Ensuring Fairness in AI-Assisted EFL Writing Assessment

Menglin Liu

School of Foreign Language, Sichuan Normal University, Chengdu, Sichuan 610066, China

Abstract. This study critically examines the role of artificial intelligence (AI) in EFL (English as a foreign) writing assessment, focusing on three key dimensions: algorithmic bias, system transparency, and student voice. Despite the potential for AI to revolutionize assessment through scalability and consistency, its rapid adoption has outpaced ethical considerations. This research underscores the importance of transparent systems to build trust and ensure fairness, as well as the necessity of incorporating student voice to enhance educational equity. Through a comprehensive literature review and case studies, this paper proposes strategies to mitigate bias, enhance transparency, and empower student participation. These strategies include diversifying training data, conducting regular fairness audits, and establishing transparent feedback channels. The findings suggest that truly effective AI assessment systems must balance technological efficiency with pedagogical considerations to support diverse EFL learners equitably.

Keywords: AI-based Essay Scoring; EFL Writing Assessment; Algorithmic Fairness; Assessment Transparency; Student Participation.

1. Introduction

The integration of AI into education has been increasingly recognized as a transformative force, with the potential to revolutionize teaching and learning processes. This shift is supported by various educational policies that emphasize the importance of leveraging technology to enhance educational outcomes. For instance, the U.S. Department of Education's Office of Educational Technology has outlined the need for exploring AI's role in education, highlighting its potential to personalize learning experiences and improve educational equity. In the realm of EFL education, AI-assisted writing assessment systems (such as the ChatGPT, Grammarly, Smart Education of China and Doubao etc.) offer unprecedented scalability and consistency, promising to provide timely and objective feedback to learners. However, the rapid adoption of these systems has outpaced critical examination of their fairness implications, raising significant ethical concerns.

The purpose of this study is to systematically investigate three fundamental aspects of equitable assessment in the context of AI-assisted EFL writing: the identification and mitigation of algorithmic bias, the establishment of transparent scoring mechanisms, and the incorporation of meaningful student participation. By addressing these dimensions, this research aims to contribute to the development of AI assessment tools that are not only technologically advanced but also ethically sound and pedagogically effective. The findings suggest that truly effective AI assessment systems must balance technological efficiency with pedagogical considerations to support diverse EFL learners equitably.

2. Literature Review

2.1 Algorithmic Bias in Writing Assessment

Contemporary scholarship identifies algorithmic bias as a systemic distortion emerging from both data limitations and design choices in AI systems [2]. Wambsganss [10] found that AI-assisted writing tools, such as ChatGPT, may transmit implicit biases through feedback, leading to a convergence in students' writing styles, thereby suppressing creativity and diversity. While no significant gender bias was identified, the standardising tendency of AI-generated feedback may undermine students' individual expression [10]. Yamashita [12], using a multilevel Rasch model, analysed bias in GPT-4o's scoring and revealed notable discrepancies across different racial and

ethnic groups: scores for Asian/Pacific Islander students were overestimated, whereas those for Hispanic/Latino students were underestimated. This bias may stem from historical patterns embedded in training data or from the algorithm's sensitivity to specific linguistic features [12]. Moreover, AI scoring tools tend to be more stringent than human raters, further highlighting their limitations in educational assessment [12].

2.2 The Imperative of Transparent Systems

Transparency in AI-enabled learning environments constitutes a multidimensional construct encompassing both technical explainability and communicative clarity [9]. Firstly, one of its key functions in educational evaluation is to ensure fairness in the assessment process. Research by Karran et al. [5] indicates that transparency significantly influences the level of trust and acceptance that students, teachers, and parents place in AI systems. It also enables educators and learners to better understand the decision-making processes of these systems, thereby enhancing their acceptance and credibility. Secondly, transparency is essential for the reliability of educational assessment. Chaudhry et al. [3] argue that transparent systems allow both educators and students to comprehend how decisions are made, which in turn improves the consistency and reliability of assessments. Transparency involves not only the openness of data but also clarity regarding the selection of AI models and the processes involved in their training. Finally, another vital role of transparency in educational evaluation is to enhance the validity of assessment. Transparent systems ensure that the feedback provided by AI is targeted and practically useful, thereby improving the overall effectiveness of the evaluation [5].

2.3 Student Voice as Pedagogical Necessity

The concept of student voice as an educational philosophy emphasises that learners should not be mere passive recipients of instruction, but active participants and collaborators in both teaching and assessment. Rudduck and McIntyre [8], through empirical research, demonstrated that "listening to students" reflections on teaching and assessment can significantly enhance learning outcomes and curriculum relevance. Fielding [4] further highlighted the transformative potential of student voice, arguing that it is not merely a form of feedback but a foundational force for promoting educational equity and systemic reform. In the context of AI-based essay scoring systems, students are often unable to effectively appeal or articulate concerns about the scores they receive. This form of technological silencing undermines students' agency and exacerbates the perceived divide between humans and technology. Drawing on Article 12 of the UN Convention on the Rights of the Child, Lundy [6] contends that students have the right to express views, to be heard, and to influence decision-making processes. Therefore, when implementing AI assessment tools, it is essential to design mechanisms that allow students to understand the scoring process, voice their opinions, and request reconsideration where appropriate, thereby ensuring meaningful participation in education.

2.4 Interdependence of Fairness Dimensions

The three dimensions of bias mitigation, transparency, and student voice form an interdependent ecosystem essential for equitable assessment. Bias identification fundamentally requires system transparency - opaque "black box" algorithms obscure potential discrimination against non-native speakers or those employing diverse linguistic styles [2]. Conversely, transparency enables student voice by providing learners with the understanding necessary to meaningfully engage with assessment outcomes [6]. Student participation then completes the cycle by surfacing hidden biases and driving demands for greater system openness [4]. This dynamic interplay suggests that fairness cannot be achieved through technical refinement alone, but requires continuous dialogue between system designers and educational stakeholders.

3. Case-Based Insights

3.1 Bias in AI scoring of Asian American Students

AI-based essay scoring systems, though praised for their efficiency and objectivity in educational contexts, have been shown to exhibit concerning biases. A large-scale study by researchers at the Educational Testing Service (ETS), as reported by Barshay [1], analysed over 13,000 essays scored by GPT-4o and found that the AI model systematically assigned lower scores to essays written by Asian American students when compared to scores given by human raters. And during the whole process of this study, only researchers know the information of the participants' race, demographics or some other private data which did not include any other secondary personnel such as the machine and human expert. And the study has showed that the human experts grade higher scores than GPT-4o by comparing their assessment which exists 0.9 distance. This disproportionate under-scoring raises critical concerns about how AI models interpret and evaluate language features that may be characteristic of Asian American students' writing styles, possibly due to training data that fails to represent such diversity adequately.

In the context of EFL education, where many learners share similar linguistic and cultural backgrounds, such bias can unfairly disadvantage students by misrepresenting their writing ability, affecting academic placement, feedback, and confidence. The AI-based essay scoring system may have stereotypes about certain groups at the initial stage of data input. For example, the system might assume by default that men are more engaged in external work, while women are mainly responsible for taking care of the family. This stereotype fails to take into account the specific circumstances of the students, which may lead to bias in the marking process.

3.2 Transparency Challenges in Automated Writing Evaluation: the Criterion Case Study

In 2015–2016, two college-level ESL writing courses at Iowa State University adopted ETS's Criterion as a formative assessment tool [7]. While the system promised instructor-free feedback on grammar and mechanics, its opaque scoring architecture quickly became a source of inequity. Criterion issued holistic scores and flagged errors without disclosing the underlying rules: error-type weightings remained undisclosed, thresholds for "passing" or "failing" a micro-feature were never published, and recall rates for categories such as preposition errors (18 %) or missing articles (37 %) were known only to the developers. Instructors received a final tally; students received a red underline and a brief suggestion—no explanation of how the algorithm arrived at its judgment, no path to contest it.

The second study in this study case simulated Criterion feedback on 20 sentences—half accurate, half deliberately flawed—to see whether ESL learners could still use the valid prompts to correct errors and whether having to judge accuracy heightened their cognitive load. It showed that without transparency documents, students from lower-proficiency backgrounds—whose error profiles diverge most from the training corpus—were disproportionately penalized. The algorithm's blind spots (e.g., under-identifying verb-form errors common among L1 Korean speakers) remained invisible to both teachers and learners, embedding historical data biases into daily classroom practice.

The criterion case study underscores the necessity for transparency in AI systems, particularly in high-stakes educational contexts. Transparency ensures stakeholders understand how AI functions and makes decisions, fostering trust and accountability. It also highlights the importance of including diverse stakeholder voices in the development and implementation of these systems to ensure fairness and equity. Conversely, deploying opaque algorithms without adequate oversight poses significant risks, especially when they have profound implications for individuals' educational and professional futures. Opaque systems can perpetuate biases and inequalities, leading to unjust outcomes that undermine the principles of fairness in education. Therefore, prioritizing transparency and stakeholder engagement is crucial for creating ethically sound and effective AI assessment tools.

3.3 Student Engagement with AI Feedback in the Chinese EFL Context

In the study by Yang, Gao, and Shen [11], student voice is critically illuminated through Chinese university students' interaction with the AI-powered writing evaluation system, Pigai. The study employed 5 students from a Chinese university in Shanghai who were required to complete an English writing task and then submitted their first draft to Pigai. They were encouraged to have multiple revisions and resubmissions, but the specific number of resubmissions was decided by individual students. The research reveals that students are not merely passive recipients of automated feedback but active participants who engage in multiple rounds of revision—up to twelve in one case—demonstrating a clear willingness to respond to, interpret, and sometimes challenge the system's suggestions. Particularly a student's (with the number 5) iterative engagement, which reflects a form of dialogic learning in which the learner attempts to understand and act on feedback despite limited guidance. However, the study also exposes key limitations in the system's ability to accommodate student voice. For example, when the system flagged a phrase such as "acquire a great deal of money award" as "Chinglish," it failed to provide contextualised explanations or culturally sensitive alternatives, leaving the student to engage in a trial-and-error process. This lack of transparency and responsiveness silences the student's linguistic and cultural expression, undermining their agency. Furthermore, the presence of self-initiated revisions—those not prompted by system feedback—signals learners' critical stance and autonomy in managing their writing development.

It can be conducted that within AI-based essay scoring systems, student voice manifests in multiple ways. It is not solely reflected by their acceptance or rejection of the feedback provided. Equally significant is their independent revision behavior, which showcases their critical thinking and autonomous learning processes as they work towards refining their writing skills. However, a critical gap exists in the current systems. Without proper mechanisms that allow students to question or appeal automated judgments, their voices are largely marginalized in the decision-making processes of these AI systems. This situation effectively silences their input and diminishes their agency in the learning and assessment journey.

This scenario underscores a crucial need in the design and implementation of AI systems in educational contexts. AI systems should not merely function as tools that deliver feedback. Instead, they should be designed to actively facilitate meaningful student participation in assessment practices. This involves creating platforms where students can voice their concerns, seek clarifications, and engage in dialogues about the feedback they receive. By incorporating such participatory elements, AI systems can better align with the principles of fairness and accountability that are fundamental to effective EFL education. This approach not only enhances the educational value of AI but also promotes a more inclusive and equitable learning environment where every student's voice is heard and valued.

4. Strategies to Address Unfairness

4.1 Mitigating Bias in Training Data and Algorithms

Bias in AI-based scoring systems often originates from imbalanced or unrepresentative training data, which can systematically disadvantage learners from particular linguistic, cultural, or demographic backgrounds. Firstly, to counteract this, training corpora must be diversified to reflect a wide spectrum of English varieties, writing styles, and cultural expressions, particularly from non-native speakers in EFL contexts. This helps the AI model learn to recognise and fairly evaluate legitimate linguistic variation rather than penalising non-standard but valid forms. Secondly, regular fairness audits should also be institutionalised, involving statistical analyses of scoring discrepancies across groups to detect and address systemic inequalities. These audits can identify whether certain types of language, vocabulary, or syntax are disproportionately penalised. In addition, the integration of human benchmarks—comparing AI scores with those from experienced human raters across demographic groups—can help calibrate algorithms and reveal misalignments. A commitment to

ongoing refinement and ethical oversight is essential to ensure that AI tools do not reinforce educational injustice under the guise of objectivity.

4.2 Improving Transparency in AI Scoring Systems

Transparency is a foundational condition for fairness in AI-based educational assessments. In the context of essay scoring, it entails making the internal workings of AI models intelligible and accessible to both educators and learners. Firstly, one critical step is the public disclosure of the criteria used by AI systems to evaluate writing quality, such as coherence, vocabulary, grammar, and organisation. Without this disclosure, students are left to interpret opaque results without a clear understanding of how their texts are judged. Furthermore, enhancing explainability through user-friendly interfaces—such as annotated feedback, confidence scores, or simplified decision paths—can help students and teachers trace how specific inputs lead to specific outputs. This supports not only learning but also trust. Another important measure is the inclusion of justifications for individual scores, enabling learners to understand not merely what was wrong or right in their writing, but why certain choices were rewarded or penalised. Transparent AI systems should also make available clear documentation of model updates, limitations, and procedures for appeal or review.

4.3 Empowering Student Voice and Participation

Ensuring fairness in AI-based assessment is not only a technical issue but also a democratic imperative. Student voice—understood as the right of learners to understand, respond to, and challenge assessment outcomes—must be actively incorporated into the design and implementation of AI scoring systems. One practical strategy is to establish transparent feedback channels through which students can appeal or question automated scores. These channels must go beyond formality and be embedded into the workflow of AI systems, allowing students to submit explanatory comments, request human review, or highlight perceived errors in feedback. Furthermore, students should be consulted during the development and refinement of AI tools, for example through participatory design processes or user-testing that captures their experiences and preferences. Such involvement can surface unintended impacts and improve system responsiveness. Equally important is fostering feedback literacy—the ability of students to interpret and act upon feedback—so that learners engage critically with AI output rather than accepting it as absolute. This includes classroom instruction that addresses the nature and limitations of AI-generated evaluation.

5. Conclusion

This study has critically examined the role of AI in EFL writing assessment, focusing on algorithmic bias, system transparency, and student voice. The central findings highlight the significant demographic biases present in current AI scoring systems, such as the systematic under-scoring of essays written by Asian American students. This underscores the urgent need for more equitable AI assessment tools. Transparency in AI systems is essential for building trust and ensuring fairness, as it allows stakeholders to understand and challenge AI-generated decisions. Furthermore, incorporating student voice into the assessment process is crucial for promoting educational equity and enhancing student engagement.

Despite these insights, the study has several limitations. The sample size and scope of the case studies may limit the generalizability of the findings. Additionally, the study primarily focused on specific demographic groups and may not fully capture the experiences of other underrepresented populations. Future research should aim to address these limitations by conducting larger-scale studies with more diverse samples. There is also a need for further investigation into the long-term impacts of AI feedback on student learning outcomes and the development of more sophisticated fairness metrics that can better account for the complexities of language and culture in EFL contexts.

In conclusion, while AI offers significant potential to enhance EFL writing assessment, it is imperative to address the ethical challenges it presents. Future research should focus on developing

AI systems that are not only technologically advanced but also ethically sound and pedagogically effective. By prioritizing fairness, transparency, and student voice, we can create AI assessment tools that support diverse EFL learners equitably, ensuring that every student has the opportunity to succeed.

References

[1] Barshay, J. Proof points: Asian American students lose more points in an AI essay grading study—but researchers don't know why. The Hechinger Report. 2024.

[2] Binns, R. Fairness in machine learning: Lessons from political philosophy. In Conference on fairness, accountability and transparency. 2018, p. 149-159.

[3] Chaudhry, M. A., Cukurova, M., & Luckin, R. (2021). A transparency index framework for AI in education. Cham: Springer International Publishing. 2022. p. 195-198.

[4] Fielding, M. (2004). Transformative approaches to student voice: Theoretical underpinnings, recalcitrant realities. British Educational Research Journal. 2004, Vol. 30(No. 2), p. 291–311.

[5] Karran, A. J., Charland, P., Martineau, J.-T., Ortiz de Guinea Lopez de Arana, A., Lesage, A. M., Senecal, S., & Leger, P.-M. (2023). Multi-stakeholder perspective on responsible artificial intelligence and acceptability in education. npj Science of Learning. 2025, Vol. 10(No. 1), p. 44.

[6] Lundy, L. 'Voice' is not enough: Conceptualising Article 12 of the United Nations Convention on the Rights of the Child. British Educational Research Journal. 2007, Vol. 33(No. 6), p. 927–942.

[7] Ranalli, J., Link, S., & Chukharev-Hudilainen, E. Automated Writing Evaluation for Formative Assessment of Second Language Writing: Investigating the Accuracy and Usefulness of Feedback as Part of Argument-Based Validation. Educational Psychology, 2017, Vol. 37(No. 1), p. 8-25.

[8] Rudduck, J., & McIntyre, D. (2007). Improving Learning through Consulting Pupils. Routledge. 2007.

[9] U.S. Department of Education, Office of Educational Technology. Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations Washington, DC: U.S. Department of Education. 2023. Information on: https://tech.ed.gov.

[10] Wambsganss, T. Hidden bias in AI writing tools: What educators need to know. SocietyByte. 2023.

[11] Yang, H., Gao, C., & Shen, H. Learner interaction with, and response to, AI-programmed automated writing evaluation feedback in EFL writing: An exploratory study. Education and Information Technologies. 2024. Vol 29(No. 6), p. 3837–3858.

[12] Yamashita, T. Exploring potential biases in GPT-4o's ratings of English language learners' essays. Language Testing. 2025, Vol 42(No. 3), p. 1–15.