AsiaCALL Online Journal
ISSN 1936-9859; https://asiacall-acoj.org
Vol. 16, No. 1, 2025
pp. 327-343

Received: 13/10/2024
Revision: 21/05/2025
Accepted: 23/05/2025
Online: 26/05/2025

Using Gemini for Formative Assessment in English Academic Writing - Critical Insights into The AI Tool's Efficacy

Nguyen Dinh Luat[1], Le Pham Thien Thu[1], Le Thi Thuy[1*]

[1]Faculty of Foreign Languages, Industrial University of Ho Chi Minh City, Vietnam
[*]Corresponding author's email: lethithuy@iuh.edu.vn
https://orcid.org/0000-0003-4416-4579
https://doi.org/10.54855/acoj.2516117

Received: 13/10/2024 Revision: 21/05/2025 Accepted: 23/05/2025 Online: 26/05/2025

ABSTRACT

The emergence of Artificial Intelligence (AI) has triggered revolutionary transformations in language teaching and learning. When it comes to academic writing, current educational practitioners must more than once wonder which AI-powered tools, among the overwhelming number mushrooming recently, can assist their learners' self-study by providing reliable and relevant feedback. This paper explores the effectiveness of Gemini, a large language model (LLM) developed by Google AI, in providing rubric-aligned commentary on student essays. The article employed a mixed-methods approach in which quantitative data are collected from academic writing samples while qualitative data are coded from Gemini-assisted feedback. Through the critical analysis of the comments provided by Gemini on twenty students' essays, against the IELTS Writing Task 2 band descriptors, Gemini's feedback tends to be more consistent when it comes to task achievement and coherence and cohesion, with rubric or band descriptors included in the prompt. Within each criterion in the rubric, the initial indicators tend to be more adequately examined. Also, paragraphing, spelling, and punctuation are the indicators that are neither consistently nor sufficiently commented on. These findings lay a foundation for language educators to evaluate the efficacy of LLM-assisted learning tools in academic writing education, paving the way for their proper application in classroom instruction.

Keywords: AI-powered tools, consistency, essay assessment, rubrics, band descriptors

Introduction

The past decade has documented remarkable technological evolution, especially after the global disruption of the COVID-19 pandemic. The contemporary academic landscape is witnessing a heightened integration between technology and teaching, learning, and assessment. Various assessment strategies, particularly those leveraging technology, are increasingly recognized as

CITATION | Nguyen, D. L., Le, P. T. T., & Le, T. T. (2025). Using Gemini for Formative Assessment in English Academic Writing - Critical Insights into The AI Tool's Efficacy. *AsiaCALL Online Journal*, *16*(1), 328-343. DOI: https://doi.org/10.54855/acoj.2516117

essential tools for reducing the teachers' workload while still providing valuable feedback for students to foster their engagement and cultivate their lifelong learning skills, as Meenakumari (2021) determined in his research that AI is mainly used in the process of formative evaluation and also for the automatic grading of students.

Among the large language models (LLMs), Gemini has emerged as one of the latest and most convenient tools to assist language teachers, especially those who teach writing skills, in assessing students' writing and empowering students in self-regulated learning by providing formative feedback aligned with the predefined International English Language Testing System (IELTS) test.

The IELTS Writing Task 2 Band Descriptors are the standardized rubric that can be used to assess students' writing proficiency. These descriptors evaluate four key criteria: Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. Task Response measures how well the essay addresses the prompt, while Coherence and Cohesion assess the organization and logical flow of ideas. Lexical Resource evaluates the range and accuracy of vocabulary used, and Grammatical Range and Accuracy determines the student's command of grammar and sentence structures. By comparing students' performance on the essay writing test with formative feedback provided by Gemini based on the IELTS Writing Task 2 Band Descriptors, educators can gauge how well students meet these criteria and identify areas for improvement.

Although several studies have been conducted to investigate the utilization of Gemini in automated essay scoring and providing timely feedback, not many have examined the consistency of Gemini in delivering formative feedback aligned with the rubric or rating scales from a standardized international test. Thus, this study investigates the efficacy and reliability of Gemini in two key roles: as a tool for teachers to provide formative feedback on academic essays based on IELTS rubrics and as a resource for language learners to engage in self-directed learning. It explores how effectively Gemini can facilitate academic writing assessment and enhance students' writing skills in alignment with a recognized international standardized test like IELTS.

Literature review

An Overview of Essay Writing in Academic Disciplines

Essay writing is a fundamental skill for students in various academic disciplines, and learners can express their feelings, ideas, and experiences and argue their perspectives (Rosenfeld, Courtney, & Fowles, 2004). Different researchers define essay writing in several educational settings and demonstrate the key aspects of essay writing, which provide a comprehensive understanding of this skill in academic contexts.

Structure of an Essay

Essays were traditionally defined as "a short piece of writing on a particular subject" (Oxford English Dictionary, 2023); however, according to Barley (2014), essay writing at a higher education level is much more than just writing on a topic; it requires a systematic approach to exploring and organizing ideas, and demonstrating critical thinking to support the writers' perspectives. The author also emphasizes the requirements of clarity, coherence, and well-structured organization, typically including an introduction, body, and conclusion. The introduction of an essay presents the topic and thesis statement, the body paragraphs provide supporting arguments and evidence, and the conclusion summarizes the main discussed points.

Similarly, Cumming et al. (2000) and Biber et al. (2004) advocate that an academic essay is a document that has a defined structure – an introduction, a body, and a conclusion. In writing academic essays, students are required to present a thesis statement and support it with details and strong evidence. Mukiminin (2012) shares a similar idea that writing is not only about grammar but also about organizing ideas, mechanics, and cohesion to make good writing. Consequently, a writer produces a text to reflect the structure of their argument, using coherence and cohesive devices to guide readers through their ideas (Swales & Feak, 2004). Therefore, it is noted that academic writing is complex, with a demanding level of clarity, precision, and critical engagement with ideas (Paltridge & Starfield, 2016).

Mastery of Essay Writing

Writing is considered one of the highest forms of performance in academic skills, reflecting a person's language competence. Nunan (1991) emphasizes that mastering this skill requires critical thinking, researching, and organizing skills, which is a significant challenge for learners. It is commonly shared among many authors that good writing goes beyond grammatical accuracy, as it involves showing clear purposes, presenting a particular point of view, and supporting it with organized and coherent information (Chenoweth & Hayes, 2001; Cumming, 2001; Sasaki, 2000; Weissberg, 2000; Wiseman, 2012). Linguistic competence is insufficient to ensure effective writing, requiring more unity, cohesion, and coherence. Therefore, it is much more challenging for language learners to create a coherent text with clarity and logical flow.

Essay Assessment: Bias and Objectivity

In second language learning, assessing writing proficiency is a critical issue that presents challenges of bias and subjectivity (Hamp-Lyons, 2003; Anderson, 2005; Brown & Abeywickrama, 2010). Several studies highlight the impact of rater bias on L2 writing assessments (Fahim & Bijani, 2011, p.1) and inconsistencies in raters' evaluations (Kondo-Brown, 2002; Schaefer, 2008). The use of standardized rubrics is a recommendation from scholars to reduce the level of bias and ensure consistent assessment among raters (Brown & Jaquith, 2007; Hamp-Lyons, 2007; Jonsson & Svingby, 2007; Aryadoust & Riazi, 2016). The reason is the clear evaluation criteria that promote reliability and validity in evaluating writing performance (Biggs & Tang, 2007; Dunsmuir & Clifford, 2003; Spurr, 2005). With the standard framework, rubrics assist raters and examiners in applying consistent standards to all writing tasks; as a consequence, they reduce subjectivity and bias in essay evaluation.

Standardized Rubrics for Essay Assessment in Academic Writing

As an invaluable tool for assessing essay writing, analytical lists or rubrics provide a well-structured framework that ensures validity and objectivity in this learning practice. Specific criteria and descriptive indicators in rubrics guide raters in evaluating students' performance. Different criteria are set relevant to an assignment, assessment, or learning outcome, and the possible levels of achievement are stated in a specific and objective way. Talevski et al. (2014) and Moskal and Leydens (2000) claim that the indicators describe the quality of students' writing that match the criteria; therefore, instructors based on the matching features to assess their students' work fairly, consistently and efficiently. Assessed by rubrics, students receive formative feedback on their strengths and weaknesses to identify the areas that need improvement and enhance their performance. The utilization of the same criteria ensures consistency in assessment among evaluators. In summary, standardized rubrics enhance precision and measurability and promote fairness and reliability of criteria for assessing essays in academic settings.

Most authors agree that a rubric typically has criteria, standards or performance levels, and descriptors. Brookhart (2013) stated that rubrics include "appropriate criteria and well-written performance descriptions." According to The University of Texas at Austin (n.d.), rubrics include "performance criteria, rating scale, and indicators". Accordingly, this study would employ the term "indicators" to refer to the core descriptors within each criterion in the IELTS Writing Band Descriptors.

Assessment in language teaching and learning

Assessment has been considered a crucial component of language teaching and learning since it serves multiple purposes. Numerous studies have attempted to confirm the assessment's key roles in teaching and learning.

Assessment Key Roles

The study by Alderson, Clapham, and Wall (1995) revealed that evaluating students' progress is vital to understanding the effectiveness of language learning programs. Teachers can determine if their teaching approaches lead to the expected objectives by comparing students' achievement to set learning-specific goals. In this case, assessment works as a measurement of the learning outcomes. This view is supported by Hughes (1989), who writes that assessment can be used to hold schools and teachers accountable for student achievement. Educators can justify the resources and time invested in language learning programs by demonstrating that students progress. Through assessments, educators can track students' progress over time to ensure the effectiveness of the present curriculum or teaching methods and adjust them if necessary. Moreover, when analyzing the assessment data, teachers can locate the areas where students need additional support (Black & William, 1998). This is not only essential for the teacher but also significant for the learners themselves. Once the learners can identify their strengths and weaknesses, they have a chance to recall their achievements and then push themselves forward to make remarkable improvements. An intriguing research study conducted by Gardner (2000) investigating the attitude and motivation in second language learning also indicated that well-designed assessments can serve as a powerful motivator by providing learners with a sense of accomplishment and progress. Learners who see themselves making strides are more likely to remain engaged and committed to their language learning goals.

Two types of assessment: summative and formative

Among the various assessment methods, there are two outstanding types: formative assessment and summative assessment. While both types aim at assessing students' performance, they have their distinct purposes and are applied at different stages of the learning process.

Bachman (1990) insisted that formative assessment offers ongoing feedback that helps learners identify strengths and weaknesses and adjust their learning strategies accordingly. In the line, Heritage (2007) also emphasizes that formative writing is a systematic process of gathering evidence about learning continuously to identify the gap between a student's current level of learning and their desired learning goal. This formative feedback informs students what the next step in learning should be and guides them forward. This kind of assessment showcases a student's progression during the learning process instead of giving the final score to certify students' final achievement as a summative assessment (Gikandi et al., 2011). As stated by this author, a summative tool is not enough to measure learners' progress because summative is often conducted at the end of the semester or an academic year to evaluate the overall achievement of the students. Feedback from the formative evaluation is regarded as qualitative as it is descriptive and focuses on the quality of writing instead of showing what is right or

wrong. Comments and suggestions guide students in revising and building their writing skills and enhancing their language proficiency in the long run.

In brief, the formative tool is less formal since it happens during the learning process to provide immediate feedback by focusing more on improvement rather than grades. This gives students more timely chances to learn from mistakes, which benefits their lifelong learning. Alternatively, formative feedback accompanies the students' learning journey by providing comments or remarks so they can self-correct on the way to self-progress.

Formative Assessment in English Academic Writing

As Black and William (2009) pointed out, formative assessment plays a crucial role in academic writing by providing students with timely feedback that helps them revise and improve their work before it is formally assessed. Ongoing and constructive feedback in formative assessment is provided during the learning process, aimed at helping students improve their writing skills. Similarly, a caution was sounded by Andrade and Cizek (2010), who noted that through formative assessment, students gain a deeper understanding of the writing process as they actively engage with feedback to enhance their writing skills. Moreover, a study by Nicol and Macfarlane-Dick (2006) found that formative assessment encourages iterative writing practices, allowing students to continuously refine their drafts, ultimately leading to higher-quality academic writing. When integrated with standardized test criteria like IELTS, formative feedback can define students' academic writing strengths, weaknesses, and common problems. This feedback can be tailored to help students focus more on their limitations in specific areas, such as task achievements, grammar, coherence, and cohesion, or lexical resources to set specific goals for their writing enhancement. By aligning formative feedback with IELTS rubrics, students can see how their skills are evolving and adjust their study strategies accordingly.

Utilizing AI-powered feedback for assessment in Academic Writing

Since its emergence, AI-powered tools have revolutionized how people learn and teach with their special functions, which involve using algorithms that can analyze data, identify patterns, and make predictions (Harry & Sayudin, 2023). AI's numerous intelligent features allow educators to personalize learning for each student and support them in leveraging more efficient learning evaluation. Furthermore, Meenakumari (2021) highlights that these algorithms can be used to assess students in a real-time environment to provide constructive and progressive feedback at scale and to drive scaffolds to students while they learn, blending assessment and learning without using the critical instructional time for evaluation. Additionally, in their work, Mizumoto and Eguchi (2023) discuss that AI-powered tools can be used as an Automated Writing Evaluation tool by providing detailed feedback to both learners and teachers with such benefits as mitigating evaluation bias as well as allowing teachers to focus on the crucial writing aspects like overall structure, coherence, content, and writing strategies. With the same sense, research by Kartika (2024) on the impact of Google Gemini feedback on writing proficiency also proves remarkable improvements in students' essay writing skills, especially in these aspects: grammar, vocabulary, coherence, and overall task achievement after applying AI's immediate and specific feedback for their revision and self-practice.

Review of previous studies

Several scholars have recently investigated the impacts of AI language models on language assessment efficacy (Mizumoto & Eguchi, 2023; Dong, 2023; Mahapatra, 2024). Mizumoto and Eguchi (2023) evaluated 12,100 essays by utilizing ChatGPT to investigate the feasibility of using an AI language model (i.e., GPT) for automatic essay scoring (AES). The study

suggests that ChatGPT can be employed as a viable alternative for automated evaluation and providing feedback on L2 writing. ChatGPT shows its effectiveness in grading students' writing and providing feedback promptly, which assists students in identifying mistakes and rewriting to correct them. In alignment with previous research, Dong (2023) advocates that AI tools improve learners' writing scores as students' performance has progressed after employing the tools for reviewing the writing paper. Furthermore, it also facilitates the teaching process by offering personalized feedback and increasing student engagement. Another study by Steele (2023, as cited in Nguyen & Pham, 2024) also points out the positives of utilising ChatGPT in academic settings in "assessing students' competencies and verifying the correctness of the subject matter" (p.61).

Additionally, Mahapatra (2024) (as cited in Truong et al.,2025) highlights that as a tool for providing formative feedback, ChatGPT can be harmoniously integrated into large writing classes. Tailored formative feedback provided by ChatGPT significantly improves students' academic writing. ChatGPT can be a reliable resource when conducting a number of time-consuming tasks in large writing classrooms, including timely input on writing organization, lexical range, and grammatical accuracy (Xiao et al., 2024).

Some recently published studies further explore specific AI tools, namely Google Gemini, in educational contexts. Lang et al. (2024) contend that GPT-4 and Gemini significantly improve students' writing skills by providing valuable feedback. Learners critically analyze their mistakes and receive suggestions to improve the quality of their writing. Kartika (2024) consistently reveals that when AI feedback by Google Gemini is integrated into educational practices, it can substantially enhance writing outcomes among language learners. Specifically, AI feedback systems, like Google Gemini, offer immediate corrections and suggestions, which can lead to improved writing skills. These systems encourage self-directed learning, allowing students to engage with their writing actively.

However, despite the increasing interest in AI-assisted feedback among researchers, limited empirical research examines Gemini's effectiveness and consistency in providing formative feedback for educators and learners that motivates further exploration into the field, especially when it comes to the inclusion of a standardized rubric or rating scale in the prompt so that the AI tools' feedback could be more criterion-oriented. The gap in the scholarly studies drives the necessity of this research topic.

Research Questions

To achieve the research objective, this study aimed to address the following questions:

1. To what extent can Gemini provide consistent feedback for each criterion?
2. To what extent can Gemini provide consistent feedback for each indicator?
3. For which indicator does Gemini tend to provide sufficient feedback?

Methods

Pedagogical Setting & Participants

The study was conducted at the Faculty of Foreign Languages at a Ho Chi Minh City university. Writing skills play a vital role in enhancing learners' competence. In the Writing 2 classes, English-majored sophomores were required to produce essays of various kinds, including advantage-disadvantage, cause-effect, and opinion essays. Students must complete an opinion essay on different topics in the final test.

Twenty essays written by English-majored sophomores were selected as the samples for analysis after being scored by the examiners. These essays were opted randomly from 14 classes of the semester and varied in the score range from 2.0 to 9.0 points.

Table 1

Number of samples for analysis

The score range was given by examiners	2.0	3.5	4.0 - 4.5	5.0 - 5.5	6.0 - 6.5	7.0 - 7.5	8.0	8.5 - 9.0	**Total**
Number of essays	1	1	3	3	4	3	2	3	**20**

Design of the Study

A quantitative approach was employed to achieve the research objectives. Qualitative data collected from Gemini's feedback was imported into an Excel sheet. After encoding the criterion and indicators, the data were transformed into a numerical form, indicating the frequency percentage and illustrating the consistent level of each criterion and the indicators. The data processing procedures are described in detail below.

Data collection and analysis

After collecting the writing samples from the students, the researchers started to conduct the following steps:

Step 1: Encoding the IELTS Writing Task 2 Band Descriptors.

The certified IELTS writing examiners used the IELTS Writing Task 2 Band Descriptors to evaluate the contestants' essay writing proficiency according to four main criteria: Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range & Accuracy. Each of these criteria assesses some specific indicators related to academic writing ability. The indicators refer to the underlying skills in writing or the competency that each criterion aims to measure. The researchers needed to define the key points in the indicators in each criterion of the band descriptors to encode them. To easily analyze the data later, the IELTS Writing Task 2 Band Descriptors were encoded as:

Table 2

The encoding table of IELTS Writing Task 2 Band Descriptors

Task Response Indicator Codes	Task Response Indicator	**Coherence & Cohesion Indicator Codes**	Coherence & Cohesion Indicator	**Lexical Resource Indicator Codes**	Lexical Resource Indicator Codes	**Grammatical Range & Accuracy Indicator Codes**	Grammatical Range & Accuracy Indicator
TR1	Prompt relevance	CC1	Organization of ideas	LR1	Vocab accuracy and flexibility	GA1	Structure variety
TR2	Position/ viewpoint	CC2	Cohesion - ideas connection (cohesive devices)	LR2	Vocabulary variety (quality/ word choice & quantity)	GA2	punctuation + grammar accuracy & appropriacy
TR3	Ideas development (lapses in content relevance and idea support)	CC3	Lapses - ideas org & connect	LR3	Spelling + word formation	GA3	error impact on communication
		CC4	Paragraphing				
		CC5	Extra indicator				

After that, all researchers worked together to agree on how to define the descriptors and indicators exactly to unify the way of working, even when they work independently.

Step 2: Applying the prompt to scan for feedback on the Gemini platform

The researchers utilized the same prompt attached to the writing samples on three different laptops at different points of time to scan for feedback from the Gemini platform. There were two contrasting ways of scanning: one with the band descriptors enclosed in the prompt and one without the band descriptors.

Figure 1

Example of the prompt used in the study

Prompt: Act as a certified IELTS writing examiner, score this student's essay according to the following IELTS Writing Task 2 band descriptors and give judgment for the score based on four criteria: Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy, as detailed in the enclosed band descriptors.

This is an essay question for the IELTS writing task 2: *[inserting each of the essay topics]*

This is the student's essay: *[inserting each of the sampling essays]*

Step 3: Using the encoding table as a tool to take notes

The encoding table of IELTS Writing Task 2 Band Descriptors was uploaded on Google Sheets with some main regulations as mentioned below:

a. If the indicator was mentioned in Gemini Feedback, would note 1
b. If the indicator was not mentioned in Gemini Feedback, would note 0

Step 4: Comparing the feedback achieved from Gemini and the key indicators

The comparison between the feedback received from Gemini and the key indicators from the Band Descriptors is noted as 1 when one of the indicators in the Band Descriptors is mentioned in Gemini's feedback and as 0 when it is not referred to. There was a norming step between the researchers to agree on comparing the relevance. The norming step helps the researchers have a consistent viewpoint about the comparison so that the encoding steps can be as objective as possible.

Step 5: Analyzing the consistency of the data:

In the context of this study, consistency refers to the similarity among Gemini's comments across the computers, in which a specific indicator is either mentioned in all the computers' comments concerning that indicator or ignored altogether.

Analyzing the consistency of the indicators for each criterion between the three different computers.

- The indicator was considered as consistently assessed when Gemini's comments from all three computers either mentioned or failed to mention the same construct.
- The indicator was regarded as inconsistently assessed when not all the computers mentioned the indicator in Gemini's comments.

Step 6: Verifying the consistent comments

The total number of consistent comments was then converted into percentages to verify the consistency of comments given by Gemini in comparison with the Band Descriptors.

The final stage is to generate a discussion from these findings.

Findings and Discussion

The results of the investigation depict significant consistency with the findings of previous studies. However, since the data collection is more criterion-oriented, with the IELTS band descriptors included in the prompts, more specific and original findings exist to explore.

The level at which Gemini can provide consistent feedback for each criterion

As mentioned earlier, the first issue to be addressed is how consistent Gemini is in providing feedback on the students' essays. In other words, the focal piece of data to be examined is the percentage of comments that are consistent across the three computers.

The table above represents the level of consistency concerning each criterion in the band descriptors. The most remarkable observation is that when the band descriptors are included in the prompt, Gemini's comments tend to be more consistent in referring to the indicators in each criterion, with the number of consistent comments accounting for 70 percent or more of all the comments offered. On the other hand, the number of consistent comments in the case of band descriptor withdrawal from the prompts reaches only 35 to 75 percent of all the comments. Another crucial tendency is that the criteria of task achievement, and coherence and cohesion show a higher level of consistency, with the percentages of consistency comments ranging from 81 to approximately 87 percent with band descriptors and 75 to 77 percent, respectively, without band descriptors. With sufficient feedback on task achievement, coherence and cohesion,

lexical resource, grammatical range and accuracy, these results reveal the similarity with what Xiao and the co-authors concluded in their study in 2024 about ChatGPT as a reliable source of input on writing organization, lexical range, and grammatical accuracy. This confirmation at the overall level continues to be repeated when a more detailed analysis is conducted on how consistent Gemini's feedback is concerning each indicator and on which indicators Gemini tends to provide sufficient feedback.

Table 3

Criterion Consistency

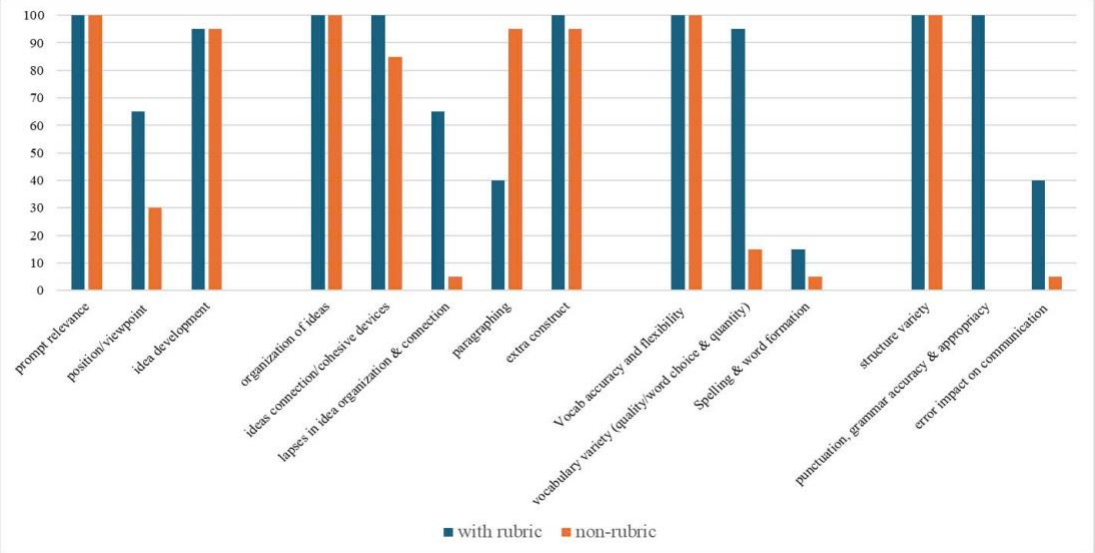
Criteria	Total number of comments/ descriptors	Number of consistent comments		Percentages	
		With band descriptors	Without band descriptors	With band descriptors	Without band descriptors
Task achievement	60	52	45	**86.7 %**	**75.0 %**
Coherence & cohesion	100	81	77	**81.0 %**	**77.0 %**
Lexical resource	60	42	24	**70.0 %**	**40.0 %**
Grammatical range & accuracy	60	45	21	**75.0 %**	**35.0 %**

The level at which Gemini can provide consistent feedback for each indicator

To evaluate the consistency of Gemini's feedback, the collected data was transformed into a chart to verify the frequency of each indicator within the band descriptors.

The graph below demonstrates the percentage of each indicator in descriptors referenced in Gemini's feedback.

Figure 2
Indicator Consistency



The data illustrate a higher consistency level in Gemini's comments when the rubric or band descriptors were included in the prompts. This is evident in higher percentages of consistent comments, including the rubric, among the prompts. The paragraphing indicator exhibits an opposite tendency, in which 95 percent of the comments provided by Gemini when the band descriptors were not included in the prompts are consistent, as opposed to only 40 percent among the comments with the band descriptors being found consistent.

Additionally, the initial indicators in each criterion of the band descriptors tend to be more consistently assessed than the other indicators, with the percentages of consistent comments reaching 100 percent, whether the band descriptors were included or excluded from the prompts. Those indicators with a percentage of consistent comments reaching under 70% include position or viewpoint, lapses in ideas organization and connection, spelling and word formation, error impact on communication, vocabulary variety (without the band descriptors), and paragraphing (with the band descriptors).

The indicators with the significant differences included those related to position or viewpoint, lapses in idea organization and connection, paragraphing, vocabulary variety, and error impact on communication, among which vocabulary variety is the one with the most dramatic difference, reaching approximately 80 percent.

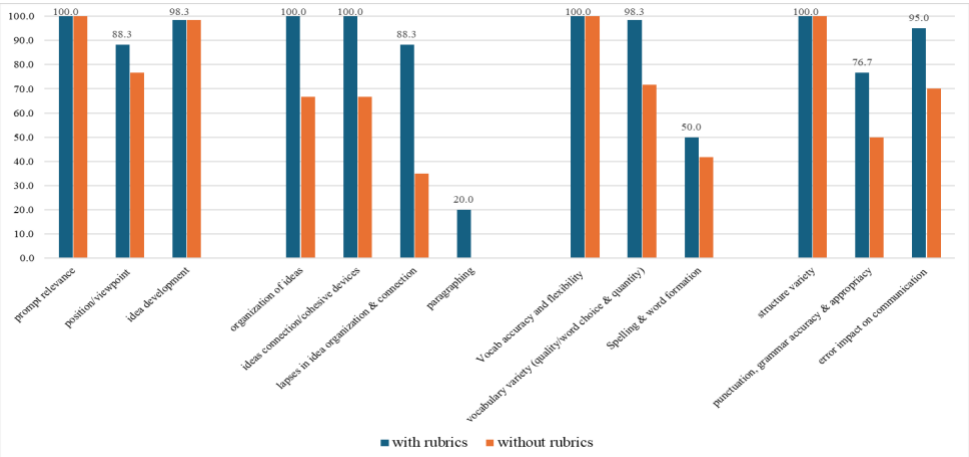
Last but not least, spelling and word formation are the indicators with the lowest level of consistency, with the percentage of consistent comments when the band descriptors included reaching only 15 percent, compared with approximately 5 percent of the consistent comments in the non-rubric group.

The indicators with Gemini's sufficient or insufficient feedback

However, the data concerning indicator consistency reveals how consistent Gemini tends to be when giving feedback on the indicators. However, what is even more significant is how sufficient the feedback on each indicator is, i.e., whether the feedback refers to the indicator, because indicator consistency means all of Gemini's feedback generated by all the computers either includes or makes no reference to the indicators. Ideally, the feedback should refer to as many indicators in each criterion as possible.

The chart below delves into the sufficiency of Gemini's feedback with reference to each indicator and compares the level of sufficiency between the prompt that included the band descriptors and those that did not.

Figure 3
Indicator Reference



As can be seen from the chart, the prompts with the Band Descriptors give feedback on the given indicators more frequently than the others. Another noticeable tendency is that the first indicator from each criterion, regardless of whether the prompts included the band descriptors or not, tends to refer to the indicator. When the prompts with the Band Descriptors are separately taken into account, it is noteworthy that while most indicators amount to 88% of reference, those related to paragraphing, spelling, word formation, punctuation, grammar accuracy, and appropriacy depict a great deal lower percentage, with paragraphing serving as an indicator with the least frequent reference.

Discussion

Addressing the same issue of how to utilise such AI-powered tools as Gemini for promoting the teaching and learning of language skills, the study differentiates itself from prior research in its investigative target.

Other previous studies explored the effectiveness of several popular AI tools, such as ChatGPT or Gemini, in enhancing students' writing skills by relying on those tools' feedback. Specifically, those authors recommended integrating the tools in teaching and learning writing skills due to their efficacy in providing useful feedback on students' work and suggestions for learners' improvement (Mizumoto & Eguchi, 2023; Dong, 2023; Mahapatra, 2024; Xiao et al., 2024; Lang et al., 2024; Kartika, 2024).

On the other hand, the current study focuses on evaluating the reliability or consistency of feedback and comments provided by Gemini. The results indicate that with carefully designed prompts and band descriptors included, Gemini gave more consistent feedback on the criteria of task achievement, cohesion and coherence than lexical range and grammatical accuracy. For each indicator, the consistent feedback was revealed in the initial indicators, whereas some inconsistencies were shown in the indicators of position viewpoints, lapses in idea organisation and connection, paragraphing, spelling and word formation, and error impacts on communication.

Conclusion

While confirming the findings of previous studies, the investigation could further research literature by indicating a higher degree of consistency in Gemini's feedback based on the criteria in the band descriptors when the rubrics are included in the prompts. Although all the four criteria show statistically acceptable levels of consistency, Task Achievement, Coherence, and Cohesion tend to be more consistent. When indicators are taken into account, the initial indicators within each criterion tend to be more sufficiently and consistently assessed, while spelling, word formation, and paragraphing are neither sufficiently nor consistently assessed, even with the band descriptors. In contrast, the position or viewpoint and lapses in idea organization and connection, though sufficiently mentioned, depict a low level of consistency.

Recommendations for instructors for formative feedback

As the findings of the study suggest, specific rubrics are essential for Gemini to provide detailed feedback. When using Gemini to obtain formative feedback, more attention should be paid to indicators such as the writer's viewpoint or position, lapses in logic/ connecting ideas, paragraphing, spelling, and word formation due to their potential inconsistencies or insufficient feedback. It is crucial to provide students with a foundational understanding of paragraphing

and grammatical issues as these issues were not fully referenced in the Gemini's comments.

Recommendations for students using Gemini for self-study

To begin the practice of using Gemini for students' self-study or feedback preparation, proper attention should be paid to the content of the prompts, in which the rubrics need to be included in the prompts and the AI's role be described with details possibly including the topic, context, expertise or purpose as well as instructions for the precise task should be given.

For the teachers to use Gemini as a source of reference for formative feedback and for students to effectively use Gemini as a self-study platform, the instructions on the writing of academic essays, especially opinion essays, should be properly provided initially. In those instructions, the knowledge of paragraphing, viewpoint or position verbalization, coherence and cohesion strategies, paragraphing, spelling, and word formation must be the focal points before the knowledge and skills of how to use rubrics, how to generate appropriate and effective prompts and how to accurately interpret Gemini's feedback are taught to ensure that they apply the tool effectively. Once those tasks are thoroughly completed, students' overall writing skills can be enhanced.

Limitations of the study

One of the limitations of the study is the relatively small sample size, consisting of just 20 students' essays. The study only focuses on evaluating the frequency of criteria and indicators referenced without verifying the accuracy of Gemini's feedback. Furthermore, the study focuses on the level of consistency and sufficiency among Gemini's comments rather than exhaustively examining the detailed content of the comments to test their reliability compared with the descriptors in the band descriptors.

Implications

The study recommends that band scores provided by Gemini should be considered as a reference only, as there is no benchmark between the band scores suggested by Gemini and the examiners' scores. Further studies could expand the sample size and deeply investigate the accuracy or reliability of Gemini's comments by comparing them with the details in learners' writing papers.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Anderson, C., & Anderson, C. (2005). *Assessing writers*. Heinemann.

Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Bachman, L. F. (1990). Fundamental considerations in language testing. In B. D. Shavelson, R. J. Sternberg, & D. P. Berliner (Eds.), *Evaluation: A comprehensive guide to theory and practice* (pp. 357-385). Kluwer Academic Publishers.

Bailey, S. (2014). *Academic writing: A handbook for international students*. Routledge.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., … & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus*. Test of English as a Foreign Language.

Black, P. J., & Wiliam, D. (1998). Inside the black box: Raising standards through assessment.

Phi Delta Kappan, 80(2), 139-148.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21, 5-31.

Brookhart, S. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson.

Cambridge University Press. (1989). *Testing and assessment in language education*. Cambridge University Press.

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, 18(1), 80-98.

Cumming, A. (2001). Learning to write in a second language: Two decades of research. *International Journal of English Studies*, 1(2), 1-23.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework*. Educational Testing Service.

Dong, Y. (2023). Revolutionizing academic English writing through AI-powered pedagogy: Practical exploration of teaching process and assessment. *Journal of Higher Education Research*, *4*(2), 52. https://doi.org/10.32629/jher.v4i2.1188

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Routledge.

Gardner, R. C. (2000). Correlation, causation, motivation, and second language acquisition. *Canadian Psychology/Psychologie Canadienne*, 41(1), 10.

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, 57(4), 2333-2351. https://doi.org/10.1016/j.compedu.2011.06.004

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In *Exploring the dynamics of second language writing* (pp. 162–189).

Harry, A., & Sayudin, S. (2023). Role of AI in education. *Interdisciplinary Journal and Humanity (INJURITY)*, 2(3), 260-268. e-ISSN: 2963-4113 and p-ISSN: 2963-3397

Hyland, K. (2019). *Second language writing* (2nd ed.). Cambridge University Press.

Kartika, S. (2024). Enhancing writing proficiency through AI-powered feedback: A quasi-experimental study using Google Gemini. *LinguaEducare: Journal of English and Linguistic Studies*, 1(2), 83–96. https://doi.org/10.63324/h6q1ak58

Lang, G., Triantoro, T., & Sharp, J. H. (2024). Large language models as AI-powered educational assistants: Comparing GPT-4 and Gemini for writing teaching cases. *Journal of Information Systems Education*, 35(3), 390-407. https://doi.org/10.62273/YCIJ6454

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, *11*(1), 9. https://doi.org/10.1186/s40561-024-00295-9

Meenakumari, J. (2021). Harnessing the power of artificial intelligence for summative and

formative assessments in higher education. *EdTechReview*. https://www.edtechreview.in/trends-insights/trends/power-of-ai-for-assessments-in-higher-ed/

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, and Evaluation*, 7(1), 10. ISSN:1531-7714

Mukminin, A. (2012). Acculturative experiences among Indonesian graduate students in U.S. higher education: Academic shock, adjustment, crisis, and resolution. *Excellence in Higher Education (EHE)*, 3(1), 14-36.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199-218. https://doi.org/10.1080/03075070600572090

Nunan, D. (1991). *Language teaching methodology: A textbook for teachers*. Prentice Hall.

Paltridge, B., & Starfield, S. (2016). *Getting published in academic journals: Navigating the publication process*. University of Michigan Press.

Rosenfeld, M., Courtney, R., & Fowles, M. (2004). Identifying the writing tasks important for academic success at the undergraduate and graduate levels. *ETS Research Report Series*, 2004(2), i-91.

Sasaki, M. (2000). Toward an empirical model of EFL writing processes: An exploratory study. *Journal of Second Language Writing*, 9(3), 259-291. https://doi.org/10.1016/S1060-3743(00)00028-X

Steele, J. L. (2023). To GPT or not GPT? Empowering our students to learn with AI. *Computers and Education: Artificial Intelligence*, *5*, 100160. https://doi.org/10.1016/j.caeai.2023.100160

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (Vol. 1). University of Michigan Press.

Talevski Dimitrija, J., Janusheva, V., & Pejchinovska, M. (2014). Formative assessment and its effects on the teaching practice. https://eprints.uklo.edu.mk/id/eprint/1172/1/Formative%20Assessment%20And%20Its%20Effects%20In%20The%20Teaching%20Practice.pdf

The University of Texas at Austin. (n.d.). *Build-rubric*. https://ctl.utexas.edu/sites/default/files/build-rubric.pdf

Truong, T. A. A., Le, H. K. N., & Nguyen, V. H. Q. (2025). English-Major Master's Students Regarding the Use of ChatGPT in Learning Research Writing at IUH. *International Journal of AI in Language Education*, *2*(1), 92-115. https://doi.org/10.54855/ijaile.25215

Vy, N., & Pham, V. P. H. (2024). AI chatbots for language practices. *International Journal of AI in Language Education*, *1*(1), 10–54855. https://doi.org/10.54855/ijaile.24115

Weissberg, B. (2000). Developmental relationships in the acquisition of English syntax: Writing vs. speech. *Learning and Instruction*, 10(1), 37-53.

Weissberg, R. (2006). 13 Scaffolded feedback: Tutorial conversations with advanced L2 writers. In *Feedback in second language writing: Contexts and issues* (p. 246).

Wiseman, C. S. (2012). A comparison of the performance of analytic vs. holistic scoring

rubrics to assess L2 writing. *International Journal of Language Testing*, 2(1), 59-92.

Xiao, C., Ma, W., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2024). From automation to augmentation: Large language models elevating the essay scoring landscape. *arXiv e-prints*. https://doi.org/10.1145/3706468.3706507

Biodata

Nguyen Dinh Luat *is a lecturer at the Industrial University of Ho Chi Minh City, Vietnam. He has been teaching English macro skills, pronunciation, linguistics, and translation to a diverse range of learners. His research interests include technology applications in language teaching, language skill development, linguistics, and language testing.*

Le Pham Thien Thu *is a lecturer at the Industrial University of Ho Chi Minh City, Vietnam. She has more than 20 years of teaching experience for a diversity of levels and areas, focusing on teaching methodology, testing and assessment, and the four macro skills. Her research interests consist of technology applications in language teaching, language skills development, and teaching methodology.*

Le Thi Thuy *is a lecturer at the Industrial University of Ho Chi Minh City, Vietnam. She has developed an interest in teaching English skills, reading, and writing to students at the tertiary level. She is passionate about researching technology integration in language instruction and learner autonomy.*