TPM Vol. 32, No. 3, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

ALGORITHMIC FAIRNESS AND PEDAGOGICAL LEGITIMACY IN AI SCORING SYSTEMS: PERSPECTIVES FROM UNIVERSITY ENGLISH WRITING IN CHINA

NANNAN ZHANG
JIANGXI UNIVERSITY OF SOFTWARE PROFESSIONAL TECHNOLOGY, EMAIL: z7726.edu.cn@outlook.com

SHUSHU HUA
ZHONGKAI UNIVERSITY OF AGRICULTURE AND ENGINEERING, EMAIL: applehua165@163.com

TINGTING ZHANG
THE EDUCATION UNIVERSITY OF HONG KONG, EMAIL: s1146946@s.eduhk.hk

YANJUN LIU
ASSUMPTION UNIVERSITY

HAILONG ZHANG
RUSSIAN PRESIDENTIAL ACADEMY OF NATIONAL ECONOMY AND PUBLIC ADMINISTRATION
EMAIL.: sharedemali@outlook.com

Abstract: As artificial intelligence (AI) technologies become increasingly integrated into higher education, AI-based writing scoring systems are gaining traction as tools to evaluate student performance efficiently. While these systems offer potential benefits such as speed and consistency, they also raise significant concerns regarding fairness, transparency, and the evolving role of teachers in the assessment process. This qualitative case study investigates how university students and English writing instructors in China perceive the use of AI scoring systems in academic writing courses. Drawing on semi-structured interviews with 13 participants, the study identifies four major themes: perceived algorithmic bias and rigidity, lack of transparency in score generation, tensions between teacher authority and AI judgment, and institutional gaps in policy and support. Findings reveal that despite some operational advantages, AI scoring systems are often viewed as pedagogically misaligned and ethically ambiguous. The study underscores the need for more robust governance mechanisms, teacher training, and transparency standards to ensure the responsible use of AI in educational assessment. It contributes to ongoing discussions on educational fairness, teacher agency, and the ethical implementation of digital technologies in classroom settings.
Keywords: AI scoring systems; educational fairness; teacher agency; Chinese higher education

1. INTRODUCTION

In recent years, artificial intelligence (AI) technologies have rapidly expanded into the domain of education, transforming the ways in which teaching, learning, and assessment are conducted (Zhai et al., 2021; Luckin & Holmes, 2016). Among these technologies, AI-assisted writing evaluation systems and ETS e-rater—have been increasingly adopted in Chinese universities as tools to assess students' English writing performance (Liang et al., 2024; Lu, 2019). These systems are often praised for their efficiency, consistency, and ability to provide instant feedback, making them particularly attractive for large-scale language courses (Xu, 2022). However, the introduction of algorithmic scoring into writing classrooms also raises a series of pedagogical and ethical questions, particularly around fairness, transparency, and the role of human judgment in assessment (Chen & Pan, 2022). As higher education institutions seek to modernize evaluation practices through digitalization, it becomes urgent to investigate how these tools are experienced, interpreted, and governed within the everyday realities of teaching and learning.

Previous research on AI-based scoring systems has primarily focused on their technical validity and scoring accuracy (Wang & Brown, 2020; Liu et al., 2022), often comparing algorithmic outputs to human ratings using correlational or statistical analyses. While such studies contribute valuable insights into performance metrics, they tend to overlook the classroom-level experiences and perceptions of the key actors involved—teachers and students. A smaller body of literature has begun to explore users' attitudes toward AI scoring (Schepman & Rodway, 2020; Stein et al., 2024), but these studies are often limited in scope, focusing narrowly on student satisfaction or usability, rather than on

pedagogical alignment, trust, or perceived fairness (Cicero et al., 2025). Furthermore, relatively little attention has been paid to the institutional governance mechanisms—or lack thereof—that shape how AI systems are integrated into assessment practices.

Despite the growing presence of AI in writing instruction, there remains a critical gap in the literature regarding the social and managerial dimensions of AI scoring in higher education. Specifically, it lack in-depth qualitative studies that examine how AI-based scoring systems are understood, contested, and managed by teachers and students in real classroom contexts (Kim et al., 2024). This gap is especially salient in the Chinese educational landscape, where digital reforms are rapidly progressing, yet institutional infrastructures for ethical implementation, policy guidance, and teacher training are still underdeveloped. Without addressing these questions, there is a risk that AI tools may inadvertently reinforce inequities, undermine trust, and reduce teachers' pedagogical agency.

The present study adopts a qualitative case study approach to explore how university teachers and students in China perceive the fairness of AI-based scoring systems in English writing courses, and how these perceptions reflect broader issues of classroom management and institutional governance. Drawing on semi-structured interviews with instructors and learners, the study contributes to the field by offering an empirically grounded understanding of fairness controversies, identifying key challenges in teacher–AI interaction, and proposing management strategies for more responsible and pedagogically aligned implementation. The study is guided by the following research questions:

1) How do university students and teachers perceive the fairness and legitimacy of AI-based writing scoring systems?
2) What challenges do these systems pose for classroom teaching, assessment, and trust?
3) How can universities better manage the implementation of AI scoring systems to promote transparency, equity, and pedagogical integrity?

2. METHOD

2.1 Research Design

This study adopts a descriptive qualitative research design to examine the fairness-related controversies and management responses surrounding the implementation of AI-assisted writing scoring systems in a Chinese higher institution. The choice of a qualitative design is based on the assumption that fairness, especially in educational assessment, is a subjective and socially constructed concept that cannot be adequately captured through numerical indicators alone (Edwards, 2020). In particular, as AI writing evaluation tools are increasingly deployed in English writing courses across Chinese universities, it becomes critical to understand how stakeholders—especially students and instructors—perceive the transparency, objectivity, and usefulness of such systems. Rather than testing a theoretical model or assessing performance accuracy of specific algorithms, this study focuses on collecting firsthand experiential data from real classroom settings where AI scoring has been implemented.

2.2 Participants and Data Collection

Participants in this study were recruited from a comprehensive public university in eastern China that has incorporated AI-based writing scoring platforms into its English language curriculum for non-English majors. A total of 98 undergraduate students and 10 college English teachers voluntarily participated in the study, all of whom had prior experience using the AI system either as writing task assessors or as learners receiving feedback. This participant pool was considered sufficient to reflect a diversity of views while allowing for manageable in-depth analysis within a single-institution case study.

Data were collected through semi-structured interviews. Semi-structured interviews were conducted with a subsample of participants who volunteered to share more detailed experiences. These included five English instructors and eight undergraduate students, selected to represent varying levels of familiarity with the AI scoring system. Each interview lasted between 20 to 30 minutes and was conducted either in person or via video conferencing. The interviews focused on three main themes: (1) participants' experiences and impressions of using the AI scoring system; (2) their perceptions of fairness and transparency; and (3) their views on how the system could be improved or better managed at the institutional level. All interviews were audio-recorded with consent and transcribed verbatim for analysis.

2.3 Data Analysis

The collected data were analyzed using a combination of basic descriptive statistics and thematic qualitative analysis, allowing for both breadth and depth in interpreting participants' views. The core of the analysis rested on the interview data, which were examined through thematic analysis following the six-step process proposed by Braun and Clarke (2006): (1) familiarization with data, (2) generation of initial codes, (3) searching for themes, (4) reviewing themes, (5) defining and naming themes, and (6) producing the report. Initial coding was performed manually by reading and re-reading the transcripts to identify recurring concepts and patterns. Examples of emergent codes included "algorithm bias," "feedback mismatch," "lack of human explanation," and "efficiency vs. accuracy." These codes were then grouped into broader themes such as "Perceived Unfairness," "Trust and Distrust in AI," "Teacher Agency

," and "Institutional Governance Gaps." Efforts were made to ensure that both student and teacher perspectives were equally represented in the final thematic structure.

3. RESULTS

3.1 Perceptions of Algorithmic Fairness and Bias

Both teachers and students expressed mixed feelings about the fairness of AI scoring. While participants acknowledged the efficiency and consistency of the system, they frequently questioned its sensitivity to the quality of thought and the depth of argumentation. Students in particular believed that the AI favored surface-level linguistic features such as word length, sentence complexity, or the presence of transitional markers, while overlooking creative or contextually appropriate expressions.

"The AI seems to reward complicated vocabulary and long sentences. It doesn't really understand if my argument makes sense."

(Student Interviewee #5)

Teachers also voiced concern that such automated evaluation systems could inadvertently reinforce formulaic writing styles. They observed that students often tried to "write for the machine," imitating the linguistic patterns that seemed to produce higher scores in previous submissions. This practice, they warned, might ultimately narrow students' expressive range and undermine the goal of fostering critical and original thinking in writing education.

"Students have learned to play the system. They insert fancy connectors and long words because they believe that's what the AI wants. It makes their essays look mechanical."

(Teacher Interviewee #2)

Overall, the interviews suggested a shared perception that the AI system's fairness was limited by its narrow evaluative scope and its inability to appreciate rhetorical nuance or conceptual innovation.

3.2 Lack of Transparency and Explainability

A second and equally prominent theme concerned the opacity of the scoring process. Both teachers and students reported that the AI system did not offer clear explanations for how scores were assigned or which linguistic features contributed most to the results. For many students, the absence of detailed feedback led to frustration and confusion.

"It just gives me a number. I have no idea what that number means or how to improve."

(Student Interviewee #1)

Teachers found themselves in a similarly difficult position. They were expected to integrate AI-generated feedback into their teaching but lacked access to the underlying scoring logic or weighting criteria.

"When students ask me why the AI gave them 76 instead of 82, I honestly cannot explain. The system doesn't tell us how it decides."

(Teacher Interviewee #4)

This lack of transparency undermined trust in the technology and limited its pedagogical usefulness. Participants repeatedly emphasized that without interpretability, AI feedback could not effectively support formative learning or fair evaluation. Instead, it risked becoming a "black box authority" — accepted because it is fast, but not respected because it is not understood.

3.3 Redefining Teacher Authority and Student Trust

The introduction of AI scoring also reshaped the traditional power relations between teachers, students, and institutional systems of evaluation. Teachers described moments of role conflict, where their professional judgment was challenged by algorithmic scores. Students, in turn, reported feeling uncertain about whose feedback to trust when discrepancies occurred between human and AI evaluations.

"I told a student their essay lacked coherence, but the AI gave them a higher score than I would have. The student said, 'Maybe the AI understands better than you.' That was awkward."

(Teacher Interviewee #3)

Such situations blurred the boundaries between technological efficiency and pedagogical authority. Some teachers expressed a sense of diminished agency, as their grading autonomy became partially outsourced to an algorithm. At the same time, a few participants acknowledged that the AI could provide a valuable "second opinion" that encouraged reflection on their own biases or grading consistency. However, the overall tone was cautious: teachers emphasized that human interpretation remains indispensable in contextualizing writing quality, cultural appropriateness, and emotional nuance—dimensions the AI cannot capture.

Students, for their part, revealed ambivalence. While they appreciated the immediacy of AI feedback, they tended to trust teachers more when seeking constructive advice for revision. The coexistence of these two evaluative authorities created confusion about the standards of "good writing" and raised new challenges for classroom management.

3.4 Institutional Gaps and the Need for Governance

results should contribute to final grades. This lack of procedural clarity led to inconsistencies across classes and departments.

"We were told to use the AI system, but no one explained its role in assessment. Some teachers use it for practice only; others use it for grading. There's no policy."

(Teacher Interviewee #1)

Students also reported uncertainty about the weight of AI scores in their course evaluation. Some expressed concern that appeal mechanisms were unavailable, leaving them powerless to contest results they believed were inaccurate. These management gaps reflected a broader issue of technological governance in higher education—where tools are deployed for efficiency without parallel investments in policy design or stakeholder communication.

"If the AI makes a mistake, who is responsible? Can I ask for regrading? There's no rule about that."

(Student Interviewee #7)

Both groups suggested that institutional management should play a more active role in establishing transparent procedures, training programs, and feedback channels to ensure that AI scoring serves educational rather than purely administrative purposes.'

4. **DISCUSSION**

This section interprets the findings in light of existing literature and theoretical concerns about fairness, technological integration, and governance in higher education. The results presented in the previous chapter reveal a complex interplay between stakeholder perceptions of fairness, the algorithmic logic of AI scoring systems, and the institutional context in which these technologies are implemented. Three key areas are discussed: (1) rethinking fairness in algorithmic evaluation, (2) explainability as a prerequisite for trust, and (3) the shifting role of teachers in AI-mediated assessment.

One of the most prominent findings from this study is the widespread concern over the fairness of AI-assisted scoring systems in college English writing courses. Participants frequently described the system as favoring surface-level linguistic markers—such as syntactic complexity, word frequency, and transitional phrases—while neglecting deeper dimensions of writing quality, including argumentative coherence, originality, and contextual appropriateness. This echoes concerns raised in earlier studies (Adorni & Piatti, 2024; Grivokostopoulou et al., 2017) that algorithmic assessment tools tend to prioritize measurable proxies over holistic writing competence. In this context, fairness is not only about consistency or objectivity, but about the extent to which the scoring system respects the diversity of linguistic expression and values higher-order thinking. As Binns (2018) notes, fairness in algorithmic decision-making must be interpreted not only as the absence of bias, but also as procedural transparency, contextual sensitivity, and user alignment.

Another central theme is the lack of transparency in how AI systems generate writing scores. Students were unable to understand how their scores were calculated, and teachers struggled to interpret or defend the outputs. This lack of explainability significantly undermined trust in the system, despite its perceived consistency. This aligns with broader debates in AI ethics about the role of explainable AI in educational contexts. As Koenecke et al. (2020) and Wassink et al. (2022) argue, explainability is crucial not only for user trust, but also for accountability, contestability, and pedagogical utility. In the classroom, scoring systems should not operate as black boxes, but as dialogic tools that support learning. Without clear feedback loops or interpretable metrics, students are left disempowered, and teachers are rendered passive enforcers of algorithmic judgments they cannot critique. This insight also underscores the distinction between automation for administration and automation for learning. The current implementation of AI scoring systems in our case institution appears to favor the former—streamlining grading—while neglecting the latter—supporting revision, reflection, and skill development. Unless this imbalance is addressed, explainability will remain a critical point of failure in the adoption of AI-based educational tools.

A third theme concerns the tensions between AI systems and teacher authority. Teachers in this study reported feeling displaced or challenged by the perceived "objectivity" of AI scores, particularly when students questioned their feedback based on discrepancies between human and machine evaluations. This phenomenon aligns with Selwyn's (2019) concept of "data-driven displacement", where educational technologies subtly reconfigure professional identities and power relations. While AI is often presented as a neutral assistant, its adoption can lead to a form of deprofessionalization if teachers are not actively involved in the design, interpretation, and integration of technological tools. As Perrotta et al. (2021) argue, the adoption of AI systems in education must be accompanied by pedagogical agency and epistemic autonomy for educators. In this study, the lack of clear institutional positioning about the respective roles of AI and human evaluation exacerbated confusion. Teachers need guidance—not just on how to use AI systems, but also on how to position themselves in relation to them, pedagogically and ethically. Without such positioning, AI scoring tools risk becoming disciplinary instruments rather than collaborative tools.

5. CONCLUSION

This study set out to investigate the fairness controversies associated with AI-based scoring systems in Chinese university English writing courses and to explore their implications for classroom practice and institutional management. Through a qualitative case study approach involving in-depth interviews with students and teachers, the research uncovered significant tensions between the technical efficiency of AI assessment tools and the pedagogical and ethical demands of higher education.

The findings highlight four major concerns. First, both students and teachers perceive AI scoring as biased toward superficial linguistic features, often overlooking content quality and rhetorical complexity. Second, the opacity of the scoring process undermines user trust and limits the tool's formative potential. Third, the presence of AI in assessment reconfigures classroom authority, creating confusion about the relative roles of teachers and algorithms. Fourth, institutional management has not kept pace with the deployment of AI technologies, resulting in policy gaps, lack of training, and absence of clear governance structures.

Taken together, these insights reveal that the integration of AI scoring into writing instruction is not merely a technological issue, but a deeply social, pedagogical, and ethical one. While AI systems can undoubtedly support assessment processes, their meaningful and equitable adoption depends on active management, stakeholder engagement, and pedagogical alignment.

6. Limitations and Directions for Future Research

This study has several limitations. First, it is based on a single institutional case, which limits the generalizability of findings. Future studies could conduct comparative research across multiple universities with differing levels of digital infrastructure or AI integration policies. Second, while the focus was on teacher and student perspectives, further research could explore the views of administrators, system designers, and policy-makers to provide a more comprehensive picture of AI governance in education.

In addition, this study focused on English writing assessment. Similar investigations could be conducted in other disciplines, such as history, philosophy, or creative writing, where subjectivity and narrative complexity present unique challenges for AI evaluation. Finally, future research might adopt longitudinal designs to track how stakeholder perceptions evolve over time as systems improve or as training increases.

REFERENCES

1. Adorni, G., & Piatti, A. (2024). Designing the virtual CAT: A digital tool for algorithmic thinking assessment in compulsory education. arXiv preprint arXiv:2408.01263.
2. Binns, R. (2022). Human Judgment in algorithmic loops: Individual justice and automated decision-making. Regulation & governance, 16(1), 197-211.
3. Brown, K., & Wang, R. C. (2020). Politics and science: The case of China and the coronavirus. Asian Affairs, 51(2), 247-264.
4. Chen, H., & Pan, J. (2022). Computer or human: A comparative study of automated evaluation scoring and instructors' feedback on Chinese college students' English writing. Asian-Pacific Journal of Second and Foreign Language Education, 7(1), 34.
5. Cicero, L., Russo, A., Di Stefano, G., & Zammitti, A. (2025). The General Attitudes towards Artificial Intelligence Scale (GAAIS): validation and psychometric properties analysis in the Italian context. BMC psychology, 13(1), 641.
6. Edwards, A. (2020). Qualitative designs and analysis. In Doing early childhood research (pp. 155-175). Routledge.
7. Grivokostopoulou, F., Perikos, I., & Hatzilygeroudis, I. (2017). An educational system for learning search algorithms and automatically assessing student performance. International Journal of Artificial Intelligence in Education, 27(1), 207-240.
8. Kim, H., Baghestani, S., Yin, S., Karatay, Y., Kurt, S., Beck, J., & Karatay, L. (2024). ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores compared to human raters. Exploring artificial intelligence in applied linguistics, 73-95.
9. Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the national academy of sciences, 117(14), 7684-7689.
10. Liang, J., Huang, F., & Teo, T. (2024). Understanding Chinese University EFL Learners' Perceptions of AI in English Writing. International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT), 14(1), 1-16.
11. Liu, M., Lv, W., Yin, B., Ge, Y., & Wei, W. (2022). The human-AI scoring system: A new method for CT-based assessment of COVID-19 severity. Technology and Health Care, 30(1), 1-10.

12. Lu, X. (2019). An empirical study on the artificial intelligence writing evaluation system in China CET. Big data, 7(2), 121-129.
13. Luckin, R., & Holmes, W. (2016). Intelligence unleashed: An argument for AI in education.
14. Perrotta, D. (2021). Universities and Covid-19 in Argentina: from community engagement to regulation. Studies in Higher Education, 46(1), 30-43.
15. Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards Artificial Intelligence Scale. Computers in human behavior reports, 1, 100014.
16. Selwyn, N. (2022). Critical data futures. Digital Society, 593-609.
17. Stein, J. P., Messingschlager, T., Gnambs, T., Hutmacher, F., & Appel, M. (2024). Attitudes towards AI: measurement and associations with personality. Scientific Reports, 14(1), 2909.
18. Wassink, A. B., Gansen, C., & Bartholomew, I. (2022). Uneven success: automatic speech recognition and ethnicity-related dialects. Speech Communication, 140, 50-70.
19. Wu, X. (2022). Dynamic evaluation of college English writing ability based on AI technology. Journal of Intelligent Systems, 31(1), 298-309.
20. Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., ... & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. Complexity, 2021(1), 8812542.