Academic Education Publishing House
Occupation and Professional Education Vol. 1 No. 9, 2024

Overview of Automated Evaluation Systems for Second Language Writing: A Study of Domestic and International Research

Qianxue Zhao
School of Humanities and Social Sciences, Xi'an Polytechnic University, Xi'an, Shaanxi, China

Abstract: **The first automated essay scoring system was developed 50 years ago. Automated writing evaluation systems are evolving into feature-rich platforms, moving beyond simple scoring mechanisms. This paper reviews and analyzes the current state of English writing automated evaluation systems through a survey of related literature and studies. It introduces common English writing evaluation systems and websites, explaining their basic principles, technical characteristics, and application directions. Additionally, it reviews recent case studies of automated scoring systems that provide feedback. This critical literature review aims to extract insights and suggests a dynamic combination of human and machine evaluation for second language writing assessment.**

Keywords: Automated Essay Scoring System; Second Language Writing; Research and Application

1. Introduction

The rapid advancement of information technology has ushered in transformative changes across various sectors, with education being one of the most impacted. In the specialized field of foreign language teaching, the integration of intelligent tools has gained considerable attention due to their potential to enhance learning outcomes and streamline teaching practices. Among these tools, Automated Essay Scoring (AES) systems stand out for their ability to automatically evaluate and provide feedback on students' writing through sophisticated algorithms. These systems, leveraging advancements in natural language processing and machine learning, have been shown to not only improve teaching efficiency but also to play a pivotal role in developing learners' writing abilities by offering timely and constructive feedback. While research and application of AES systems have reached a relatively advanced stage internationally, the exploration and adoption of these technologies in China remain in their early phases, facing challenges related to linguistic diversity and contextual adaptability.

This paper aims to provide a comprehensive review and analysis of AES-related literature, focusing on both international and domestic developments. To establish a foundation, we will begin by defining the key concepts and scope surrounding AES systems. Following this, an in-depth examination will cover the main components of AES research, including model design, feedback mechanisms, and applications within various educational contexts. A critical evaluation will highlight existing research gaps and assess the impact of AES systems across different educational stages and language backgrounds. By examining the integration of human and machine evaluation, this paper seeks to propose pathways for the further advancement of AES technology in second language writing instruction, ultimately aiming to bridge the gap between current global practices and China's evolving educational landscape.

2. Overview of Automated Evaluation Systems for Second Language Writing

2.1 Literature Review

2.1.1 Automated essay scoring systems

The development of automated English writing evaluation systems began in the 1960s in the United States with the creation of the Project Essay Grade (PEG) system, which marked the inception of Automated Essay Scoring (AES) systems. PEG, developed by Ellis Page and his team at Duke University in 1966, was trained by analyzing samples of manually scored essays, establishing a multiple regression equation based on language features and writing scores to evaluate essays[1]. The PEG scoring process is divided into two phases: the training phase and the evaluation phase. In the training phase, human scorers analyze and score 10 to 40 essays,

identifying 30 relevant variables. Using these variables as predictors, a regression prediction model is constructed, with Beta weights applied to each predictor variable. In the evaluation phase, each essay's variables are computed based on prior results, and using the weighted coefficients from the earlier phase, a regression prediction formula is applied to derive the essay's score. PEG has been used for exam scoring for many years; its fundamental concept is straightforward and computationally simple, with many studies demonstrating high correlation with human scoring.

The second phase, occurring in the 1990s, saw the emergence of three writing automated evaluation systems. E-rater utilizes statistical analysis, vector space modeling, and natural language processing techniques to assess the linguistic quality and structure of essays. IntelliMetric, as a machine learning system, simulates human thought processes, integrates expert intelligence, understands language, and evaluates essays according to English characteristics. Its scoring accuracy shows a 97% to 99% consistency rate with expert scores. IEA consumes relatively low computational resources and provides writers with quick feedback while detecting plagiarism.

The third phase, at the beginning of the 21st century, built upon the earlier scoring systems with the development of automated scoring systems like My Access!, Criterion, Writing Roadmap, and Holt Online Essay Scoring. Among these, My Access! and Criterion have been extensively studied. The core scoring concept of My Access! aligns closely with that of IntelliMetric, differing mainly by providing a more organized feedback and analysis report environment for students, which aids in improving writing skills. Table 1 provides a comprehensive comparison of automated evaluation systems for second language writing, highlighting each system's developer, key features, accuracy, and language support.

Table 1. Overview of the Automatic Evaluation System for L2 Writing

System	Developer/ Organization	Key Features	Accuracy	Language Support
PEG	Ellis Page, Duke University	Uses multi dimension regression to evaluate language features and identify factors affecting writing scores	Correlated with human scoring	--
E-rater	Jill Burstein, ETS	Statistical analysis, latent space modeling, natural language processing; evaluates quality and structure of essays	97% consistency (GMAT essays)	Primarily English
IntelliMetric	Vantage Learning	Al, natural language processing, and statistical analysis; analyze more than 30 features	97%-99% consistency	English,Spanish, Portuguese, Hindi, etc.
IEA	Thomas Landauer, University of Colorado	Based on Latent Semantic Analysis (LSA), quick feedback and plagiarism detection	85%-91% consistency (GMAT essays)	--
My Access!	Similar to IntelliMetric	Provides quick feedback on organizational structure and reports writing performance	--	--
Criterion	Integrates E-rater and Critique technology	Provides essay scoring and grammar error diagnostics	--	--

2.1.2 Model research

Since 2016, the availability and utilization of datasets in Automated Essay Scoring (AES) research have paved the way for significant advancements in model development, with deep neural networks quickly rising as the preferred approach. This shift has established deep learning as the mainstream research trend in AES, yet it has also brought to light certain gaps in model depth and refinement. While current models are effective to a certain extent, they often lack the detailed capacity to accurately analyze and interpret complex linguistic features, particularly in diverse educational settings. These limitations point to a need for continued innovation, involving more specialized models

and rigorous validation methods, to enhance AES models' adaptability and precision.

A notable contribution to this field is the framework proposed by Hussein and colleagues, where they adopted the model developed by Taghipour et al.[2] for training on the ASAP dataset. This foundational model, although straightforward and representative, comes with significant limitations, primarily due to its inability to incorporate more sophisticated architectures such as recurrent neural networks (RNNs) or transformers. These advanced models are particularly effective for handling intricate text structures and capturing subtle language patterns, which are essential in accurately assessing student writing. The reliance on a simpler model limits the ability to produce a detailed representation of student essays, resulting in a single evaluative score derived from a basic interpretation of the text, thus oversimplifying the evaluation. By integrating more advanced components, such as RNNs and transformer architectures, future AES models could achieve substantial improvements in analytical depth, interpretative accuracy, and overall robustness, moving closer to providing nuanced evaluations that better capture the complexity of student language and thought (see Figure1).

In the model by Mathias et al., trait scores were obtained using the model developed by Dong et al.[3]. They employed Convolutional Neural Networks (CNN) and attention layers to generate representation vectors for each word and sentence, achieving good performance in overall ASAP scoring (see Figure 2). However, while this approach shows promising results on the ASAP dataset, its generalization ability and adaptability to varied text types, writing styles, or genres remain unverified. Further research is necessary to confirm whether this model can maintain accuracy across diverse datasets and effectively evaluate texts with different structural or linguistic characteristics.

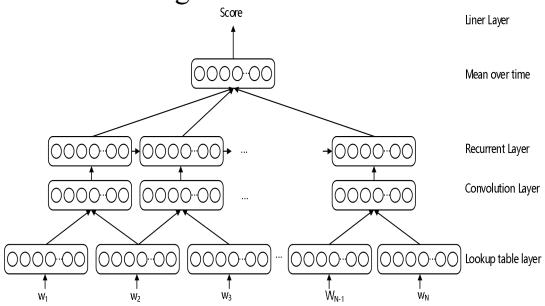

Figure 1. Model Framework

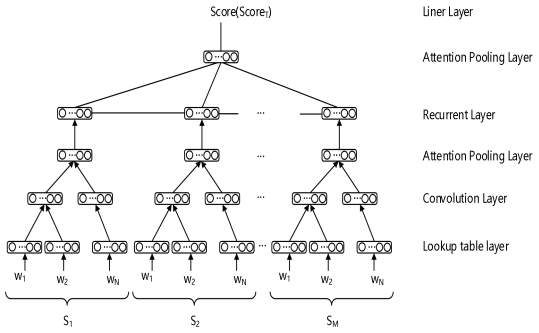

Figure 2. Model Framework

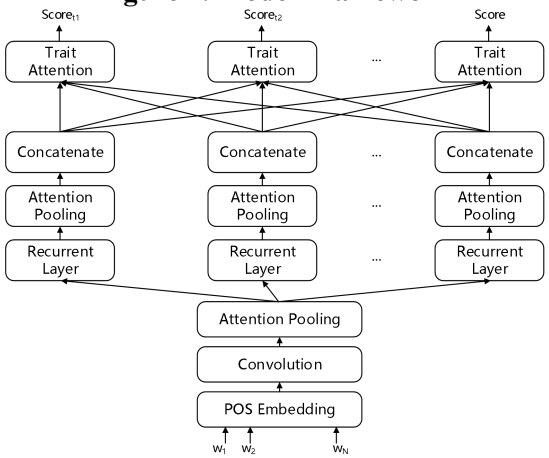

Figure 3. Model framework

In their model research, Ridley et al.[4] introduced modifications to enhance the feature-sharing structure, specifically adjusting the lower half of the model to improve shared representation. In contrast, the upper half of the model maintains independence in calculations, allowing each feature to be processed separately before any interaction. Following these independent computations, attention layers are added to capture the relationships between traits, enabling a more refined understanding of inter-feature connections. This combination of feature-sharing adjustments in the lower half with independent processing and attention-based integration in the upper half is considered a balanced and effective strategy for enhancing model interpretability and accuracy (see Figure 3).

In their model research, Ridley et al.[4] introduced modifications to enhance the feature-sharing structure, specifically adjusting the lower half of the model to improve shared representation. In contrast, the upper half of the model maintains independence in calculations, allowing each feature to be processed separately before any interaction. Following these independent computations, attention layers are added to capture the relationships between traits, enabling a more refined understanding of inter-

feature connections. This combination of feature-sharing adjustments in the lower half with independent processing and attention-based integration in the upper half is considered a balanced and effective strategy for enhancing model interpretability and accuracy (see Figure 3).

2.1.3 Application research

This study reviewed the existing literature on Automated Essay Scoring (AES) systems for second language writing by searching terms such as "AES," "Automated Essay Scoring," "Automated Essay Grading," "Automatic Evaluation Scoring," and "Automatic Essay" using Google Scholar, Web of Science, Wiley, Proquest, and ERIC. The application of AES systems in the field of educational technology has attracted widespread attention and research. The literature on feedback provided by AES generally shows that it can effectively assist students in writing. Studies indicate that AES can help students improve their writing errors by providing immediate and personalized feedback[5][6][7]. However, although AES is considered beneficial for enhancing students' writing skills, there is still a lack of research on the acceptance of feedback from these tools by students and the applicability of the tools in different cultural contexts[8]. Moreover, these studies often do not consider individual differences among students, such as how language proficiency and writing style affect the utility of the tools[9]. Therefore, future research needs to explore how these factors influence the educational effectiveness of AWE tools.

Research in language testing mainly focuses on the consistency and accuracy of scoring systems. For example, Yao[10] found that the Pigai automated scoring system has moderate correlations with human scoring in certain aspects, particularly in identifying punctuation and grammatical errors, but performs inadequately with more complex language structures, such as pronoun usage. Additionally, Huang and Wilson[11], through a long-term tracking study, examined how AWE tools affect students' English writing skills, finding that this type of feedback can improve students' writing skills over the long term. These studies reveal the potential and limitations of automated scoring tools in practical applications, highlighting that scoring systems need further development to meet diverse testing requirements and cultural backgrounds.

In the educational context, AES systems are widely used across different educational stages, with higher usage rates among non-native speakers. Studies by Yao[10]and Huang and Wilson[11] emphasize the broad applicability of AES systems and their role in enhancing the writing skills of non-native speakers. Furthermore, studies exploring how technological tools support student learning demonstrate the positive impact of these tools on improving students' self-regulated learning abilities and overall learning outcomes. Research indicates that using learning management systems and AWE tools can significantly enhance students' self-regulation in learning and writing skills[12]. However, although these technological tools have been shown to aid educational practice, studies often fail to adequately explore how to effectively integrate these tools into instructional design to meet diverse educational needs[13]. Therefore, future research needs to focus more on the long-term effects of using technological tools and their impact on different student groups. These findings collectively highlight the significant value of AES systems in education, particularly in providing writing support for non-native English learners. This section reviewed 14 studies that investigated the application research of different AWE systems. Table 2 shows these studies.

Table 2. Application Research on Writing AES System

Content	Quantity	Research Review
Automatic Feedback	7	Alharbi, 2023 Liu et al., 2024 Zhai & Ma, 2021 Zhao, Li & Feng, 2023 Yao, 2024 Mohsen & Alshahrani, 2019 Lee, 2020
Writing Test	2	Yao, 2024 Huang & Wilson, 2021
Education Field	9	Alharbi, 2023 Wang et al., 2024 Ke, Carlile, Gurrapadi & Ng, 2018 Ngo, Chen, & Lai, 2022 Wei, Wang & Dong, 2023 Wilson & Roscoe, 2020 Xu & Zhang, 2022 Mohsen & Alshahrani, 2019 Lee, 2020

2.2 Review of Domestic Research

Compared to international research, studies on AES systems in China are relatively lagging. Chinese scholars have generally focused on the following areas: validity and reliability studies of the systems themselves[14][15]. Li and Tian[16] conducted an empirical analysis of the scoring reliability of the iWrite 2.0 English writing teaching and assessment system, comparing manual scores with machine scores across four dimensions: language, content, structure, and technical standards, to explore scoring consistency. Bai[17] assessed the scoring validity of Pigai in formative assessments, finding it has high generalization and inferential validity for scoring reading reflections but lower validity for narrative writing, recommending caution when using Pigai for creative texts.

Secondly, the impact of automated writing evaluation systems on English writing. Hu[18] found that multiple self-revisions by students using online automated essay scoring systems can significantly improve essay scores, although the number of revisions is not correlated with the improvement in scores. Cao[19] explored the impact of Pigai feedback on the syntactic complexity of writing by non-English major students, finding that system feedback can positively enhance the syntactic complexity of second language learners, with significant differences in syntactic complexity observed among students of different language proficiency levels. Li[20], through text analysis and interviews, investigated how the diverse feedback from AES systems affects the quantity, type, and effectiveness of students' essay revisions, finding that this diverse feedback positively influences students' revisions.

Research on the feedback nature of AES systems in writing. Li[21], through a semester-long survey and analysis, explored the acceptance of teacher electronic feedback by college students of different English proficiency levels, finding that teacher feedback tends to focus on content for high-level students and grammar for low-level students, with low-level students generally showing higher acceptance of feedback than high-level students. Li and Wang[22] explored strategies to improve feedback effectiveness in teaching, including integrating different types of feedback, considering individual learner factors, and adopting diverse feedback modes.

3. Conclusion

While AES systems are widely used globally, their acceptance and implementation effectiveness vary significantly across countries and regions. These differences are mainly driven by educational policies, technological infrastructure, the technological proficiency of teachers and students, and cultural attitudes toward educational technology. In developed countries such as the United States and the United Kingdom, AES systems have developed and been applied rapidly. These countries generally possess strong technological infrastructures and high investment in educational technology. For example, in the United States, many schools and universities have integrated AES systems like Turnitin and Grammarly into daily teaching activities to support academic integrity and the development of writing skills. These systems are not only used for scoring but also for providing feedback and guidance on students' writing development. In contrast, in many developing countries, the application of AES systems faces more challenges. The lack of adequate technological infrastructure and uneven distribution of educational resources have hindered the widespread adoption and effective application of AES systems. Furthermore, the acceptance of new technologies and technological proficiency among teachers and students may also be lower. These factors limit the potential of AES systems in educational reform and academic assessment.

Through comparative analysis of AES applications across different countries and regions, it is evident that AES systems offer valuable tools for enhancing writing instruction and assessment. However, their success relies on various factors, including technological infrastructure, supportive educational policies, cultural acceptance, and adequate economic backing. Effective adaptation of these systems to diverse educational settings requires a careful balance of these elements to fully harness their potential. Future research should delve deeper into strategies for optimizing AES in varied educational environments, addressing unique local challenges, and ensuring these systems can achieve widespread educational impact globally.

References

[1] Page, E.B. Grading essays by computer: Progress report. In Educational Testing Service (Ed.), Proceedings of the Invitational Conference on Testing

Problems. New York City: Princeton, NJ: Educational Testing Service, 1966:87 -10.
[2] Taghipour, K., & Ng, H.T. A neural approach to automated essay scoring. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 1882-1891.
[3] Dong, F., Zhang, Y., & Yang, J. Attention-based recurrent convolutional neural network for automatic essay scoring. Proceedings of the 21st Conference on Computational Natural Language Learning (CONLL 2017), 2017: 153-162.
[4] Ridley, R., He, L., Dai, X., Huang, S., & Chen, J. Automated cross-prompt scoring of essay traits. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35: 13745-13753.
[5] ALharbi, W. AI in the foreign language classroom: A pedagogical overview of automated writing assistance tools. Education Research International, 2023: 1-15.
[6] Wei, P., Wang, X., & Dong, H. The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: a randomized controlled trial. Frontiers in Psychology, 2023, 14: 1249991.
[7] Zhao, J., Li, Y., & Feng, W. Investigating the validity and reliability of a comprehensive essay evaluation model of integrating manual feedback and intelligent assistance. International Journal of Emerging Technologies in Learning (Online), 2023, 18(4): 248-261.
[8] Mohsen, M. A., & Abdulaziz, A. The effectiveness of using a hybrid mode of automated writing evaluation system on EFL students' writing. Teaching English with Technology, 2019, 19(1): 118-131.
[9] Burstein, J. The E-rater scoring engine: Automated essay scoring with natural language processing. Shermis, M. D., & Burstein, I. Automated essay scoring: a cross-disciplinary perspective. Mahwah, NJ: Lawrence Erlbaum Associates, 2003.
[10] Yao, D. Automated writing evaluation for ESL learners: A case study of Pigai system. Journal of Asia TEFL, 2021, 18(3): 949-958.
[11] Huang, Y., & Wilson, J. Using automated feedback to develop writing proficiency. Computers and Composition, 2021, 62.
[12] Wang, I. X., Wu, X., Coates, E., Zeng, M., Kuang, J., Liu, S., ... & Park, J. Neural Automated Writing Evaluation with Corrective Feedback. arXiv preprint arXiv:2402.17613, 2024.
[13] Xu, J., & Zhang, S. Understanding AWE feedback and English writing of learners with different proficiency levels in an EFL classroom: A sociocultural perspective. The Asia-Pacific Education Researcher, 2022, 31(4): 357-367.
[14] Ke, Z., Carlile, W., Gurrapadi, N., & Ng, V. Learning to Give Feedback: Modeling Attributes Affecting Argument Persuasiveness in Student Essays. IJCAI, 2018: 4130-4136.
[15] Ngo, T. T. N., Chen, H. H. J., & Lai, K. K. W. The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. Interactive Learning Environments, 2022: 1–18.
[16] Li Yanling, Tian Xiachun. Reliability Study of iWrite 2.0 Online English Composition Scoring. Modern Educational Technology, 2018, 28(2): 75-80.
[17] Bai Lifang. Validity Study and Usage Suggestions of Automated Scoring Systems in Formative Testing. Contemporary Foreign Language Education, 2023(0): 17-28.
[18] Hu Xuewen. The Impact of Online Composition Self-Revision on College Students' English Writing Results. Foreign Language Educational Technology, 2015(3): 45-49.
[19] Cao Ting. The Impact of Feedback from Automated Composition Scoring Systems on the Syntactic Complexity of Students at Different Levels: A Case Study of Pigai. Examination and Evaluation (College English Research Edition), 2018(5): 94-101.
[20] Li Guangfeng. A Study on the Impact of Multiple Feedback Based on Automated Evaluation Systems on English Composition Revision. Foreign Language Teaching, 2019, 40(4): 72-76.
[21] Li Yihua. An Investigative Analysis of Learners' Acceptance of Teachers' Electronic Feedback in English Writing. Language Education, 2016, 4(1): 34-41.
[22] Li Guangfeng, Wang Nannan. Research on Feedback Effectiveness: Influencing Factors and Teaching Strategies. Theory and Practice of Education, 2018, 38(19): 57-60.