Journal of Applied Linguistics Studies, Vol.4, No.1, 2024: 177-187
https://jals.aliabad.iau.ir
ISSN: 2820-9974

A Review of the Use of Artificial Intelligence and Automated Writing Evaluation Systems as Sources of Provision of Feedback in Assessing Writing

Samira Gharekhani[1], Seyyed Hassan Seyyedrezaei[1*]

[1]Department of English Language Teaching, Ali.C., Islamic Azad University, Aliabad Katoul, Iran.

E-mail: Samiragharekhani606@gmail.com

*Corresponding author's Email: S.h.seyyedrezaei@iau.ac.ir

Received: 30-01-2025, Accepted: 18-03-2025

ABSTRACT

Due to the novelty of the field of the artificial intelligence (AI) and automated writing evaluation systems (AWES) as sources of provision of feedback in language teaching and the need to collect different positive and negative findings in the field into one major study, the current systematic review examines literature on current notions, terminologies, methodologies, and designs as well as the main findings in the use of AI and AWES for providing feedback in writing assessment. Utilizing search strategies for keywording and mapping, the study explored common themes in the articles to address its research questions. Analyzing the general themes revealed that all the articles employed quantitative methods with a quasi-experimental design. Moreover, various notions and terminologies were used to capture the primary findings in the field. Based on the findings, the study emphasized the need for researchers to adopt more interactive designs to further investigate the potential of AI and automated writing evaluation systems in providing feedback in writing assessment. It also called for additional studies to review related domain findings to offer more productive and practical outcomes for educational settings.

KEYWORDS: Artificial Intelligence (AI), Automated Writing Evaluation Systems (AWES), Provision of Feedback, Technology, Writing Assessment

INTRODUCTION

The use of Artificial Intelligence (AI) and Automated Writing Evaluation Systems (AWES) in academic writing as a feedback tool has been specifically acknowledged in academic settings. Various studies have investigated the reproducibility and quality of feedback produced by AWES and AI models like ChatGPT compared to human-written abstracts. Dugan et al. (2023) focused on the rise of AI content detection in research articles, emphasizing the application of models like GPT-3.5 and GPT-4. Kim et al. (2021) evaluated the reproducibility of structured articles produced by ChatGPT and Bard, stressing adherence to journal guidelines. Ayers (2023) carried out a single-blind analysis comparing the quality of human versus AI-generated literature, finding notable similarities. Zribi and Smaoui (2021) examined thematic progression in texts created by both humans and AI, demonstrating ChatGPT's capability to mimic human speech patterns. Collectively, these studies highlight AI's potential to generate feedback that closely matches human-written content, offering valuable insights into the evolving field of AI-assisted academic writing.

Research on feedback in writing improvement and the effectiveness of human versus AI-generated feedback presented a diverse array of findings and perspectives. This body of work highlighted a key distinction between the immediate enhancement of a particular piece of writing through revisions and the long-term development of a writer's overall skills. Research by Ashwell (2000) and Bitchener (2008) indicated that feedback can effectively improve subsequent drafts of the same work. However, the broader impact on overall writing proficiency, assessed through various instances of writing, presents a more complex picture. The effectiveness of

feedback seems to depend on the learner's developmental readiness, as suggested by Pienemann (1989), with several researchers (Bitchener & Knoch, 2010b; Ferris, 2002, 2003; Ferris & Hedgcock, 2005; Goldstein, 2004; Hyland, 2009; Storch, 2010) advocating for tailored feedback.

The role of AI in providing feedback has been examined in numerous studies. Wu et al. (2024) found that Intelligent Personal Assistants (IPAs) significantly enhance language skills through interactive dialogue. Escalante et al. (2023) and Dai et al. (2023) reported that AI feedback can be as effective as, or even more efficient than, human feedback in specific contexts. These findings suggest that while AI feedback matches human feedback in terms of clarity and precision, it is particularly appreciated for its time-saving benefits. Additionally, Mizumoto and Eguchi (2023) highlighted the reliability and accuracy of AI feedback, underscoring its potential for writing assessment, although its effectiveness in improving BR abstract writing skills remains underexplored.

REVIEW OF THE LITERATURE

In an important study in the field, Carter and Absalom (2023) reviewed recent research using corpora, with a focus on written registers in academic contexts. He examined the various applications and uses of corpora in writing classrooms, citing several applied linguistics studies, including those by Zhang et al. (2021), which have employed diverse research methodologies over the past two decades to analyze a wide range of academic texts. These methodologies provide reliable and generalizable findings that effectively describe these registers and often have direct pedagogical applications.

Escalante et al. (2023) conducted a two-phase study on English language students. In phase one, they assessed the learning outcomes of 48 students over six weeks by having them write a 300-word science text. The experimental group received feedback from ChatGPT (GPT-4), while the control group received feedback from a human teacher. Results showed no significant difference in language development between the two types of feedback, with the primary advantage of AI feedback being time-saving for teachers. In phase two, they explored students' preferences for feedback types. About half of the students preferred teacher feedback for its emotional benefits and improvement of speaking skills, while the other half favored AI feedback for its clarity and precision in improving writing skills.

Dai et al. (2023) investigated the use of ChatGPT for providing corrective feedback on undergraduate students' writing skills. They found that AI-generated feedback was more readable and accurate than teacher feedback, though teacher feedback was considered more efficient in certain situations. Mizumoto and Eguchi (2023) examined ChatGPT's performance in essay evaluation, feeding it 12,100 essays by non-native English writers and evaluating its rubric-based feedback and scores against benchmark levels. Their results confirmed the reliability and accuracy of ChatGPT's feedback. These studies highlight the feasibility and reliability of using AI tools like ChatGPT for writing assessment, although the role and effectiveness of ChatGPT feedback in improving writing skills remain unexplored.

Wu et al. (2024) studied the impact of Intelligent Personal Assistants (IPAs) on Mandarin learners in the second grade of primary school, focusing on their interaction and the acquisition of listening and speaking skills during this critical learning period. The study concluded that students had more conversations, used more interactive strategies, and found better topics when interacting with IPAs compared to classmates. Overall, conversations with IPAs enhanced students' listening and speaking skills. To address the need identified by these studies, the present research aims to answer the following questions:

What are the main methodologies, findings, and challenges in the use of AI and AWES for providing feedback in writing assessment?

DESIGN

The present systematic review investigates existing literature on current concepts, terminologies, methodologies, designs and findings in the field of AI and AWES systems as a source of feedback to written works, offering valuable insights. The study employed keywording and mapping search strategies to select 15 key articles from leading journals. The common themes of these studies were then analyzed and presented to address the research questions.

Inclusion/exclusion Criteria:

The research work is related to the AI and AWES systems as a source of feedback to written works.

The publication includes state of the art on digital competence.

Papers published between 2015 and 2024 (to include the recent developments in the field).

The papers are written in English.

They are published both in national and overseas journals.

SEARCH STRATEGY

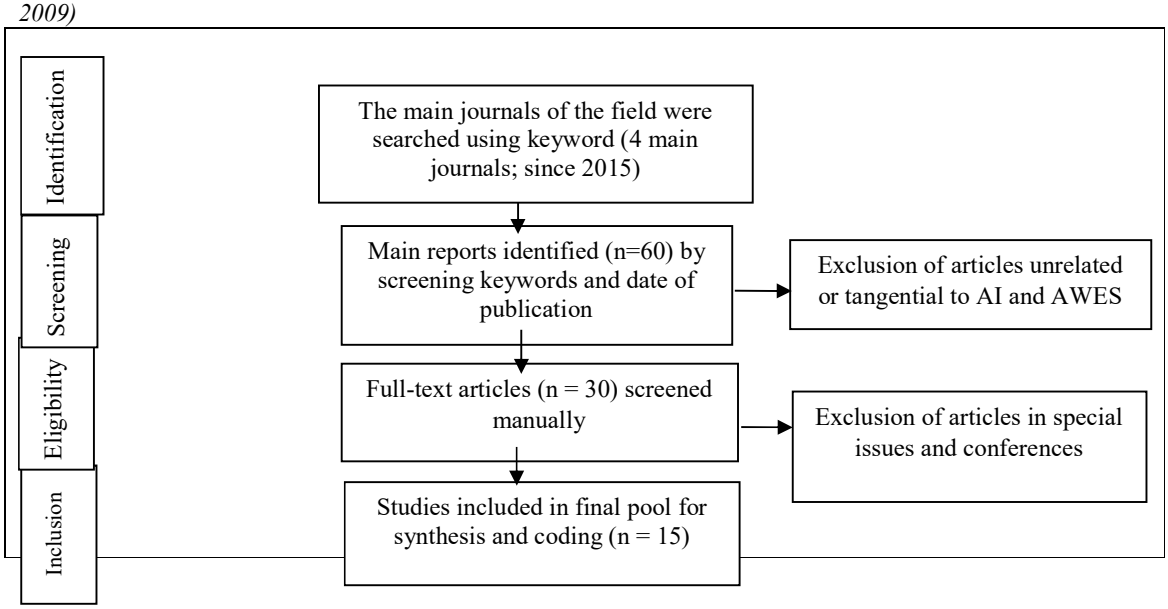
The search for the most relevant, precise, and reliable findings was conducted in the main journals of the field using keywording and mapping techniques. Advanced search engines within these primary publications in the field were utilized, with the following keywords:

- Methodology and Design
- Notions and Terminology
- The Main Findings

PRISMA Chart of the Study

Based on the outlined criteria and a thorough analysis of approximately 60 relevant articles by two researchers, a total of 15 studies were ultimately selected for review. The steps and process for identifying and selecting suitable articles are depicted in the PRISMA Chart (Figure 1), originally adapted from Moher et al. (2009).

Figure 1.
The PRISMA Chart of the processes for the creation of the report pool (originally adopted from Moher et al., 2009)



Content Analysis and Coding Process

After collecting the necessary data, content analysis was performed using thick description and qualitative interpretation with MAXQDA software. The aim was to validate the results by diversifying data related to the relevant themes under the thematic headings in the results section. The data analysis was carried out by the Validity Committee, consisting of three education experts who wrote, grouped, coded, identified themes, and interpreted the findings, with two raters ensuring the reliability of the process, who employed a coding scheme

alongside themes identified using MAXQDA Software (2012) to analyze the final articles. The process involved keywording and mapping strategies, first selecting prominent journals in the field and then narrowing their content to align with the study's main themes. To ensure the thoroughness of the analysis, the raters effectively utilized coding books and features within the MAXQDA Software to screen and code both quantitative and qualitative data. The themes that emerged were coded collaboratively by the two raters, who then calculated inter-coder reliability coefficients to validate the consistency and reliability of their coding and interpretations. The result was a coefficient of 0.91, significantly exceeding the conventional threshold of 0.70 set by Bos (1989).

RESULTS

The qualitative data of the study was presented based on themes that emerged from the content analysis. According to this analysis, the main notions, terminologies, methodologies, and designs as well as the main findings in the use of Artificial Intelligence and automated writing evaluation systems for providing feedback in writing assessment were divided into main themes and categories, which will be elaborated upon.

METHODOLOGY AND DESIGN OF THE REVIEWED STUDIES

The first emergent theme was related to the methodology and design. Regarding the methodology and design of the articles, all of the 15 studies followed Quasi-Experimental design with quantitative methods. The summary of this theme is presented in the following table.

Table 1.

Designs and Methodologies of the Reviewed Studies

Methodology	Quantitative
Number	15
Percentage	100
Design	Quasi-Experimental
Number and percentage of articles in Design	15 (100%)

The first reviewed article which belonged to Dai et al. (2023), utilized a systematic and comprehensive approach to vision-language instruction tuning based on the pretrained BLIP-2 models. They gathered 26 publicly available datasets and transformed them into instruction tuning format. The work by Escalante et al. (2023) conducted two longitudinal studies; the first study examined learning outcomes of 48 university English as a New Language (ENL) learners over six weeks, comparing feedback from ChatGPT (GPT-4) and human tutors. Second study analyzed perceptions of 43 ENL learners who received feedback from both sources. Mizumoto and Eguchi (2023) utilized the GPT-3 text-davinci-003 model to automatically score 12,100 essays from the ETS Corpus of Non-Native Written English (TOEFL11). The scores were then compared to benchmark levels to evaluate accuracy and reliability. Dasgupta et al. (2018) presented a deep convolution recurrent neural network (DCRNN) model for automatic essay scoring (AES). The model incorporated complex linguistic, cognitive, and psychological features to enhance essay evaluation.

Dong and Zhang (2016) conducted a study that employed a recurrent neural network (RNN) for automatic feedback generation in language learning. The model was trained on a large dataset of learner essays, focusing on evaluating content, structure, and language use. Taghipour and Ng (2016) developed a neural approach to automated essay scoring (AES) using deep neural networks, specifically focusing on convolutional and recurrent neural networks (CNNs and RNNs). The model was trained on the ASAP-AES dataset. Liu and Kunnan (2016) investigated the application of the WriteToLearn automated writing evaluation (AWE) system on Chinese undergraduate English majors' essays. The essays were scored by both human raters and the WriteToLearn system.

Geckin (2023) compared the scores assigned by a large language learning model (ChatGPT-3.5) and human raters for second-language academic writing. The research involved first-year college students (n=43) who completed a paragraph writing task. Kim et al. (2021) compared the intra- and inter-rater reliability of a deep learning model and human examiners for detecting laryngeal penetration or aspiration in videofluoroscopic

swallowing studies (VFSS). The test dataset consisted of 173 video files. Zhang et al. (2022) explored the clinical value of AI-assisted bone age assessment (BAA) among children with growth hormone deficiency (GHD). It involved 52 children and 290 radiographs, comparing AI-assisted evaluations with those of senior pediatric endocrinologists. Chen and Pan (2022) compared the effectiveness of an automated evaluation scoring system (Aim Writing) and human instructors' feedback on Chinese college students' English writing. The research involved a mixed-methods approach, including surveys and performance assessments.

Algburi (2024) investigated the impact of automated writing feedback (AWE) on students' writing in ESL/EFL contexts. It involved analyzing the quality of writing and learning outcomes when AWE was integrated with human feedback. Aldosemani (2023) reviewed the effectiveness, impact, and pedagogical implications of Automated Writing Evaluation (AWE) in English as a Foreign Language (EFL) contexts. It involved analyzing 16 studies that met specific inclusion criteria. Nguyen (2023) explored the efficacy of ChatGPT in language teaching by evaluating its feedback on ten writing essays from advanced English students. The feedback was compared to that of an experienced senior teacher. The study employed a qualitative approach, including in-depth interviews with the teacher. Alikaniotis et al. (2016) introduced a model using Long-Short Term Memory (LSTM) networks for Automated Text Scoring (ATS). The model learned word representations based on their contribution to the text's score.

Overall, the review of the studies showed that a majority of the studies apply quantitative methods in order to investigate the provision of feedback by AI and AWES systems. These findings suggest that researchers try to explore the way the intervention by the researchers can result in the efficacy of the feedback systems, or to investigate the way such systems can have different impacts on the participants, due to their different feedback mechanisms.

NOTIONS AND TERMINOLOGY

The studies introduced different notions and terminology which are reviewed here;

The first reviewed article by Dai et al. (2023), included the following Notions and Terminology: Key concepts include vision-language instruction tuning, instruction-aware Query Transformer, and zero-shot performance. Key concepts of the work by Escalante et al. (2023) included Automated Writing Evaluation (AWE), generative AI tools, ChatGPT (GPT-4), and learner perceptions. Mizumoto and Eguchi's (2023) concepts included Automated Essay Scoring (AES), natural language processing (NLP), transformer-based large language models, and linguistic features.

Key concepts of the study by Dasgupta et al. (2018) included Automated Essay Scoring (AES), deep convolution recurrent neural network (DCRNN), linguistic features, and psychological features. Dong and Zhang (2016) also presented Recurrent Neural Network (RNN), Automatic Feedback Generation, Learner Essays, and Content Evaluation. Key concepts of Taghipour and Ng (2016) included Automated Essay Scoring (AES), Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and ASAP-AES dataset.

Liu and Kunnan (2016) also introduced Automated Writing Evaluation (AWE), WriteToLearn, human raters, and error feedback. Geckin (2023) presented key concepts that included Automated Writing Evaluation (AWE), generative AI tools, ChatGPT-3.5, and second-language writing assessment. Key concepts of Kim et al. (2021) included Automated Detection, Deep Learning Model, Videofluoroscopic Swallowing Studies (VFSS), and Reliability Assessment. Zhang et al.'s (2022) key concepts included Automated Bone Age Assessment (BAA), Artificial Intelligence (AI), Growth Hormone Deficiency (GHD), and Radiographic Analysis. Chen & Pan (2022) also presented Automated Evaluation Scoring (AES), Human Feedback, English as a Foreign Language (EFL), and Hybrid Feedback Models. Algburi (2024) introduced key concepts including Automated Writing Evaluation (AWE), ESL/EFL contexts, human feedback, and writing quality. Key concepts of Aldosemani (2023) included Automated Writing Evaluation (AWE), Natural Language Processing (NLP), EFL contexts, and pedagogical implications. Key concepts of Nguyen (2023) included Automated Feedback, ChatGPT, English Writing Assessment, and Teacher Comparison. Alikaniotis et al. (2016) also introduced notions like Automated Text Scoring (ATS), Long-Short Term Memory (LSTM) networks, word representations, and text scoring. Generally, the studies utilized different key concepts, ranging from recurrent terms in the field (e.g., Automated Writing Evaluation; AWE and Natural Language Processing; NLP), to more recent and innovative

trends and notions (e.g., Automated Detection and ChatGPT). These findings also show the timeline of the development from more traditional studies working on NLP to more recent forms of AI systems.

INFLUENTIAL FINDINGS

The first reviewed article which belonged to Dai et al. (2023) demonstrated that Instruct BLIP achieved state-of-the-art zero-shot performance across various tasks, outperforming BLIP-2 and larger models like Flamingo. It also showed improved performance when finetuned on individual downstream tasks.

Escalante et al.'s (2023) study showed no significant difference in learning outcomes between AI-generated and human feedback. Study revealed a near even split in preference for AI-generated versus human-generated feedback, with both forms showing clear advantages. Mizumoto and Eguchi (2023) found that AES using GPT-3 has a certain level of accuracy and reliability, suggesting that AI language models can effectively support human evaluations. It also highlighted the potential of linguistic features to enhance scoring accuracy.

Dasgupta et al. (2018) demonstrated that augmenting traditional word/sentence embeddings with qualitative feature vectors improved the accuracy of AES. This approach provided a more comprehensive evaluation of essay quality. Dong & Zhang (2016) found that the RNN model could effectively generate feedback that was comparable to human feedback in terms of accuracy and usefulness. This approach demonstrated the potential of AI to assist in language learning by providing consistent and timely feedback.

The study by Taghipour and Ng (2016) demonstrated that the neural approach to AES achieved competitive performance compared to traditional methods, highlighting the potential of deep learning techniques in improving the accuracy and reliability of essay scoring. Liu and Kunnan (2016) found that WriteToLearn was more consistent but highly stringent compared to human raters. It also had difficulty detecting errors related to articles, prepositions, word choice, and expression.

Geckin (2023) found a slight to fair but significant level of agreement between the scores assigned by ChatGPT-3.5 and two human raters. It suggested that combining AI-generated scores with human scores could provide reliable assessments. Kim et al. (2021) found that the deep learning model achieved almost perfect intra-rater reliability and moderate to substantial inter-rater reliability compared to human examiners. This suggests that AI can reliably detect laryngeal penetration or aspiration in VFSS videos.

Zhang et al. (2022) found that AI assistance significantly improved the accuracy and consistency of junior pediatric endocrinologists' evaluations, reducing errors and variability. This suggests AI's potential to enhance clinical assessments. Chen and Pan (2022) found that while Aim Writing provided instant and corrective feedback, it was insufficient to meet all students' needs. The results suggested that a hybrid model combining AES and human feedback was more effective in improving students' writing skills.

Algburi (2024) found that AWE can improve the quality of writing and learning outcomes if it is integrated with and supported by human feedback. It also provided recommendations for further research to enhance AWE tools. Aldosemani (2023) found that AWE can enhance EFL student writing skills, with varying effectiveness based on student proficiency levels. It provides quality feedback and can be a reliable and valuable tool, but human intervention is essential to maximize its outcomes and mitigate limitations.

Nguyen (2023) found a significant similarity between the grades assigned by ChatGPT and the senior teacher, suggesting that ChatGPT can be a useful tool for reducing teachers' workload. However, it also highlighted the importance of human expertise for comprehensive feedback. Alikaniotis et al. (2016) demonstrated that the LSTM-based model achieved excellent results compared to similar approaches. It also introduced a method for identifying text regions that the model found more discriminative.

Overall, the studies resulted in influential findings, which underscored the advantages of the AI and AWES systems. They indicated that AI provides consistent scoring, unlike human raters, who might be influenced by subjective biases or fatigue, thereby enhancing the reliability of writing assessments. They also highlighted the advantage of immediate feedback, as AI systems can deliver real-time feedback to students, allowing for revisions before final submission, creating a dynamic learning environment.

However, they shed light on the disadvantages and shortcomings of such feedback systems. For instance, these tools lacked nuance, as AI systems may struggle to evaluate creativity, voice, and emotional resonance, which are crucial for effective communication. Human raters can provide qualitative feedback that AI might miss. Moreover, they showed that while AI systems offered consistent feedback, they were often more stringent than human raters, indicating a potential mismatch in grading standards. They also pointed out that AI can introduce biases in its training data, potentially disadvantaging certain student groups if these biases are present in the data used to train AI raters.

DISCUSSIONS

Overall, the review of the studies showed that the use of AI and AWES in writing assessment is well-established, with promising results reported in the use of algorithms for scoring student writing which is consistent with different studies (e.g., Dikli, 2006; Refaat et al., 2012). Commercially available AI products use specific scoring algorithms to evaluate essays for tests like TOEFL, IELTS, or GMAT (Wong & Bong, 2021; Mizumoto & Eguchi, 2023). Such high-stake tests have undeniable impact on student motivation (Rezaeian et al., 2020). As educational institutions and teachers seek more efficient and effective methods for evaluating student writing, AI integration offers a promising alternative to traditional human raters (Dasgupta et al., 2018; Dong & Zhang, 2016). AI raters, powered by machine learning algorithms and natural language processing (NLP), evaluate various aspects of writing, including grammar, coherence, style, and content relevance (Song & Song, 2023; Taghipour & Ng, 2016). Machine learning allows AI systems to learn from previously scored essays, identifying patterns that correlate with high or low scores, thereby enhancing the accuracy of the AI rater over time. AI can also benefit from Automated Essay Scoring (AES) systems, which provide numerical scores based on predefined rubrics tailored to specific writing tasks or educational standards (Jackaria et al., 2024).

These reviewed studies showed that AWES and AI raters offer several advantages, including scalability, which allows them to process a vast number of essays much faster than human raters, making them highly useful in large-scale assessments or standardized testing which is supported by Geckin (2023). Kim et al. (2021) also noted that AI provides consistent scoring, unlike human raters, who might be influenced by subjective biases or fatigue, thereby enhancing the reliability of writing assessments. Ayers (2023) highlighted the advantage of immediate feedback, as AI systems can deliver real-time feedback to students, allowing for revisions before final submission, creating a dynamic learning environment. Additionally, AI tools are cost-effective; they reduce the need for extensive human rater involvement, lowering costs for large educational institutions (Zhang et al., 2022).

However, the study showed that AWES and AI tools also have limitations and challenges. Lloyd et al. (2022) pointed out that these tools lack nuance, as AI systems may struggle to evaluate creativity, voice, and emotional resonance, which are crucial for effective communication. Human raters can provide qualitative feedback that AI might miss. Liu and Kunnan (2016) found that while AI systems like WriteToLearn offered consistent feedback, they were often more stringent than human raters, indicating a potential mismatch in grading standards. Chan et al. (2022) pointed out that AI can introduce biases in its training data, potentially disadvantaging certain student groups if these biases are present in the data used to train AI raters. Another issue with AI raters is their "black box" nature, as described by Kural et al. (2022), which can make it difficult for educators and students to understand how scores are derived, leading to mistrust in AI evaluations.

Additionally, there were contradictory findings in studies' findings and their writing instruction and evaluation, particularly regarding the integration of AWES and AI with human expertise. Some studies highlight the benefits of combining AI and human feedback to improve writing skills and learning outcomes. Escalante (2023) found that students who preferred AI-generated feedback appreciated its clarity and specificity in enhancing their writing. Algburi (2024) also demonstrated that Automated Writing Evaluation (AWE) can improve writing quality and learning outcomes when combined with human feedback. Wei (2023) discussed how AI technologies can aid English as a Foreign Language (EFL) writing by offering personalized and contextualized feedback, thus improving writing instruction quality.

However, the findings of the studies showed that ethical considerations remain a significant challenge. Humphry and Heldsinger (2020) raised concerns about the role of technology in education, particularly regarding data privacy and the potential over-reliance on automated systems. Additionally, transparency and trustworthiness issues persist, with many educators hesitant to fully embrace AI due to a lack of understanding of its rating criteria. Some studies suggest that the reliability of scores assigned to written work improves when human raters collaborate with AI assessment tools, with a high correlation observed between their scores.

Despite these findings, different studies have underscored the influential role of technologies in directing the teaching and assessment objectives. Iranbakhsh and Seyyedrezaei (2011) highlighted the role of IT in scholarly communication or scientific collaboration and Seyyedrezaei (2015) believed that IT should be an integral part of teaching and assessment. Moreover, Seyyedrezaei et al. (2016) indicated that innovative technologies like MALL which can enhance vocabulary acquisition should be used in language teaching. Hence, using available resources

strategically is a main characteristic of good language learners (Maftoon & Seyyedrezaei, 2012). Therefore, more studies in other disciplines using various methodology and design are needed in order to better broaden our knowledge of the impact of

CONCLUSION AND IMPLICATIONS

Overall, the review of the studies showed that the integration of AWES and AI raters in writing assessments holds significant promise for educational institutions by offering scalability, consistency, and immediate feedback. However, it is vital to address challenges related to nuance and bias. Hybrid models that combine AI and human evaluations can improve the overall quality of writing assessments. With technological advancements, continuous research and dialogue are necessary to ensure AI raters effectively support student learning and development. The study also emphasized the need for researchers to adopt more interactive designs to further investigate the potential of AI and automated writing evaluation systems in providing feedback in writing assessment.

The review indicates that both human and AI rating tools have their own pros and cons for language learners' writing assessments. To mitigate the limitations of each, researchers have explored hybrid models that integrate both AI and human evaluations. The comparison between human and AI raters in writing assessment has become a crucial research area, especially as educational institutions increasingly adopt technology and different terminology in their assessment practices. This discussion focuses on the effectiveness, reliability, and pedagogical implications of using AI to evaluate written work, particularly in second-language learning contexts.

In spite of the fact that more studies with innovative designs and methodology in the field are needed to improve the quality of such feedback, one thing is for sure. Human feedback still holds irreplaceable value, and the significance of human feedback alongside AWE and AI systems is well-documented. This balanced approach could lead to a more comprehensive and effective writing assessment system.

The review underscores AI's role as a complementary tool, enhancing the quality of writing instruction and evaluation through a hybrid model, preserving human interaction at the forefront. It can be inferred that automated writing feedback is not only efficient in reducing teachers' workload but also provides immediate feedback to students, making it a practical addition to writing instruction. However, integrating human and AI evaluations leads to more reliable scoring of written work, with a high correlation between human raters and AI raters. This hybrid approach appears to offer a balanced and effective solution, which can finally develop on the strengths of both AI and human evaluators.

One of the primary pedagogical implications of AI feedback is its capacity to provide personalized learning experiences. AI systems can analyze substantial amounts of data regarding students' individual performance and learning styles, enabling the customization of educational content and feedback that cater to specific needs. The provision of rapid and tailored feedback allows educators to support varied learning paces, potentially leading to greater student engagement and improved outcomes. Furthermore, AI-generated feedback can shift the traditional dynamics of evaluator-evaluated relationships, promoting a more student-centered approach that prioritizes learners' agency in assessing their progress and understanding. This transformation can empower students to perceive feedback as a collaborative tool rather than merely an evaluative measure, enhancing their critical thinking and reflective skills.

Moreover, the implementation of AI in educational contexts necessitates significant infrastructural and pedagogical adaptations. Educational institutions must cultivate the technological foundation to support AI tools effectively, thereby ensuring that educators are equipped to leverage these advancements without compromising teaching quality. Future studies can utilize AWE and AI systems to study and review any significant improvement in other language skills (e.g., listening, speaking, grammar, and vocabulary) when used in conjunction with human feedback to see if combining human and analytics feedback can boost student engagement and performance. Furthermore, ethical considerations must be foregrounded as educational stakeholders navigate the implications of AI.

REFERENCES

Algburi, E. (2024). Combination of awe (criterion) feedback with the process approach and its impact on EFL writing content/idea development and organization. *International Journal of Academic Research in Progressive Education and Development, 13*(1). https://doi.org/10.6007/ijarped/v13-i1/20082

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Volume 1 Long Papers* (pp. 715-725). Stroudsburg: Association for Computational Linguistics.

Ashwell, T. (2000). Patterns of teacher response to student writing in a multiple-draft composition classroom: Is content feedback followed by form feedback the best method?. *Journal of second language writing*, 9(3), 227-257.

Ayers, M. (2023). Human versus machine. *Journal of Clinical Engineering, 48*(3), 130-138. https://doi.org/10.1097/jce.0000000000000603

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of second language writing*, *17*(2), 102-118.

Bitchener, J., & Knoch, U. (2010). The contribution of written corrective feedback to language development: A ten-month investigation. *Applied linguistics*, *31*(2), 193-214.

Carter, A., & Absalom, M. (2023). Giving Students the Tools: Looking at Teaching and Learning using Corpora. *The EuroCALL Review*, *30*(1), 52-62.

Chan, K., Bond, T., & Yan, Z. (2022). Application of an automated essay scoring engine to English writing assessment using many-facet Rasch measurement. *Language Testing, 40*(1), 61-85. https://doi.org/10.1177/02655322221076025

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.

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023, July). Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 323-325). IEEE.

Dasgupta, T., Naskar, A., Saha, R., & Dey, L. (2018). Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 93-102). Stroudsburg: Association for Computational Linguistics.

Dikli S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment, 5*(1), 1-36.

Dong, F., & Zhang, Y. (2016). Automatic features for essay scoring—an empirical study. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1072-1077). Stroudsburg: Association for Computational Linguistics.

Dugan, L., Ippolito, D., Kirubarajan, A., Shi, S., & Callison-Burch, C. (2023). Real or fake text?: investigating human ability to detect boundaries between human-written and machine-generated text. *Proceedings of the Aaai Conference on Artificial Intelligence, 37*(11), 12763-12771. https://doi.org/10.1609/aaai.v37i11.26501

Ferris, D. R. (2002). Treatment of error in second language student writing. University of Michigan press.

Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Routledge.

Ferris, D. R., & Hedgcock, J. S. (2023). *Teaching L2 composition: Purpose, process, and practice*. Routledge.

Geckin, V. (2023). Assessing second-language academic writing: ai vs. human raters. *Journal of Educational Technology and Online Learning, 6*(4), 1096-1108. https://doi.org/10.31681/jetol.1336599

Goldstein, L. M. (2004). Questions and answers about teacher written commentary and student revision: Teachers and students working together. *Journal of second language writing*, *13*(1), 63-80.

Humphry, S., & Heldsinger, S. (2020). A two-stage method for obtaining reliable teacher assessments of writing. *Frontiers in Education, 5*. https://doi.org/10.3389/feduc.2020.00006

Hyland, K. (2009). Academic discourse: *English in a global context Continuum*.

Hyon, S. (1996). Genre in three traditions: Implications for ESL. *TESOL quarterly*, *30*(4), 693-722.

Jackaria, P. M., Hajan, B. H., & Mastul, A. R. H. (2024). A comparative analysis of the rating of college students' essays by ChatGPT versus human raters. *International Journal of Learning, Teaching and Educational Research*, *23*(2), 478-492.

Kim, Y., Kim, H., Park, G., Kim, S., Choi, S., & Lee, S. (2021). Reliability of machine and human examiners for detection of laryngeal penetration or aspiration in videofluoroscopic swallowing studies. *Journal of Clinical Medicine, 10*(12), 2681. https://doi.org/10.3390/jcm10122681

Kural, M., Jin, J., Fürbass, F., Perko, H., Qerama, E., Johnsen, B., … & Beniczky, S. (2022). Accurate identification of eeg recordings with interictal epileptiform discharges using a hybrid approach: artificial intelligence supervised by human experts. *Epilepsia, 63*(5), 1064-1073. https://doi.org/10.1111/epi.17206

Liu, S. and Kunnan, A. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: a case study of write to learn. *Calico Journal, 33*(1), 71-91. https://doi.org/10.1558/cj.v33i1.26380

Lloyd, S., Beckman, M., Pearl, D., Passonneau, R., Li, Z., & Wang, Z. (2022). Foundations for ai-assisted formative assessment feedback for short-answer tasks in large-enrollment classes.. https://doi.org/10.52041/iase.icots11.t3c3

Lui, S., & Kunnan, A. J. (2016). Investigating the application of automated writing assessment to Chinese undergraduate English majors: A case study of WriteToLearn. *Computer Assisted Language Instruction Consortium, 33*, 71-91.

Maftoon, P., & Seyyedrezaei, S. H. (2012). Good Language Learner: A Case Study of Writing Strategies. *Theory & Practice in Language Studies (TPLS)*, *2*(8).

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050.

Nguyen, T. (2023). Exploring the efficacy of ChatGPT in language teaching. *Asiacall Online Journal, 14*(2), 156-167. https://doi.org/10.54855/acoj.2314210

Pienemann, M. (1989). Is language teachable? Psycholinguistic experiments and hypotheses. *Applied linguistics*, *10*(1), 52-79.

Refaat, M. M., Ewees A. A., & Eisa, M. M. (2012). Automated assessment of students' Arabic free text answers. *International Journal of Intelligent Computing and Information Science, 12*(1), 213-222.

Rezaeian, M., Seyyedrezaei, S. H., Barani, G., & Seyyedrezaei, Z. S. (2020). An investigation into Iranian non-English PhD students' perceptions regarding learning as an educational consequence of EPT. *Iranian Journal of Learning and Memory*, *3*(10), 71-80.

Seyyedrezaei, S. H. (2015). Improving E-Assessment and E-Learning in Language Learning and Teaching Using Information Technology. In *Proceedings of International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE)* (p. 87). Sofia, Bulgaria.

Seyyedrezaeia, S. H., Kazemib, Y., & Shahhoseinic, F. (2016). Mobile Assisted Language Learning (MALL): An accelerator to Iranian language learners' vocabulary learning improvement. *International Journal of research in Linguistics, language Teaching and Testing*, *1*, 7-13.

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, 1260843.

Storch, N. (2010). Critical feedback on written corrective feedback research. *International Journal of English Studies*, *10*(2), 29-46.

Taghipour K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882-1891). Stroudsburg: Association for Computational Linguistics.

Wei, P. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: a randomized controlled trial. *Frontiers in Psychology, 14.* https://doi.org/10.3389/fpsyg.2023.1249991

Wong, W. S, & Bong, CH (2021). Assessing Malaysian University English Test (MUET) Essay on Language and Semantic Features Using Intelligent Essay Grader (IEG). *Pertanika Journal of Science & Technology, 29*(2), 919-941.

Wu, J., Li, Y., Zhou, J., & Chen, S. (2024). The impact of intelligent personal assistants on Mandarin second language learners: interaction process, acquisition of listening and speaking ability. *Computer Assisted Language Learning*, 1-26.

Zhang, L., Chen, J., Hou, L., Xu, Y., Liu, Z., Huang, S., … & Liang, L. (2022). Clinical application of artificial intelligence in longitudinal image analysis of bone age among GHD patients. *Frontiers in Pediatrics, 10.* https://doi.org/10.3389/fped.2022.986500

Zhang, Y., Zhang, J., & Zhang, X. (2021). Chinese IELTS test takers' perceptions of computer- mode IELTS: A mixed-methods study. *International Journal of Applied Linguistics and English Literature Studies, 7*(2), 46-58. Https://doi.org/10.3923/ijalelsv7n2p46

Zribi, R., & Smaoui, C. (2021). Automated versus Human Essay Scoring: A Comparative Study. *International Journal of Information Technology and Language Studies (IJITLS)*, *5*(1), 62-71. International Scientific Indexing (ISI). (2024). Author guidelines. https://isindexing.com/isi/detailsss.php?id=13820&page=guide