92	.68	.03	1.00	.0	1.02	.6
M	(Count: 21)	3.15	3.17	.00	.03	.98	−1.0	.99	−.3
S.D.	(Population)	.21	.22	.60	.00	.07	3.4	.07	3.1
S.D.	(Sample)	.24	.25	.70	.00	.08	4.0	.08	3.6

RMSE .03 Adj S.D. .60 Separation 21.55 Reliability 1.00.
Fixed (all same) chi-squared: 1869.3 d.f.: 3 significance: .00.

Table A.5
All Humans (pooled) vs. GPT: group-level rater severity means with standard errors.

Group	Mean (logits)	SE	95% CI L	95% CI U
All Humans (N = 20)	2.64	0.01	2.62	2.66
GPT	1.82	0.07	1.68	1.96

Note. The All Humans mean is the inverse-variance-weighted average across 20 human raters ($w_i = 1/\mathrm{SE}_i^2$), with $SE(\bar{M}) = \sqrt{1/\sum_i w_i}$. GPT values are from the rater measurement report. CIs are Wald intervals: $\bar{M} \pm 1.96 \times SE(\bar{M})$. Values rounded to two decimals.

Table A.6
Pooled contrast on rater severity (logits): All Humans vs. GPT.

Contrast	Δ	SE(Δ)	t	95% CI L	95% CI U	p
All Humans–GPT	−0.82	0.07	−11.49	−0.96	−0.68	< 0.001

Note. Δ is the difference in mean severity logits (All Humans minus GPT). $SE(\Delta)$ was computed as $\sqrt{SE^2_{\text{All Humans}} + SE^2_{\text{GPT}}}$. CIs are Wald intervals. Values rounded to two decimals.

Table A.7
Rater × criterion bias: full list with raw p and BH-adjusted q (two-tailed).

Rater	Group	Criterion	Bias	t	p_{raw}	BH-q
GPT	GPT	Content	+0.58	+4.64	3.48×10^{-6}	2.93×10^{-5}
GPT	GPT	Grammar	+0.17	+1.25	0.211	0.347
GPT	GPT	Organization	−0.36	−2.69	0.0071	0.0241
GPT	GPT	Vocabulary	−0.45	−3.38	0.00073	0.00318
Rater 1	NES	Content	−0.02	−0.16	0.873	0.923
Rater 1	NES	Grammar	+0.01	+0.10	0.922	0.948
Rater 1	NES	Organization	−0.38	−2.93	0.0034	0.0124
Rater 1	NES	Vocabulary	+0.38	+3.00	0.0027	0.0105
Rater 2	NES	Content	+0.37	+2.92	0.0035	0.0128
Rater 2	NES	Grammar	+0.14	+1.12	0.263	0.407
Rater 2	NES	Organization	−0.71	−5.54	3.02×10^{-8}	6.35×10^{-7}
Rater 2	NES	Vocabulary	+0.19	+1.55	0.121	0.222
Rater 3	NES	Content	−0.56	−4.47	7.82×10^{-6}	5.03×10^{-5}
Rater 3	NES	Grammar	+0.50	+3.94	8.18×10^{-5}	5.02×10^{-4}
Rater 3	NES	Organization	−0.40	−3.14	0.00168	0.00722
Rater 3	NES	Vocabulary	+0.47	+3.74	0.000182	0.00100
Rater 4	NES	Content	+1.12	+8.76	$< 10^{-15}$	$< 10^{-15}$
Rater 4	NES	Grammar	−0.61	−4.53	5.90×10^{-6}	4.50×10^{-5}
Rater 4	NES	Organization	−0.16	−1.25	0.211	0.347
Rater 4	NES	Vocabulary	−0.44	−3.35	0.000802	0.00349
Rater 5	NES	Content	+0.75	+5.91	3.42×10^{-9}	1.15×10^{-7}
Rater 5	NES	Grammar	−0.36	−2.73	0.00635	0.0236
Rater 5	NES	Organization	−0.17	−1.34	0.180	0.314
Rater 5	NES	Vocabulary	−0.26	−2.03	0.0423	0.116
Rater 6	NES	Content	+0.38	+2.96	0.00309	0.0116
Rater 6	NES	Grammar	−0.02	−0.12	0.906	0.941
Rater 6	NES	Organization	−0.46	−3.64	0.000269	0.00116
Rater 6	NES	Vocabulary	+0.10	+0.82	0.411	0.534
Rater 7	NES	Content	+0.31	+2.48	0.0131	0.0442
Rater 7	NES	Grammar	+0.10	+0.76	0.446	0.567
Rater 7	NES	Organization	−0.53	−4.11	3.96×10^{-5}	2.08×10^{-4}
Rater 7	NES	Vocabulary	+0.11	+0.88	0.378	0.501
Rater 8	NES	Content	+0.17	+1.37	0.170	0.298
Rater 8	NES	Grammar	+0.07	+0.56	0.576	0.668
Rater 8	NES	Organization	−0.47	−3.70	0.000213	0.000913
Rater 8	NES	Vocabulary	+0.22	+1.78	0.0752	0.160
Rater 9	NES	Content	−0.29	−2.31	0.0210	0.0633
Rater 9	NES	Grammar	−0.01	−0.07	0.946	0.963
Rater 9	NES	Organization	+0.30	+2.39	0.0168	0.0528
Rater 9	NES	Vocabulary	−0.01	−0.05	0.960	0.974
Rater 10	NES	Content	+0.69	+4.84	1.30×10^{-6}	1.21×10^{-5}
Rater 10	NES	Grammar	+0.08	+0.64	0.522	0.631
Rater 10	NES	Organization	+0.08	+0.64	0.522	0.631
Rater 10	NES	Vocabulary	−0.74	−5.88	4.10×10^{-9}	1.15×10^{-7}
Rater 11	NNES	Content	−0.19	−1.54	0.122	0.223
Rater 11	NNES	Grammar	−0.22	−1.69	0.0908	0.181
Rater 11	NNES	Organization	+0.46	+3.68	0.000237	0.00105
Rater 11	NNES	Vocabulary	−0.06	−0.47	0.640	0.714
Rater 12	NNES	Content	−0.22	−1.72	0.0853	0.175
Rater 12	NNES	Grammar	−0.31	−2.29	0.0222	0.0666
Rater 12	NNES	Organization	+0.55	+4.38	1.19×10^{-5}	1.01×10^{-4}
Rater 12	NNES	Vocabulary	−0.05	−0.40	0.690	0.754
Rater 13	NNES	Content	−0.42	−3.29	0.000999	0.00421
Rater 13	NNES	Grammar	+0.57	+4.38	1.19×10^{-5}	1.01×10^{-4}
Rater 13	NNES	Organization	+0.11	+0.84	0.400	0.518
Rater 13	NNES	Vocabulary	−0.22	−1.68	0.0925	0.182
Rater 14	NNES	Content	−0.15	−1.18	0.238	0.377
Rater 14	NNES	Grammar	−0.08	−0.62	0.536	0.639
Rater 14	NNES	Organization	+0.21	+1.67	0.0944	0.185
Rater 14	NNES	Vocabulary	+0.01	+0.11	0.913	0.945
Rater 15	NNES	Content	−0.35	−2.76	0.0058	0.0214
Rater 15	NNES	Grammar	−0.24	−1.83	0.0668	0.146
Rater 15	NNES	Organization	+0.55	+4.36	1.30×10^{-5}	1.05×10^{-4}
Rater 15	NNES	Vocabulary	+0.03	+0.23	0.820	0.896
Rater 16	NNES	Content	−0.20	−1.60	0.109	0.208
Rater 16	NNES	Grammar	−0.13	−1.02	0.307	0.462
Rater 16	NNES	Organization	+0.46	+3.63	0.000276	0.00118
Rater 16	NNES	Vocabulary	−0.13	−1.01	0.312	0.467
Rater 17	NNES	Content	−0.15	−1.20	0.229	0.366
Rater 17	NNES	Grammar	−0.32	−2.36	0.0183	0.0568
Rater 17	NNES	Organization	+0.38	+3.00	0.0027	0.0105
Rater 17	NNES	Vocabulary	+0.07	+0.52	0.600	0.683
Rater 18	NNES	Content	−0.61	−4.85	1.23×10^{-6}	1.15×10^{-5}
Rater 18	NNES	Grammar	+0.49	+3.89	0.00010	0.000603

(*continued on next page*)

Table A.7 (*continued*).

Rater	Group	Criterion	Bias	t	p_{raw}	BH-q
Rater 18	NNES	Organization	+0.01	+0.06	0.955	0.970
Rater 18	NNES	Vocabulary	+0.10	+0.76	0.446	0.567
Rater 19	NNES	Content	−0.76	−5.46	4.76×10^{-8}	8.00×10^{-7}
Rater 19	NNES	Grammar	+0.19	+1.37	0.170	0.298
Rater 19	NNES	Organization	−0.16	−1.13	0.258	0.404
Rater 19	NNES	Vocabulary	+0.73	+5.27	1.36×10^{-7}	1.91×10^{-6}
Rater 20	NNES	Content	−0.49	−3.81	0.000139	0.000820
Rater 20	NNES	Grammar	−0.13	−0.93	0.352	0.510
Rater 20	NNES	Organization	+0.67	+5.17	2.34×10^{-7}	2.81×10^{-6}
Rater 20	NNES	Vocabulary	−0.03	−0.23	0.819	0.895

Notes. Two-tailed $p_{raw} = 2\{1 - \Phi(|t|)\}$ under large-sample normal approximation. BH uses Benjamini–Hochberg across all 84 cells (family defined as all rater–criterion tests). Highlighted cells in the main text heatmap use the dual rule (pre-BH); after BH ($q < .05$) the flagged set was unchanged.

Table A.8
Pearson correlations between GPT and individual human raters.

Rater	Content	Organization	Grammar	Vocabulary
1	0.51	0.65	0.37	0.53
2	0.17	0.64	0.26	0.50
3	0.54	0.64	0.28	0.44
4	0.12	0.62	0.58	0.70
5	0.41	0.65	0.61	0.70
6	0.40	0.64	0.31	0.53
7	0.31	0.63	0.28	0.38
8	0.23	0.62	0.28	0.49
9	0.61	0.63	0.62	0.61
10	0.65	0.58	0.60	0.62
11	0.52	0.55	0.46	0.63
12	0.62	0.56	0.45	0.62
13	0.51	0.49	0.43	0.50
14	0.58	0.62	0.50	0.62
15	0.48	0.56	0.44	0.63
16	0.57	0.53	0.45	0.59
17	0.56	0.58	0.49	0.62
18	0.62	0.65	0.71	0.74
19	0.22	0.04	0.05	0.00
20	0.36	0.36	0.33	0.35
GPT_avg	0.449	0.562	0.424	0.540

Note. The final row (GPT_avg) reports the mean correlation between GPT and each of the 20 human raters.

Data availability statement

The dataset, including the GPT-model coding, the scores from 20 human raters and the GPT model, as well as the essay responses rated by these participants, is available on Google Drive. The data can be accessed at https://drive.google.com/drive/folders/1fRVUr8S-veXXeUkoQDWnpUjOnI163hyM?usp=share_link Access will be granted upon reasonable request to the author.

References

[1] Burstein J. The E-rater® scoring engine: Automated essay scoring with natural language processing. In: Shermis MD, Burstein JC, editors. Automated essay scoring: a cross disciplinary approach. Mahwah, NJ: Lawrence Erlbaum Associates Publishers; 2003, p. 113–21.

[2] Hamp-Lyons L. Fourth generation writing assessment. In: Silva T, Matsuda PK, editors. On second language writing. Routledge; 2012, p. 117–27.

[3] Knoch U. Diagnostic assessment of writing: A comparison of two rating scales. Lang Test 2009;26(2):275–304.

[4] Lumley T, McNamara TF. Rater characteristics and rater bias: Implications for training. Lang Test 1995;12(1):54–71.

[5] Schaefer E. Rater bias patterns in an EFL writing assessment. Lang Test 2008;25(4):465–93.

[6] Linacre JM. Rasch models from objectivity: A generalization. Proc Int Object Meas Work 1989.

[7] Kim AY, Gennaro Kd. Scoring behavior of native vs. non-native speaker raters of writing exams. Lang Res 2012;48(2):319–42.

[8] Kondo-Brown K. A FACETS analysis of rater bias in measuring Japanese second language writing performance. Lang Test 2002;19(1):3–31.

[9] Wigglesworth G. Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. Lang Test 1993;10(3):305–19.

[10] Wigglesworth G. Patterns of rater behaviour in the assessment of an oral interaction test. Aust Rev Appl Linguist 1994;17(2):77–103.

[11] Escalante J, Pack A, Barrett A. AI-generated feedback on writing: Insights into efficacy and ENL student preference. Int J Educ Technol High Educ 2023;20(1):57.

[12] Hoang GTL, Kunnan AJ. Automated essay evaluation for English language learners: A case study of MY access. Lang Assess Q 2016;13(4):359–76.

[13] Hussein MA, Hassan H, Nassef M. Automated language essay scoring systems: A literature review. PeerJ Comput Sci 2019;5:e208.

[14] Atkinson J, Palma D. An LLM-based hybrid approach for enhanced automated essay scoring. Sci Rep 2025;15(1):14551.

[15] Shin D, Lee JH. Exploratory study on the potential of ChatGPT as a rater of second language writing. Educ Inf Technol 2024;1–23.

[16] Yavuz F, Çelik Ö, Yavaş Çelik G. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. Br J Educ Technol 2024;00:0–17, URL https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13494.

[17] Banihashem SK, Kerman NT, Noroozi O, Moon J, Drachsler H. Feedback sources in essay writing: peer-generated or AI-generated feedback? Int J Educ Technol High Educ 2024;21(1):23.

[18] Saricaoglu A, Bilki Z. The capacity of ChatGPT-4 for L2 writing assessment: A closer look at accuracy, specificity, and relevance. Annu Rev Appl Linguist 2025;1–21.

[19] Kim K, Lee JH, Shin D. The potential advantages of using an LLM-based chatbot for automated writing evaluation for English teaching practitioners. Lang Learn Technol 2025;29(1):1–12, URL https://hdl.handle.net/10125/73628.

[20] Yancey KP, Laflair G, Verardi A, Burstein J. Rating short L2 essays on the CEFR scale with GPT-4. In: Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023). 2023, p. 576–84.

[21] Liu Y, Qi H, Lu X. Enhancing GPT-based automated essay scoring: the impact of fine-tuning and linguistic complexity measures. Comput Assist Lang Learn 2025;1–20.

[22] Aydın B, Kışla T, Tan Elmas N, Bulut O. Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays. System 2025;133:103784. http://dx.doi.org/10.1016/j.system.2025.103784.

[23] Yamashita T. Exploring potential biases in GPT-4o's ratings of English language learners' essays. Lang Test 2025;42(3):344–58.

[24] Manning J, Baldwin J, Powell N. Human versus machine: The effectiveness of ChatGPT in automated essay scoring. Innov Educ Teach Int 2025;62(5):1500–13. http://dx.doi.org/10.1080/14703297.2025.2469089.

[25] Uyar AC, Büyükahıska D. Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. Int J Assess Tools Educ 2025;12(1):20–32.

[26] Jiao H, Song D, Lee W-C. Comparing human and AI rater effects using the many-facet rasch model. 2025, ArXiV. Org.

[27] Wang Y, Huang J, Du L, Guo Y, Liu Y, Wang R. Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. Comput Educ: Artif Intell 2025;100481.

[28] Engelhard Jr G. The measurement of writing ability with a many-faceted Rasch model. Appl Meas Educ 1992;5(3):171–91.

[29] McNamara TF. Measuring second language performance. Longman; 1996.

[30] Linacre JM. A user's guide to FACETS: Rasch-model computer programs [Software manual]. Chicago: Winsteps.com; 2004.

[31] Mizumoto A, Eguchi M. Exploring the potential of using an AI language model for automated essay scoring. Res Methods Appl Linguist 2023;2(2):100050.

[32] Kohnke L, Moorhouse BL, Zou D. ChatGPT for language teaching and learning. Relc J 2023;54(2):537–50.

[33] Connor-Linton J. Crosscultural comparison of writing standards: American ESL and Japanese EFL. World Englishes 1995;14(1):99–115.
[34] Shi L. Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. Lang Test 2001;18(3):303–25.

Hsiang-Ning Wu holds a masters degree in Linguistics of Fu Jen Catholic University, Taiwan. She has extensive experience in teaching English as a Foreign Language (EFL) and has participated in interdisciplinary research projects that combine linguistics with computational techniques. She is passionate about improving the reliability and validity of language evaluations through the integration of AI tools, while also ensuring that these technologies complement the nuanced insights provided by human raters.

Man-Ni Chu is a linguist, teaching at the Graduate Institute of Cross-Culture Studies MA. Program in Linguistics, Fu Jen Catholic University, Taiwan. Her interests are in experimental phonetics, especially using quantitative methodology to solve the puzzles of sound changes. A computational modeling of language processing or language acquisition is her another attempts.

Jia-Lien Hsu is a Professor of Department of Computer Science and Information Engineering at Fu Jen Catholic University, Taiwan. His research interests include data engineering, music informatics, medical informatics.