AI or Human? A Mixed-Methods Study on Feedback Effectiveness in IELTS Academic Writing Task 1

Yue,Xiwen* Zhou,Nan

Glasgow College, University of Electronic Science and Technology of China (UESTC), Chengdu, Sichuan, 611731, China

Abstract: This study investigates the comparative impacts of AI-generated and human rater feedback on IELTS Academic Writing Task 1 performance among Chinese EFL learners, addressing growing interest in the pedagogical potential and limitations of automated writing evaluation. Using a convergent parallel mixed-methods design, ten undergraduate students were assigned to either an AI (ChatGPT-4) or human feedback group. Each student completed an initial writing task, received formative feedback, and revised their work accordingly. Quantitative results showed significant gains across all criteria: AI feedback led to the highest improvement in grammatical range (+27.6%) and task achievement (+15.2%), while human feedback produced greater gains in lexical resource (+16.7%), coherence (+12.5%), and overall performance (Cohen's d = 4.03). Thematic analysis of feedback and learner reflections showed that AI provided consistent, efficient surface-level correction, while human raters offered nuanced, exam-aligned guidance on rhetorical and discursive features. Learners expressed a strong preference for a hybrid feedback model. These findings support a learner-centered, integrated approach that leverages the complementary strengths of automation and human expertise to enhance performance in high-stakes academic writing assessment.

Keywords: IELTS academic writing; ChatGPT-4; Human vs AI feedback; Human vs AI raters; Mixed-methods research; Chinese EFL engineering students; L2 writing assessment

DOI: 10.62639/sspjiss01.20250206

1. Introduction

As large language models like ChatGPT-4 become increasingly embedded in second language assessment, their potential to support high-stakes academic writing—such as IELTS Task 1—demands critical scrutiny. While AI-generated feedback offers rapid, consistent support in grammar and lexical accuracy, it often lacks the discourse-level depth and contextual sensitivity of human raters (Dikli & Bleyle, 2014; Naismith et al., 2023). This study investigates how feedback from AI and certified IELTS examiners differs in focus and effectiveness, and how each influences learner performance and perception. Drawing on a mixed-methods design involving Chinese EFL undergraduates, the research contributes to a deeper understanding of AI's pedagogical role in L2 writing assessment and feedback.

(Manuscript NO.: JISS-25-6-62029)

Corresponding Author

Yue,Xiwen (1997-), female, Han nationality, native of Sichuan. She holds a Master's degree in Applied Linguistics from University College London and currently works at the Glasgow College, University of Electronic Science and Technology of China. Her research interests include language testing and assessment, business communication, technology - enhanced language learning, and multimodality & material development.

About the Author

Zhou,Nan (1992-), female, Mongolian nationality, native of Inner Mongolia. She holds a Master of Science degree in TESOL from the University of Edinburgh and currently works at the Glasgow College, University of Electronic Science and Technology of China. Her research interests include language testing and assessment, business communication, and material development.

2. Literature Review

(1) From rule-based AES to transformer models

The first generation of Automated Essay Scoring (AES) engines e.g., e-rater V.2;(Attali & Burstein, 2006) relied on a small, transparent feature set and a single scoring model applicable across prompts, achieving reliability comparable to trained human raters. Subsequent psychometric work showed that machine–human correlations can match inter-rater human correlations, yet subgroup differences (gender, ethnicity, L1) remain a persistent equity concern (Bridgeman et al., 2012; Weigle, 2013).

Deep-learning approaches have since surpassed earlier pipelines. A Bi-LSTM + RoBERTa architecture yielded higher Quadratic Weighted Kappa than prior systems and, critically, captured cohesion features overlooked by traditional NLP indices (Beseiso et al., 2021). These gains underline AES's value in large-scale assessment, where time, cost, and rater-fatigue pressures are acute (Stevenson & Phakiti, 2019; Shermis & Burstein, 2003).

(2) Reliability, objectivity and fairness

Meta-analytic reviews conclude that AES offers greater score consistency and freedom from idiosyncratic rater bias than human marking (Kim et al., 2024). Nonetheless, equity studies report modest yet non-negligible divergences across demographic subgroups (Bridgeman et al., 2012), underscoring the need for continuous bias diagnostics when deploying AES in high-stakes contexts such as IELTS.

(3) Automated Writing Evaluation (AWE) and feedback depth

Advances in AES have been paralleled by Automated Writing Evaluation systems that generate formative comments. While AWE can accelerate revision cycles, instructor–AWE comparisons reveal that human feedback remains richer in rhetorical and discourse-level guidance (Dikli & Bleyle, 2014). Cotos (2014) argues that genre-specific calibration is essential if AWE is to meet advanced L2 needs, a point echoed in classroom research agendas (Warschauer & Ware, 2006).

(4) Large language models as raters and tutors

The release of ChatGPT has catalysed a new research wave. GPT-3 achieved a 54 % exact-agreement rate with TOEFL benchmarks (Mizumoto & Eguchi, 2023). GPT-4, when given calibration exemplars, now approaches state-of-the-art AWE accuracy, yet agreement still fluctuates by L1 backgrounds and task types (Yancey et al., 2023). Regarding feedback, ChatGPT produces fluent, grammatically oriented comments that align with instructor focus on clarity and flow (Dai et al., 2023; Naismith et al., 2023). However, mixed-methods evidence shows it struggles with content relevance, logical progression, and integration of multiple data sources—critical skills for IELTS Task 1 (Kim et al., 2024; Xia et al., 2024).

(5) Gaps in IELTS-specific research

Despite abundant AES work on TOEFL and placement essays, little is known about (a) how ChatGPT 4.0-generated formative comments compare with certified IELTS examiners' feedback, and (b) how learners translate each feedback type into measurable band-score gains. Prior IELTS studies have typically treated AI as a scorer, not as an interactive feedback provider. Moreover, quantitative score shifts rarely triangulate with learners' perceptions, leaving the pedagogical impact of AI feedback under-documented.

(6) Rationale for the present study

Building on the reliability of modern AES and the emerging yet incomplete evidence on LLM feedback, the current research adopts a convergent parallel mixed-methods design to examine ten Chinese undergraduates' IELTS Task 1 performance and perceptions after receiving either ChatGPT-4 or human-examiner feedback. By integrating band-score trajectories with thematic analysis of feedback content and learner reflections, the study responds

directly to the literature's calls for:

1) Fine-grained feedback comparison (surface vs. discourse focus);

2) Learning-outcome evidence of learning outcomes beyond one-time score gains—i.e., whether feedback leads to measurable improvement in subsequent writing.

3) Learner-centered perspectives on AI versus human coaching in high-stakes L2 writing.

In doing so, it extends prior AES validity work (Attali & Burstein, 2006; Beseiso et al., 2021) and addresses the pedagogical questions raised by AWE scholarship (Cotos, 2014; Warschauer & Ware, 2006) within the under-researched IELTS writing context.

3. Methodology

(1) Research design

This study employed a convergent parallel mixed-methods design (Creswell & Clark, 2017), which enabled simultaneous collection and integration of quantitative and qualitative data. This approach allowed direct comparison and triangulation of writing performance changes, feedback characteristics, and learner perceptions—supporting a comprehensive understanding of the impact of feedback type. The design was selected to address three core research questions:

1) How does feedback from AI and human raters differ in focus and depth?

2) How does each type of feedback promote students' IELTS writing performance?

3) How do learners perceive and respond to AI and human feedback?

(2) Participants

Ten first year EFL undergraduate students at a Sino-foreign English-Medium Instruction (EMI) university in China voluntarily participated in the study. Participants were evenly assigned to either the AI or human feedback group. All students had IELTS-equivalent proficiency levels ranging from Band 5.0 to 6.5 and had prior instruction on IELTS writing conventions. The cohort was relatively homogeneous in institutional background. Informed consent was obtained, and all data were anonymized.

(3) Materials

The writing task was adapted from an official IELTS Academic Task 1 prompt, requiring a description and comparison of trends in a line graph depicting participation in various social activities over a 20-year period. Human feedback was based on the official IELTS Writing Band Descriptors (Cambridge Assessment English, 2020) and provided by three certified IELTS examiners. AI feedback was generated by ChatGPT-4.0 using structured prompts aligned to each IELTS criterion, such as: "Evaluate this writing in terms of coherence and cohesion using IELTS criteria."

(4) Procedure the study consisted of three phases

1) Initial Writing and Feedback Phase: All students completed the same Task 1 prompt under timed (20-minute), exam-like conditions. Initial drafts (Essay A) were independently scored by three human raters using the IELTS rubric. Raters were blind to group assignments. Inter-rater reliability was calculated using intraclass correlation coefficients (ICC). The AI group received automated feedback via ChatGPT-4.0 with consistent prompting protocols.

2) Feedback Reception and Revision Phase: Participants received feedback from either the AI or human raters. Each student had 48 hours to revise their writing based on the feedback received. Revisions (Essay B) were guided

by the students' interpretation and application of the feedback.

3) Post-Feedback Evaluation and Reflection: Revised drafts were again evaluated by the same raters to ensure score comparability. Additionally, students completed a digital open-ended reflection questionnaire to assess their perceptions of the feedback's usefulness, clarity, and impact on their revision process.

(5) Data collection and analysis

Quantitative Analysis:

- Pre- and post-feedback scores were collected for all five IELTS criteria: Task Achievement (TA), Coherence and Cohesion (CC), Lexical Resource (LR), Grammatical Range and Accuracy (GR), and Overall Band.
- Inter-rater reliability was calculated using a two-way random effects ICC model (A,1) for absolute agreement.
- Paired sample t-tests were conducted to evaluate performance changes, and Cohen's d was used to estimate effect sizes.

Qualitative Analysis:

- Thematic analysis was conducted on AI and human feedback comments as well as student reflection responses using NVivo 12 Pro.
- Feedback comments were deductively coded according to IELTS rubric domains (TA, CC, LR, GR).
- Student reflections were inductively coded to capture emergent themes related to feedback clarity, perceived usefulness, revision strategies, and feedback preference.
- Two coders conducted iterative reliability checks, achieving inter-coder agreement above $\kappa = 0.85$.

4. Results and Discussions

(1) Quantitative analysis

1) Inter-rater reliability analysis for the human rater group

TABLE I Pre-feedback Human rater IRR analysis in assessing IELTS writing task 1

	ICC (A, 1)	95% Confidence Interval	F(4, 8)	Sig.	Range	IRR
TA	0.812	[0.105, 0.979]	5.059	P=0.025	<=1	Robust consistency
CC	0.931	[0.682, 0.992]	19	P=0.001	<=1	Robust consistency
LR	0.805	[0.191, 0.977]	5.714	P=0.012	<=1	Acceptable consistency
GR	0.931	[0.682, 0.992]	19	P<0.001	<=1	Robust consistency

A two-way random effects intraclass correlation coefficient (ICC) with absolute agreement was computed to assess consistency among three raters across four writing assessment criteria (N = 5 students). As Table I shows, all ICC values represent high consistency among the human raters, but distinct reliability patterns for different criteria needs to be carefully catered. Among the four marking criteria, CC and GR demonstrated excellent inter-rater reliability (ICC > 0.90), supporting the validity of composite scores. While significant rater bias was detected (all $p<0.05$), its practical impact is minimized through score aggregation. Moreover, the composite average scores of the four criteria equals the average overall scores of the three raters, again indicating very good reliability of the band scores given. However, for TA and LR criteria, the significant F-tests suggest additional rater training to address systematic biases. Moreover, widened confidence intervals (particularly for TA and LR) reflect precision limitations inherent in our small student cohort (N=5). Future studies should expand sample sizes to obtain more stable

estimates. Nonetheless, the high composite reliability and the acceptable range(R<=1) justifies using averaged scores in the subsequent analysis.

TABLE II Pre-feedback human rater group band scores

	TA Composite	CC Composite	LR Composite	GR Composite	Composite Average	Overall
HRS1	5.33	5.00	5.33	5.33	5.24	5
HRS2	5.67	5.33	5.00	5.00	5.5	5.5
HRS3	6.33	6.00	6.00	6.00	6.08	6
HRS4	6.67	6.33	6.67	6.33	6.5	6.5
HRS5	7.00	7.00	6.00	7.00	6.75	7

Inter-rater reliability for the post-feedback scoring (as seen in Table III) witnessed similar trend, which remained consistently strong across all criteria. For Task Achievement (TA), the intraclass correlation coefficient (ICC) reached 0.882 (95% CI [0.325, 0.987]), indicating excellent agreement when averaging raters' scores. Coherence and Cohesion (CC) demonstrated the highest reliability (ICC = 0.900, 95% CI [0.553, 0.989]), with confidence intervals fully above the 0.75 acceptability threshold. Both Lexical Resource (LR; ICC = 0.861, 95% CI [0.399, 0.984]) and Grammatical Resource (GR; ICC = 0.882, 95% CI [0.399, 0.984]) also maintained robust agreement levels. Notably, all criteria exhibited maximal score differences ≤1 point between raters, confirming high practical consensus. Significant F-tests (TA: $F(4,8) = 7.0$, $p = 0.01$; CC/LR/GR: $F(4,8) = 10.0$, $p = 0.003$) indicated persistent systematic rater biases, though their limited magnitude (constrained within 1-point ranges) suggests negligible operational impact on final score determination. Again, composite scores were used in the following analysis (as seen in Table IV).

TABLE III Post-feedback Human rater IRR analysis in assessing IELTS writing task 1

Criteria	ICC (A, 1)	95% Confidence Interval	F(4, 8)	Sig.	Range	IRR
TA	0.882	[0.325, 0.987]	7	P=0.01	<=1	Robust consistency
CC	0.900	[0.553, 0.989]	10	P=0.003	<=1	Robust consistency
LR	0.861	[0.399, 0.984]	10	P=0.003	<=1	Very good consistency
GR	0.882	[0.399, 0.984]	10	P=0.003	<=1	Robust consistency

TABLE IV Post-feedback human rater group band scores

	TA Composite	CC Composite	LR Composite	GR Composite	Composite Average	Overall
HRS1	6.00	6.00	6.33	6.00	6.08	6
HRS2	7.00	6.33	6.33	6.33	6.5	6.5
HRS3	7.00	7.00	6.67	7.00	6.92	7
HRS4	7.67	7.00	7.67	7.67	7.5	7.5
HRS5	7.33	7.00	8.00	7.67	7.5	7.5

2) Human rater feedback impact on writing performance in IELTS writing task 1

TABLE V Paired T-test for human rater group pre- and post-feedback band scores

Criterion	Pre M±SD	Post M±SD	Mean Δ	Δ%	95% CI for Δ	t(4)	p-value	Cohen's d
TA	6.20±0.67	6.80±0.63	+0.60	+9.7%	[0.12, 1.08]	3.54	.024*	1.58
CC	5.93±0.75	6.67±0.47	+0.74	+12.5%	[0.20, 1.28]	4.16	.014*	1.86
LR	6.00±0.58	7.00±0.76	+1.00	+16.7%	[0.28, 1.72]	4.00	.016*	1.79
GR	5.93±0.83	6.93±0.72	+1.00	+16.9%	[0.46, 1.54]	5.37	.006**	2.40
Overall	6.00±0.79	6.90±0.55	+0.90	+15.0%	[0.59, 1.21]	9.00	<.001***	4.03

All writing dimensions demonstrated statistically significant gains ($p<.05$), with the largest improvements in lexical resource (LR: +16.7%) and grammatical resource (GR: +16.9%). This confirms human raters provide actionable feedback which effectively addressed both linguistic and rhetorical aspects. Task Achievement had the smallest gain (+9.7%). This may suggest conceptual writing elements require longer development cycles or these elements are expressed in diverse ways from different raters which make it less easy to be turned into actionable plans. Effect sizes (Cohen's d) ranged from 1.58 (TA) to 4.03 (Overall), exceeding conventional thresholds for large effects ($d>0.8$). The exceptionally large overall effect ($d=4.03$) suggests feedback induced transformative improvements. However, the small sample ($n=5$) increases Type II error risk. Replication with larger samples is recommended to confirm effects. Narrow confidence intervals for GR ([0.46, 1.54]) and Overall ([0.59, 1.21]) indicate precise estimation of improvement magnitude. Wider intervals for LR ([0.28, 1.72]) reflect greater variability in vocabulary acquisition. Moreover, low standard deviations in post-scores ($SD\leq0.76$) suggest feedback helped standardize performance, though small sample size limits generalizability.

3) AI group feedback impact on writing performance in IELTS writing task 1

TABLE X Paired T-test for AI group pre- and post-feedback band scores

Criterion	Pre M±SD	Post M±SD	Δ	%Δ	Improvement Pattern
TA	6.60±0.89	7.60±0.55	+1.00	+15.2%	Largest gains: S3/S4 (+2)
CC	6.20±0.84	6.80±0.84	+0.60	+9.7%	Uniform growth (4/5 students)
LR	6.00±1.00	7.00±0.71	+1.00	+16.7%	S5 showed ceiling effect (7→8)
GR	5.80±0.84	7.40±0.55	+1.60	+27.6%	Breakthrough gains: S1/S2 (+2)
Overall	6.00±0.79	7.15±0.55	+1.15	+19.2%	All students improved ($\Delta\geq0.5$)

Quantitative analysis demonstrates significant writing proficiency gains following AI-generated feedback, with notable improvement differentials across criteria. Post-feedback scores increased across all dimensions, with Grammatical Resource (GR) showing the most substantial gains (+1.6 points), followed by Task Achievement (TA) and Lexical Resource (LR) (+1.0 each). Coherence and Cohesion (CC) exhibited more modest growth (+0.6 points), aligning with patterns observed in human-rated assessments. Specifically, the 27.6% GR improvement suggests AI feedback most effectively targets syntactic features. S1 and S2 demonstrated particularly dramatic gains (5→7), indicating AI excels at correcting mechanical errors. TA improvements (+15.2%) reveal AI's strength in structural feedback. S3 and S4 achieved excellent TA scores (8/9) post-feedback, showing effective remediation of task fulfillment issues. While LR improved overall, S5 (already strong) approached maximum scorable range (8/9), suggesting AI feedback has diminishing returns for advanced writers. The smallest CC gains (+9.7%) mirror human-rater findings, confirming discourse organization remains the most 'difficult to improve' dimension regardless of feedback source.

4) Comparative analysis of AI vs. human raters' feedback in promoting writing performance

Both AI and human feedback significantly improved writing proficiency but showed distinct patterns. AI excelled in Grammar (GR: +27.6% vs. humans' +16.9%) and Task Achievement (TA: +15.2% vs. +9.7%), indicating strengths in syntactic error diagnosis and structural guidance. Humans were superior for Lexical Resource (LR: +16.7% vs. AI's +15.2%), suggesting nuanced vocabulary support. Coherence (CC) gains were modest with both. Efficacy seems to correlate with proficiency: AI showed large gains for struggling writers but diminishing returns for advanced students. Humans achieved greater score standardization (post-feedback SD≤0.76) and a large overall effect (d=4.03). A pedagogical gain, therefore, could be leveraging AI's scalability for foundational skills (especially for lower-proficiency writers) and reserve human expertise for advanced lexical/conceptual refinement. There were concerns, however, related to AI's potential leniency bias in self-rated efficacy and human feedback's variable actionable guidance (wider LR CIs), as well as constrained generalizability due to the small sample (n=5), particularly affecting human effect size interpretation. Studies of larger sizes are needed to validate differential effects.

(2) Thematic analysis of rater feedback (AI vs. Human)

Following Braun and Clarke's (2006) framework, feedback was thematically coded into IELTS writing criteria. Both AI and human raters addressed core dimensions, but diverged notably in specificity, depth, and pedagogical orientation.

1) Task Achievement (TA)

AI feedback frequently prompted structural clarifications such as, "Add a one-sentence summary of the overall trend," or "Specify time span and units." One learner was advised to replace "went viral" with "rose steadily from 2000 to 2020," resulting in clearer temporal expression. Human raters reinforced these points but extended further by integrating genre conventions. They advised stating the chart type and scope (e.g., "a line graph showing trends in Melbourne from 2000–2020") and flagged vague expressions like "a lot" or "very few," suggesting more academic alternatives such as "moderate growth" or "remained consistently low."

2) Coherence and Cohesion (CC)

AI emphasized paragraph logic and connector use. Typical prompts included "Group stable lines in one paragraph" and "Avoid repeating 'however.'" One student restructured their essay by organizing paragraphs by trend type (increase, decline, stability). Human raters focused more on rhetorical flow. For instance, they recommended splitting oversized paragraphs and using cohesive devices like "by contrast" or "similarly" to improve transitions and guide the reader.

3) Lexical Resource (LR)

AI flagged vague or informal lexis, recommending terms like "rose sharply" instead of "got higher" and "surged" instead of "people liked it more." These suggestions led students to adopt more precise and formal vocabulary. Human raters offered more contextualized lexical refinement. Examples included substituting "went up again" with "rebounded" or replacing casual expressions like "cooling off" with "declined gradually." Human feedback also frequently addressed tone and register, highlighting the need for academic phrasing and varied expressions.

4) Grammatical Range and Accuracy (GRA)

AI efficiently corrected tense shifts and subject-verb agreement errors (e.g., "participants increase" → "participants increased"). It suggested passive constructions for objectivity ("the number was recorded") and encouraged varied structures. One student's score increased markedly after applying structures like, "This trend was observed across all categories." Human raters validated these suggestions but pushed for more syntactic sophistication. They recommended complex structures such as embedded clauses or inversion, e.g., "Not only did participation rise, but

it also outpaced other categories by 2015."

5) Comparative language

Both rater types highlighted the need for comparison—a core IELTS Task 1 requirement. AI provided frames like "X overtook Y by 2020" or "X remained higher than Y." Human raters, however, encouraged deeper comparative logic: e.g., "Although X started lower, it exceeded Y by the end of the period." This encouraged students to integrate temporal contrast and cross-category insights, rather than merely listing data.

6) Summary of rater differences

AI feedback was concise, systematic, and excelled in surface-level corrections—grammar, vocabulary, and structural prompts—delivered through clear, actionable directives. Human feedback was more interpretive and individualized, offering nuanced guidance on genre framing, discourse organization, and rhetorical clarity. While both overlapped on language accuracy, human raters scaffolded higher-order revisions, including logical restructuring and argument shaping. Together, they demonstrate complementary pedagogical strengths.

(3) Learner perception and uptake

1) Quantitative analysis of student responses

Student perceptions reveal a clear preference for human feedback, rated as more useful (4/5 "very" or "extremely useful" vs. AI's 3/5 "moderately useful") and significantly more detailed ("very detailed" by the majority vs. AI's "moderately/slightly detailed"). Human feedback was also uniquely linked to increased confidence (all 5 students) and perceived strength in improving organization. While AI feedback was seen as clear (4/5) and similarly actionable, its perceived value was proficiency-dependent: lower-level students (≤5.5) found it more useful and reported stronger language gains (4/5 cited "language" improvement), aligning with its focus on grammar/vocabulary/sentence structure. However, higher-level students perceived less improvement from AI, and its impact on task achievement (noted by only 1 student) and confidence (1 student) was limited. A notable disconnect exists for task achievement, where AI drove significant performance gains but limited student recognition. The small sample (n=5) warrants caution in generalizing these perceptions.

2) Learner perceptions and responses to feedback

① Perceived usefulness and limitations

All five students who received AI feedback described it as "clear" and "fast." Typical comments such as *"replace got higher with rose significantly"* or *"check subject–verb agreement in lines 3–4"* were praised for pinpointing problems instantly. Nonetheless, every AI-group participant also noted that the system rarely explained *why* a change mattered—one remarked, "It fixed the error, but I still didn't understand the rule." In the human-feedback group, every student highlighted the "why." Teachers not only flagged informal phrasing such as *a lot* but also supplied alternatives (e.g., *experienced moderate growth*) and linked them explicitly to band-score descriptors. One learner likened the feedback to "a mini lesson on genre expectations," echoing Ferris's (1997) observation that teacher commentary not only corrects errors but also guides genre-specific revision strategies.

② Uptake and revision strategies

AI-group revisions remained largely sentence-level. All five participants corrected tense errors (*participants increases → participants increased*) and adopted AI templates for comparisons (*X overtook Y by 2020*), but structural changes were minimal. In contrast, four of the five students guided by teachers undertook global revisions. Examples include rewriting the introduction to add an overview sentence, splitting an over-long paragraph, and inserting connectors such as *in contrast*, and replacing colloquialisms (*cooling off → declined gradually*). These deeper changes mapped directly onto higher gains in coherence and lexical resource.

③ Preferences for feedback types

Within the AI group, three students favored a hybrid approach— "AI for quick fixes, teachers for higher bands"—while two preferred teacher input exclusively, citing stronger alignment with IELTS criteria. Among the teacher-feedback cohort, four students endorsed human guidance alone, emphasizing its depth and individualization; the remaining student viewed AI as a useful proof-reading supplement but not a standalone resource.

④ Perceived value and recommendations

All teacher-feedback students said they would recommend examiner comments to peers because they "explain what the test wants." Four AI-group students would recommend ChatGPT-4 only as an auxiliary tool for grammar and vocabulary, cautioning that "it is not enough for Band 7."

⑤ Summary

Together, these accounts underscore a complementary division of labor. AI feedback excels in rapid, surface-level correction, whereas human feedback drives genre awareness, rhetorical depth, and higher-order restructuring. Learners therefore advocate a blended model that harnesses AI efficiency while retaining expert judgement for substantive revision—echoing the quantitative finding that grammar and task-structure gains are strongest under AI, whereas lexical growth and overall standardization peak under human guidance.

(4) Integrated discussion: how the numbers and narratives converge

Both strands of evidence point to a *complementary division of labor* between AI-generated and human feedback. Quantitatively, paired-sample tests confirmed sizeable post-revision gains for *all* criteria in both groups, yet the growth profiles diverged: ChatGPT-4 produced the steepest improvement in Grammatical Range & Accuracy (+27.6 %) and boosted Task Achievement by +15.2 %, whereas human examiners triggered the largest lift in Lexical Resource (+16.7 %) and delivered the most standardized overall bands (SD ≤ 0.76; d = 4.03).

Thematic analysis supports these quantitative trends. AI feedback operated with high consistency and accuracy, flagging surface-level errors (e.g., tense, S–V agreement, generic phrasing) and offering clear directives such as "insert an overview" or "replace got higher with rose significantly." These targeted edits prompted rapid revision cycles and reinforced mechanical accuracy—consistent with findings from Cotos (2014), who argued that automated feedback systems outperform humans in linguistic consistency. However, as Weigle (2013) warned, such systems often lack the rhetorical sensitivity needed for nuanced development. This was evident in human feedback, which focused more on global coherence, genre expectations, and discourse markers (e.g., "split paragraph," "clarify trend framing"), thereby explaining the sharper LR gains and greater organizational confidence in that group.

Learner perceptions triangulate these effects. Lower-proficiency writers particularly benefited from AI's clear and systematic grammar cues—aligning with Mizumoto and Eguchi's (2023) observation that AES tools are most effective at the foundational proficiency band. However, higher-band students reported plateauing with AI and valued human feedback for offering deeper rhetorical insights and task-specific refinement. Notably, Coherence and Cohesion remained the least improved area in both groups (~10%), with students consistently describing it as the "hardest to fix." This convergence of statistical and experiential data underlines the challenges of teaching discourse structure—a concern echoed across L2 writing research.

Taken together, the evidence supports a hybrid pedagogical model: AI should be deployed at scale for rapid, surface-level correction, especially effective for grammatical accuracy and structural clarity; human raters should be reserved for higher-order feedback on lexical choice, discourse coherence, and genre alignment. Such a model would not only optimize instructional resources but also address varying learner needs across proficiency levels in a differentiated and equitable manner.

5. Conclusion

This study set out to fill three gaps highlighted in recent AES research: **(a)** the near-absence of IELTS-specific evidence on AI feedback, **(b)** the lack of fine-grained, side-by-side analyses of AI- versus examiner-generated comments, and **(c)** the scarcity of learner-centred data linking feedback uptake to measurable score change. By employing a convergent parallel mixed-methods design with ChatGPT-4 and certified IELTS examiners, we showed that AI feedback excels at rapid grammatical repair and task-structure clarification, whereas human feedback delivers deeper lexical and discourse guidance—patterns triangulated quantitatively (differential effect sizes across criteria) and qualitatively (learner revisions and perceptions). The results therefore move the field beyond TOEFL-centric scoring correlations toward actionable guidance on matching feedback source to learner need in high-stakes IELTS writing.

Limitations should guide interpretation and future work. The sample (N = 10, one Sino-foreign university) was small and homogeneous; only a single Task-1 prompt and one ChatGPT prompting protocol were tested; improvements were tracked across one 48-hour revision cycle; and post-revision scores, while reliably double-rated, relied on the same examiner pool. Larger, proficiency-diverse cohorts, multiple prompts, varied prompt-engineering strategies, and longitudinal follow-ups are needed to confirm external validity and retention effects. Even with these constraints, our findings offer a solid empirical rationale for a hybrid pedagogy: deploy scalable AI for foundational accuracy and structural scaffolding, while reserving examiner expertise for nuanced lexical and discourse refinement essential to upper-band IELTS success. Future implementations could embed such hybrid feedback loops into IELTS preparation curricula, offering differentiated, scalable writing support that adapts to learners' evolving needs.

References

[1] Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning and Assessment*, *4*(3), Article 3. https://ejournals.bc.edu/index.php/jtla/article/view/1650.

[2] Beseiso, M., Alzubi, O. A., & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education*, *33*(3), 727–746. https://doi.org/10.1007/s12528-021-09283-1.

[3] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*. https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

[4] Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country. *Applied Measurement in Education*, *25*(1), 27–40. https://doi.org/10.1080/08957347.2012.635502.

[5] Cambridge Assessment English. (2020). IELTS Writing Band Descriptors (Public Version). https://www.cambridgeenglish.org.

[6] Cotos, E. (2014). Automated Writing Evaluation. In E. Cotos (Ed.), *Genre-Based Automated Writing Evaluation for L2 Research Writing: From Design to Evaluation and Enhancement* (pp. 40–64). Palgrave Macmillan UK. https://doi.org/10.1057/9781137333377_3.

[7] Creswell, J. W., & Clark, V. L. P. (2017). Designing and conducting mixed methods research. Sage publications.

[8] Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.-S., Gašević, D., & Chen, G. (2023). Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, 323–325. https://doi.org/10.1109/ICALT58122.2023.00100.

[9] Dikli, S., & Bleyle, S. (2014). Automated Essay Scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, *22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006.

[10]Ferris, D. R. (1997). The influence of teacher commentary on student revision. *TESOL Quarterly*, 31(2), 315–339. https://doi.org/10.2307/3588049.

[11]Kim, H., Baghestani, S., Yin, S., Karatay, Y., Kurt, S., Beck, J., & Karatay, L. (2024). ChatGPT for Writing Evaluation: Examining the Accuracy and Reliability of AI-Generated Scores Compared to Human Raters. In C. Chapelle, G. Beckett, & J. Ranalli (Eds.), *Exploring AI in Applied Linguistics* (pp. 73–95). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.06.

[12]Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050.

[13]Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 394–403). Association for

Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.32.
[14]OpenAI. (2025). *ChatGPT (Version 4.0)* [Web application]. https://chat.openai.com/chat.
[15]Shermis, M. D., & Burstein, J. C. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.
[16]Stevenson, M., & Phakiti, A. (2019). Automated feedback and second language writing. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 125–142). Cambridge University Press. https://doi.org/10.1017/9781108635547.009.
[17]Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 157–180. https://doi.org/10.1191/1362168806lr190oa.
[18]Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *18*(1), 85–99. https://doi.org/10.1016/j.asw.2012.10.006.
[19]Xia, W., Mao, S., & Zheng, C. (2024). *Empirical Study of Large Language Models as Automated Essay Scoring Tools in English Composition__Taking TOEFL Independent Writing Task for Example* (No. arXiv:2401.03401). arXiv. https://doi.org/10.48550/arXiv.2401.03401.
[20]Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating Short L2 Essays on the CEFR Scale with GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 576–584). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.49.