Irish Journal of Technology Enhanced Learning
Vol 7, Issue 2
Special Issue: The Games People Play: Exploring Technology Enhanced Learning Scholarship & Generative Artificial Intelligence

The Potential Use of Generative AI in ESL Writing Assessment: A Case Study of IELTS Writing Tasks

Tianhe Sun [*]
Beijing New Oriental School, Beijing, China

Abstract

This paper investigates the potential use of generative AI in ESL Writing Assessment, particularly in IELTS Writing Tasks. The first part introduces the author's relationship with generative AI in the educational setting, and why the author chooses this topic. This section also contains information about generative AI like ChatGPT, and the prompts employed to generate the article. The second part is the article created by ChatGPT, which compares the strengths and weaknesses of generative AI in assessing students' essays. The third part is a critical reflection of how generative AI performs in creating such an academic paper. It can indeed help researchers to write an informative article more efficiently. But it is noticeable that some specific and important information could be missing and the references are likely to be false ones. Educators should be aware of the strengths and weaknesses of generative AI tools to make the most of it.

1. Introduction

Education over the past decade has witnessed the emergence of innovative tools and technologies, like generative Artificial Intelligence (generative AI) (Bozkurt et al. 2023), employed in English language instruction (Sindermann et al. 2021, Sharadgah et al. 2022, Fitria 2023). How these tools perform in English language teaching and learning, and whether they are effective or not in second language acquisition, have drawn my great interest in my profession. As someone who has spent years helping students prepare for English proficiency exams, I have observed the struggles and challenges they face, particularly in the writing component. These challenges include language errors, lack of understanding of appropriate writing styles, and issues related to coherence and cohesion in their essays, while IELTS requires test-takers to exhibit a high degree of language proficiency, accuracy, and the ability to construct coherent and cohesive essays within strict time constraints. I have always sought ways to enhance my teaching methods and help students overcome common language errors that hinder their performance in IELTS writing tasks. For example, I usually need to review each student's essay manually and provide feedback within 72 hours after class. This can be heavy workload when there is a large number of students. Now with generative AI, I see the great potential for it to help identify common error types to assist students in improving the accuracy of their language in a more efficient way. It can also be used to improve the flexibility and complexity of their language while mastering the appropriate writing styles demanded by language proficiency tests.

[*] Corresponding author: suntianhe@hotmail.com

Therefore, I reckon the potential for AI to provide tailored, data-driven feedback and support for students is an exciting prospect.

Many existing studies have only focused on conventional educational technologies, such as video and audio, computers, tablets, and visual classroom, but among all these various advancements, I believe the incorporation of generative AI into language teaching and learning contexts can open up a world of possibilities. This article aims to explore the potential use of Artificial Intelligence (AI) in ESL writing assessment, with a specific focus on IELTS writing tasks.

Employing a generative AI tool to produce this article can help to understand the strengths and weaknesses of this tool in assessing students' essays and its potential in assisting leaners in developing their writing skills. Generative AI is a subfield of AI that refers to systems capable of generating new and original content, such as text and images. A good example is the GPT (Generative Pre-trained Transformer) models developed by OpenAI, which are particularly renowned for their ability to generate human-like text. ChatGPT is a chatbot that is trained on the GPT language model. It was first launched on November 30, 2022. ChatGPT enables conversational interactions with humans, offering various options for desired length, style, level of detail, and language. OpenAI has released multiple versions of GPT models. GPT-4, the most recent and advanced version, was released on March 14, 2023. It is available for premium ChatGPT users under the commercial name "ChatGPT Plus," costing $20 per month. The position paper in the next section was generated by ChatGPT, based on GPT-3.5. This version allows any user to sign up and have free access to it, thereby affordable for all students.

The prompts used for producing the paper are as follows:

(1) You are an English teacher who are responsible for assessing students' IELTS writing tasks. You plan to write an academic article in the field of AI essay assessment tools and IELTS writing tasks. Please first provide an overview of IELTS writing tasks, AEA tools, and a specific description of criteria for IELTS writing tasks. References should be listed at the end of the article.
(2) Please provide an analysis, which includes: (1) comparations of the precision, effectiveness and efficiency of manual review and AI review in assessing students' essays, and (2) critical evaluation the strengths and weaknesses of manual review and AI review in assessing students' essays. References should be listed at the end of the article.
(3) Building upon the analysis, please offer some recommendations for teachers when incorporating generative AI in the assessment of students' essays. References should be listed at the end of the article. Please write an article of 2000-3000 words following this format.

2. AI Generated Position Paper

2.1 Introduction

The International English Language Testing System (IELTS) is one of the most widely recognized and respected English language proficiency tests worldwide. It assesses the language skills of non-native English speakers, including listening, reading, writing, and speaking. Among these skills, IELTS writing tasks play a crucial role in evaluating an individual's ability to communicate effectively in written English. In recent years, the integration of Artificial Intelligence (AI) essay assessment tools has gained traction as a means of automating and enhancing the evaluation of IELTS writing tasks. This article provides an overview of IELTS writing tasks, explores the role of AI essay assessment tools, and discusses specific criteria used for evaluating IELTS writing tasks.

2.1.1 Overview of IELTS Writing Tasks

IELTS writing tasks consist of two sections: Task 1 (Academic and General Training) and Task 2 (Academic and General Training). These tasks assess the candidate's ability to express their thoughts and ideas clearly, coherently, and cohesively in written English. The specific nature of the tasks varies between the Academic and General Training modules:

1. Academic Writing Task 1: In this task, test-takers are presented with a visual representation of information, such as a graph, chart, table, or diagram. They are required to write a descriptive report summarizing and comparing the information presented.
2. Academic Writing Task 2: In this task, candidates are presented with an essay prompt that requires them to present an argument, provide evidence, and reach a conclusion on a given topic.
3. General Training Writing Task 1: Test-takers are presented with a situation, often in the form of a letter, and are required to respond to it using a formal or semi-formal tone.
4. General Training Writing Task 2: Similar to the Academic Writing Task 2, this task requires candidates to write an essay, but the topics are generally related to everyday life and are more personal in nature.

2.1.2 IELTS Writing Task Assessment Criteria

The assessment of IELTS writing tasks is based on four key criteria, which are used by human examiners as well as AI essay assessment tools:

1. Task Response: This criterion assesses the extent to which a candidate addresses the specific task and stays relevant to the topic. It also evaluates whether the response is well-structured with a clear introduction, body, and conclusion.
2. Coherence and Cohesion: This criterion evaluates the organization of ideas within the response. It assesses the use of paragraphing, logical sequencing of information, and the appropriate use of cohesive devices like linking words and phrases.
3. Lexical Resource: Lexical resource assesses a candidate's vocabulary and their ability to use a range of words and phrases effectively. This includes the appropriateness of vocabulary in the given context and the avoidance of repetitive language.
4. Grammatical Range and Accuracy: Candidates are assessed on their ability to use a variety of grammatical structures accurately. This includes sentence structure, verb tenses, and the avoidance of major grammatical errors.

2.1.3 AI Essay Assessment (AEA) Tools

AI essay assessment tools are software applications that leverage Natural Language Processing (NLP) and machine learning algorithms to evaluate written responses in a manner similar to human examiners. These tools are designed to provide standardized and objective assessments of writing tasks. AEA tools are advantageous for IELTS writing task assessment in several ways:

1. Consistency: AI assessment tools provide consistent evaluation, reducing the potential for human bias or subjectivity in grading.
2. Efficiency: AI tools can assess a large number of essays quickly, significantly reducing the time required for evaluation.
3. Instant Feedback: Test-takers can receive immediate feedback on their writing, aiding in their language learning process.
4. Objective Scoring: AI tools follow predefined assessment criteria, ensuring objectivity in scoring.

The integration of AI essay assessment tools in evaluating IELTS writing tasks offers great promise for standardization and efficiency in the assessment process. While these tools can provide valuable insights, it's crucial to remember that they are most effective when working in conjunction with human examiners who provide expertise, context, and the ability to interpret nuanced responses. The four key criteria used in IELTS writing task assessment—Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy—remain fundamental to the evaluation process, whether performed by humans or AI. In the dynamic landscape of language assessment, AI essay assessment tools are an important development that can help maintain the integrity and reliability of IELTS testing.

2.2 Analysis

The assessment of students' essays is a critical component of education, offering insights into their language proficiency and critical thinking skills. Traditionally, essays have been assessed through manual review, but the advent of Artificial Intelligence (AI) has introduced automated essay assessment tools. This analysis compares the precision, effectiveness, and efficiency of manual and AI essay assessment, critically evaluating their strengths and weaknesses.

2.2.1 Comparative Analysis

1. Precision:
Manual Review: Human assessors bring subject expertise and a nuanced understanding of context to their evaluations, allowing them to assess the subtleties in students' writing. This leads to precise and contextually informed grading.
AI Review: AI tools can provide consistent and objective evaluations but may struggle with nuance and contextual understanding. They rely on predefined criteria, which can result in less precision when assessing more complex or creative essays.

2. Effectiveness:
Manual Review: Human reviewers can provide valuable qualitative feedback, highlighting a student's strengths and weaknesses, and offering guidance for improvement. They can also adapt their assessment criteria to suit the specific context of an assignment.

AI Review: AI assessment tools can be highly effective for large-scale, standardized evaluations, ensuring a uniform and unbiased approach. They offer immediate feedback, which can be useful for student learning. However, they may struggle with providing qualitative feedback or assessing more creative or unconventional essays effectively.

3. Efficiency:

Manual Review: Human assessment is time-consuming, and the process may vary in terms of the time required for evaluation. This can be a limitation, especially for large classes or standardized tests with many essays to grade.

AI Review: AI essay assessment tools are highly efficient, capable of evaluating a large volume of essays quickly. This speed reduces the time and resources needed for assessment significantly.

2.2.1 Critical Evaluation

2.2.1.1 Manual Review

1. Strengths of Manual Review:
- Contextual Understanding: Human assessors bring contextual and subject-specific knowledge to the evaluation process, enabling them to appreciate nuance and creativity in essays.
- Qualitative Feedback: Manual assessment allows for the provision of qualitative feedback, which can be invaluable for students' learning and improvement.
- Adaptability: Human reviewers can adapt their criteria to suit the specific context and objectives of an assignment.

2. Weaknesses of Manual Review:
- Bias and Subjectivity: Manual review can be influenced by individual bias and subjectivity, leading to inconsistent grading.
- Time-Consuming: Manual assessment is time-intensive, making it impractical for large-scale assessments.

2.2.1.2 AI Review

1. Strengths of AI Review:
- Consistency: AI tools provide consistent, objective, and unbiased evaluations across a large number of essays.
- Efficiency: AI assessment tools significantly reduce the time and resources required for assessment, making them well-suited for large-scale evaluations.
- Immediate Feedback: AI tools offer immediate feedback, allowing students to learn from their mistakes quickly.

2. Weaknesses of AI Review:
- Lack of Contextual Understanding: AI tools may struggle with nuanced or context-dependent assessments, particularly in creative or unconventional essays.
- Inability to Provide Qualitative Feedback: AI tools are limited in their ability to provide in-depth qualitative feedback, which can hinder students' development.
- Rely on Predefined Criteria: AI tools rely on predefined assessment criteria, which may not account for variations in assignments or nuanced responses.

In conclusion, the choice between manual and AI essay assessment depends on the specific goals and context of evaluation. Manual review excels in precision, contextual understanding, and qualitative feedback, but it is time-consuming and subject to bias. AI review offers efficiency, consistency, and immediate feedback, but it may struggle with context-dependent assessments and lacks the ability to provide qualitative feedback. Ultimately, a hybrid approach that combines the strengths of both manual and AI assessments may provide a well-rounded and effective solution for evaluating students' essays.

2.3 Discussion

In the ever-evolving landscape of education, Artificial Intelligence (AI) has made significant inroads, and one of its noteworthy applications is in the assessment of students' essays. While AI can streamline the grading process, it is essential for teachers to navigate this integration thoughtfully to ensure it complements their educational objectives. This article offers recommendations for teachers when incorporating generative AI in the assessment of students' essays.

2.3.1 The Role of Generative AI in Essay Assessment

Generative AI, a subfield of AI, is designed to produce human-like text and, in the context of essay assessment, can be employed to both generate prompts and evaluate written responses. It leverages Natural Language Processing (NLP) and machine learning algorithms to understand and respond to text in a human-like manner. When used in the assessment of essays, generative AI can bring several benefits, including increased efficiency, consistency, and objectivity.

2.3.2 Recommendations for Teachers

1. Understand the Capabilities and Limitations of Generative AI:
To effectively incorporate generative AI into the assessment process, teachers must have a comprehensive understanding of the capabilities and limitations of the technology. While AI can provide quick and objective evaluations, it may struggle with nuanced or creative responses and lacks the ability to provide qualitative feedback.

2. Set Clear Objectives:
Before utilizing generative AI for essay assessment, define clear objectives. Consider what you aim to achieve by using AI, such as reducing grading time, ensuring consistent evaluations, or providing immediate feedback to students.

3. Select the Right AI Tools:
Choose AI assessment tools that align with your objectives and the specific needs of your students. Ensure that the tools have a proven track record in education and offer features that suit your assessment requirements.

4. Hybrid Approach:
Rather than replacing manual assessment entirely, consider a hybrid approach. Use AI for initial assessments to expedite the grading process and identify common issues, while retaining manual review for more complex essays requiring nuanced evaluation.

5. Train Students on AI Assessment:
Familiarize your students with the AI assessment process. Explain how AI tools work and the criteria they use for evaluation. This transparency can demystify the process and help students understand how to improve their writing.

6. Customize Assessment Criteria:
If the AI tool allows for customization, tailor the assessment criteria to match the specific objectives of your assignments. This customization can help ensure that the AI assessments align with your educational goals.

7. Provide Clear Prompts:
When using generative AI to generate prompts for assignments, ensure that the prompts are clear and well-structured. Ambiguous or poorly constructed prompts can lead to misinterpretation and inaccurate assessments.

8. Combine AI and Human Expertise:
Encourage collaboration between AI and human assessors. AI can handle routine, repetitive tasks, while human assessors can offer their expertise in assessing creativity, complex arguments, and context-specific nuances.
9. Evaluate the Feedback Loop:
Assess the effectiveness of the feedback loop provided by AI. Ensure that the feedback generated by the tool is meaningful and actionable for students. It should help them understand their strengths and weaknesses.

10. Stay Informed and Adapt:
The field of AI is constantly evolving. Stay informed about the latest developments in AI assessment tools and be willing to adapt your approach as new technologies emerge.

Generative AI has the potential to revolutionize the assessment of students' essays by providing efficiency, consistency, and objectivity. However, to ensure its effective incorporation into the educational process, teachers must carefully consider their objectives and the specific needs of their students. By understanding the capabilities and limitations of generative AI, selecting the right tools, and promoting transparency in the assessment process, educators can harness the power of AI to enhance their teaching and provide valuable insights to their students. In the ever-changing landscape of education, the thoughtful integration of generative AI can help teachers meet the demands of the digital age while maintaining the quality and rigor of essay assessment.

References:

1. Bridgeman, B., Trapani, C. S., & Attali, Y. (2012). Comparison of automated scoring of writing quality to human scoring. Educational Measurement: Issues and Practice, 31(3), 23-31.
2. British Council. (n.d.). IELTS. Retrieved from https://www.britishcouncil.org/ielts
3. Cohen, J. (2018). AI and the Future of Testing. Educational Measurement: Issues and Practice, 37(3), 27-38.
4. Elliott, R., Briller, V., & Leslie, S. J. (2018). The Next Generation of Automated Essay Scoring. Educational Measurement: Issues and Practice, 37(3), 6-20.
5. IELTS. (n.d.). The IELTS Scoring Process. Retrieved from https://www.ielts.org/about-the-test/how-ielts-is-scored/the-ielts-scoring-process
6. Shermis, M. D., & Burstein, J. (2013). Handbook of Automated Essay Evaluation. Routledge.
7. Vantage Learning. (n.d.). AI Scoring Engine. Retrieved from https://www.vantagelearning.com/our-products/ai-scoring-engine/
8. Perelman, L. (2012). The Implications of the Babel Generator for Computer Scoring of Essays. Assessing Writing, 17(3), 234-243.

3. Critical Reflection

The article consists of three sections, so it took ChatGPT three times with three different prompts to generate the whole article. The time it took to generate these three parts with ChatGPT was as follows: 18 seconds, 15 seconds, and 17 seconds, respectively. Each time, the generative AI could provide 4 to 5 references. However, within all these references, only one of them is correct with authors and their publication matched (Shermis & Burstein 2013). Another useful reference has the right names of authors, but the title of their article is incorrect (Bridgeman & Attali 2012). The correct title is 'Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country'. Other authors and publication are all false ones. It is also noticeable that websites where generative AI can retrieve information may have denied access (https://www.britishcouncil.org/ielts) or cannot be found (https://www.ielts.org/about-the-test/how-ielts-is-scored/the-ielts-scoring-process & https://www.vantagelearning.com/our-products/ai-scoring-engine/). This problem has widely occurred to many scholars using ChatGPT-3.5.

The introduction generated by ChatGPT gives enough information on IELTS and its writing tasks. Readers can easily get an idea of what is IELTS and what writing tasks test-takers need to finish. However, when it comes to the criteria of assess writing tasks, only basic information has been generated. There are four major shortages: (1) the criteria do not apply to both tasks of IELTS writing section. Since task 1 and task 2 (both in Academic, and General Training) have different descriptions of criteria, especially in the first one (it is called 'task achievement' in task 1 but 'task response' in task 2), it is necessary for the candidates to notice the difference to be better prepared. (2) each of the four key criteria is only briefly introduced without specific instruction. In this case, detailed descriptions matched with different bands will be useful for test-takers to know their current level and what requirements they need to meet to achieve a certain band. (3) Some newly published up-to-date descriptions of assessment are missing in the generated article. For example, the new criterion concerning Lexical Resource assesses whether the candidate's response contains topic-specific items. This key information cannot be found in this part of article, which only follows the old criterion given in the Official Guide of IELTS published 9 years ago (Cullen & Jakeman 2014). (4) In spite of four key criteria, other criteria including underlength, memorised, off-topic, are not mentioned at all.

As an educator and researcher, I would include these missing parts to provide all the necessary information about the criteria of IELTS writing tasks. It is only when someone has a full understanding of these criteria that he or she will be able to know what to focus when practicing writing skills. For example, some candidates will focus on using synonyms of some common words to avoid repetition in their essay. When writing a topic on whether university should accept equal numbers of male and female students in every subject, they would wrongly use 'receive' to substitute 'accept'. If they understand that IELTS writing tasks encourage candidates to include topic-specific items in the essay, they could use words like 'admit' or 'enroll'.

The generated article also concerns the advantages of AI Essay Assessment (AEA) Tools in both introduction and analysis. This makes it quite repetitive and makes the article difficult to follow. It is only necessary to analyse the advantages and disadvantages in the comparative analysis, so information regarding this in the introduction can be deleted for good.

The discussion in this article has great implications for educators in English language instruction area. It gives useful advice for teachers to make the most of generative AI in their teaching and assessment. By following these recommendations, a teacher should be able to greatly improve his or her teaching efficiency and gain a deeper insight into the future development of education in this digital age. This could be particularly helpful to those who have just started their teaching path. In old age, such advice and insights were either given by experienced teachers in teacher training sessions or gained independently through years of teaching practice. Now with the use of generative AI, it is easier for these teachers to get a good start and more likely to keep on the right track. For experienced teachers, these recommendations could also be enlightening. For example, the customization of assessment criteria would be very useful when giving one-on-one lessons, as some of these learners have their own specific goals, so the assessment criteria should be customized accordingly.

Overall, it was not difficult to write a long scholarly paper with the help of generative AI like ChatGPT, which can be informative for someone to have an overview of a certain topic. But we should also be aware of the fact that the generated article may lack specific information and the sources could be questionable in some cases.

References

Bozkurt, A., Xiao, J., Lambert, S., Pazurek, A., Crompton, H., Koseoglu, S., Farrow, R., Bond, M., Nerantzi, C., Honeychurch, S., Bali, M., Dron, J., Mir, K., Stewart, B., Costello, E., Mason, J., Stracke, C., Romero- Hall, E., Koutropoulos, A., . . . Jandrić, P. (2023). Speculative futures on ChatGPT and generative artificial intelligence (AI): A collective reflection from the educational landscape. *Asian Journal of Distance Education*, 18(1), 53-130. https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/709

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, 25(1), 27-40. https://doi.org/10.1080/08957347.2012.635502

Cullen, P., French, A., & Jakeman, V. (2014). *The official Cambridge guide to IELTS student's book with answers with DVD-ROM*. Cambridge University Press.

Fitria, T. N. (2023, March). Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. *ELT Forum: Journal of English Language Teaching*, *12*(1), 44-58. https://doi.org/10.15294/elt.v12i1.64069

Sharadgah, T. A., & Sa'di, R. A. (2022). A systematic review of research on the use of artificial intelligence in English language teaching and learning (2015-2021): What are the current effects? *Journal of Information Technology Education, 21,* 337–377. https://doi.org/10.28945/4999

Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., Sariyska, R., Stavrou, M., Becker, B., & Montag, C. (2021). Assessing the attitude towards artificial intelligence: Introduction of a short measure in German, Chinese, and English language. In *KI. Künstliche Intelligenz (Oldenbourg)* (Vol. 35, Issue 1, pp. 109–118). Springer Berlin Heidelberg. https://doi.org/10.1007/s13218-020-00689-0