PASAA Journal
Volume 70, January – June 2025, 340–370

PASAA

AI-assisted L2 Assessment: A Biblio-Systematic Analysis

Ecem Kopuz[a]*, Galip Kartal[b]

[a] City University of New York, The Graduate Center, Speech-Language-Hearing Sciences, New York, USA

[b] Necmettin Erbakan University, English Language Teaching, Konya, Türkiye

**Corresponding author: ekopuz@gradcenter.cuny.edu*

Article information	
Abstract	The developments in artificial intelligence (AI) have significantly transformed second language (L2) learning and assessment, and the role of AI technologies in L2 assessment have been investigated in recent research. This study presents a biblio-systematic analysis of AI-assisted L2 assessment. Using both systematic analysis and bibliometric research approaches, the study analyzed 57 SSCI-indexed articles to address participants, research methods, research foci, AI technologies employed, as well as the effectiveness, advantages, and challenges of AI in L2 assessment. Furthermore, bibliometric analysis was conducted via co-occurrence and co-citations analyses using VOSviewer. Findings have indicated that AI tools, such as automated scoring systems and natural language processing technologies, are predominantly used in writing and speaking assessments. These tools offer personalized feedback, enhance learner motivation, and provide scalable solutions for large-scale evaluations. Despite the positive impact on engagement and efficiency, challenges remain, including technical limitations, data privacy concerns, and the need for more balanced datasets. The study also highlights the intellectual foundations of the field, mapping key authors and influential papers. This study contributes to the growing body of literature by offering a biblio-systematic analysis of AI's role in L2 assessment and identifying areas for future investigation.

E-ISSN: 2287-0024

Keywords	artificial intelligence (AI), L2 assessment, bibliometric analysis, co-citation
APA citation:	Kopuz, E. & Kartal, G. (2025). AI-assisted L2 Assessment: A Biblio-Systematic Analysis. *PASAA, 70,* 340–370.

1. Introduction

The emergence of generative AI (GenAI) technology increased the interest in its application to L2 assessment (Ding & Zou, 2024). With this technology, evaluations are now more precise and adaptive to individual learners' needs. AI-enhanced assessments change the degree of the challenge in response to the learner's performance, providing different levels of difficulty based on the current level of achievement. The GenAI supported tools help generate learner-centered materials, support personalized instruction, and provide individualized feedback (Bonner et al., 2023; Caines et al., 2023). This aspect of flexibly is rather useful in L2 learning since the students are at different proficiency levels.

To better evaluate the state of research on AI-supported second language assessment, this study adopted a biblio-systematic approach. Such an analysis provides insights into major research trends, influential studies, and emerging topics in the field. Examining publication patterns helps identify shifts in research paradigms and the frequency of studies on AI-assisted second language assessment. The approach used in this study also addressed the dominant aspects of AI in L2 assessment and its evolving trajectory. Understanding the current landscape and anticipated developments in the field allows researchers to further advance AI-assisted L2 assessment. Through systematic analysis, this study aimed to identify overarching trends, frequently used key terms, and thematic shifts in the research, ultimately offering a comprehensive overview of the field's progression.

1.1 AI and L2 Assessment

The rapid advancements in AI-assisted L2 assessment have undeniably transformed the field, reshaping assessment methods and enhancing assessment accuracy and efficiency. By leveraging advanced algorithms in GenAI, machine learning (ML) and natural language processing (NLP), L2 assessment has become more efficient, precise, and adaptive (Ding & Zou, 2024). AI-driven systems offer automated scoring, automated writing evaluation (AWE), and AI-driven adaptive feedback technologies in computer-assisted language learning (CALL), significantly improving the accuracy and personalization of language assessments (Kenshinbay & Ghorbandordinejad, 2024). AI-driven systems also minimize subjectivity, providing more accurate and adaptable evaluations while offering detailed insights into learners' language skills (Wei, 2023). AI's ability to process vast amounts of linguistic data enables more effective language proficiency analysis and error correction, allowing learners to receive immediate, context-aware feedback. Automated feedback systems not only demonstrate linguistic errors in real time but also offer adaptive metalinguistic explanations that promote self-directed learning, ultimately improving L2 writing accuracy (Barrot, 2021). Moreover, modern AI-powered language assessment tools further support personalized learning by dynamically adjusting assessment complexity based on learners' proficiency levels, making the evaluation process more interactive and data-driven (Caines et al., 2023).

In recent years, there has been growing interest in the way AI tools are included into language learning and evaluation. Studies demonstrating their ability to improve learner self-assessment, enable automated grading, and raise the quality of comments have received increasing attention. Algaraady and Mahyoob (2023), for example, investigated ChatGPT's ability to spot and evaluate writing mistakes among EFL students, therefore confirming its value in raising writing correctness and learning results. Song and Song (2023) also looked at how AI-assisted language learning might affect students' writing. Research on the use of AI in writing evaluation is quite widespread. Examining the effects of peer feedback supported by AI on university students' writing quality, Guo et al. (2024) found that AI-driven interventions greatly improved writing skills. Moreover, Evaluating ChatGPT's dependability as an automated essay grading method, Bui and Barrot (2024) found that its scores matched those of human assessors, therefore verifying its relevance in the classroom. Nazari et al. (2021) underlined even more the need of AI-driven writing aids in enhancing correctness and involvement in higher education. When reviewing Automated Writing Evaluation (AWE) systems including Grammarly, Pigai, and Criteria, it could be seen that these tools offer instantaneous feedback and point out areas needing improvement (Ding & Zou, 2024). It is also evident in the literature that AI helps improve academic writing and encourages introspection. According to Liu et al. (2024), AI tools inspire reflective thinking in academic communication, raising cognitive engagement and the quality of the work produced. Similarly, Shen et al. (2023) showed that AI-generated feedback improved student engagement and performance in writing activities, implying its superiority over conventional approaches in some circumstances.

AI-assisted L2 assessments now extend beyond written evaluations, offering accurate analysis of spoken responses, including grammar, coherence, word choice, and even accent. Language assessments driven by machine learning can be developed quickly while maintaining validity, reliability, and security, aligning with high-stakes English exams (Settles et al., 2020). Notably, high-stakes tests such as the Pearson Test of English (PTE) and the Duolingo English Test (DET) utilize these AI-driven technologies to deliver prompt, comprehensive, and highly accurate evaluations. As AI continues to evolve, its role in L2 assessment will likely expand, further refining the efficiency, reliability, and accessibility of language evaluation frameworks. These advancements position AI as a transformative tool in language assessment, making evaluation processes more precise, responsive, and learner focused. Moreover, AI tools can process enormous amounts of linguistic data and find patterns, including typical learner mistakes. Currently, this datafocused approach helps educators create more effective and precise assessments, learning tools, and resources. It also provides opportunities for interaction practice and integrates realistic communicative language use that tests both accuracy and communicative competence.

AI tools have shown promise in assessing oral proficiency during speaking tests. According to Jin and Fan (2023), higher degrees of involvement with AI systems lead to more accurate competence evaluations, due to increased test-taker participation in AI-mediated assessments. Emphasizing AI's ability to analyze spontaneous speech with human-level accuracy, Al-Ghezi et al. (2023) examined automatic speaking assessment systems for Finnish and Swedish L2 learners. Nonetheless, as Voss et al. (2023) underlined—who explored the ethical consequences of AI in language testing and the need of openness in AI-driven assessments—questions about justice and authenticity still exist.

1.2 The Present Study

Although AI improves language learning and assessment—especially in providing instantaneous, regimented feedback and supporting learner autonomy—in-depth reviews are limited. There are some meta-analyses (e.g., Huang et al., 2024), systematic reviews (e.g., Gao et al., 2024; Meniado, 2023), technology reviews (e.g., Osawa, 2023), and a bibliometric analysis of general English language assessment (Yang & Wang, 2025). However, there appears to be a lack of a comprehensive mapping of only AI-assisted L2 assessment research, suggesting the need for both a systematic review and a bibliometric analysis. These two methods complement each other. A systematic analysis provides a structured assessment of research themes, methodologies, and findings (Roa & Halim, 2024), while a bibliometric analysis provides a picture of the field with maps (Kartal & Yeşilyurt, 2024). In addressing this gap, the current study aimed to seek answers to the following research questions:

1. Who were the participants in the AI-assisted L2 assessment research?
2. What were the research methods in the AI-assisted L2 assessment research?

3. What were the research foci and what were the AI technologies adopted for them?
4. How effective were the AI tools, and what were the advantages and challenges in the AI-assisted L2 assessment research?
5. What were the key authors, references, and sources (bibliometric indicators) that have contributed significantly to AI-assisted L2 assessment research?

2. Methodology

The present research used both systematic and bibliometric analyses to analyze AI-assisted L2 assessment. The former involved 57 research articles whose contents were extracted and analyzed, as suggested by Roberts et al. (2017). This approach yielded data regarding research foci, types of research, sample, AI tools, benefits derived from using AI, and challenges encountered in AI-assisted L2 assessment. The latter aimed at discovering emerging trends and the intellectual structure in AI-assisted L2 assessment. VOSviewer software was used to perform two bibliometric analyses, namely co-citation and co-occurrence analysis (van Eck & Waltman, 2023). The co-citation analysis focused on cited references and sources to identify seminal works as well as key publication outlets, thereby shedding light on the intellectual foundation and scholarly standing of AI-assisted L2 assessment research. Second, a keyword co-occurrence analysis was done to shed light on patterns and relationships among commonly mentioned terms or constructs in the field. This visualization represented the scientific maps of AI-assisted L2 assessment research, as it illustrated semantic relationships and approaches among the components.

2.1 Data Collection

The dataset consisted of journal articles extracted from the Social Sciences Citation Index (SSCI), one of several databases included in the Web of Science Core Collection. This selection was made for two predominant reasons. The first was to meet the wider objective of repositioning WoS as a leading global resource for scientific citation search, discovery, and analytical information (Li et al., 2018). The second, more topic-specific reason, was that SSCI allowed for the most encompassing view of available research articles published by high-ranking academic journals (Duman et al., 2014). Table 1 shows the search terms used in WoS.

Table 1

Search terms in Web of Science

Category	Search Terms
Assessment	assessment, e-assessment, evaluation, test, rubric, washback, backwash, exam, exams, grading, feedback, portfolio, e-portfolio
AI	artificial intelligence, AI, machine intelligence, artificial neural network, machine learning, deep learning, natural language processing, robotics, thinking computer systems, evolutionary computation, hybrid intelligent system, expert system, intelligent tutoring system, intelligent agent, ChatGPT, chatbot, advanced language model, Generative Pre-trained Transformer, GPT
L2 Research	foreign language learning, second language learning, foreign language teaching, second language teaching, second language acquisition, second language assessment, EFL, ESL, ELT, L2, TEFL, TESL, second language, foreign language, applied linguistics, TESOL, IELTS, TOEFL

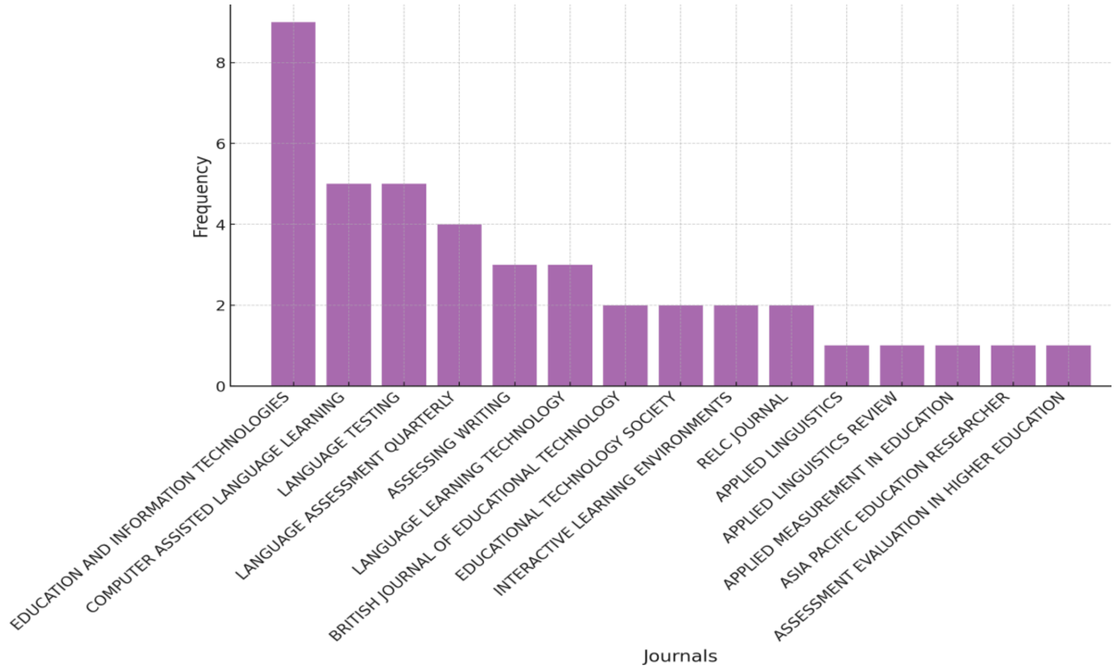
The directed search described in Table 1 generated an initial number of 2,169 results, which were gathered in July 2024. After filtering only for articles, the number decreased to 1,471. Then, choosing five fields (Educational Research, Linguistics, Language & Linguistics, Educational Psychology, and Education & Educational Research), the number went down to 386. Further filtering for SSCI-indexed articles reduced the total number to 250. Of those, narrowing down to

English-only articles resulted in 245 entries. After a manual review by the researchers, 57 articles specifically on L2 AI-enhanced assessment remained.

Documents other than journal articles (e.g., books, book chapters, and conference proceedings) were excluded because journals and their articles play a crucial role in the dissemination of L2 research (Plonsky & Derrick, 2016). The categories were selected in line with the objectives of the present study. Since L2 teaching and applied linguistics were addressed, the categories were limited to the above five categories. The top 15 journals based on the number of publications are shown in Figure 1.

Figure 1

The top 15 Journals



2.2 Coding Scheme for the Systematic Analysis

The coding scheme utilized Li's (2022) study to systematically analyze the trends and impacts of AI in second language assessment. It focused on understanding participant demographics, research methodologies, AI applications, and the overall effectiveness of AI. Regarding the participants' information, the

AI-assisted L2 assessment research was categorized into target languages, educational levels, and sample sizes. Examples of the educational levels included, but were not limited to, elementary school, secondary school, higher education, and adult education. Sample sizes included small, medium, medium to large, large, and not specified. Regarding the research methods used in AI-assisted L2 assessment research, both general and specific categorizations were identified. The general research methods included quantitative, qualitative, mixed-methods, and systematic review. Specific research methods focused on research design such as surveys, experimental designs, action research, case studies, interviews, and system development or evaluation. The foci of the research were divided into two main categories: assessment focus and learner perception.

As for assessment foci, the language skills or areas targeted in the AI-assisted assessment research were listening, speaking, reading, writing, and general language proficiency. Regarding the perception of the learners, research captured participants' attitudes toward AI-assisted L2 assessments in terms of satisfaction, motivation, anxiety, self-regulation, autonomy, and perceived usefulness. The adopted AI technologies were classified into the type of AI tool or system used in L2 assessments. These included machine learning algorithms, NLP tools, automated scoring systems, intelligent tutoring systems, speech recognition technologies, virtual platforms, adaptive testing systems, and other unspecified technologies. The effectiveness of tools, advantages, and challenges of AI-assisted L2 assessments were measured through the results, discussions, and other relevant sections of the selected articles. The effectiveness was coded as positive, negative, neutral, or unspecified to represent the impact of AI on the outcomes of assessments. Advantages were summarized as efficiency, scalability, immediate feedback, and personalization.

2.3 Bibliometric Analysis

Extensive cleaning on the data was conducted, which involved removing duplicates, rectifying errors, and standardizing the presentation of data. For example, consolidating "artificial intelligence" and "AI" under the single abbreviation "AI." Subsequently, we used VOSviewer to plot conclusions that would map to current research topics, trends, and patterns (van Eck & Waltman, 2023). In this vein, VOSviewer was broadly employed in the visualization of co-citation maps, network diagrams, and other visual patterns that exhibited interactions within different keywords and concepts in the literature, using visual indicators like circles or lines in two or three dimensions (Markscheffel & Schröter, 2021). We generated a keyword map of terms and ideas interrelated within the field of language assessment using VOSviewer. This helped identify keywords and their changes over time, giving a grasp of the topics and trends in the literature.

3. Findings

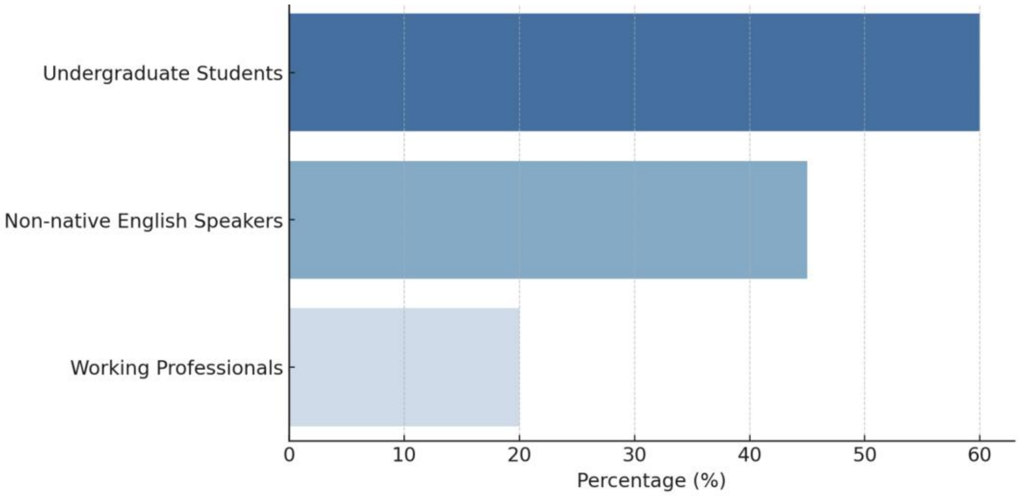
Findings are organized in line with the systematic review and bibliometric analyses conducted in the study. In other words, participants, research methods, and AI technologies are presented first. Then, bibliometric indicators are given. Finally, interpretations of the findings are presented after each section.

3.1 Participants in AI-assisted L2 Assessment Research

The sample sizes in the reviewed studies varied widely, ranging from small to large samples, reflecting the diverse research scopes and contexts. This variability underscored the broad applicability of AI technologies in evaluating L2 proficiency across different populations and settings. Analysis of the participants is illustrated in Figure 2.

Figure 2

Distribution of the Participants



As the figure shows, undergraduate students made up approximately 60% of the participants, with around 5,070 involved, predominantly from humanities, social sciences, and engineering disciplines. The prominence of this group was likely due to the accessibility of university populations and the importance of L2 proficiency in higher education. Non-native English speakers represented about 45% of the total sample, including over 3,800 participants from varied linguistic backgrounds such as Chinese, Spanish, Arabic, and French, with proficiency levels ranging from beginner to advanced. This substantial inclusion emphasized the global focus on English as a lingua franca and the need for effective AI-based assessment tools.

About 1,200 working professionals—from sectors like business, healthcare, and technology—made up roughly 20% of the study total. Professionals were the focus, as the use of AI-assisted L2 assessments beyond academic settings—such as for professional advancement and workplace communication—was attracting increasing interest. Most of these studies examined how AI may support continuous language learning. Many studies applied innovative AI techniques, including machine learning and natural language processing, to analyze trends in language ability, looking at more than 10,000 language samples gathered from online platforms and standardized tests. Emphasizing adult learners, the participants—with a mean age of 24—had ages ranging from 18 to 55. Reflecting an inclusive study approach to gathering various student experiences in L2 proficiency, the sample was balanced in gender distribution, with women making up 52% and men 48%.

3.2 Research Methods in AI-assisted L2 Assessment Research

The systematic review of AI-assisted L2 assessment research methods revealed a diverse range of approaches that reflected the field's complexity and adaptability (Figure 3).

Figure 3

The Research Methods

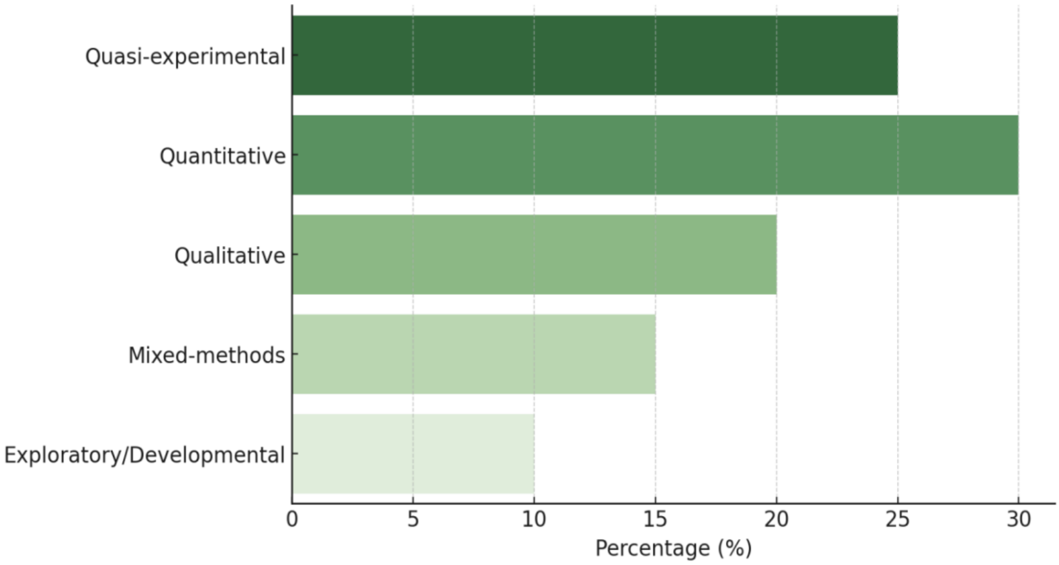


Figure 3 shows that quasi-experimental designs, such as pretest-posttest studies and randomized control trials (RCTs), were prominent in 25% of the studies, underscoring a rigorous evaluation framework for AI tools on language learning outcomes. Quantitative methods appeared in 30% of the studies, focusing on reliability, validity, and the comparison between AI-driven and traditional assessments, a trend supported by research on the statistical rigor and assessment reliability in AI-augmented educational environments. Qualitative approaches were used in 20% of the studies, employing case studies and reflective discussions to explore participants' experiences with AI tools, reflecting that qualitative insights were crucial for understanding the nuances of AI's impact on language learning. Mixed-methods, combining qualitative and quantitative techniques, were employed in 15% of the studies, integrating advanced statistical methods and NLP tools. In general, this methodological approach is widely recognized for its ability to provide a comprehensive understanding of educational outcomes, as noted in studies that leveraged mixed methods to examine the multifaceted impacts of AI in language instruction. Finally, exploratory and developmental research, comprising 10% of the studies, utilized design-based research (DBR) to innovate and refine AI assessment tools.

3.3. Research Foci and Adopted AI Technologies

Our analysis revealed that AI was used to assess language skills and other language related areas. The findings on the research foci and AI technologies are summarized in Table 2.

Table 2

Research foci and the AI tools

Research Foci	Specific AI tool
Writing	AVA, Eva, ChatGPT, TAALED, TAALES, TAASSC, Google Bard, Pigai, Criterion, DeepL Translator, Grammarly, Notion AI, Wordtune, XLM-RoBERTa, NLTK, SpaCy, EssayCritic, e-rater®
Automated item generation	ChatGPT, Perplexity AI, genQue
Anxiety	Pigai, Criterion
Motivation	Pigai, Criterion®
Grammar	Google Dialogflow
Feedback	Custom Chatbot (built using Chatfuel), Notion AI, Pigai, Wordtune, Grammarly, EssayCritic

Research Foci	Specific AI tool
Speaking	Wav2Vec2, Alexa, Google Speech-to-Text, SpeechRaterSM, Coh-Metrix, TAALES, LIWC
Vocabulary	NLTK, SpaCy, NLTK, SpaCy
Reading	BERT (Spanish, Multilingual, English), Skip-Thought
Cognitive and emotional engagement	LAIX

The analyses showed that writing assessment was one of the most frequently explored topics. AI assessment tools such as AVA, Eva, ChatGPT, TAALED, TAALES, TAASSC, and Google Dialogflow were utilized to gauge or enhance L2 learners' writing capabilities. These tools furnished instant, automated feedback on key dimensions of writing like lexical diversity, complexity, and coherence. For the most part, L2 writing assessment seemed to have found a strong ally in AI. A secondary focus was on automated item generation for language assessments, where AI tools such as ChatGPT, Perplexity AI, and genQue were employed. These tools assisted with the generation of diverse, adaptive, and responsive language test items that were finely tuned to the performance of the learners. This means that the test items were aligned well with the learners and that any scale established for the item also served as an appropriate scale for them. The use of these tools, as with any test generation tool, served to improve the scalability and reliability of language assessments.

Another area of emphasis in L2 learning and testing environments was anxiety. AI tools such as Pigai and Criterion were found to assess and address learner anxiety. These tools provided personalized feedback and real-time guidance. These systems' immediate, non-judgmental feedback fostered a less stressful learning experience. Regarding motivation, tools such as Pigai and Criterion also played a significant role. They kept learners engaged and motivated to improve themselves by supplying continuous, real-time feedback. The capability of AI to deliver individualized learning pathways ensured that students received relevant feedback that directly addressed their learning needs, thereby maintaining their motivation throughout the assessment process. Finally, for grammar assessment, Google Dialogflow was utilized to detect and correct grammatical errors. This tool provided adaptive feedback that aideds learners in improving their grammatical accuracy over time.

3.4 Tool Effectiveness, Advantages, and Challenges

Nearly half of the studies reported positive outcomes which included improved feedback quality, enhanced student engagement, and support for continuous learning. Only a small portion of the studies noted negative effects, such as inconsistent scoring and the limited ability of AI tools to capture the complexity of language use. However, some studies did not specify the effectiveness of AI. Regarding advantages, AI tools frequently provided immediate feedback, significantly enhancing student engagement and supporting ongoing learning. Other noted benefits included improved writing abilities, increased learner autonomy, and the capacity to facilitate multiple revisions using comprehensive linguistic resources.

Despite these benefits, several challenges were identified. Technical limitations of AI tools, the need for more balanced datasets to ensure accurate generalization across different learner contexts, and concerns about the potential reduction in critical thinking and writing autonomy due to over-reliance on AI were highlighted. Additionally, methodological issues such as short experimental durations and small sample sizes were noted, along with the limited applicability of some AI tools to broader contexts. AI-assisted language assessments have demonstrated significant potential, but addressing the highlighted challenges is crucial for maximizing their effectiveness in language education.

3.5 Bibliometric Indicators

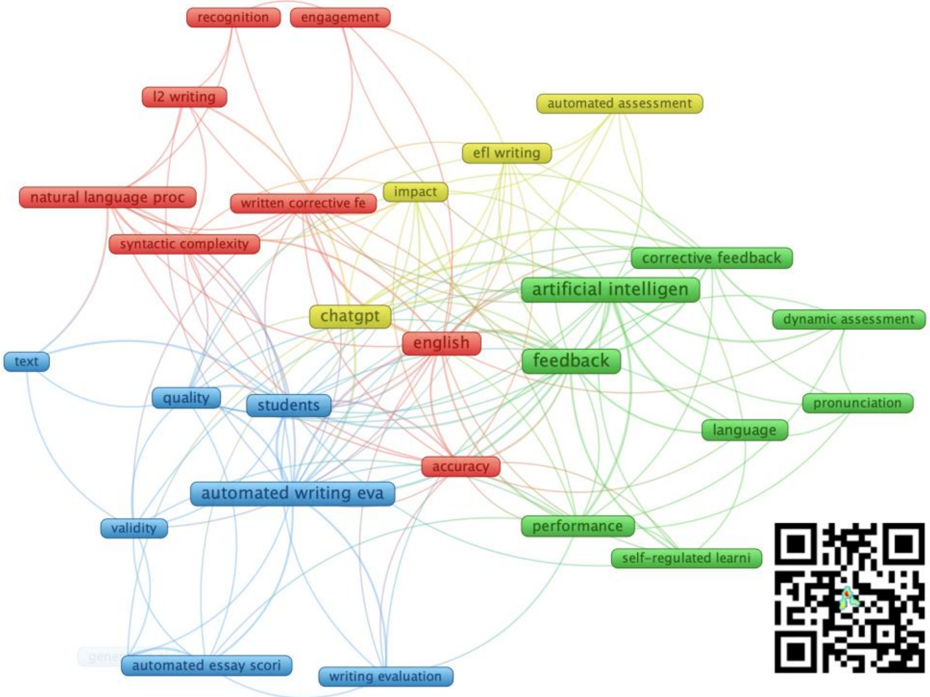
Co-occurrence and co-citations analyses were conducted using VOSviewer. While the co-occurrence analysis focused on identifying research topics and trends in the field, the co-citations analysis uncovered influential authors and articles.

3.5.1 Co-occurrence Analysis

The co-occurrence of the keywords is illustrated in Figure 4. The minimum number of occurrences of a keyword was set to three, resulting in 28 keywords and four clusters.

Figure 4

Co-occurrence Network Map



Note: Scan the QR code to zoom in on the map for a detailed view of all keywords, occurrences, links, and total link strengths.

The figure demonstrates the essential topics and trends that defined the field of research on AI-assisted L2 evaluation. The topics included in this study were the prominent role of AI technologies in delivering feedback (green cluster), the focus on automated writing assessment (blue cluster), the growing significance of sophisticated AI models such as ChatGPT (yellow cluster), and the continuous investigation of learner engagement and perceptions (red cluster).

The green cluster brought to the fore the critical role of AI in giving feedback. An increased interest in AI-based feedback systems, therefore, reflected the growing importance of AI in delivering feedback, not only in language learning in general but especially in L2 contexts. They helped enhance L2 learning outcomes through the delivery of personalized, quick, and adaptive feedback, which was in line with the current research trend, which increasingly employs AI to improve the accuracy, efficiency, and effectiveness of language assessment.

AWCF is more effective than traditional techniques in enabling learners to locate and correct their errors, due to the comprehensive and timely feedback it provides. Writing evaluation AI technologies, therefore, provide learners with enhanced feedback on grammar, syntax, and coherence, which has led to quantifiable improvements in their writing proficiency. The other proposals mentioned include AI for dynamic assessment, which contains assessment and training with computer-based feedback adapted to the learner's current level. Studies analyzed in the present study that compared dynamic assessment to conventional approaches of feedback found that AI-supported dynamic assessment led to better process writing outcomes, especially in the facilitation of learner development due to scaffolded support. AI-driven feedback mechanisms were widely recognized as a potentially integral part of language learning. They boasted several advantages, including providing high-quality, adaptive, and immediate feedback. This feedback helped learners get rid of their errors and thus improved their L2 writing.

Another noteworthy result (blue cluster) was the prevalence of "automated writing evaluation." Such a finding highlighted the increasing use of AI technology to assess writing assignments In fact, the emphasis on writing evaluation was remarkable since it tackled one of the most difficult areas of language assessment: delivering consistent and unbiased assessments of written outputs. AI solutions demonstrated their efficacy in this aspect, as AI integration in writing assessment facilitated learners' skill development by allowing for several edits and delivering immediate feedback.

The integration of more sophisticated language models, such as ChatGPT, had transformed the teaching of a language that was previously considered exclusively for L2 settings. ChatGPT was increasingly being found in association with terms such as written corrective feedback, EFL writing, and its overall influence on individualized learning. Research analyzed in this study indicated that more sophisticated tools were being leveraged to create language materials, simulate dialogues, and offer personalized learning experiences, which benefited both learners and teachers. This reflected a paradigm shift in the application of AI in the language acquisition sphere because these models offered interactivity and personalization that were never achievable before, allowing for much more realistic and stimulating language practice. The role of ChatGPT in finding contextually appropriate responses made it an important L2 assessment tool. However, while ChatGPT could spot surface-level writing errors, it was much less effective at detecting deeper issues like pragmatics.

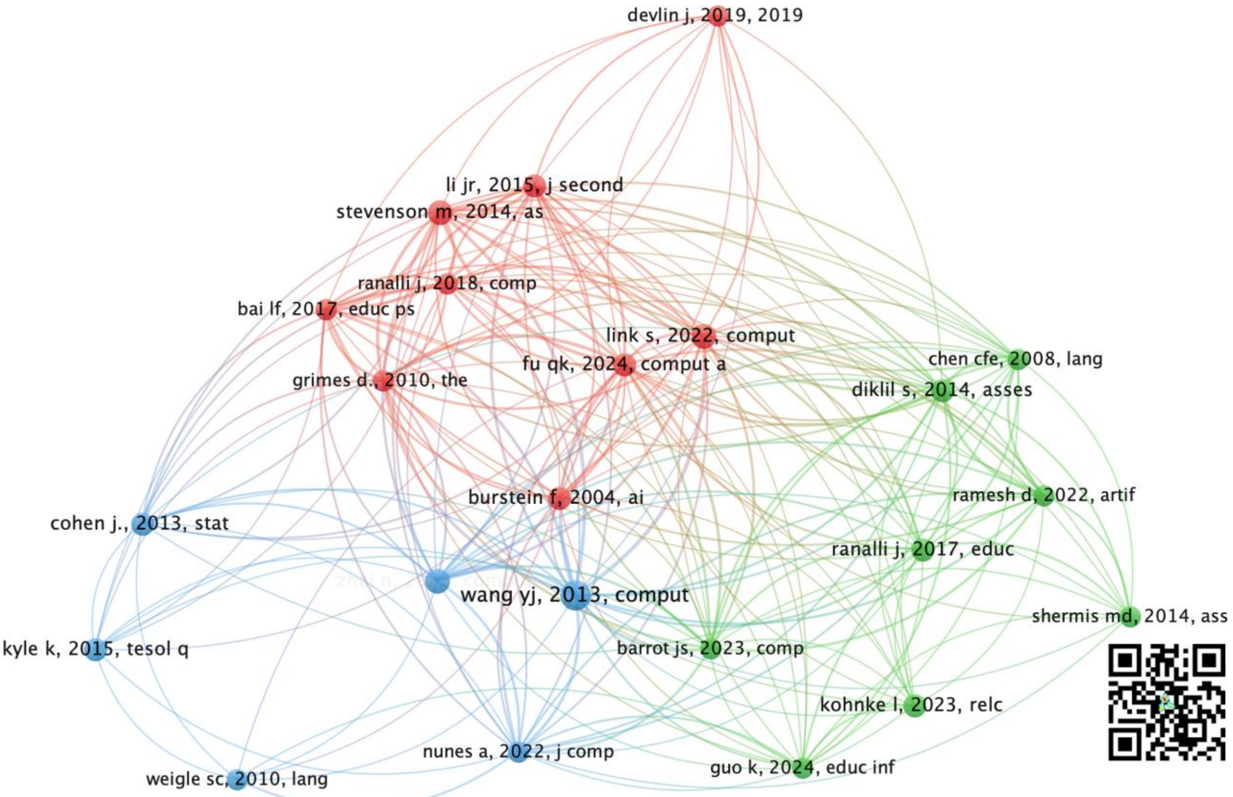
The red cluster encompassed the themes of "L2 writing" and "student engagement," which underscored the importance of exploring how learners interacted with AI tools and their attitudes toward these technologies. The successful integration of AI in language learning environments hinged on these crucial factors, as they directly affected students' willingness to engage with the technology and, ultimately, their learning outcomes. This emphasis on student perceptions signified a broader recognition that the success of AI in education was not solely about the technology itself, but about how learners viewed its role and how they incorporated it into their learning experiences.

3.5.2 Reference Co-citation Analysis

The co-citation analysis of the references is given in Figure 5. The minimum number of citations of a cited reference was set to 5. The map displaying the co-citation analysis offered a thorough representation of the intellectual framework in the field. This methodology examined the frequency of citations of certain references across several studies, helping to reveal influential publications and authors who had a substantial impact. Through the analysis of these co-citation relationships, a deeper comprehension of the fundamental and emergent concepts that influenced this swiftly developing domain could be achieved (van Eck & Waltman, 2023).

Figure 5

Network Maps: (a) Co-citation Links of Cited References.



Note: Scan the QR code to zoom in on the map for a detailed view of all references, links, citations, and total link strengths.

At the center of the red cluster were fundamental studies that largely informed the realization of AI in language assessment. Outstanding works of this cluster included Devlin et al. (2019), who proposed the introduction of BERT, which is considered a milestone in natural language processing. Added contributions came from Li et al. (2015), particularly in second language learning and assessment. Strong co-citation tied within this cluster suggested that these studies were often cited together, thus demonstrating their impact on the advancement of AI-based language assessment tools and methodologies. Another prominent cluster in green was formed of pioneering works in developing automated scoring systems and AI-powered feedback systems. A prominent contribution by Shermis et al. (2014) was crucial for the emergence of AES and the embedding of AI in educational contexts. The strong equivalency within this cluster was indicative of the applied approach to AI in the assessment of language competencies, especially in writing and feedback systems. The blue cluster in the network centered on the foundational statistical and methodological underpinnings to forge the area of AI and language assessment research. Influential references, such as Cohen (2013) on statistical power analysis, were essential for establishing the reliability and accuracy of AI tools. These methods flagged the importance of an impressive methodological foundation, including stringent statistical procedures for obtaining validation for AI tools and ensuring effectiveness within language assessment. These subsequently remained the bedrock upon which all AI applications in education continued to be developed.

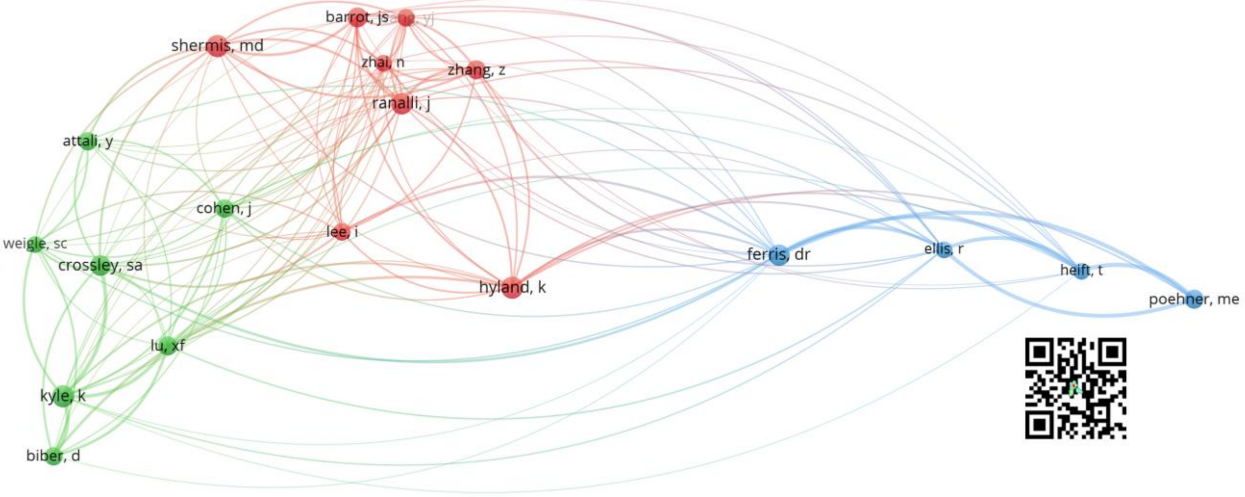
The co-citation map revealed substantial interconnectedness between these clusters, particularly between the red and green clusters. This interrelationship suggested a strong link between foundational AI research and its practical implementations in L2 assessment. The adoption of sophisticated AI models like BERT and GPT in language evaluation was becoming more prevalent (Wang et al., 2022), reflecting a shift toward advanced, scalable assessment methods. This trend demonstrated that research in AI-assisted language assessment was rooted in a robust blend of theoretical and applied studies. The groundwork laid by earlier research on NLP and automated scoring paved the way for current and future innovations in AI-driven evaluation systems. As the field progressed, these intellectual connections continued influencing developments in both AI technology and its educational applications.

3.5.3 Author Co-citation Analysis

The author's co-citation analysis is shown in Figure 6. The minimum number of citations for a source was set to 10. Method: association strength, full counting.

Figure 6

Co-citation Links of Authors



Note: Scan the QR code to zoom in on the map for a detailed view of all authors, links, citations, and total link strengths.

The figure shows the presence of three separate groups of authors, each making unique contributions to certain thematic domains in the field of AI-driven language assessment. The primary cluster consisted of writers such as Barrot, Hyland, and Shermis, who had made substantial progress in the advancement and utilization of AI tools for writing and feedback in L2 contexts. Their studies were based on the application of AI to elevate the quality and uniformity of evaluative aspects, targeting the vicinity of writing, where automated technologies serve large-scale assessments with scalable solutions. The second cluster consists of influential scholars like D. Biber, J. Cohen, and S.A. Crossley-whose contributions have played a pivotal role in creating the methodological and statistical underpinnings that work toward the validity of AI methods in language assessment. The emphasis on thorough statistical skill and linguistic analysis in this cluster places importance on the development of effective and fair assessment tools. The applicability of this cluster is demonstrated in research on dynamic assessment frameworks. The third cluster includes authors like R. Ellis, D.R. Ferris, and T. Heift, who examine the conjunction between AI and pedagogy. Here, the focus is on feedback and dynamic assessment. Their research adds a perspective on how the incorporation of AI into language teaching affects assessment.

4. Discussion

The findings of the presnet study havesuggested that AI-assisted L2 assessment research spans from evaluations involving undergraduates, professionals, and English-as-a-second-language students, illustrating that such assessments are broadly applicable to a variety of learner types. Given their accessibility and the growing presence of AI technologies in higher education contexts, it is not surprising to see a lot of attention on undergraduate learners. However, such a focus raises questions about the generalizability of the findings to younger learners, working professionals, and other underrepresented groups. AI technology is a shifting landscape, and the current research does not focus on K-12 students enough to address whether AI-based assessment applications will accurately adapt to K-12 language learners who are still early in the language acquisition process. The possibility of AI-based assessments meaningfully supporting the above depends, however, on another critical dimension, which is the potential impact such assessments will have on learners of different linguistic backgrounds. Though AI-based tools promise individualized feedback, the appropriateness of these tools in other cultural and educational settings is still largely unexamined (Guglielmi, 2008). As a result, inequities about the use of AI-led assessment also exist since learners who can speak less commonly taught (lesser-taught) languages will not be able to use the same level of AI-led assessment (Bui & Barrot, 2024). Therefore, further research is needed on the availability and fairness of using AI assessments in diverse L2 learning contexts.

Dominating the study were quasi-experimental, quantitative, qualitative, and mixed-methods research approaches to AI-assisted L2 assessment studies. While the substantial fraction of experimental designs indicates a high priority given to understanding how AI affects learning outcomes, the relatively low inclusion of qualitative studies points to the potential neglect of a deeper understanding of how such tools are integrated within the learning experience (Wei, 2023). Although quantitative studies offer convincing and sometimes descriptive evidence about the effectiveness of AI, qualitative approaches are crucial as they show the perspectives, opinions, and involvement of learners (Shen et al., 2023). It is also worth noting the absence of longitudinal studies that track the long-term effects of AI-generated evaluations on language fluency. Most studies were short-term interventions and hence have yielded further gaps in the generalizability of the impact of AI tools on learner autonomy and language retention in the long run (Firat, 2023). Moreover, the heavy dependence on small-scale studies indicates the necessity for evidence generated from larger, more representative samples to verify that results are generalizable across diverse educational environments. Future studies may involve larger mixed-methods designs in different contexts to reconcile statistical significance with learner-centric insights (Jia et al., 2022).

It was also found in this study that AI tools are mostly used for writing and speaking assessment, automated item generation, anxiety and motivation, grammar checking, feedback, and vocabulary acquisition. This aligns with findings of previous studies (Circi et al., 2023; Prasetyo et al., 2020). The commonly used tools for L2 writing assesment are AVA, ChatGPT, and Grammarly. ChatGPT not only provides automated feedback but also gives feedback on writing quality (Zadorozhnyy & Lai, 2023). Today's AI assessment tools may encourage formulaic writing, which can limit learners' ability to develop their own writing styles (Ding &

Zou, 2024). The co-citation analysis of the current study revealed key research areas such as automated feedback, L2 writing assessment, and AI-driven engagement strategies. ChatGPT and other LLMs were found to be influential. Several studies support this finding (see among others Amin, 2023; Ding & Zou, 2024); however, the diversity of AI tools should be increased (de la Vall & Araya, 2023).

AI is increasingly being explored as a tool for evaluating speaking skills, with models like Wav2Vec2 and SpeechRaterSM showing promise in assessing pronunciation, fluency, and coherence. Also, tools such as Speeko and Call Annie provide feedback on general speaking proficiency. However, they are limited in assessing pragmatic competence and natural conversational skills (Jin & Fan, 2023). A key limitation of AI-driven speech assessment is its tendency to overlook communication's cultural and contextual nuances, which are essential in real-world language use (Al-Ghezi et al., 2023). This problem can be solved by focusing on AI's ability to assess discourse competence and cross-cultural communication skills (Voss et al., 2023).

It is evident in the literature that AI tools help improve feedback quality, boost learner engagement, and foster learner autonomy (Liu et al., 2024). However, there are some challenges, such as bias, authenticity, data privacy, and inconsistencies (Ding & Zou, 2024; Voss et al., 2023). Some AI tools may not represent diverse learner populations (Osawa, 2023). In other words, AI tools may result in unfair assessments. One reason for this, especially in foreign language contexts, is the diverse linguistic patterns of learners. This may not be represented by the AI tools, which are generally trained with native use of the language. These challenges may be addressed via policies on the ethical use of AI in assessment (Voss et al., 2023). Also, the growing reliance on AI in assessment raises concerns about its limitations. Although AI offers efficient assessment tools, educators should use it as a complement rather than a replacement for human evaluators (Coskun & Alper, 2024).

5. Conclusion

This biblio-systematic analysis provided an overview of AI-assisted L2 assessment research. Our study examined the research on AI tools in language assessment and produced both systematic and bibliometric findings. AI was found to mainly contribute to providing feedback and assessing L2 writing skills. The study also revealed the importance of understanding how learners perceive these tools to facilitate L2 assessment. Generative AI such as ChatGPT has the potential to reshape not onl how language is taught and learned but also how it is being assessed. While the positive impact of AI in L2 assessment is evident, the present study also highlighted challenges, particularly the necessity for more diverse methodological approaches and balanced datasets to ensure AI tools function effectively across various contexts.

This study had some limitations. Only SSCI-indexed articles were retrieved. Therefore, not all AI-assisted L2 assessment articles were analysed. Further studies may utilize Emerging Social Sciences Citation Index (ESCI), the Arts and Humanities Citation Index (AHCI), Scopus, and other databases to broadening the scope of the review to gain additional insights and emerging trends. Moreover, future review articles should consider comparing effect sizes of the experimental research. Future studies should investigate the long-term effectiveness, fairness, and ethical concerns of AI-assisted assessments, particularly regarding bias, data privacy, and equitable access. Pedagogically, AI offers opportunities for personalized feedback and adaptive assessment, but its limitations in evaluating creativity, discourse coherence, and pragmatic competence highlight the need for human oversight. Educators should be trained to effectively integrate AI tools while ensuring that AI serves as an augmentation rather than a replacement for human evaluation. A balanced approach that prioritizes fairness, validity, and pedagogical effectiveness is essential for the sustainable integration of AI in L2 assessment.

6. About the Authors

Ecem Kopuz is a Science Fellow and doctoral researcher in the Ph.D. Program in Speech-Language-Hearing Sciences at the CUNY Graduate Center, City University of New York. She is also an active researcher and lecturer with a strong background in applied linguistics and cognitive science. Her interdisciplinary research focuses on the neurobiology of learning and the neurophysiology of language development, with an emphasis on individual differences in language learning and processing. By integrating methods from neurolinguistics, electrophysiology (EEG/ERP), and cognitive psychology, her work aims to uncover how cognitive and neural mechanisms shape language acquisition and use across diverse learner populations.

Galip Kartal is an associate professor at Necmettin Erbakan University, Türkiye, specializing in English Language Teaching. He holds advanced degrees in the field and focuses on innovative teaching methods, vocabulary education, corpus-assisted language teaching, and AI integration in language teaching. Additionally, he has presented at numerous international conferences, contributing to discussions on teacher training, digital pedagogy, and AI-driven language assessment.

7. References

Algaraady, J., & Mahyoob, M. (2023). ChatGPT's capabilities in spotting and analyzing writing errors experienced by EFL learners. *Arab World English Journal*, *9*, 3–17. https://doi.org/10.24093/awej/call9.1

Al-Ghezi, R., Voskoboinik, K., Getman, Y., Von Zansen, A., Kallio, H., Kurimo, M., Hildén, R. (2023). Automatic speaking assessment of spontaneous L2 Finnish and Swedish. *Language Assessment Quarterly, 20*(4–5), 421–444. https://doi.org/10.1080/15434303.2023.2292265

Amin, M. (2023). AI and ChatGPT in language teaching: Enhancing EFL classroom support and transforming assessment techniques. *International*

Journal of Higher Education Pedagogies, *4*(4), 1–15. https://doi.org/10.33422/ijhep.v4i4.554

Barrot, J. (2021). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning, 36*(4), 584–607. https://doi.org/10.1080/09588221.2021.1936071

Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology*, *23*(1), 23–41. https://doi.org/10.56297/bkam1691/wieo1749

Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, *30*, 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, Ø. E., Yuan, Z., Elliott, M., Moore, R., Bryant, C., Rei, M., Yannakoudakis, H., Mullooly, A., Nicholls, D., & Buttery, P. (2023). *On the application of large language models for language teaching and assessment technology.* ArXiv. https://doi.org/10.48550/arXiv.2307.08393

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education, 8*, Article 858273. https://doi.org/10.3389/feduc.2023.858273

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Coskun, T., & Alper, A. (2024). Evaluating the evaluators: A comparative study of AI and Teacher Assessments in Higher Education. *Digital Education Review*. https://doi.org/10.1344/der.2024.45.124-140.

Crossley, S., & Kyle, K. (2018). Assessing writing with the tool for the automatic analysis of lexical sophistication (TAALES). *Assessing Writing*. https://doi.org/10.1016/J.ASW.2018.06.004.

de la Vall, R. R. F., & Araya, F. G. (2023). Exploring the benefits and challenges of AI- language learning tools. *International Journal of Social Sciences and*

Humanities Invention, 10(01), 7569–7576. https://doi.org/10.18535/ijsshi/v10i01.02

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. ArXiv. https://doi.org/10.48550/arXiv.1810.04805.

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies, 29*, 14151–14203. https://doi.org/10.1007/s10639-023-12402-3

Duman, G., Orhon, G., & Gedik, N. (2015). Research trends in mobile assisted language learning from 2000 to 2012. *ReCALL, 27*(2), 197–216. https://doi.org/10.1017/S0958344014000287

Firat, M. (2023). What ChatGPT means for universities: Perceptions of scholars and students. *Journal of Applied Learning and Teaching, 6*(1), 57–63. https://doi.org/10.37074/jalt.2023.6.1.22

Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C., & Srinivasa, A. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 100206. https://doi.org/10.1016/j.caeai.2024.100206

Guglielmi, R. (2008). Native language proficiency, English literacy, academic achievement, and occupational attainment in limited-English-proficient students: A latent growth modeling perspective. *Journal of Educational Psychology, 100*, 322–342. https://doi.org/10.1037/0022-0663.100.2.322

Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *The Internet and Higher Education, 63*, 100962. https://doi.org/10.1016/j.iheduc.2024.100962

Huang, X., Xu, W., Li, F., & Yu, Z. (2024). A meta-analysis of effects of automated writing evaluation on anxiety, motivation, and second language writing skills. *The Asia-Pacific Education Researcher*, 1–20.

Jia, F., Sun, D., Ma, Q., & Looi, C. (2022). Developing an AI-based learning system for L2 learners' authentic and ubiquitous learning in English language. *Sustainability*, *14*(23). https://doi.org/10.3390/su142315527

Jin, Y., & Fan, J. (2023). Test-taker engagement in AI technology-mediated language assessment. *Language Assessment Quarterly, 20*, 488–500. https://doi.org/10.1080/15434303.2023.2291731

Kartal, G., & Yeşilyurt, Y. E. (2024). A bibliometric analysis of artificial intelligence in L2 teaching and applied linguistics between 1995 and 2022. *ReCALL*, *36*(3), 359–375. doi:10.1017/S0958344024000077

Kenshinbay, T., & Ghorbandordinejad, F. (2024). Exploring AI-driven adaptive feedback in the second language writing skills prompt. *EIKI Journal of Effective Teaching Methods*, *2*(3). https://doi.org/10.59652/jetm.v2i3.264

Li, K., Rollins, J., & Yan, E. (2018). Web of Science use in published research and review papers 1997–2017: A selective, dynamic, cross-domain, content-based analysis. *Scientometrics, 115*(1), 1–20. https://doi.org/10.1007/s11192-017-2622-5

Li, R. (2022). Research trends of blended language learning: A bibliometric synthesis of SSCI-indexed journal articles during 2000–2019. *ReCALL, 34*(3), 309–326. https://doi.org/10.1017/S0958344021000343

Liu, F., Jiang, Y., Lai, C., & Jin, T. (2024). Teacher engagement with automated text simplification for differentiated instruction. *Language Learning & Technology, 28*(2), 163–182. https://hdl.handle.net/10125/73576

Markscheffel, B., & Schröter, F. (2021). Comparison of two science mapping tools based on software technical evaluation and bibliometric case studies. *COLLNET Journal of Scientometrics and Information Management, 15*(2), 365–396. https://doi.org/10.1080/09737766.2021.1960220

Meniado, J. C. (2023). The impact of ChatGPT on English language teaching, learning, and assessment: A rapid review of literature. *Arab World English Journal*, *14*(4). https://dx.doi.org/10.24093/awej/vol14no4.1

Nazari, N., Shabbir, M., & Setiawan, R. (2021). Application of artificial intelligence-powered digital writing assistant in higher education:

Randomized controlled trial. *Heliyon, 7*. https://doi.org/10.1016/j.heliyon.2021.e07014

Osawa, K. (2023). Integrating automated written corrective feedback into e-portfolios for second language writing: Notion and Notion AI. *RELC Journal*, *55*(3), 881–887 https://doi.org/10.1177/00336882231198913

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal, 100*(2), 538–553. https://doi.org/10.1111/modl.12335

Prasetyo, S. E., Adji, T. B., & Hidayah, I. (2020). Automated Item Generation: Model and Development Technique. *Proceedings of the 7th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2020*. 64–69. https://doi.org/10.1109/ICITACEE50144.2020.9239243.

Roa, A., & Halim, S. (2024). The impact of AI-powered software on second language (L2) writing: A systematic literature review. *Research and Innovation in Applied Linguistics-Electronic Journal*, *2*(2). https://doi.org/10.31963/rial.v2i2.4801.

Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin*, *143*(2), 117. https://doi.org/10.1037/bul0000088

Saeidnia, H., Hosseini, E., Abdoli, S., & Ausloos, M. (2024). *Unleashing the power of AI. A systematic review of cutting-edge techniques in AI-enhanced Scientometrics, Webometrics, and Bibliometrics.* ArXiv. https://doi.org/10.1108/LHT-10-2023-0514.

Settles, B., Hagiwara, M., & LaFlair, G. (2020). Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics, 8*, 247–263. https://doi.org/10.1162/tacl_a_00310

Shen, C., Shi, P., Guo, J., Xu, S., & Tian, J. (2023). From process to product: Writing engagement and performance of EFL learners under computer-generated feedback instruction. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1258286

Shermis, M. D. (2014). State-of-the-art automated essay scoring: A United States demonstration and competition, results, and future directions. *Assessing Writing*, 20, 53–76. https://doi.org/10.1016/j.asw.2013.04.001

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.* https://doi.org/10.3389/fpsyg.2023.1260843

van Eck, N. J., & Waltman, L. (2023). *VOSviewer manual: Manual for VOSviewer version 1.6.20*. Leiden University.

Voss, E., Cushing, S. T., Ockey, G. J., & Yan, X. (2023). The use of assistive technologies including generative AI by test takers in language assessment: A debate of theory and practice. *Language Assessment Quarterly, 20*(4–5), 520–532. https://doi.org/10.1080/15434303.2023.2288256

Wang, Y., Wang, C., Li, R., & Lin, H. (2022). *On the use of Bert for automated essay scoring: Joint learning of multi-scale essay representation.* ArXiv. https://doi.org/10.48550/arXiv.2205.03835

Wei, L. (2023). Artificial intelligence in language instruction: Impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1261955

Yang, Z., Wang, P. (2025). Current status and research trend of English language assessment: a bibliometric analysis. *Lang Test Asia*, *15*, 11. https://doi.org/10.1186/s40468-024-00317-w

Zadorozhnyy, A., & Lai, W. (2023). ChatGPT and L2 written communication: A game-changer or just another tool? *Languages.* https://doi.org/10.3390/languages9010005