Language Testing in Asia

RESEARCH **Open Access**

Assessing the reliability and relevance of DeepSeek in EFL writing evaluation: a generalizability theory approach

Huixin Gao[1], Harwati Hashim[1*] and Melor Md Yunus[1]

*Correspondence:
harwati@ukm.edu.my

[1] Education, National University of Malaysia, Kuala Lumpur, Malaysia

Abstract

This study explored the potential of DeepSeek to contribute to English writing assessment via generalizability (G-) theory and qualitative feedback analysis. Specifically, it assessed the reliability of holistic scores and qualitative feedback generated by DeepSeek-V3 and DeepSeek-R1 for essays written in English as foreign language (EFL) learners, and compared these scores with those assigned by four college English teachers. The data consisted of 92 College English Test Band 4 (CET-4) essays written by non-English majors at a university in Heilongjiang Province, China. All the essays were holistically scored by DeepSeek-V3, R1, and four college English teachers. In addition, all three groups of raters provided qualitative feedback on content, language use, organization, and coherence. G-theory analysis revealed that the scoring reliability of DeepSeek-V3 was consistently lower than that of DeepSeek-R1 and the teacher raters; however, DeepSeek-R1 demonstrated consistently higher reliability coefficients than the teachers did. The qualitative feedback analysis indicated that both DeepSeek-V3 and R1 consistently provided more relevant feedback on the EFL essays than did the teacher raters. Furthermore, DeepSeek-V3 and R1 were equally relevant across the content, language use, organization, and coherence aspects of the essays, whereas the teacher raters generally focused more on language use but provided less relevant feedback on content, organization, and coherence. Consequently, DeepSeek-V3 and R1 could be useful AI tools for enhancing EFL writing assessments. The implications of adopting DeepSeek for classroom writing assessments are discussed.

Keywords: DeepSeek, EFL writing assessment, Generalizability theory, Qualitative feedback

Introduction

The advent and implementation of Artificial Intelligence (AI) have fundamentally transformed pedagogical paradigms and knowledge acquisition processes, revolutionizing teaching, learning, and assessment through instant feedback and personalized learning experiences (Chen et al., 2020; Kamalov et al., 2023; Pedro et al., 2019). The potential of AI in language education is gaining increasing recognition, particularly in EFL contexts, where it offers innovative approaches to teaching, learning, and assessment (Koraishi, 2023; Mohamed, 2024; Zhai & Wibowo, 2023). Among various AI tools, the launch of

DeepSeek-V3 in December 2024 has drawn increasing attention because of its open access, multilingual capabilities, and applicability in educational settings (Sallam et al., 2025). V3, which is designed to perform tasks such as text generation, summarization, and open-domain conversation, has strong potential in diverse linguistic settings (DeepSeek-AI et al., 2024). Building on this foundation, the release of DeepSeek-R1 in January 2025 further shifted academic attention by explicitly prioritizing reasoning and analytical precision (He et al., 2025; Mercer et al., 2025). Unlike its predecessors, R1 incorporates reinforcement learning and domain-specific fine-tuning strategies to enhance reasoning performance, internal consistency, and task-specific accuracy, making it particularly suitable for tasks that demand higher-order cognitive processing (DeepSeek-AI et al., 2025).

Assessing EFL learners' writing tasks is inherently complex and time-consuming, requiring significant expertise and resources (Neal, 2011; Tayyebi et al., 2022). To address these challenges, EFL educators and researchers continually seek innovative solutions to streamline the assessment process while ensuring the fairness and quality of educational outcomes. As generative AI continues to evolve, AI-powered tools have shown increasing promise for enhancing writing assessment and feedback in EFL contexts (Bucol & Sangkawong, 2024; Ebadi & Bashir, 2021; Zawacki-Richter et al., 2019). Among these tools, DeepSeek has become a widely discussed option in EFL writing assessment. By early 2025, it had been embraced by 1,045 universities across China and had begun to be integrated into university-level teaching practices (CERNET Authentication & Resource Sharing Infrastructure, 2025). While DeepSeek has attracted increasing interest, it shows promise as an AI-tool for supporting scalable and consistent formative feedback in EFL contexts, particularly in large classes where teacher workload makes individualized feedback impractical.

Despite recent advancements in AI-assisted writing tools (Guo, 2024; Yavuz et al., 2025), a research need remains to understand the specific applications and effectiveness of DeepSeek in scoring EFL essays and providing formative feedback on content, organization, language use, and coherence. Addressing this gap requires evaluating the performance of different DeepSeek versions, as V3 emphasizes fluency and speed, whereas R1 prioritizes reasoning accuracy and scoring consistency (DeepSeek-AI et al., 2024, 2025; He et al., 2025). Comparing the two can help identify their strengths and inform the pedagogical use of AI in EFL writing assessment.

Therefore, this study aims to address this gap by investigating how DeepSeek-V3 and R1 can support EFL teachers in essay evaluation. Specifically, it examines the models' ability to generate reliable scores and provide relevant qualitative feedback, offering insights into how DeepSeek may enhance EFL writing assessment practices.

Literature review

The nature of EFL writing assessment

Writing assessment requires balancing quantitative evaluation with qualitative instructional feedback to support learners' writing development effectively (Lee, 2017; Pearson, 2022). Traditional approaches for scoring EFL essays typically involve one or more human raters, who are invited to score the essays either holistically or analytically (Barkaoui, 2008; Li & Huang, 2022; Wang & Xie, 2022). However, the scores that human raters assign to EFL essays may fluctuate due to various factors, such as linguistic backgrounds, interpretative frameworks, and tolerance for errors (Barkaoui, 2008; Neittaanmäki & Lamprianou, 2024). Therefore, raters are a potential source of measurement error that can affect score reliability (Li, 2022; Neittaanmäki & Lamprianou, 2024).

Reliability is a fundamental requirement for valid and effective EFL writing assessment (Li, 2022). According to the UK's Office of Qualifications and Examinations Regulation (Ofqual, 2013), reliability refers to the consistency of assessment outcomes across different testing occasions, regardless of when or by whom the test is administered. Inter-rater reliability in EFL writing assessment has been the focus of extensive research, with findings providing practical implications for improving scoring consistency and assessment quality (Chen et al., 2022; Huang, 2008).

Furthermore, EFL writing assessments require raters to provide qualitative feedback on the language, content, organization, and coherence of each essay (Li & Huang, 2022). Ideally, this feedback should be clear, relevant, and supportive of students' revision efforts. However, its formative potential in enhancing learners' metalinguistic awareness, rhetorical skills, and self-regulated revision is often overlooked in rubric-based evaluations (Brookhart, 2017; Hyland & Hyland, 2006; Yu & Liu, 2021). To address these pedagogical shortcomings, the present study proposes a tripartite feedback framework underpinned by three core principles: contextualized applicability (Lee, 2017; Rae & Cochrane, 2008), cognitive accessibility (Deane et al., 2008), and metalinguistic scaffolding (Boggs, 2019).

Contextualized applicability emphasizes feedback alignment with specific rhetorical contexts and audience expectations, helping learners produce discipline-appropriate texts (Rae & Cochrane, 2008). In the context of argumentative essays, this may involve guiding students to strengthen claims using discipline-specific evidence or revising thesis statements to align with established discourse conventions. Cognitive accessibility ensures that feedback matches learners' linguistic and cognitive readiness, simplifying complex linguistic terminology to improve understanding and use of feedback (Deane et al., 2008). To enhance feedback effectiveness, technical jargon should be minimized and replaced with familiar expressions. For example, instead of "syntactic inversion," the term "word order" may be used to facilitate comprehension and uptake among novice writers. Metalinguistic scaffolding, meanwhile, fosters learner autonomy by prompting self-reflection on language choices and encouraging independent error correction through guided prompts or coded annotations (Bitchener & Ferris, 2012; Boggs, 2019). In practice, it is often realized through coded annotations or prompts—such as "awk" (awkward expression) or "revise for clarity"—which direct learners' attention to specific linguistic problems and invite them to refine their drafts through self-regulated analysis and revision. Together, these principles operationalize feedback as a pedagogical tool that bridges assessment and instruction, fostering learner agency and deeper engagement with the writing process. Clarifying these fundamental principles of effective feedback and the measurement challenges associated with EFL writing assessment provides the necessary foundation for examining the reliability and validity of AI-supported assessment methods, such as the AI-based approach explored in this study.

Advantages of using generalizability theory in EFL writing assessment

Originating from Cronbach (1972), G-theory provides a powerful alternative to Classical Test Theory (CTT) by offering a more holistic and effective method for assessing and enhancing test reliability. Unlike CTT, which provides only a single reliability score, G-theory explicitly distinguishes how much of the variance in scores comes from raters (e.g., some raters being stricter than others), task difficulties (e.g., different essay prompts), and the interactions among these factors (Brennan, 2001; Linacre, 1993). This analytic precision is particularly valuable in EFL writing assessment, where variability introduced by subjective ratings and uneven task demands can compromise score fairness (Vispoel et al., 2019; Webb et al., 2006).

G-theory's analytical process unfolds in two distinct stages. First, a G-study identifies and quantifies the contribution of each assessment facet, such as raters and tasks. A decision study (D-study) examines possible modifications to the assessment design, including increasing the number of raters or applying standardized scoring criteria, to determine the most effective configuration (Briesch et al., 2014; Cardinet et al., 2011). For example, Huang (2012) demonstrated that doubling the number of rating sessions substantially reduced rater-related variance by 38% in EFL writing assessments.

Moreover, G-theory enhances fairness by explicitly identifying biases and irrelevant factors, ensuring that scores reflect students' true writing abilities rather than extraneous influences (Chen et al., 2022). Research by Zhang and Zhang (2022) further highlights that G-theory-informed assessment designs help narrow achievement gaps among different student groups, significantly improving assessment equity.

Overall, G-theory's systematic and transparent approach to identifying and controlling measurement error makes it uniquely suited as the theoretical framework for this study, enabling a fairer, more reliable, and more precise evaluation of learners' writing performance. Thus, applying G-theory in this research facilitates a rigorous evaluation of the reliability and generalizability of AI-generated feedback relative to human ratings.

The benefits of AI in EFL writing assessment

The integration of AI into EFL writing assessment has significantly enhanced evaluative precision, adaptive feedback, and pedagogical scalability. Transformer-based AI models, such as ChatGPT-4, demonstrate near-human or superior performance in assessing syntactic complexity, lexical diversity, and grammatical accuracy (Dong, 2024; Lim et al., 2023). Advanced Natural Language Processing (NLP) algorithms offer precise and timely corrective feedback on linguistic features such as article usage and collocation errors, with greater depth and specificity than traditional methods (Chen & Zhang, 2022).

Empirical evidence underscores AI's capacity not only to improve assessment precision but also to enhance learners' writing proficiency and engagement. Tsai et al. (2024), in a longitudinal intervention with Chinese English majors, reported significant improvements ($p < 0.01$) in vocabulary use, grammatical accuracy, and textual coherence after AI-assisted revisions. Similarly, Stevenson and Phakiti (2019) reported that students interacting with AI-driven platforms increased self-initiated revisions by 42%, highlighting AI's role in promoting iterative and autonomous learning processes. Additionally, multimodal AI tools leveraging sentiment analysis and discourse modelling effectively assess higher-order competencies, including rhetorical coherence and argumentative structure (Koltovskaia, 2020). Meta-analytic studies confirm these advantages, reporting an average 0.61 standard deviation improvement in writing outcomes using AI-mediated assessment compared with traditional methods (Wilson et al., 2021).

Despite these promising results, the sustainable integration of AI technologies poses several ongoing challenges. Key concerns include ensuring transparency and interpretability of scoring algorithms, balancing the efficiency of automated assessments with the nuanced judgments provided by human evaluators, and addressing ethical issues such as academic misconduct and fairness in AI-supported assessment environments (Zou & Huang, 2023). For example, Zheng et al. (2024) identified four critical determinants of AI acceptance among 620 Chinese undergraduates: performance expectancy, effort efficiency, social influence, and intrinsic motivation.

AI technologies demonstrate transformative potential in EFL writing assessment through enhanced diagnostic precision, learner-centered feedback loops, and scalable pedagogical interventions. However, sustainable integration requires addressing persistent challenges: (1) establishing transparent algorithmic auditing protocols, (2) balancing automated efficiency with human evaluative nuance, and (3) developing ethical frameworks to mitigate the risks of academic misconduct. Future research should consider prioritizing cross-cultural validations and longitudinal impact assessments to optimize AI's role in global language education ecosystems. Considering these benefits and challenges, the present study specifically examines the reliability and effectiveness of DeepSeek feedback in EFL writing assessment.

Research gaps and research questions

Although previous studies have explored the transformative potential of AI in EFL writing assessment (Wilson et al., 2021; Zheng et al., 2024), comprehensive studies comparing the performance of DeepSeek-V3 and R1 in this context are lacking. While other AI tools, such as ChatGPT, have received considerable research attention, DeepSeek remains relatively underexplored, highlighting a clear gap in the current literature. While various AI models have been applied to essay scoring, comparative research evaluating DeepSeek's performance against human raters remains limited, particularly in terms of scoring reliability and qualitative feedback. Additionally, there is a need to evaluate how these AI tools measure up to human teacher raters in terms of content, language use, organization and relevance of feedback. Given the limited research available on DeepSeek, systematically evaluating its reliability and feedback quality compared to human raters will contribute valuable insights to both AI-enhanced assessment practices and pedagogical applications.

This study aims to address these gaps by exploring the following two research questions:

> RQ1: What is the reliability of the holistic scores assigned to EFL essays by DeepSeek-V3 and R1 compared to human raters?

RQ2: How relevant is the qualitative feedback provided by DeepSeek-V3 and R1 on EFL essays compared with the feedback from teacher raters?

Methods

Research methodology

This study employs an explanatory sequential mixed-method design (Creswell, 2021; Plano Clark, 2017) to assess DeepSeek's scoring reliability and accuracy, with a focus on its comparison to teacher raters in an EFL context. The design was conducted in two phases to assess the accuracy of AI-based scoring and identify its limitations in evaluating specific aspects of writing.

In the first phase, a quantitative analysis was conducted to compare the scores assigned by DeepSeek-V3, R1, and the four teacher raters to a sample of EFL essays from a single class, all of which were of similar levels according to their Gaokao scores. This phase aimed to establish a baseline for evaluating scoring accuracy, with particular attention given to how well the AI models align with human raters. In the second phase, qualitative analysis is employed to examine the reasons behind discrepancies in scoring and feedback, providing a deeper understanding of the content, language use, organization, and coherence feedback offered by both AI models compared with human raters. This sequential approach allows for a comprehensive evaluation of DeepSeek-V3 and R1, shedding light on their strengths and limitations in providing accurate and relevant assessments of EFL writing.

The participants

The research sample comprised 92 non-English major undergraduates (gender-balanced: 47 males, 45 females) and four college English teachers from a university in Heilongjiang Province, China. The four English teachers, each with over 15 years of experience in teaching college-level English writing, also served as official raters for the national CET-4 writing part. In addition to the teacher raters, DeepSeek-V3 and R1 were employed as AI-based raters in this study. These AI models were selected for their ability to provide detailed and contextually relevant feedback on English writing, which has been shown to support more accurate and consistent assessments than traditional methods do (Albuhairy & Algaraady, 2025; Nguyen, 2025). Their integration into the rating process aimed to examine whether such models could match or exceed human raters in evaluating EFL learners' written performance.

Data collection and rating procedures

Data collection

This study involved both quantitative scoring and qualitative feedback to evaluate EFL writing performance (Li et al., 2014). The writing task was adapted from an actual CET-4 prompt, requiring participants to write a 180-word essay taking a clear stance on whether granting children more freedom enhances their independence and confidence. All the essays were completed in class under the supervision of the instructor to ensure authenticity and standardization. The prompt was presented with the following instructions:

> *For this part, you are allowed 30 minutes to write an essay that begins with the sentence "Nowadays parents are increasingly aware that allowing kids more freedom to explore and learn on their own helps foster their independence and boost their confidence". You can make comments, cite examples or use your personal experiences to develop your essay. You should write at least 150 words but no more than 200 words (not including the sentence given).*

The assessment process involved comparing four aspects of writing—content, language use, organization, and relevance—across three rater groups: DeepSeek-V3, R1, and four college English teachers.

In the first phase, the four college English teachers, all with prior experience as CET-4 writing raters, independently assessed the essays on the basis of the official CET-4 writing scoring criteria (see Appendix 11). They provided both holistic scores and qualitative feedback across the four dimensions. Before scoring, they participated in a brief offline training session that included reviewing the scoring criteria and discussing five representative writing samples from the *College English Test Band 4 Syllabus (2016 Revised Edition)* (National Education Examinations Authority, 2016). These samples, rated at 14, 11, 8, 5, and 2 points, represented a full range of proficiency levels and served as reference anchors for score calibration.

In the second phase, DeepSeek-V3 and R1 were trained via the same training materials and procedures. Specifically, both versions were fine-tuned with these five benchmark samples to internalize the scoring standards and associate specific language features with corresponding proficiency levels. This training process mirrored that of the human raters, ensuring alignment with the CET-4 scoring rubric. Following the training, DeepSeek-V3 and R1 evaluated the same set of essays using procedures aligned with those of the human raters. Each essay was first assigned a holistic score, followed by qualitative feedback just as the human raters had done. To assess reliability and stability, the evaluation by DeepSeek-V3 and R1 was repeated four times, with a three-day interval between each round.

Data rating

Teacher ratings A total of 92 students were expected to submit their essays. However, after the initial screening, eight essays were excluded: three students did not submit their essays, three submitted blank papers, and two displayed a high degree of similarity in content, raising concerns about potential plagiarism or collaboration. Consequently, 84 valid essays were retained for rating and further analysis. To control for potential rater severity or leniency, the raw scores were adjusted via the Many-Facet Rasch Model (MFRM). A two-facet model was applied, incorporating examinees and raters as facets. This method enables the calibration of rater effects, providing adjusted scores that more accurately reflect students' actual writing performance. Additionally, MFRM analysis assessed the quality of the rating process through rater fit statistics (infit

Table 1 Process of removing misfitting essays to ensure the reliability of teacher ratings

	Mean-square range for "rater" infit/outfit	Exclusion criteria applied to improve "examinee"
After the first estimation	.53–1.61	5 essays with mean-square values and Z-standardized fit statistics above 2.0 (indicating severe underfit) were removed. Two essays with no text were also excluded
After the second estimation	.63–1.47	3 additional essays were removed using the same criteria
After the third estimation	.62–1.34	No essays had Z-standardized fit statistics above 2.0, though some still had mean-square values above 2.0. Therefore, the removal of misfitting essays was discontinued

The mean squares near 1.0 indicate little distortion of the measurement system. The acceptable upper and lower bounds of infit/outfit values are usually set from.7 to 1.3; however, Wright and Linacre (1994) state that mean-square values from.5 to 1.5 are also productive for measurement

and outfit), ensuring that the final ratings were consistent and reliable. These calibrated scores served as the benchmark for evaluating the scoring accuracy of DeepSeek.

Table 1 provides an overview of the process of removing misfitting essays to ensure the reliability of the teacher ratings.

DeepSeek ratings The same set of 84 essays was also rated by DeepSeek-V3 and R1. To ensure comparability with human scoring, DeepSeek followed the same procedure as the teacher raters did. Each essay was input and rated individually, following the same step-by-step scoring process used by the human raters. To further align with real-world classroom applications, the temperature setting for DeepSeek was intentionally left at the default level. This decision aimed to replicate authentic classroom evaluation conditions, where instructors are likely to use the system without adjusting advanced parameters. The temperature parameter in generative AI controls the randomness of output; lower values produce more focused and consistent responses, whereas higher values lead to more diverse and creative outputs (Mozaffar et al., 2022). Maintaining the default setting ensures that the AI-generated scores reflect realistic usage scenarios in educational contexts (Denny et al., 2023). Although rating justifications were not explicitly requested, DeepSeek automatically generated explanatory statements alongside each score.

Data analysis

The operational reliability of DeepSeek's algorithmic scoring system was rigorously validated through computational concordance analysis, which employs inter-rater consistency metrics to quantify the precision of the evaluation outcomes. To assess the consistency of DeepSeek ratings with teacher ratings, several inter-rater reliability measures, including adjacent agreement, Cohen's kappa, and Krippendorff's alpha, were calculated.

Next, to identify the writing aspects influencing DeepSeek's accuracy, the rationale behind its scoring decisions was analyzed alongside the scores under the "with writing prompt and source text" condition. Four researchers qualitatively examined these scores. Two essays where DeepSeek assigned a higher score than teacher raters, and two where it assigned a lower score, were randomly selected from the pool of essays with discrepant ratings. The researchers collaboratively reviewed the essay responses and DeepSeek's rationale, encouraging a discussion to identify writing aspects that the AI system might struggle to assess accurately.

Quantitative analysis was conducted within the framework of G-theory via GENOVA (Crick & Brennan 1983), a program designed to estimate variance components and standard errors in balanced designs. Specifically, random effects G-studies were performed for person-by-DeepSeek rater ($p \chi r_D$), and person-by-DeepSeek R1 rater ($p \chi r_{R1}$), and person-by-teacher rater ($p \chi r_t$) random effects G-studies were executed, followed by person-by- DeepSeek rater ($p \chi r_D$), person-by-DeepSeek R1 rater ($p \chi r_{R1}$), and person-by-teacher rater ($p \chi r_t$) random effects D-studies, with *G*- and *Phi*-coefficients calculated. Because the aim of this study was to compare the score variability and reliability among DeepSeek, DeepSeek R1, and teacher raters, only the facet of the rater was considered in the G-theory analyses.

Qualitative feedback from each rater type (DeepSeek-V3, R1, and teacher raters) was analysed at two levels. Initially, the feedback was independently color-coded, sorted, and organized by the researchers, and then grouped into themes through collaborative discussion (Creswell, 2021). A three-level coding scheme was applied to categorize the feedback into content, language, organization, and coherence. The grouped feedback was further quantified for descriptive statistical analysis, with major themes within each domain identified.

Results

Table 2 presents the interrater reliability analysis between DeepSeek (V3 and R1 versions) and four teacher raters. The results show strong agreement across multiple metrics. Exact agreement reached 61.6%, indicating moderate alignment under strict scoring criteria, whereas adjacent agreement (allowing for neighboring score tolerances) reached 97.9%, reflecting nearly perfect consistency in practical scoring scenarios. Cohen's weighted kappa (0.734) confirmed substantial nonrandom agreement between DeepSeek and the teacher raters, and Krippendorff's alpha (0.868) indicated "almost perfect" reliability, accounting for multiple raters and ordinal scales. These findings suggest that DeepSeek performs comparably to teachers in scalable scoring tasks, particularly when minor score variations are permissible. However, they also highlight areas for refinement to improve accuracy under strict assessment conditions.

Previous studies have noted the consistency in the feedback rationale between DeepSeek-V3, R1, and teacher raters, with both focusing on aspects such as clarity, flow, and structure (Dai et al., 2023; Naismith et al., 2023). In light of this, we conducted a qualitative analysis of four essays with discrepant ratings and examined the rationales provided by DeepSeek. Our analysis revealed that DeepSeek exhibited limitations in identifying content-related issues in these essays.

Table 2 Interrater reliability between teacher raters and DeepSeek

	With writing prompt and source texts
Exact agreement	0.616
Adjacent agreement	0.979
Cohen's kappa (linear weighted)	0.734
Krippendorff's alpha	0.868

Table 3 Random effects person-by-rater G-studies results

Rater type	Sources of variability	DF	σ2	%	Standard error
DeepSeek-V3	p	83	1.1608	64.6	0.1785
	r_{V3}	3	0.1620	2.8	0.1406
	pr_{V3}	249	0.5290	33.6	0.0472
	Total	335	1.8518	100	1.8979
DeepSeek-R1	p	83	3.6138	88.2	0.5544
	r_{R1}	3	0.0712	1.7	0.0631
	pr_{R1}	249	0.4109	10.1	0.0367
	Total	335	4.0959	100	0.5592
Teachers	p	83	3.6802	77.6	0.5648
	r_t	3	0.1338	4.3	0.1191
	pr_t	249	0.8375	18.1	0.0748
	Total	335	4.6514	100	1.3332

Table 4 Random effects person-by-rater D-studies results

Number of persons (*p*)	Number of raters(*r*)	DeepSeek-V3		DeepSeek-R1		Teachers	
		G	Phi	G	Phi	G	Phi
84	1	0.67	0.65	0.89	0.88	0.80	0.79
84	2	0.77	0.72	0.94	0.92	0.89	0.88
84	3	0.82	0.79	0.95	0.93	0.92	0.90
84	4	0.89	0.87	0.96	0.96	0.94	0.93
84	5	0.91	0.89	0.96	0.96	0.95	0.94
84	6	0.93	0.92	0.97	0.97	0.96	0.95
84	7	0.93	0.93	0.97	0.97	0.97	0.96
84	8	0.94	0.94	0.98	0.98	0.97	0.96
84	9	0.95	0.95	0.99	0.98	0.98	0.97
84	10	0.96	0.96	0.99	0.99	0.98	0.97

Holistic score reliability of DeepSeek-V3, R1, and teacher raters

Generalizability theory (G-theory) was employed to evaluate score variability and identify the contribution of variance components to total score variance. The results from the G- and D- studies are summarized in Tables 3 and 4, respectively.

As shown in Table 3, the outcomes from the three G-studies conducted for DeepSeek versions V3, R1, and teacher raters, respectively. In all three cases, the object of measurement (i.e., person, *p*), representing the students' writing abilities, emerged as the most significant source of variance. Specifically, for DeepSeek version V3 scoring, the variance component associated with person accounted for 64.6% of the total variance. For DeepSeek R1 scoring, the same variance component explained 88.2% of the total variance. For teachers' scoring, the person variance component explained 77.6% of the total variance. The person variance component is the desired variance because it is believed that students' writing abilities vary from person to person (Brennan, 2001; Huang, 2008, 2012).

Furthermore, as shown in Table 3, for the DeepSeek V3 scoring scenario, the residual variance component explained 33.6% of the total variance. It includes variability due to the interactions between facets and other unmeasured errors. Over 32% of the unexplained variance indicates hidden facets, which were not considered in the design but might have affected the scoring of the students' essays (Brennan, 2001; Huang, 2008, 2012). For DeepSeek R1 and teacher raters, the same variance component explained 10.1% and 18.1% of the toral variance respectively.

Finally, as shown in Table 3, in all three cases, the rater variance component (i.e., r_{V3}, $r_{R1,}$ and r_t), reflecting the differences in scoring leniency or stringency the among raters, was the least significant source of variance. It explained 2.8%, 1.7%, and 4.3% of the variance for DeepSeek V3, R1, and teacher raters, respectively, suggesting that the three types of raters scored students' writing tasks fairly consistently.

The results of the three D-studies highlight an overall increase in the reliability of holistic scores from DeepSeek V3, teachers, and DeepSeek R1 (see Table 4). Table 4details the reliability measures for the three types of raters, utilizing the G-coefficients (for norm-referenced interpretations) and Phi-coefficients (for criterion-referenced interpretations). These coefficients were calculated under conditions from one to ten raters for each essay.

Table 4 presents the results of D-studies, highlighting how increasing the number of raters improves reliability. For single-rater scoring, the G-coefficients for DeepSeek-V3 and R1, and the teacher raters were 0.67, 0.89, and 0.80, respectively, with corresponding Phi-coefficients of 0.65, 0.88, and 0.79. When the number of raters increased to three, there was a notable improvement in reliability for three types of raters. Specially, the G-coefficients increased to 0.82, 0.95, and 0.92 for DeepSeek V3, R1, and teacher raters, respectively, while Phi-coefficients increased to 0.79, 0.93, and 0.90, respectively.

In comparison, the DeepSeek V3 score presented lowed reliability coefficients than did the teacher scores; however, the DeepSeek R1 score presented higher reliability coefficients than did the teacher score. These findings suggest that DeepSeek V3 shows potential as a supplementary tool to support teachers in evaluating EFL essays. DeepSeek R1 outperformed teacher raters in scoring consistency, highlighting its potential as a viable alternative for EFL essay evaluation in instructional settings.

In comparison, DeepSeek-V3 exhibited slightly lower reliability than the teacher raters, whereas DeepSeek-R1 demonstrated higher reliability coefficients. These findings suggest that although DeepSeek-V3 shows promise in enhancing EFL writing assessments, DeepSeek-R1 may provide an even more effective AI tool, potentially replacing human raters in scoring EFL essays.

Relevance of qualitative feedback across DeepSeek-V3, R1, and teacher raters

The qualitative feedback provided by DeepSeek-V3, DeepSeek-R1, and the teacher raters was analyzed via both quantitative and qualitative approaches. The quantitative analysis assessed the relevance of the feedback based on mean scores and standard deviations, whereas the qualitative analysis identified recurring themes across the feedback from different raters. The results are summarized in Tables 5 and 6.

As shown in Table 5, DeepSeek-R1 provided the highest number of relevant feedback points across all categories: content (mean =4.88, SD =2.620), language use (mean

Table 5 Descriptive results of the qualitative feedback

Levels	Rater groups	N of sample	Number of relevant feedbacks*		
			Total	Mean	SD
Content	DeepSeek-V3	84	404	4.81	2.607
	DeepSeek-R1	84	410	4.88	2.620
	Teachers	84	89	1.06	0.798
Language	DeepSeek-V3	84	1262	15.05	6.202
	DeepSeek-R1	84	1323	15.75	6.507
	Teachers	84	249	2.97	1.497
Organization	DeepSeek-V3	84	163	1.94	2.113
	DeepSeek-R1	84	335	3.99	1.404
	Teachers	84	166	1.97	0.598
Coherence	DeepSeek-V3	84	158	1.88	1.103
	DeepSeek-R1	84	232	2.76	1.201
	Teachers	84	88	1.05	0.497

*Relevant feedback refers to the degree to which the feedback is perceived as clear, comprehensible, and pedagogically supportive, allowing students to engage in revising their EFL writing (Black & Wiliam, 1998; Carless et al., 2011; Hattie & Timperley, 2007)

=15.75, SD =6.507), organization (mean =3.99, SD =1.404), and coherence (mean =2.76, SD =1.201). DeepSeek-V3 followed closely in most areas: content (mean =4.81, SD =2.607), language use (mean =15.05, SD =6.202), organization (mean =1.94, SD =2.113), and coherence (mean =1.88, SD =1.103). In contrast, teacher raters provided the fewest relevant feedback points across all categories: content (mean =1.06, SD =0.798), language use (mean =2.97, SD =1.497), organization (mean =1.97, SD =0.598), and coherence (mean =1.05, SD =0.497).

These results highlight significant variation in feedback effectiveness across the three rater groups, with DeepSeek-R1 offering the most comprehensive and relevant feedback, followed by DeepSeek-V3. The teacher raters provided the least detailed feedback overall.

To further evaluate feedback effectiveness, major themes related to content, language use, organization, and coherence were identified and categorized. These themes, presented in Table 6, shed light on the specific areas where students need the most improvement and demonstrate how feedback from each rater group addresses these writing aspects. The findings provide valuable insights into the strengths and limitations of feedback from both DeepSeek and teacher raters, helping to assess their utility in improving student writing.

This thematic analysis revealed that DeepSeek-V3 and R1 predominantly addressed content-related issues, such as the clarity of the argument and the need for better evidence and support. Notably, DeepSeek-V3 highlighted concerns about the lack of argument clarity, whereas DeepSeek-R1 identified gaps in counterarguments and persuasiveness. In contrast, teacher raters focused more on language-related issues, particularly grammar and syntax errors, with less attention given to the deeper structural or content-related aspects of the essays. Teacher feedback also emphasized problematic sentence structures and issues with word choice, including the influence of Chinese thinking. DeepSeek-R1 emerged as the most comprehensive rater, offering substantial feedback on content, language, and coherence, indicating its greater potential for enhancing student writing than do both DeepSeek-V3 and teacher raters.

Table 6 Themes of the qualitative feedback

Level of coding scheme	Major themes	DeepSeek-V3	DeepSeek-R1		Teachers
Content	• Depth & specificity	• Lack depth and specificity in arguments	Advantages	➢ Clear theme	✓ Central ideas are not clearly articulated
			Disadvantages	➢ Arguments lack depth, vagueness in expression	
	• Clarity & support	• Lack of clarity in argument structure	Advantages	➢ Conveys the core idea	✓ Arguments are poorly substantiated and lack support
			Disadvantages	➢ Insufficient evidence and logical support	
	• Engagement and persuasiveness	——	Advantages	——	——
			Disadvantages	➢ Absence of counterarguments, not evident persuasiveness	
Language	• Grammar & syntax errors	——	Advantages	——	✓ Consistent errors in grammar and verb tense usage
			Disadvantages	➢ Frequent grammatical and syntactical errors	
	• Sentence structure and complexity	——	Advantages	——	✓ Sentence structures are overly complex and unclear
			Disadvantages	➢ Monotonous sentence structure: mostly simple sentences	
	• Word choice & vocabulary	——	Advantages	➢ Incorrect collocations	✓ Incorrect word choice, influence from first language
			Disadvantages	➢ Repetitive vocabulary	
	• Spelling & typos	——	Advantages	——	✓ Spelling and typographical errors detract from clarity
			Disadvantages	——	
Organization	• Structural concerns	——	Advantages	➢ Uses appropriate argument structure with clear subpoints	✓ Logical flow and organization are lacking
			Disadvantages	➢ Conclusion is missing, development of ideas is weak	
Coherence	• Cohesion & flow	• Disjointed ideas, lack of logical transitions	Advantages	➢ Paragraphs follow a logical order	✓ Coherence is minimal, logical connectives are missing
			Disadvantages	➢ Unclear organizational structure, poor sequencing	

The following section presents specific examples of qualitative feedback provided by DeepSeek-V3, DeepSeek-R1, and teacher raters on Essay #4 (see above).

Example

Parents nowadays are becoming increasingly aware of giving their children more freedom to explore and learn, helps cultivate their independence and enhance their self-confidence.

Giving children more freedom and allowing them to explore and learn on their own has many benefits. For example, firstly, it can promote the comprehensive development of a child's body and mind. By constantly engaging in activities and experiencing, children can acquire skills to use their bodies, which also enables them to have a healthy mind and full emotions, enabling them to develop both physically and mentally. Secondly, cultivate children's ability for independent thinking. Children gradually develop a spirit of active thinking through free play, which helps them face the challenges of independent life in the future. Thirdly, possess qualities of confidence and self-improvement. Children who receive respect and freedom will gradually believe in their abilities, and they will demonstrate the quality of self-trust and self-improvement in everything they do. And so on.

In short, there are many benefits to giving children freedom to explore and learn on their own. Parents should give their children freedom and make them better!

Content

The feedback from all three groups identified common issues in argument development and support in Essay #4. These concerns primarily involve: (a) insufficient depth and specificity, (b) unclear arguments and inadequate supporting evidence, and (c) limited engagement and persuasiveness.

First, all the raters consistently noted a lack of depth and specificity, suggesting that incorporating concrete examples and personal anecdotes could enhance argumentative strength. Second, the teacher raters emphasized that the essays frequently deviated from the main topic or failed to clearly convey a central idea. This concern was echoed by both DeepSeek models, which highlighted vague arguments and insufficient supporting evidence. Finally, DeepSeek-R1 specifically pointed out the absence of counterarguments, describing the essays as lacking convincing, balanced perspectives and recommending more compelling and thoroughly developed arguments.

For example, DeepSeek-V3 noted, "The essay addresses the importance of freedom for children's development, covering physical, mental, and emotional aspects." DeepSeek-R1 provided more detailed feedback: "The essay addresses various aspects of child development, including physical, mental, and emotional dimensions. However, it primarily mentions general benefits without concrete examples such as specific activities such as sports or creative projects. To improve, consider incorporating real-life scenarios—for instance, allowing children to participate in team sports fosters teamwork and resilience." A teacher rater also highlighted structural issues, stating: "The central theme of the article is unclear; although divided into three paragraphs, the content is fragmented and disorganized."

Language use

The feedback on language use identified several issues impacting the clarity and coherence of the essays, including (a) grammatical and syntactical errors, (b) sentence structure and complexity, (c) word choice and vocabulary, and (d) spelling and typographical mistakes.

All the rater groups highlighted significant grammatical issues, such as incorrect verb tenses and article usage, which hindered the clarity and readability of the essays. The sentence structure also raised concerns. DeepSeek-V3 revealed a lack of sentence complexity, whereas DeepSeek-R1 and the teacher raters noted problematic sentence structures and incoherent phrasing. Word choice issues were prevalent across all groups, with feedback noting awkward phrasing and limited vocabulary. The teacher raters also highlighted the influence of the Chinese language patterns, reflecting a crosslinguistic challenge. Finally, both DeepSeek-R1 and the teacher raters identified spelling and typographical errors, underscoring the need for better proofreading.

DeepSeek-V3: "The essay demonstrates several language-related issues, including grammatical inaccuracies, syntactical challenges, and imprecise vocabulary choices, which hinder clarity. For example, some sentences are either unclear or fragmented, as seen in the phrase, "Giving children more freedom and allowing them to explore and learn on their own has many benefits," which could be revised for greater precision. Additionally, there are occurrences of awkward phrasing, such as "enabling them to develop both physically and mentally," which might be more effectively articulated as "promoting their physical and mental development." Moreover, the lack of transitional phrases between ideas detracts from the overall coherence of the text, and the conclusion is abrupt, lacking sufficient depth. Last, the essay exhibits repeated use of terms and a limited range of vocabulary, restricting the expression of ideas."

DeepSeek-R1 provided the following feedback on language use for the same essay: "The essay contains several language errors that hinder clarity and understanding. These include grammatical mistakes, such as "graduate of my university," "May be most students who like me," and "I was often talk," as well as incorrect word choices, such as "go to walk by you" and "getting by under false pretences." Additionally, awkward phrasing, such as "we have a same answer" and "concider it will be what" further complicates comprehension. These issues significantly affect the coherence of the essay and make it difficult for the reader to follow the argument. To enhance the essay, it is recommended to focus on maintaining consistent verb tenses, using the correct prepositions, and expressing ideas more clearly. Moreover, simplifying the language and avoiding unnecessary complexity can help improve the overall clarity."

"There are several language errors, most of which are significant grammatical issues that impair sentence correctness. For example, the sentence "Giving children more freedom and allowing them to explore and learn on their own has many benefits" misuses the verb "allowing," as it should be in a parallel structure with "giving." Additionally, "By constantly engaging in activities and experiencing" has an incomplete structure, lacking a clear subject and verb agreement. Phrases such as "is go to work" and "I was often talk about" exhibit multiple verb errors in one sentence, affecting overall coherence and clarity. These issues need careful revision to enhance sentence accuracy and clarity."

Organization

The feedback on organizations highlighted several concerns, including (a) cohesion and flow, and (b) structural issues.

All the rater groups noted the essays' lack of coherence, with some describing them as disjointed or barely coherent. Specific issues included missing transitions and unclear topic sentences, suggesting difficulties in logically connecting ideas. Additionally, all the raters noted problems with the overall logical flow. DeepSeek-R1 specifically mentioned the failure to establish a coherent structure and reach a meaningful conclusion, emphasizing challenges not only with the sequencing of ideas but also with effectively concluding and reinforcing the arguments.

The following specific feedback was provided by DeepSeek-V3, R1, and teachers regarding the organization of Essay #4 (see above):

> DeepSeek-V3: "The essay lacks a clear organizational structure, with ideas presented in a loosely connected manner. Transitions between points are weak, making it difficult to follow the logical flow of arguments. Additionally, the abrupt conclusion fails to reinforce the central argument effectively. The lack of transitional phrases and structured progression between ideas further contributes to the essay's overall incoherence."
>
> DeepSeek-R1: "The organization of the essay lacks clarity, making it difficult to follow the progression of ideas. While the essay adopts a "general-specific-general" structure with clearly defined points (first, second, third), there are inconsistencies in the way ideas are developed and linked. The absence of effective transitions between these ideas causes the essay to feel fragmented and unclear."
>
> Teacher Rater: "The essay's structure is unclear and disjointed, with ideas presented inconsistently. The lack of a proper conclusion leaves the argument unresolved. An effective argumentative essay requires a clear viewpoint and a coherent structure that supports the central argument."

Coherence

Coherence emerged as a critical area of concern among all the raters, primarily involving (a) cohesion and flow and (b) structural concerns. Across all groups, essays were frequently described as disjointed or marginally coherent, largely owing to missing transitions and unclear topic sentences that disrupted the logical progression between ideas. Additionally, structural weaknesses and illogical sequencing of ideas significantly impact essay quality. DeepSeek specifically emphasized the essays' failure to maintain coherence and provide meaningful conclusions, noting difficulties in idea sequencing and insufficient reinforcement of arguments.

Feedback examples from DeepSeek-V3, R1, and the teacher raters on the coherence of Essay #4 illustrate these points:

> DeepSeek-V3: "The transitions between paragraphs are abrupt, with a lack of linking words such as "for example" or "furthermore". Some ideas feel disconnected, especially the shift from "freedom" to "self-confidence," with no clear explanation of how one leads to the other."
>
> DeepSeek-R1: "The organization is hard to follow. Ideas are presented in a disjointed manner, with an abrupt transition from personal reflections to family discussions. The conclusion is vague ("So my choice is indeterminate"), leaving the reader unsure

of the stance. A clear structure-introduction, body paragraphs focused on specific reasons or examples, and a definitive conclusion-would greatly improve coherence."

One teacher rater: "The essay lacks clear organization, with an abrupt transition between ideas. The conclusion is vague, leaving the reader uncertain about the final stance. A well-defined introduction, focused body paragraphs, and a concise conclusion summarizing the main points are necessary for improvement."

The consistent feedback from DeepSeek-V3, DeepSeek-R1, and teacher raters highlights significant challenges in content development, language use, and organization. These findings underscore the need for targeted instructional interventions in grammar, vocabulary, argumentative structure, and essay organization to improve students' overall writing clarity and persuasiveness.

Discussion

This study adds to the growing body of research on AI-driven assessment tools, especially the DeepSeek platform, in the evaluation of EFL writing. By assessing the reliability of DeepSeek's scoring system and the relevance of its qualitative feedback, this study contributes to the ongoing exploration of AI's role in writing assessment (Chen et al., 2020; Kamalov et al., 2023). The findings highlight how AI technologies, such as DeepSeek, can enhance writing evaluation by providing both reliable scoring and constructive feedback that support student performance (Wilson et al., 2021).

In line with the trends observed in prior studies such as Jiang et al. (2023) and Guo (2024), our findings indicate a notable improvement in DeepSeek's capabilities. Earlier interactions of the platform, similar to the findings of Stevenson and Phakiti (2019), showed more variability, whereas the current version demonstrated higher reliability and closer alignment with teacher ratings. This suggests that AI tools, such as DeepSeek, have undergone substantial refinement, reinforcing Lim et al. (2023) conclusion that AI can complement and even enhance traditional assessment methods.

Comparative analyses of other AI tools such as ChatGPT (Bucol & Sangkawong, 2024; Dong, 2024; Lim et al., 2023; Tsai et al., 2024) show that while ChatGPT primarily offers linguistic-focused feedback, emphasizing improvements in vocabulary use and grammatical accuracy (Tsai et al., 2024; Zou & Huang, 2023), DeepSeek provides an evaluation that spans multiple dimension of writing, including structural coherence, content organization, and genre-specific rhetorical expectations. This holistic feedback approach may extend beyond surface-level corrections toward deeper, multidimensional skill development, supporting a broader range of writing competencies that conventional AI tools have often been criticized for overlooking (Bucol & Sangkawong, 2024; Zou & Huang, 2023).

From the perspective of G-theory, which views score reliability as the extent to which observed performance generalized across facets such as raters, tasks, and persons, the findings indicate strong reliability along the rater facet. In the G-study, person variance accounted for the majority of the total variance-64.6% for DeepSeek-V3, 88.2% for DeepSeek-R1, and 77.6% for teacher raters-suggesting that score differences mainly reflected students' true writing ability. Moreover, rater variance was minimal (2.8%, 1.7%, and 4.3%, respectively), indicating high scoring consistency across four teachers and DeepSeek.

Although each essay was rated only once by DeepSeek-R1, it still achieved a G-coefficient of 0.89 in the D-study, exceeding those of both DeepSeek-V3 (0.67) and the teacher raters (0.80). This finding demonstrates that DeepSeek-R1 maintains strong reliability even under minimal rating conditions. These findings suggest that DeepSeek-R1 provides scores that are consistent with human judgment and robust enough to generalize across raters. However, as the study was based on a single task and a relatively homogeneous sample, generalizability across tasks and learner populations remains limited.

Our study also supports Jiang et al. (2023), who highlighted AI's potential in aiding the writing process. We found that both versions of DeepSeek offered comparatively more comprehensive feedback than teacher raters in this study. Notably, DeepSeek-R1 provided a more balanced evaluation across language, content, and organization, while teacher raters tended to focus primarily on language-related aspects. This observation aligns with that of Brookhart (2017), who emphasized the need for more holistic feedback from educators.

Overall, the findings suggest a potentially expanding role for AI, particularly advanced models such as DeepSeek-R1, in enhancing assessment practices in EFL writing contexts. This aligns with Zhang and Zhang (2022), who argued for the importance of tools that motivate students and improve learning outcomes. By automating initial assessments and delivering continuous feedback, AI tools like DeepSeek enable educators to focus on more personalized instruction, a benefit also discussed by Zou and Huang (2023) in the broader context of AI applications in education.

Conclusion

The findings of this study highlight the pedagogical implications of DeepSeek's algorithmic capabilities in refining assessment methodologies for English writing, particularly regarding the reliability and relevance of DeepSeek-V3 and R1. Empirical data demonstrate that DeepSeek-R1 achieved higher generalizability coefficients than human raters, suggesting that algorithmic scoring systems may help reduce rater subjectivity in writing assessment contexts. Moreover, both versions of DeepSeek provided more comprehensive feedback across content, language use, organization, and coherence, offering broader coverage than typically observed in teacher feedback.

These findings indicate pedagogical implications for language assessment practice. The increased reliability of AI scoring can contribute to fairer assessments by enhancing consistency and reducing rater subjectivity. The comprehensive and multidimensional feedback offered by DeepSeek enables students to address a wider range of issues in their writing, from content development to organizational structure. For educators, integrating DeepSeek into assessment practices can complement traditional methods, offering AI-supported insights that enrich instruction and promote deeper writing awareness. Additionally, this integration may enable teachers to devote more attention on personalized guidance and targeted instruction.

Despite these contributions, several limitations should be acknowledged. First, the sample size was relatively small and drawn from a single university, limiting the generalizability of the findings. Second, the evaluation was based on a single writing task sourced from CET-4, which constrains the applicability of the results to other writing prompts, genres or task types. Third, all the teacher raters were from the same institution, potentially introducing institutional bias due to differences in grading practices across educational contexts. Finally, the study examined only two versions of DeepSeek, and the findings may not extend to other AI-driven assessment tools.

Future research should expand to more diverse learner populations, task types, and AI systems. Such studies would offer a more comprehensive understanding of the scalability and generalizability of AI-assisted writing assessment across different educational contexts. Despite the current limitations, the findings underscore the transformative potential of AI tools such as DeepSeek in enhancing English writing assessment. DeepSeek can support writing instruction by enhancing scoring reliability, offering more detailed feedback, reducing teacher workload, and facilitating personalized learning. Nonetheless, it is not without limitations. Nonetheless, it is not without limitations. While it performs well in surface-level language correction and structural analysis—such as detecting grammatical errors and offering foundational feedback—it may struggle to capture the nuances of complex arguments or fully contextualize specific responses. Recognizing these limitations is essential to ensuring that AI tools complement, rather than replace, human judgment in the evaluation of student writing.

This study provides an early exploration of DeepSeek's writing assessment performance via G-theory, which is more commonly applied in human-rated assessments. While DeepSeek demonstrated strong reliability and informative feedback on source-based prompts, it may still face challenges in interpreting nuanced arguments or fully contextualizing student responses. To maximize its effectiveness, future tool development should focus on improving prompt design and aligning feedback generation more closely with human evaluation standards, ensuring that AI functions as a pedagogical support rather than a substitute for expert judgement.

Appendix

Level	档次描述（中文）	**Description (English)**
14	译文准确表达了原文的意思。译文流畅，结构清晰，用词恰切，基本无语言错误，仅有个别小错。	The translation accurately conveys the meaning of the original text. The translation is fluent, well-structured, and uses appropriate vocabulary with almost no language errors, only a few minor mistakes.
11	译文基本表达了原文的意思。结构较清晰，语言通顺，但有少量语言错误。	The translation generally conveys the meaning of the original text. The structure is clear, the language is smooth, but there are a few language errors.
8	译文勉强表达了原文的意思。译文流畅度差，语言错误相当多，其中有些是重复错误。	The translation barely conveys the meaning of the original text. The fluency is poor, and there are many language errors, some of which are repeated.
5	译文仅表达了部分原文的意思。译文意思模糊，有较多的严重语言错误。	The translation conveys only part of the original meaning. The translation is vague, with many serious language errors.
2	除个别词语或句子，译文基本没有表达原文的意思。	Except for a few words or sentences, the translation hardly conveys the meaning of the original text.

Abbreviations

G-theory	Generalizability Theory
EFL	English as a Foreign Language
CET-4	College English Test Band 4
AI	Artificial Intelligence
CTT	Classical Test Theory
D-study	Decision Study
NLP	Natural Language Processing
MFRM	Many-Facet Rasch Model

Authors' contributions

The first author managed the research process, including the collection and analysis of data. The second author was responsible for coordinating the research planning and execution, as well as modifying the manuscript. The third author contributed to the coordination of the research activities and assisted in the planning and execution. All authors contributed to the manuscript and approved the submitted version.

Funding

The authors did not receive any funding for this research.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

This study was approved by the Faculty of Education, Universiti Kebangsaan Malaysia (UKM). It was granted permission to collect data for research purposes related to the study. All necessary approvals for data collection have been obtained from UKM.

Consent for publication

All authors have reviewed and approved the manuscript for publication.

Competing interests

The authors declare no competing interests.

Received: 8 March 2025 Accepted: 20 May 2025
Published online: 19 June 2025

References

Albuhairy, M. M., & Algaraady, J. (2025, February 2). *DeepSeek vs. ChatGPT: Comparative efficacy in reasoning for adults' second language acquisition analysis* [Preprint]. SSRN. https://ssrn.com/abstract=5137935

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* [Doctoral dissertation, University of Toronto]. TSpace. http://hdl.handle.net/1807/117862

Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing*. Routlegde.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Boggs, J. A. (2019). Effects of teacher-scaffolded and self-scaffolded corrective feedback compared to direct corrective feedback on grammatical accuracy in English L2 writing. *Journal of Second Language Writing, 46*, Article 100671. https://doi.org/10.1016/j.jslw.2019.100671

Brennan, R. L. (2001). *Variability of statistics in generalizability theory*. Springer.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology, 52*(1), 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

Brookhart, S. M. (2017). *How to give effective feedback to your students (2nd)*. Alexandria, VA: ASCD

Bucol, J. L., & Sangkawong, N. (2024). Exploring ChatGPT as a writing assessment tool. *Innovations in Education Teaching International*, 1–16. https://doi.org/10.1080/14703297.2024.2363901

Cardinet, J., Johnson, S., & Pini, G. (2011). *Applying generalizability theory using EduG*. New York, NY: Routledge

Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education, 36*(4), 395–407. https://doi.org/10.1080/03075071003642449

CERNET Authentication and Resource Sharing Infrastructure. (2025). *Schools and institutions (IdP)*. CERNET. https://www.carsi.edu.cn/IdPlist.html

Chen, B., & Zhang, J. (2022). Pre-training-based grammatical error correction model for the written language of Chinese hearing impaired students. *IEEE Access, 10*, 35061–35072. https://doi.org/10.1109/ACCESS.2022.3159676

Chen, D., Hebert, M., & Wilson, J. (2022). Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application. *American Educational Research Journal, 59*(6), 1122–1156. https://doi.org/10.3102/00028312221106773

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510

Creswell, J. W. (2021). *A concise introduction to mixed methods research (2nd ed.)*. Thousand Oaks, CA: SAGE Publications

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA*: A generalized analysis of variance system (ACT Technical Bulletin No. 43). Iowa City, IA: American College Testing

Cronbach, L. J. (1972). *The dependability of behavioral measurements: Theory of Generalizability for scores and profiles*. New York, NY: Wiley

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. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 323–325). IEEE. https://ieeexplore.ieee.org/document/10260740

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). Cognitive models of writing: Writing proficiency as a complex integrated skill. *ETS Research Report Series, 2008*(2), i–36. https://doi.org/10.1002/j.2333-8504.2008.tb02141.x

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R.,…Bi, X. (2025). *Deepseek-R1: Incentivizing reasoning capability in llms via reinforcement learning*. arXiv. https://doi.org/10.48550/arXiv.2501.12948

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B.,…Ruan, C. (2024). *DeepSeek-V3 technical report*. arXiv. https://doi.org/10.48550/arXiv.2412.19437

Denny, P., Khosravi, H., Hellas, A., Leinonen, J., & Sarsa, S. (2023). *Can we trust AI-generated educational content? Comparative analysis of human and AI-generated learning resources.* arXiv. https://doi.org/10.48550/arXiv.2306.10509

Dong, L. (2024). Exploring the interplay between writing feedback perception and lexical complexity among Chinese university students: A latent profile analysis and retrodictive qualitative modeling study. *Reading Writing, 37*(10), 2687–2706. https://doi.org/10.1007/s11145-023-10489-1

Ebadi, S., & Bashir, S. (2021). An exploration into EFL learners' writing skills via mobile-based dynamic assessment. *Education and Information Technologies, 26*, 1995–2016. https://doi.org/10.1007/s10639-020-10348-4

Guo, K. (2024). EvaluMate: Using AI to support students' feedback provision in peer assessment for writing. *Assessing Writing, 61*, Article 100864. https://doi.org/10.1016/j.asw.2024.100864

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

He, T., Li, H., Chen, J., Liu, R., Cao, Y., Liao, L., Zheng, Z., Chu, Z., Liang, J., Liu, M., & Qin, B. (2025). *A survey on complex reasoning of large language models through the lens of self-evolution* [Preprint]. ResearchGate. https://doi.org/10.13140/RG.2.2.23943.30886

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments?—A generalizability theory approach. *Assessing Writing, 13*(3), 201–218. https://doi.org/10.1016/j.asw.2008.10.002

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17*(3), 123–139. https://doi.org/10.1016/j.asw.2011.12.003

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

Jiang, Z., Xu, Z., Pan, Z., He, J., & Xie, K. (2023). Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning. *Languages, 8*(4), 247. https://doi.org/10.3390/languages8040247

Kamalov, F., Santandreu Calonge, D., & Gurrib, I. (2023). New era of artificial intelligence in education: Towards a sustainable multifaceted revolution. *Sustainability, 15*(16), 12451. https://doi.org/10.3390/su151612451

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing, 44*, Article 100450. https://doi.org/10.1016/j.asw.2020.100450

Koraishi, O. (2023). Teaching English in the age of AI: Embracing ChatGPT to optimize EFL materials and assessment. *Language Education Technology*, *3*(1). https://langedutech.com/letjournal/index.php/let/article/view/48

Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts*. Singapore: Springer

Li, J., & Huang, J. (2022). The impact of essay organization and overall quality on the holistic scoring of EFL writing: Perspectives from classroom English teachers and national writing raters. *Assessing Writing, 51*, Article 100604. https://doi.org/10.1016/j.asw.2021.100604

Li, W. (2022). Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: Insights from many-facet Rasch measurement. *Reading Writing, 35*(10), 2409–2431. https://doi.org/10.1007/s11145-022-10279-1

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System, 44*, 66–78. https://doi.org/10.1016/j.system.2014.02.007

Lim, K., Song, J., & Park, J. (2023). Neural automated writing evaluation for Korean L2 writing. *Natural Language Engineering, 29*(5), 1341–1363. https://doi.org/10.1017/S1351324922000298

Linacre, J. M. (1993, April). Generalizability theory and many-facet Rasch measurement. *Annual meeting of the American Educational Research Association*, Denver, CO, United States. https://eric.ed.gov/?id=ED364573

Mercer, S., Spillard, S., & Martin, D. P. (2025). *Brief analysis of DeepSeek R1 and its implications for Generative AI* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2502.02523

Mohamed, A. M. (2024). Exploring the potential of an AI-based Chatbot (ChatGPT) in enhancing English as a Foreign Language (EFL) teaching: Perceptions of EFL faculty members. *Education Information Technologies, 29*(3), 3195–3217. https://doi.org/10.1007/s10639-023-11917-z

Mozaffar, M., Liao, S., Xie, X., Saha, S., Park, C., & Cao, J.,…Gan, Z. (2022). Mechanistic artificial intelligence (mechanistic-AI) for modeling, design, and control of advanced manufacturing processes: Current state and perspectives. *Journal of Materials Processing Technology, 302*, Article 117485. https://doi.org/10.1016/j.jmatprotec.2021.117485

Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 394–403). Association for Computational Linguistics. https://aclanthology.org/2023.bea-1.32/

National Education Examinations Authority. (2016). *College English Test Band 4 and Band 6 Syllabus (2016 Revised Edition)*. National Education Examinations Authority. https://cet.neea.edu.cn/html1/folder/16113/1588-1.htm

Neal, M. R. (2011). Writing assessment and the revolution in digital texts and technologies. Teachers College Press.

Neittaanmäki, R., & Lamprianou, I. (2024). All types of experience are equal, but some are more equal: The effect of different types of experience on rater severity and rater consistency. *Language Testing in Asia, 41*(3), 606–626. https://doi.org/10.1177/02655322241239362

Nguyen, K. V. (2025). The use of generative AI tools in higher education: Ethical and pedagogical principles. *Journal of Academic Ethics*, 1–21. https://doi.org/10.1007/s10805-025-09607-1

Ofqual. (2013). *Introduction to the concept of reliability*. GOV. UK. https://www.gov.uk/government/publications/reliability-of-assessment-compendium/introduction-to-the-concept-of-reliability

Pearson, W. S. (2022). The mediating effects of student beliefs on engagement with written feedback in preparation for high-stakes English writing assessment. *Assessing Writing, 52*, Article 100611. https://doi.org/10.1016/j.asw.2022.100611

Pedro, F., Subosa, M., Rivas, A., & Valverde, P. (2019). *Artificial intelligence in education: challenges and opportunities for sustainable development* (Working Papers on Education Policy). UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000366994

Plano Clark, V. L. (2017). Mixed methods research. *The Journal of Positive Psychology, 12*(3), 305–306. https://doi.org/10.1080/17439760.2016.1262619

Rae, A. M., & Cochrane, D. K. (2008). Listening to students: How to make written assessment feedback useful. *Active Learning in Higher Education, 9*(3), 217–230. https://doi.org/10.1177/1469787408095847

Sallam, M., Al-Mahzoum, K., Sallam, M., & Mijwil, M. M. (2025). DeepSeek: Is it the end of generative AI monopoly or the mark of the impending doomsday? *Mesopotamian Journal of Big Data*, *2025*, 26–34. https://doi.org/10.58496/MJBD/2025/002

Stevenson, M., & Phakiti, A. (2019). Automated feedback and second language writing. In F. Hyland & K. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 125–142). Cambridge University Press.

Tayyebi, M., Abbasabady, M. M., & Abbassian, G.-R. (2022). Examining classroom writing assessment literacy: A focus on in-service EFL teachers in Iran. *Language Testing in Asia, 12*(1), 12. https://doi.org/10.1186/s40468-022-00161-w

Tsai, C.-Y., Lin, Y.-T., & Brown, I. K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. *Education Information Technologies, 29*, 22427–22445. https://doi.org/10.1007/s10639-024-12722-y

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2019). Using generalizability theory with continuous latent response variables. *Psychological Methods, 24*(2), 153. https://doi.org/10.1037/met0000177

Wang, Y., & Xie, Q. (2022). Diagnostic assessment of novice EFL learners' discourse competence in academic writing: A case study. *Language Testing in Asia, 12*(1), 47. https://doi.org/10.1186/s40468-022-00197-y

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 reliability coefficients and generalizability theory. *Handbook of Statistics, 26*, 81–124. https://doi.org/10.1016/S0169-7161(06)26004-8

Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education, 168*, Article 104208. https://doi.org/10.1016/j.compedu.2021.104208

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology, 56*(1), 150–166. https://doi.org/10.1111/bjet.13494

Yu, S., & Liu, C. (2021). Improving student feedback literacy in academic writing: An evidence-based framework. *Assessing Writing, 48*, Article 100525. https://doi.org/10.1016/j.asw.2021.100525

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1–27. https://doi.org/10.1186/s41239-019-0171-0

Zhai, C., & Wibowo, S. (2023). A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university. *Computers Education: Artificial Intelligence, 4*, Article 100134. https://doi.org/10.1016/j.caeai.2023.100134

Zhang, X. S., & Zhang, L. J. (2022). Sustaining learners' writing development: Effects of using self-assessment on their foreign language writing performance and rating accuracy. *Sustainability, 14*(22), 14686. https://doi.org/10.3390/su142214686

Zheng, Y., Wang, Y., Liu, K. S.-X., & Jiang, M. Y.-C. (2024). Examining the moderating effect of motivation on technology acceptance of generative AI for English as a foreign language learning. *Education Information Technologies*, 1–29. https://doi.org/10.1007/s10639-024-12763-3

Zou, M., & Huang, L. (2023). To use or not to use? Understanding doctoral students' acceptance of ChatGPT in writing through technology acceptance model. *Frontiers in Psychology, 14*, 1259531. https://doi.org/10.3389/fpsyg.2023.1259531

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.