Proceedings of ICELIA 2027 Symposium: Modern Education Strategies: Innovative Pedagogies and Digital Learning
DOI: 10.54254/2753-7048/2027.NL37942

Generative AI Tools for Performance Assessment of University Students: Effects, Challenges, and Countermeasures

Linyun Tang

College of Education, Zhejiang University of Technology, Hangzhou, China
2350795985@qq.com

Abstract. Against the backdrop of educational evaluation reform, Generative Artificial Intelligence (GAI) has been increasingly applied to performance assessment, yet its actual effects and challenges remain underexplored. This study surveyed 50 undergraduate students to examine their perceptions of GAI-based performance assessment across five dimensions: evaluation efficiency and feedback timeliness, accuracy and fairness, appropriateness of scoring criteria and weight distribution, stability and trust, and deep learning and innovation. Results show that over 70% of respondents recognized GAI's advantages in scoring speed and feedback timeliness, and most regarded its scoring criteria as clear and consistent. However, nearly 75% believed GAI failed to identify innovative highlights, over 70% perceived a tendency to emphasize form over content, and 44% reported that such preferences constrained their creative expression. Respondents also raised concerns about incorrect information, technical instability, and the lack of emotional interaction compared with human evaluation. These findings suggest that GAI should act as an auxiliary tool rather than a full replacement for teacher evaluation. To address current limitations, it is necessary to improve GAI's deep evaluation and fact-checking capabilities, establish human-machine collaborative review mechanisms, and cultivate students' critical thinking. This study provides implications for integrating GAI into university performance assessment.

Keywords: Generative Artificial Intelligence, performance assessment, university students, human-machine collaboration, educational evaluation reform

1. Introduction

With ongoing educational reform and the implementation of key evaluation policies, the importance of educational evaluation has been further emphasized. In this context, performance assessment and process-oriented assessment have become important means of cultivating students' core competencies. Traditional evaluation methods tend to emphasize memorization and comprehension, limiting assessment of problem-solving, innovation, communication, and collaboration abilities. Although performance assessment evaluates students' comprehensive abilities in authentic contexts, its implementation in higher education still faces challenges such as complex task design, heavy

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2027.NL37942

teacher workload, and insufficient consistency in evaluation. With the rapid development of Generative Artificial Intelligence (GAI), GAI has been increasingly used in education, with applications extending to areas such as teaching evaluation, learning feedback, and performance assessment.

In recent years, GAI represented by ChatGPT has developed rapidly, offering new technical support for addressing the aforementioned challenges. Existing studies have explored the application of GAI in education; however, most current research remains theoretical or is limited to preliminary exploration in specific disciplinary courses. In particular, empirical studies on the effects of using GAI for performance assessment among university students and the related student-perceived challenges are still relatively limited.

Therefore, this study employs a questionnaire survey to examine university students' perceptions of the effects and practical challenges of GAI-based performance assessment, and to propose optimization suggestions from the following dimensions: evaluation efficiency and feedback timeliness; accuracy and fairness; appropriateness of scoring criteria and weight distribution; stability and trust; and perceptions of deep learning and innovation.

2. Literature review

2.1. Definition of GAI

Generative Artificial Intelligence (GAI) refers to a type of artificial intelligence capable of generating original content by learning from large-scale datasets [1]. Unlike traditional rule-based or statistics-based AI approaches, GAI has the ability to create new artifacts. By learning from vast amounts of data, GAI understands inputs and generates new content based on these data, for example, GPT can generate text, images, music, etc.

2.2. Performance assessment

Performance assessment refers to an evaluation method that creates real or simulated contexts, requiring students to apply their knowledge and skills to complete specific tasks, thereby examining their mastery of knowledge and the development of complex abilities such as problem-solving, communication, collaboration, and critical thinking. Compared with traditional examinations, performance assessment places greater emphasis on process-oriented evaluation and can reflect, to a certain extent, students' ability to comprehensively apply knowledge to solve practical problems [2]. Compared with traditional evaluation methods, performance assessment focuses more on students' overall performance in specific tasks and represents a development in educational evaluation from single-outcome-based assessment toward diversified evaluation [3].

2.3. Current research on GAI application in performance assessment

The application of GAI in education continues to expand, and related studies have begun to focus on its value in performance assessment. For example, Sun Wei et al. explored the use of GAI in performance assessment design in geography teaching and found that it could, to some extent, reduce teachers' repetitive workload [4]. Wang Jing et al., from the perspective of human-machine collaboration, explored the application of GAI in evaluating students' scientific inquiry abilities [5]. Pan Ting et al. explored the application of GAI in computational thinking instruction, providing relevant practical examples for GAI's involvement in educational activities [6]. Overall, existing

DOI: 10.54254/2753-7048/2027.NL37942

research has begun to focus on human-machine collaboration models, but relatively limited attention has been paid to university students' actual perceptions after using GAI-based performance assessment.

3. Research methods

3.1. Participants and design

This study adopted convenience sampling to recruit current undergraduate students via an online survey platform. A total of 55 questionnaires were collected, and after excluding invalid responses, 50 valid responses were retained. The questionnaire was adapted from relevant literature and existing scales to fit GAI-based scoring contexts. It employed a 5-point Likert scale and covered five dimensions: evaluation efficiency and feedback timeliness, accuracy and fairness, appropriateness of scoring criteria and weight distribution, stability and trust, and perceptions of deep learning and innovation.

3.2. Questionnaire description

The evaluation efficiency and feedback timeliness dimension assessed the speed of GAI scoring feedback and the operational process (4 items), focusing on perceived efficiency and convenience. The accuracy and fairness dimension measured participants' recognition of the accuracy and fairness of GAI scoring. The scoring criteria and weight distribution dimension investigated evaluations of the scoring items and weight settings. An attention-check item was included to identify invalid or mechanical responses. The stability and trust dimension primarily measured students' perceptions of the consistency of GAI scoring, with three items in total; Item 13 was reverse-scored to identify negative perceptions of fluctuations in GAI scoring. Finally, the deep learning and innovation dimension examined whether GAI-based performance assessment could identify students' innovative highlights, whether there was a tendency to emphasize form over content, and students' perceptions of negative learning effects regarding innovation content. In addition to the above scale items, the questionnaire included one open-ended question to understand other potential advantages or problems participants perceived with GAI in performance assessment.

3.3. Data processing

After data collection, data cleaning was conducted based on attention-check items and response patterns to exclude invalid questionnaires. Based on the valid sample, descriptive statistical analyses were conducted using Wenjuanxing and Python to describe sample characteristics and the distribution of responses across the five dimensions. Since the primary purpose of this study was to provide a preliminary depiction of participants' perceptions, no complex inferential statistical analyses such as t-tests were conducted. However, descriptive comparisons across different grade levels and gender groups were used to summarize perceptual differences among groups.

4. Questionnaire results

4.1. Basic sample characteristics

A total of 50 valid questionnaires were obtained. In terms of grade distribution, third-year students accounted for over 70%, second-year students accounted for 10%, fourth-year students accounted

DOI: 10.54254/2753-7048/2027.NL37942

for 16%, no first-year students participated, and 2% preferred not to disclose their grade information. In terms of gender distribution, all second-year respondents were male; among third- and fourth-year respondents, males accounted for over 70% and females approximately one quarter.

Over 70% of respondents received GAI scoring feedback once a month or less, less than 10% received it twice or more per week, and nearly one quarter received it once a week, indicating relatively low overall usage frequency. Further comparison of usage frequency across grades revealed some differences. Among third-year respondents, nearly 80% reported a frequency of once a month or less; among fourth-year respondents, over 60% reported once a week; and all second-year respondents reported once a month or less. Overall, respondents from different grades showed some variation in GAI scoring feedback usage frequency, with upper-grade students tending to use it more frequently.

4.2. Results by dimension

This survey examined respondents' evaluations of GAI scoring across five dimensions: evaluation efficiency and feedback timeliness, accuracy and fairness, the appropriateness of scoring criteria and weight distribution, stability and trust, and perceptions of deep learning and innovation.

4.2.1. Evaluation efficiency and feedback timeliness

In the evaluation efficiency and feedback timeliness dimension, 74% of respondents selected "agree" or "strongly agree" regarding scoring speed, 20% remained neutral, and 6% selected "disagree" or "strongly disagree." For the statement "GAI can reduce delayed score release," 70% agreed, 18% remained neutral, and only 12% disagreed or strongly disagreed.

Regarding time savings and feedback content, 64% of respondents agreed that GAI scoring saves time and provides relatively detailed feedback. 30% remained neutral, and only 6% selected "disagree" or "strongly disagree." For the statement "the operation process of GAI scoring is more cumbersome," responses were relatively dispersed: 36% disagreed, 36% remained neutral, and 28% agreed, indicating no overall consensus.

4.2.2. Accuracy and fairness

In terms of accuracy, 56% of respondents agreed that GAI scoring could adequately reflect task completion, 34% remained neutral, and 10% disagreed or strongly disagreed. Regarding whether GAI scoring met their own evaluation expectations, 56% selected "agree" or "strongly agree," 30% remained neutral, and 14% selected "disagree" or "strongly disagree." In terms of fairness, 66% agreed that GAI scoring was fair, 22% remained neutral, and 12% disagreed or strongly disagreed. Overall, respondents' evaluations of GAI scoring accuracy and fairness were relatively positive, though some remained reserved.

4.2.3. Appropriateness of scoring criteria and weight distribution

Regarding GAI scoring criteria and weight distribution, respondents reported relatively positive perceptions of the clarity and consistency of the scoring criteria and the appropriateness of the weight distribution. Over 70% acknowledged the evaluation items and weight distribution set by GAI, 20% remained neutral, and only 10% expressed disagreement. Overall, attitudes toward the clarity of scoring criteria and weight distribution were relatively positive.

DOI: 10.54254/2753-7048/2027.NL37942

4.2.4. Stability and trust

Most respondents believed that GAI could maintain relatively consistent scoring for similar assignments, with about 20% noting some volatility and 14% expressing doubt about the scoring results.

4.2.5. Perceptions of deep learning and innovation

Nearly 75% of respondents believed that GAI struggled to identify innovative elements in student work, while only 4% disagreed or strongly disagreed. Meanwhile, over 70% believed that GAI scoring tends to emphasize form over content depth, and 44% indicated that this preference has already influenced their creative and thinking processes.

4.3. Differences by grade and gender

Respondents from different grades varied somewhat in their overall evaluations of GAI's participation in performance assessment. Second-, third-, and fourth-year students showed relatively similar levels of acceptance of GAI's positive effects, with third-year students slightly more positive. Negative perceptions, such as GAI scoring preferences and insufficient innovation recognition, tended to be more evident among upper-grade students, showing a mild upward trend. No notable gender differences were found, although male respondents were slightly more accepting than female respondents. These group differences, however, need further verification with a larger sample.

4.4. Open-ended questions

Feedback from open-ended questions indicated that the main advantages mentioned by respondents centered on efficiency improvement and developmental expectations. On the one hand, some respondents believed that GAI could rapidly generate evaluation feedback, help reduce teachers' repetitive evaluation workload, and accommodate large-scale assessment needs. On the other hand, some considered GAI's involvement in educational evaluation as aligned with future development trends.

The main concerns raised by respondents focused on three areas: technical limitations, evaluation effectiveness, and user experience. At the technical level, respondents mentioned issues such as "possible generation of incorrect information," inadequate handling of details, limited adaptability to paper-based assignments, and system lag or unavailability. In terms of evaluation effectiveness, some thought that GAI's scoring criteria and results tended to be template-driven, potentially leading students to conform to set templates to obtain higher scores, thereby restricting personalized expression. Additionally, compared with teacher evaluation, some respondents noted that GAI lacks human appreciation, commentary, and direct interaction. A few respondents worried that GAI involvement might increase learning or work pressure.

5. Discussion

5.1. The role of GAI in performance assessment

Based on the questionnaire results, the advantages of GAI-based performance assessment are mainly concentrated in three areas: evaluation efficiency, feedback timeliness, and standardization of scoring criteria. First, respondents valued efficiency and feedback speed most. Over 70% agreed that

DOI: 10.54254/2753-7048/2027.NL37942

GAI scoring was faster than manual grading and reduced delays in score release. When there are many tasks and clear criteria, GAI can handle repetitive work and speed up feedback.

Second, respondents also appreciated the clear scoring criteria, reasonable weight distribution, and consistency. Most found the scoring items clear, the weight distribution appropriate, and the scoring consistent for similar tasks. With explicit criteria, GAI can generate standardized scores across preset dimensions, helping maintain consistency.

Third, respondents generally accepted GAI's role and even saw it as an important direction for educational evaluation, as indicated by open-ended feedback. Therefore, GAI is best used as an auxiliary tool rather than a full replacement for teachers. Its strengths lie in efficiency, timely feedback, and standardized criteria, while creativity, complexity, and personalized performance still require teachers' judgment.

5.2. Practical challenges of GAI in performance assessment

Based on the questionnaire results, the challenges of applying GAI to performance assessment are mainly concentrated at three levels: limitations in evaluation capacity, insufficient technical maturity, and lack of humanistic value.

GAI still struggles with in-depth evaluation and recognizing innovation. Over 70% of respondents felt that GAI-based scoring emphasized form over content, relying too much on formatting and keywords while ignoring content depth. Moreover, nearly 80% believed that GAI failed to identify their innovative elements, suggesting that GAI is not yet capable of recognizing complex features such as personal expression, innovative viewpoints, and content depth. This tendency has already had negative effects: nearly half of respondents admitted adjusting their expression styles to match GAI's preferences, and upper-grade students perceived this issue more strongly.

The technical reliability of GAI still needs improvement. Open-ended feedback mentioned issues such as generating incorrect information, poor compatibility with paper-based assignments, and system lag or crashes. Some respondents also noted that replacing human evaluation with GAI would reduce opportunities for direct communication, feedback, and emotional interaction with teachers. Compared with human evaluation, GAI still has limitations in personalized communication, emotional responsiveness, and targeted guidance.

5.3. Practical implications

To address these shortcomings in deep evaluation, innovation recognition, and technical reliability, GAI's evaluation capabilities and technical functions should be further enhanced. Its ability to recognize complex logic, content depth, and innovative viewpoints should be improved to reduce over-reliance on surface features such as formatting and keywords; fact-checking capabilities should also be strengthened to reduce the likelihood of generating incorrect information. Additionally, GAI's multimodal adaptability should be improved to accommodate various performance assessment scenarios, such as paper-based assignments, images, and videos.

Given these limitations, GAI should not simply replace teacher evaluation. A human-machine collaborative mechanism is more appropriate. GAI can handle preliminary scoring and generate initial feedback, while teachers review innovative, complex, or contentious content and retain personal comments and interaction with students. This balance preserves both GAI's efficiency and the human value of teacher evaluation. In addition to optimizing technology and evaluation mechanisms, students' capacity to use GAI should also be developed. Critical thinking is particularly

DOI: 10.54254/2753-7048/2027.NL37942

important; students should be guided to make rational judgments about GAI-generated scores and feedback, identify possible information biases and logical loopholes, and avoid excessively catering to GAI's evaluation preferences, thereby protecting their personalized expression and creative space.

6. Conclusion

This study examined the challenges and countermeasures of using GAI tools for performance assessment of university students. Through a questionnaire survey, it explored whether GAI could save time, produce accurate and reasonable scoring, and demonstrate stability in this context. The results show that GAI offers clear efficiency advantages, and most respondents viewed its scoring items, weight distribution, and trustworthiness positively. However, challenges are equally evident: GAI tends to emphasize form over content, making it difficult to identify innovative highlights. Technical immaturity and the lack of humanistic interaction further limit evaluation quality. Therefore, optimization should aim to leverage strengths, compensate for weaknesses, and promote human-machine collaboration. Specifically, deep evaluation and fact-checking capabilities need improvement; human review should remain part of the process; and students' critical thinking needs to be developed to support a healthier human-machine collaborative evaluation system.

This study has certain limitations. First, the sample was drawn mainly from a single university and was relatively small, so the validity of the conclusions needs further verification. Second, the questionnaire data were based on students' subjective perceptions and lacked objective measures. Future research should expand the sample scope and incorporate experimental designs to further explore the impact of GAI-based performance assessment on students.

References

[1] World Economic Forum. (2023). Top 10 emerging technologies of 2023.
[2] Zhao, D. C., & Lu, M. Z. (2004). *New curriculum and student evaluation*. Higher Education Press.
[3] Zhang, Z. F., & Xu, S. Q. (2014). *30 new explorations in educational evaluation*. East China Normal University Press.
[4] Sun, W., Wang, Y. T., Zhu, C. X., et al. (2025). Exploration of generative artificial intelligence in performance assessment in geography teaching: Taking "properties of seawater" as an example. *Geography Education*, (4), 53–56.
[5] Wang, J., Shen, S. W., & Chen, X. F. (2025). Research on performance assessment of students' thinking ability from the perspective of human-AI collaboration: Taking scientific inquiry ability as an example. *Journal of Distance Education, 43*(6), 101-110. https://doi.org/10.15881/j.cnki.cn33-1304/g4.2025.06.012
[6] Pan, T., Tang, G., & Zhang, Y. J. (2024). Generative AI empowers computational thinking teaching. In *Proceedings of the 2024 Computational Thinking and STEM Education Seminar and Bebras China Community Annual Working Conference* (pp. 191–199). School of Educational Technology, Beijing Normal University. https://doi.org/10.26914/c.cnkihy.2024.009013