

The use of Artificial Intelligence in English language assessment: Empirical evidence and future directions

Dennis Alonzo^{1#}, Jan Michael Vincent Abril¹, and Cherry Zin Oo²

¹*School of Education, University of New South Wales, Australia*

²*Myanmar Imperial College, Myanmar*

#Corresponding author: d.alonzo@unsw.edu.au

(Submitted: 5 March 2025; Accepted: 1 September 2025)

 @Dennis_A_Alonzo | @janmicahelabril | @CherryZinOo

Abstract

Artificial intelligence (AI) is increasingly recognised as a useful tool for assessment, but its specific role in English language assessment remains unexplored. We conducted a scoping review of peer-reviewed studies published between 2011 and 2025, following the framework by Arksey and O'Malley (2005). Results show seven main ways AI is used in English assessment: generating test items, using chatbots for conversation practice, marking and scoring, supporting self-assessment, enabling adaptive testing, giving instant feedback, and recognising speech. AI helps make assessments more efficient, keeps students engaged, and supports more personalised learning. We also found that AI is helping teachers develop flexible teaching practices, meet standards, and manage assessments more effectively. This review highlights both the potential and the challenges of using AI in English language assessment and calls for more research to support its responsible and effective use in schools.

Keywords: artificial intelligence, assessment, macro skills, English language, outcomes

Introduction

The application of artificial intelligence (AI) in education has significantly changed the landscape of learning and teaching. It offers advanced ways of personalising learning, providing virtual and augmented reality activities (Cui, 2022; Lampropoulos, 2023), offering flexibility and autonomy (Xia, et al., 2023), and supporting data-driven teacher decision-making (Hwang, et al., 2020), among other benefits. Many educational institutions have used them as an integral part of their processes and procedures. Although there have been claims that AI has positively impacted learning and teaching (Alam, 2022; Sun, et al., 2021), its application in assessment is limited. Review studies focusing on the use of AI have been published in general education (Chiu, et al., 2023), early childhood education (Su, et al., 2023), assessment methods (Martínez-Comesaña, et



This publication is covered by a Creative Commons Attribution 4.0 International license. For further information please see: <http://creativecommons.org/licenses/by/4.0/>.

al., 2023), and student assessment (González-Calatayud, et al., 2021). While there is a consensus that AI can be used in assessing knowledge and skills, there is a lack of a coherent knowledge base on how and why it is applied in English language assessment. This paper reviews the existing literature to provide a comprehensive overview of research on using AI in English language assessment. Due to the prominence of English language teaching across English-speaking and non-English-speaking countries, we focus on this key learning area. English language assessment encompasses complex and multifaceted skills (e.g., reading, writing, speaking, listening). As a domain, it provides a substantial amount of data that can be utilised to train AI models capable of identifying patterns and making predictions about English language proficiency. Teachers find assessment time-consuming and repetitive as a task, which may well be suited for automation. AI has the potential to make assessment more efficient, consistent, personalised, adaptive, and accessible (Minn, 2022). Moreover, AI can be programmed to consider the complex nature of the English language, diverse skills to assess, and cultural nuances and sensitivity (Alghamdy, 2023).

Additionally, we focus on English language assessment because the role of assessment in the classroom is pivotal in ensuring effective learning and teaching (Alonzo, 2020; Black, 2017; Hattie, 2008). Although there are debates and competing conceptualisations about assessment (perspectives on the key elements of assessment (Baroudi, 2007), including its nature and process (Bennett, 2011), the distinction between summative and formative assessment (Sadler, 1989), and the appropriateness of the use of assessment types (Wiliam, 2011), there is an agreement that if assessment and assessment data are used by students and teachers to inform learning and teaching decisions, it will improve student outcomes (Alonzo & Loughland, 2022; Black, 2017). However, teachers often struggle with confidence and capability in assessment design and implementation (Davison & Michell, 2014; Loughland & Alonzo, 2019).—Thus, many school assessment reforms are implemented (Oo, et al., 2023), including preparing in-service teachers (Oo, et al., 2022; Oo, Alonzo & Davison, 2023). However, teachers complain that assessment-related workload, ranging from developing to marking and reporting, impedes their classroom practices (Stacey, et al., 2022). With the functionalities of AI, such as automated scoring, adaptive learning, and real-time feedback, AI may offer opportunities for enhancing both the efficiency and effectiveness of teacher assessment practices. Yet, the integration of AI into assessment is underexplored. This paper responds to that gap by critically mapping existing applications of AI in English language assessment. By doing so, we aim to inform future research and practice, highlighting not only what is possible but also what is pedagogically desirable and just.

Methods

A scoping review, following Arksey and O'Malley (2005) was considered an appropriate strategy to explore what research has already been undertaken into using AI in English language assessment in schools. The rationale for using a scoping review was to identify knowledge gaps, scope and reflect on the body of literature, and clarify concepts with a forward-looking view to inform practice and future research. Many papers used a scoping review (Alonzo, et al., 2023; Ellis, et al., 2020). This approach provided a framework for identifying and reflecting on the

literature and a way to identify the AI technologies used in English language assessment. The following describes each stage of the scoping review.

Stage 1: Identifying the research question

Initially, a research question was devised to guide the scoping review focused on the research topic *Use of AI in English Language Assessment*. This question was:

1. What does the literature say about using AI in English language assessment?

As patterns and themes began to emerge from the literature during the initial stages of the scoping review, it became clear that a revision of the research question was necessary, and additional questions were added to better represent the context and challenges that arose as the scoping process progressed. The final research questions are:

- (1) How have **AI technologies** been used in English language assessment in schools?
- (2) What are the **outcomes and challenges** reported about using AI in English language assessment in schools?
- (3) What **recommendations or implications** for practice and theory have been outlined by the extant literature for using AI in English language assessment in schools?

Stage 2: Identifying relevant studies

To undertake the scoping review, ProQuest, Scopus and Web of Science databases were used to search the literature for keywords and terms associated with AI, language assessment, and schools. As the literature search was based on the discipline of education, each of the databases allowed the researchers to conduct a search across a large selection of journals. The search also included terms that are associated with A/L practices. Each search was confined to articles published from 2011 to 2025, to examine AI technologies associated with assessment in English.

Search terms comprised artificial intelligence and English language and literacy assessment (such as reading, writing, vocabulary, and speaking/oral assessment). The search terms combinations are shown in Table 1. A manual search was conducted in journals where articles would most likely be published. Two journals were included: *Computer and Education: Artificial Intelligence* and *British Journal of Education Technology*. Three articles were added from this search.

Stage 3: Study selection

From the combined database and manual search, 52 articles were potentially related to the research topic. A closer examination of each article's content and research focus revealed that 18 articles were irrelevant to the research questions, and these were removed from the sample. Although some articles discussed AI and language assessment, these studies were omitted because they did not explicitly focus on English language assessment.

Table 1. Search Terms for Databases

Electronic database	Keyword/Syntax	Number
Proquest (ERIC and Education database)	noft(Artificial Intelligence OR AI) AND (Language OR English or vocabulary OR speaking OR oral OR reading) AND (assessment OR exam* OR formative OR summative OR test)	2,311
Scopus	TITLE-ABS-KEY ((Artificial Intelligence OR AI) AND (Language OR English or vocabulary OR speaking OR oral OR reading) AND (assessment OR exam* OR formative OR summative OR test)))	1,714
Web of Science	TS-((Artificial Intelligence OR AI) AND (Language OR English or vocabulary OR speaking OR oral OR reading) AND (assessment OR exam* OR formative OR summative OR test)))	1,000

Stage 4: Data charting

Once the articles were identified, the focus and theme for each article were categorised and compiled. This allowed the researchers to analyse the focus and findings of each reported research area and group the articles into categories according to themes. Keywords identified by the authors in each publication were used as themes. These keywords were compiled and presented in Table 2.

Table 2: Authors' Keywords

Authors	Keywords
Almegren, et al. (2025)	no keywords in the published paper
An, et al. (2022)	English teacher, foreign language learning, behavioral intention, AI, middle school
Bezirhan & von Davier (2023)	AI-generated reading passages; automated item generation; large language models; natural language processing; reading assessment
Bulut & Yildirim-Erbasli (2022)	reading comprehension, natural language processing, automatic item generation, language modeling, text generation
Chen & Lee (2011)	no keywords in the published paper
Chomphoooyod, et al. (2023)	AI-based learning; automatic question generation; education technology; language learning; multiple-choice question
Derakshen & Ghiasvand (2024)	artificial intelligence, ChatGPT, educational technology

Ericsson & Johansson (2023)	conversational AI; dialogue-based computer-assisted language learning; longitudinal educational experience; spoken dialogue system; virtual human
Gayed, et al. (2022)	CALL, AI in education, L2 writing, cognitive load, AI agent
Ghafouri, et al. (2024)	ChatGPT, EFL, writing instruction, self-efficacy, language learning
Hannah, et al. (2022)	no keywords in the published paper
Jamshed, et al. (2024)	ChatGPT application, personalized feedback, targeted corrections
Jeon (2021)	AI; chatbots; cognitive load theory; dialogflow; dynamic assessment; vocabulary learning
Kumar & Boulanger (2020)	automated essay scoring; deep learning; explainable AI; feedback; learning analytics; rubric; sharp; trust
Lee, et al. (2023)	AI; learner-generated context; learner-generated content; intelligent computer-assisted language learning; development research; secondary education
Liu, et al. (2023)	AI; automatic written feedback; english as foreign language; language learning; peer assessment
Liu & Wang (2024)	AI tools; critical thinking; educational technology; EFL learners; English literature classes; intervention study
Moghadam, et al. (2024)	Micro-break activities; e-learning systems; emotion regulation; evolutionary algorithms; genetic algorithm; AI
Moorhouse & Kohnke (2024)	Generative AI; initial language teacher education; ChatGPT; teacher educators
Nazaretsky, et al. (2022)	no keywords in the published paper
Ormerod, et al. (2022)	artificial intelligence; assessment; education; neural networks; short answer

Peng, et al. (2023)	data science applications in education; distance education and online learning; evaluation methodologies; improving classroom teaching; secondary education
Rahimi, et al. (2017)	automatic essay assessment; analytical writing in response to text; evidence; organization; task-dependent; feedback; natural language processing
Rodriguez-Barrios et al. (2021)	artificial intelligence; bayesian networks; education; reading comprehension
Safi, et al. (2023)	autism spectrum disorder, virtual voice assistants, language skills, social skills, artificial intelligence, children
Sargazi Moghadam, et al. (2023)	artificial intelligence; e-learning systems; emotion regulation; evolutionary algorithms; genetic algorithm; micro-break activities
Srinivasan & Murthy (2021)	cooperative/collaborative learning; distance education and online learning; elementary education; improving classroom teaching; teaching/learning strategies
Tafazoli (2024)	GenAI; ChatGPT; educational challenges; English language teachers
Wilson, et al. (2021)	Automated feedback; automated writing evaluation; automated essay scoring; writing assessment; educational technology
Xia, et al. (2023)	AI; K-12; prior knowledge, self-determination theory, self-regulated learning
Wei, et al. (2023)	automated writing evaluation, L2 writing skills; EFL learners; writing self-efficacy; complex script
Zhang (2023)	English composition; automatic scoring; artificial intelligence; text matching degree; natural language processing
Zhang & Han (2021)	educational psychology; reading aloud training; self-efficacy; mental disorder in English learning; English learning anxiety
Zhao et al. (2023)	artificial intelligence; automated writing assessment; cross-modal matching; picture-cued writing

The categories used to summarise the articles were outlined using the following groupings:

- Details (author, year of publication)
- Focus (what the article is about)
- Key Assessment Concepts (list of assessment concepts)
- Key Themes (relationship between assessment & AI)

Once each article had been coded, patterns and themes which determined the categorising of findings.

Stage 5: Collating, summarising, and reporting

Following *Stage 4*, findings were examined in relation to the existing themes in the literature (Arksey & O'Malley, 2005). This approach allowed the findings to be synthesised as a narrative and presented as identifying effective AI and their functionalities, outcomes reported, and implications.

Results

We present our answers to our research questions in the succeeding sections.

1. How have AI technologies been used in English language assessment in schools?

There are seven major categories of AI use in English language assessment.

Automated test-item generation

AI is used for automatic test item generation (Choi, et al., 2024; Derakhshan & Ghiasvand, 2024). Bezirhan and von Davier (2023) used Instruct GPT and Bulut and Yildirim-Erbasli (2022) applied GPT3 and T5 to generate reading comprehension items. Both studies have found that AI-generated content is comparable to human-written passages. Chomphooyod, et al. (2023) used Text-to-Text Transformer to generate items for assessing grammar, demonstrating the potential of AI to generate high-quality multiple-choice questions. Choi, et al. (2024) found that generative AI and large language models can automate the creation of targeted grammar assessments. Although these studies present positive outcomes, the validity of AI-generated items across diverse student profiles or educational contexts warrants further investigation. Moreover, while generative AI enhances scalability, there is limited discussion about item bias, cultural appropriateness, or the risks of over-reliance on algorithmically generated content in high-stakes assessments.

Conversational agents for speaking practice

Seven AI-powered conversational agents are increasingly used in English language assessment to support the development of students' speaking skills. Ericsson and Johansson (2023)

demonstrated that the Enskill Spoken dialogue system effectively maintained engagement and fostered positive emotional responses, supporting the use of AI for sustained speaking practice in a low-anxiety learning environment. Other technologies, such as the Human Emotion Recognition System (HERS) have also been employed to reduce language anxiety by interpreting physiological responses during speech (Chen & Lee, 2011). Hannah, et al. (2022) utilised automated speech recognition (ASR) systems (e.g., Google Cloud, Microsoft Azure, and Kaldi) to explore various speech restrictions and cognitive load conditions, thereby enhancing assessment design. Chatbot-Assisted Dynamic Assessment (CA-DA), as studied by Jeon (2021), offered differentiated support and was more effective than traditional methods in promoting vocabulary acquisition. Xia, et al. (2023) demonstrated that AI chatbots can guide English learning while providing automated scoring, thereby further supporting formative assessment practices. Similarly, Almegren, et al. (2025) found that AI-powered chatbots, such as Copilot and PI AI, support interactive language practice through pronunciation feedback and role-playing. These AI-driven conversational agents provide students with continuous, personalised speaking practice and immediate feedback, enabling them to refine their pronunciation and conversational skills in a low-stress environment (Lee, et al., 2024).

Critically, while conversational agents offer scaffolded, low-stress environments for practice, few studies examine the depth and quality of language output they elicit. Moreover, concerns about assessment validity and teacher oversight are insufficiently addressed. There is also an implicit assumption that chatbots are culturally neutral and universally effective, which may not hold in multilingual or diverse classrooms.

Automated marking and scoring

Eight AI technologies were reported for automatic essay scoring. Wei, et al. (2023) highlighted the use of Automated Writing Evaluation (AWE) tools such as Grammarly and ChatGPT, which leverage Natural Language Processing (NLP) to assess grammar, vocabulary, coherence, and organisation. Similarly, Jamshed, et al. (2024) reported on advancements in AI-driven scoring systems that provide real-time, scalable feedback on student writing. Kumar and Boulanger (2020) introduced the Suite of Automatic Linguistic Analysis Technology (SALAT which evaluates grammar, sentiment, cohesion, lexical diversity, and syntactic complexity, offering both holistic and rubric-aligned scores. Nazaretsky, et al. (2022) utilised AI Grader to automate constructed response assessment, achieving high scoring accuracy. Ormerod, et al. (2022) applied a short answer scoring engine capable of evaluating short responses in large-scale assessments, reporting performance above human-level accuracy. Rahimi, et al. (2017) developed a task-dependent automatic scoring model for evaluating evidence and organisation in essays, aligning with rubrics to provide targeted feedback, such as suggesting more specific use of evidence or pointing to alternative sources. Zhang (2023) and Zhao, et al. (2023) also applied NLP for large-scale essay assessment, which can efficiently assess a large number of essays with consistent accuracy, saving time for teachers. In addition to writing, ReadToMe dashboards (Srinivasan & Murthy, 2023) and Netica (Rodriguez-Barrios, et al., 2021) were used for automatic scoring in

reading assessments. Both technologies showed high accuracy and efficiency in assessing students' reading comprehension and fluency. These tools collectively demonstrate the potential of AI to enhance assessment reliability, reduce teacher workload, and deliver timely, actionable feedback to students. However, these efficiencies raise critical issues. Some studies, such as Rahimi, et al. (2017), attempted to align AI scoring with human rubrics, but the depth of alignment with curriculum goals and formative assessment intentions remains unclear. While AI may efficiently detect surface-level features, its capacity to assess higher-order thinking, argument quality, or creativity is still limited. Moreover, the risk of students learning to "game the system" through formulaic responses remains largely unaddressed.

For self-assessment

Two studies highlighted AI tools for self-assessment. ChatGPT and Generative AI support autonomous learning by offering immediate, personalised feedback and generating customised content aligned with students' proficiency levels (Derakhshan & Ghiasvand, 2024). Lee, et al. (2023) emphasise the potential of Learner-Generated Contexts (LGC) to facilitate self-directed learning even with minimal teacher support. These AI tools were developed to develop students' agency, empowering them to assess their progress, engage in learning activities at their own pace, and take ownership of their learning, promoting confidence and sustained motivation. While these tools promote student independence, the accuracy of self-assessment, especially among novice students, is not critically examined. There is also a lack of analysis on how students interpret and act on AI feedback, which is a key factor in making self-assessment pedagogically meaningful. Teacher mediation or guidance appears to be a missing but essential component in most implementations.

Adaptive testing

Five AI technologies were reported for adaptive testing. Pandarova, et al. (2019) employed a Language Proficiency Assessment tool that incorporated Dynamic Difficulty Adaptation (DDA) to predict item difficulty through linguistic analysis, thereby enabling semi-automated scoring and providing insights into learning target complexity. It could provide valuable insights into the relative difficulty of learning targets. Sargazi Moghadam, et al. (2023) combined adaptive learning platforms with sentiment analysis to adjust task difficulty based on students' emotional states (e.g. break activities into smaller chunks). Srinivasan and Murthy (2021) utilised the ReadToMe dashboard, a multisensory computer-assisted language learning (CALL) system that has been shown to enhance knowledge retention and improve reading comprehension scores by 20 – 40%. Additionally, AI tools like ChatGPT and other Generative AI systems contribute to adaptive testing by generating responsive, personalised assessments (Derakhshan & Ghiasvand, 2024). These technologies also support teachers in developing cost-effective, differentiated testing systems tailored to students' individual learning needs (Moorhouse & Kohnke, 2024).

Adaptive testing clearly supports personalised learning paths, but ethical concerns about data privacy, algorithmic transparency, and test fairness are notably absent. Additionally,

while emotional states were used to adapt tasks in some studies, these affective dimensions were not validated with psychological or pedagogical frameworks, raising questions about reliability and unintended consequences.

Immediate feedback delivery

Sixth, five AI technologies were used to deliver immediate feedback. Tools such as ReadToMe (Srinivasan & Murthy, 2021), MI Write (Wilson, et al., 2021), Grammarly (Wei, et al., 2023) and NLP-based applications such as ChatGPT (Derakhshan & Ghiasvand, 2024; Ghafouri, et al., 2024; Jamshed, et al., 2024; Zhang, 2023) provide personalised, timely feedback that supports students in revising their work. GenAI tools also assist teachers in offering detailed assessments on language use, including expressions and sentence patterns (Choi, et al., 2024). Additionally, Zhang and Han (2021) used a Chatbot to assist students in assessing and improving their vocabulary and grammar, thereby enhancing language accuracy and promoting student autonomy.

Despite AI's promising features for immediate feedback, these feedback systems often lack context sensitivity. They identify linguistic issues, but not content-specific or discipline-based conventions. Furthermore, there is limited critical analysis of how students use the feedback. The absence of teacher feedback integration in most studies is also problematic.

Speech recognition

Zhang and Han (2021) demonstrated that AI chatbots using speech recognition helped students improve vocabulary and grammar. Although the technology is promising, its analytical validity, whether it accurately assesses speaking proficiency across varied accents and dialects, is not well addressed. The broader pedagogical value of such tools depends on their cultural inclusivity and responsiveness to diverse speech patterns, which future research must interrogate.

2. What are the outcomes and challenges reported about using AI in English language assessment in schools?

Studies on integrating AI technologies in English language assessment report various outcomes and challenges. The six key themes are: adaptive practices, compliance and accountability, efficient assessment administration, enhancing motivation and engagement during assessment, and pedagogical functions.

Creating adaptive practices

First, evidence shows that AI technologies offer functionalities for creating adaptive instructional practices tailored to individual student profiles. Some AI technologies can generate personalised test items and lesson materials. AI-based tool employing a dynamic difficulty adaptation (DDA) framework to generate English grammar exercises. The tool adjusts the sequence and complexity

of tasks based on students' progress, matching the task difficulty with individual students' proficiency. These capabilities highlight the potential of AI to support differentiated instruction and assessment. In a related study, Ericsson and Johansson (2023) explored the integration of a spoken dialogue system (SDS) chatbot to scaffold language learning. They have shown that student-chatbot interactions can effectively support adaptive scaffolding by aligning feedback to students' prior knowledge and interests. Furthermore, generative AI tools such as ChatGPT present opportunities for developing cost-effective, personalised assessment systems. These tools not only assist teachers in designing adaptive rubrics but also promote greater objectivity in evaluation processes (Derakhshan & Ghiasvand, 2024). Collectively, these studies demonstrate the capacity of AI to personalise instructional support, thereby enhancing the precision and responsiveness of educational delivery.

Connecting both the automated generation of test items and learning materials, Lee, et al. (2023) underscored the potential of AI to automate Learner-Generated Context-based (LGCB) learning experiences. By using such system, students can pursue personalised goals by creating and using their English content. This student agency reflects the adaptive affordances of AI. Similarly, Srinivasan and Murthy (2021) found that an AI-based multisensory platform improved English reading and comprehension by introducing adaptive personalisation at both individual and group levels, enhancing instructional effectiveness without altering pedagogy or content. Additionally, AI technologies can scaffold learning by offering graduated support across tasks, functioning as a "more knowledgeable other" in Vygotsky's Zone of Proximal Development (ZPD), thereby promoting students' cognitive development and critical thinking (Liu & Wang, 2024).

More sophisticated AI technologies used for adaptive practices include automating the identification of students' emotions and determining adaptive learning paths based on the emotions. Modelling effect (i.e., considering a student's emotional or motivational state) should also be considered in designing learning activities (Chen & Lee, 2011). Using an algorithm framework utilising emotion ontology, Sargazi Moghadam, et al. (2023) demonstrated that Genetic Algorithm and Kot's emotion sets can help determine students' emotions during learning and recommend adaptive micro-break activities to students based on their preferences to return them to an optimal learning condition. This points to the critical role emotions play in the learning process, i.e., when students are in their optimal emotional status, they can learn better. Additionally, teachers can utilise micro-break activities to adjust their teaching approach and tailor it to students' emotional needs.

While AI technologies are promising for adaptive practices that enhance learning, measures are not in place to ensure the ethical use of AI. For example, for automated test items and learning materials development, a supportive atmosphere for active content sharing (OER) and Chatbot interaction, especially with younger students, is raised as a concern. Previous studies have emphasised the need for humans to lead interaction between students and the system, especially when managing the quality of content, verifying the accuracy of information, and even encouraging students to overcome challenges in assessments (Jeon, 2021; Lee, et al., 2023). Additionally, some technological issues may exist in the system, such as optimisation and

convenience problems with the passage and video wizards, difficulty operating some modules on a smartphone, and a nonintuitive UI for creating content and then a learning plan (Lee, et al., 2023; Srinivasan & Murthy, 2021).

Meeting accountability and compliance

AI-based technologies in assessment offer tools for generating reliable outcome measures aligned with accountability standards (Wei, et al., 2023; Sun, et al., 2022; Long & Lin, 2024).

Automatic essay assessment (AEA) systems, such as Response to Text Assessment (RTA), evaluate analytical writing and reading comprehension (Rahimi, et al. 2017). Similarly, MI Write, developed for English Language Arts (ELA) assessment, supports accountability by integrating demographic and state test data to produce reliable outcome measures (Wilson, et al., 2021). The system generates anonymised datasets, including survey responses and writing prompt data, enhancing transparency and compliance. Alongside RTA, MI Write exemplifies how AI technologies can standardise assessments and facilitate data-driven decision-making aligned with accountability standards. The designs of both RTA and MI Write demonstrate how the integration of AI can standardise and support data-driven accountability. However, scholars stress the need for ethical safeguards. Ng, et al. (2022) emphasise the importance of human-centred considerations (e.g., fairness, accountability, transparency, ethics) in guiding responsible AI use. Such principles address gaps in institutional guidelines for AI adoption in assessment (Moorhouse & Kohnke, 2024; Ghafouri, et al., 2024) and help mitigate concerns around data privacy and security (Jamshed, et al., 2024; Hockly, 2023).

Improving efficiency

Improving efficiency across various aspects of assessment is one of the outcomes of this review. AI technologies enhance teacher performance by automating grading, feedback, and evaluation tasks, improving efficiency in assessment and instruction (Zawacki-Richter, et al., 2019). AI Tools can significantly reduce workload and save time in lesson planning, marking, and proofreading, allowing teachers to focus on instructional quality (Derakhshan & Ghiasvand, 2024; Choi, et al., 2024; Tolstykh & Oshchepkova, 2024). These tools streamline processes, reduce delays, and deliver immediate, actionable feedback.

AI-based scoring systems such as Automated Essay Scoring and Automated Writing Evaluation have optimised writing instruction and assessment processes by automating feedback on grammar, vocabulary, structure, and content organisation. AWE tools offer immediate, personalised feedback, aiding faster learning and error correction. Liu et al. (2023) found AWE reduced teachers' workload in EFL contexts, while Zhang (2023) highlighted its efficiency in improving writing quality through feedback mechanisms and analytics capabilities.

Furthermore, such technology can reduce staffing and material resources, while also ensuring impartiality in scoring, especially in large-scale assessments. AI can enhance the efficiency of existing algorithms used in content and test generation. Bezirhan and von Davier (2023) suggest that utilising GPT-based models can reduce costs and streamline automated

passage creation for large-scale reading assessments. Peng, et al. (2023) found that combining ICT-related features with a Random Forest model enabled optimised predictions of students' reading outcomes in blended learning. Ghafouri, et al. (2024) note that tools like ChatGPT, improve instructor productivity by automating tasks such as essay scoring, thereby increasing overall instructional efficiency and supporting more scalable and responsive educational practices.

Despite the positive outcomes of AI-based English grading systems, studies tend to agree on the limitations of automated scoring. Primarily, the limitation of AI technologies in terms of contextual understanding and ethical considerations necessitates a balanced approach by integrating AI with human expertise. The overall quality of writing assessment and teaching is improved by leveraging the strengths of both AI and human judgment (Liu, et al., 2023). However, teachers should check the accuracy and reliability of AI-generated content (Bezirhan & von Davier, 2023).

Increasing student motivation and engagement

By offering personalised learning experiences, facilitating engagement with content, and providing real-time feedback, AI technologies have the potential to contribute to enhancing student motivation (Shaikh, et al., 2023; Shruti, et al., 2025).

Evidence from various studies demonstrates that AI-based technologies play a crucial role in enhancing student motivation through embedding different interactive and dynamic features (Liu & Wang, 2024; Shaikh, et al., 2023). First, user-friendly interfaces can enhance students' ability to understand and use AI technologies, making them more engaged. In AI writing systems such as the Automatic Scoring System for English Composition, picture-cued writing that combines visual stimulation and real-world situations helps students acquire a better understanding of language knowledge (Zhang, 2023; Zhang & Zou, 2021). Similarly, digital story writing with AI allows students to create AI-driven solutions by applying prior knowledge, observing authentic problems, and applying AI concepts in specific scenarios (Liu, et al., 2023; Wilson, et al., 2022).

Second, AI technologies enhance student motivation by accommodating diverse learning styles and supporting multidimensional skill development (Bewersdorff, et al., 2023; Lee, et al., 2022; Pandorava, et al., 2019; Rodriguez-Barrios, et al., 2021). User-friendly learning management systems and embedded videos encouraged students to study English independently with minimal effort (Lee, et al., 2022). This learner-generated content can promote active engagement.

Third, AI-based assessment technologies can actively engage students by adapting to individual learning characteristics. Rodriguez-Barrios, et al. (2021) employed a Bayesian network model to evaluate the relationship between language skills and students' learning styles, pace, speed, and reading comprehension. They found that AI-driven speed-reading tools increased the motivation of young students and improved their reading outcomes. Similarly, Pandorava, et al. (2021) emphasise the importance of adaptive selection and sequencing of exercise items in

ensuring that students receive the appropriate assessment stimulus. As such, applying a mechanism in assessment further increases achievement, motivation, engagement, and reduces test anxiety compared with non-adaptive tests (Moghadam, et al., 2024). These adaptive methods contribute to a more responsive and emotionally supportive learning environment (Long & Lin, 2024; Wang, 2024; Wiyaka, et al., 2024).

Lastly, providing real-time feedback fosters a supportive learning environment that enhances student motivation (Wei, et al., 2023; Zhao, et al., 2023). Wilson, et al. (2021) found that students using the AI-based writing tool MI Write improved their writing skills and motivation, particularly among younger students who responded more positively than older students. Similarly, AI-facilitated Peer Assessment (PA) encourages students to compare their work with that of their peers and provide constructive comments and suggestions, thereby boosting motivation and engagement (Liu, et al., 2023). During the peer assessment process, positive peer comments, especially on writing strengths, can inspire revision efforts (Hattie & Timperley, 2007; Shintani, 2016) and cultivate a collaborative climate that promotes higher-quality work and sustained engagement (Shih, 2011).

3. What recommendations or implications for practice and theory have been outlined by the extant literature for using AI in English language assessment in schools?

The studies reported in this paper outline key recommendations to further expand the evidence base related to the use of AI in English assessment. There are seven broad categories of recommendations, grouped into three broader themes as shown in Figure 1.

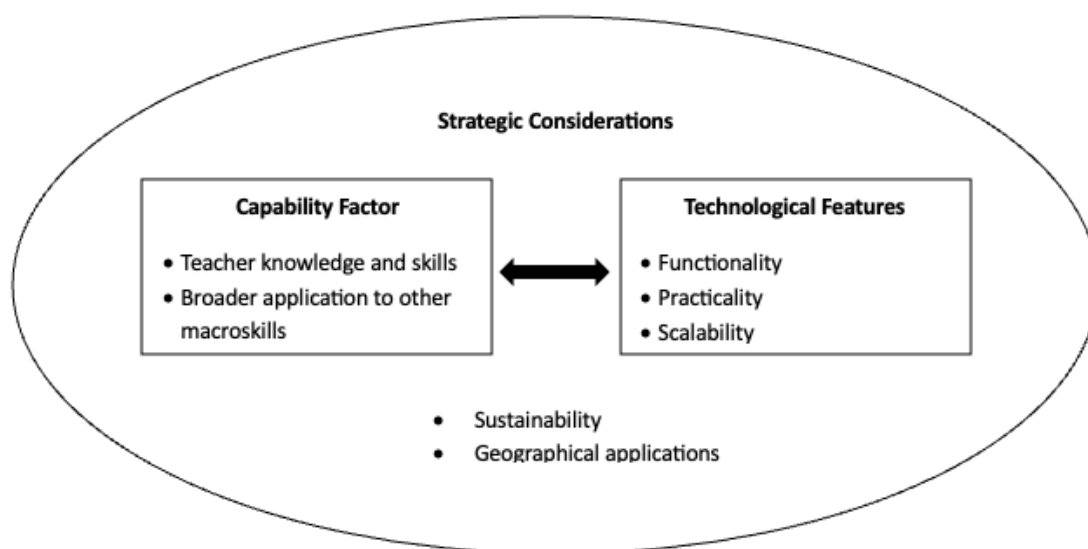


Figure 1: Key areas for further investigation to expand the evidence base of using AI in assessment

Broader application to other macros kills

A prominent recommendation across the sources is the need for broader application of AI tools

to encompass all language macro-skills (reading, writing, listening, and speaking), moving beyond singular foci to offer comprehensive support (Almegren, et al., 2025; Shaikh, et al., 2023). While large language models (LLMs) can support the automatic generation of stories and items, they are not yet fully equipped to assess complex reading skills (Bulut & Yildirim-Erbasli, 2022). LLMs enable teachers to quickly generate authentic materials, reducing the need to search for existing resources. However, these outputs often require human oversight, as they may contain semantic errors, text that appears to be grammatically correct but nonsensical (Moorhouse & Kohnke, 2024). Moreover, AI-generated items may not effectively measure higher-order reading skills such as inferencing, analysis, and critique. Thus, further research is needed to examine the long-term impacts of AI-assisted language learning, particularly chatbot-supported instruction, across all aspects of language acquisition (Wiyaka, et al., 2024).

Sustainability

Ensuring the sustainability of AI-based English language assessment practices requires a good balance between teacher expertise and technological innovation, cultivating a supportive school environment, and focusing on long-term implementation strategies. While AI tools can ease teacher workload and reduce stress, research highlights the necessity of maintaining human-AI collaboration in content creation and evaluation to ensure quality, alignment with learning goals, and factual accuracy (Almegren, et al., 2025; Bulut & Yildirim-Erbasli, 2022; Moorhouse & Kohnke, 2024).

Long-term sustainability also depends on embedding AI technologies within existing pedagogical practices rather than treating them as temporary innovations (Wei, et al., 2023; Wiyaka, et al., 2024). For instance, integrating AI-based reading comprehension tools into current curricula can minimise disruption and enhance usability (Srinivasan & Murthy, 2021). Similarly, Lee, et al. (2023) underscore the importance of a learner-generated context (LGC) approach, promoting student autonomy and active engagement with content and peers. However, such systems require ethical oversight, including protections for intellectual property, identity protection, and the promotion of responsible content sharing.

Supporting systems are vital to sustaining AI in educational settings, especially in resource-constrained environments (Gayed, et al., 2022). Scholars recommend the development of clear policies and guidance at the school level (Lee, et al., 2023; Liu & Wang, 2024; Long & Lin, 2024), along with long-term plans for professional development to equip teachers with necessary competencies and avoid resource fatigue (Bulut & Yildirim-Erbasli, 2022; Choi, et al., 2024; Ghafouri, et al., 2024). Peng, et al. (2023) further emphasise the need for stakeholder collaboration with school leaders, teachers, and parents to create environments conducive to effective AI integration.

Ethical considerations are central to sustainable AI use. Ng, et al. (2022) found that students' perceptions of AI use, particularly in creative tasks like story writing, are shaped by the ethical framing of these technologies. For broader sustainability, several sources advocate for the seamless integration of AI into established curricula, the implementation of pilot programs to

assess effectiveness, and the promotion of autonomous learning models (Lee, et al., 2022).

Practicability

The practicability of AI technologies in English assessment lies in harmonising technical functionality with pedagogical design to enhance student learning engagement (Choi, et al., 2024; Esfandiari & Allaf-Akbary, 2024; Tafazoli, 2024). A practical implementation requires adaptive frameworks that align AI tools with instructional goals and student needs (Jamshed, et al., 2024; Liu & Wang, 2024).

For example, Lee, et al.'s (2023) proposed a Learner-Generated Context model where AI platforms match students with relevant materials and peers based on shared interests. There is a proposed "content-sharing rule adjustments" that will be translated into user-friendly features like anonymous commenting, collaborative editing technology, and the ability to delete content at any time if the content has not yet been shared. However, they emphasised the integration of Open Educational Resources (OERs) to supplement learner-generated content and minimise copyright infringement concerns.

In writing assessment, Wilson et al. (2021) highlight that even AWE systems can provide multimodal feedback (e.g., written, audio, visual), thereby supporting diverse students and improving outcomes, teachers must balance accuracy with interpretability of feedback. They suggest scaffolding revision support and tailoring feedback to student needs, particularly for younger students. Teachers should guide them to act on feedback to improve their outcomes.

Scalability

AI-based technologies offer potential for content generation for assessment tasks, making them more efficient and objective. Bezirhan and von Davier (2023) used GPT-3, combined with well-designed prompts and human editing, to generate high-quality reading passages for assessment. This can significantly reduce the time and resources needed for item generation, increasing scalability in assessment development. This is consistent with Zhang (2023) recommendation that the ability of AI technologies to generate immediate feedback, consistent scoring, and detailed analytics can be scaled up to benefit teachers, students, and the entire education systems.

However, for effective use of AI, the AI-human interaction is emphasised for optimal scalability (Bezirhan & von Davier, 2023; Zhang, 2023). Human judgment is critically essential for ensuring the quality, fairness, and ethical considerations of AI-generated content and assessments. The balanced approach can maximise the strengths of both technologies, leading to sustainable and scalable assessment practices.

Knowledge and skills of teachers

Findings are limited in terms of what specific knowledge and skills teachers need to acquire to successfully implement AI technologies in English assessment. The focus of these studies is on the features and affordances of AI use. For example, Hannah, et al. (2022) found that teachers

need more awareness and training because existing automated speech recognition systems can measure reading skills such as phonological awareness and oral reading fluency that teachers could use to conduct universal screening, inform their teaching, and track students' progress. Similarly, Ericsson and Johansson (2023) recommend incorporating teachers' perspectives and utilising SDS dashboards to monitor students' progress in speaking skills. While information from this feature can provide helpful information for teachers to adjust their pedagogical support accordingly, the skills that teachers need are not explicitly discussed. Integrating ICT can be a challenge for teachers lacking proficiency, and they should be supported in developing digital competencies (Shruti, et al., 2025).

In addition, Wilson, et al. (2021) found that teachers may lack sufficient technological, pedagogical, and content knowledge (TPACK; Mishra & Koehler, 2006) to effectively implement AI technologies. Sufficient TPACK would enable teachers to understand how to enact instructional best practices that emphasise frequent practice (Graham, et al., 2012) with the affordances of AI. In addition, An, et al. (2022) found that teachers with higher Technological Knowledge about AI-based language applications tend to continue to use AI to support their teaching and assessment practices. Thus, teachers should be trained to leverage AI tools effectively, with professional development focusing on AI literacy (Liu & Wang, 2024).

Functionality

To improve the technical functionality and design, several studies recommend refining AI tools to better support students. This includes improving algorithms for differentiating assessment based on students' ability levels (Jamshed, et al., 2024) and monitoring cognitive and emotional responses to assessment to better support them (Long & Lin, 2024). In addition, further investigation of AI features such as word suggestion and reverse translation is needed (Gayed, et al., 2022)

Lee, et al. (2023) recommended that AI technologies should have content-sharing rules using OER and address technological issues such as graphics user interface and automatic text caption or translation. This will allow content creators to delete their content at any time if it has not yet been shared to address concerns among students unfamiliar with sharing their work. OERs could be helpful in promoting active content sharing because they carry a lower risk of copyright infringement concerning content use.

Technological issues of the system: optimisation and convenience problems with the passage and video wizards; difficulty operating some modules on a smartphone; a nonintuitive UI for creating content and then a learning plan; and auto-captioning functions not transcribing English conversations accurately. To solve these issues, it is necessary to improve the performance of the wizards, develop a more mobile-friendly UI, and enhance the interface design of the learning management module. Furthermore, the system can benefit from a human assistant who can help correct errors in automatically generated captions or text translation.

Geographical applications

One interesting finding relates to geographical gaps. Over the last ten years of research on the

intersection between AI and English assessment, relatively large studies were reported in 2023 ($n = 12$), followed by 2024 ($n = 7$), 2022 ($n = 4$), and 2021 ($n = 5$). In previous years, only one to two studies were reported. There is also evidence for geographical gaps as most studies were from China ($n = 6$), USA ($n = 4$), Iran ($n = 4$), Taiwan ($n = 2$), Hong Kong ($n = 2$), India ($n = 2$), and one each from Canada, Germany, Israel, Korea, Japan, Malaysia, Mexico, Pakistan, Sweden, and UAE. The absence of studies from other countries suggests that the use of AI for assessment purposes is not prevalent in certain regions. Studies across many countries are needed to explore the contextual factors that influence the use of AI for English assessment.

Discussion and implications

Our paper aims to explore the use of AI technologies in English language assessment. Using a scoping review to answer our research questions, we argue that AI is reshaping English language assessment by enhancing efficiency and enabling new pedagogical possibilities. Five noteworthy findings are presented.

First, our paper has provided an overview of the AI technologies used in English language assessment. The seven broad categories or typology (automated test item generation, conversation agent, marking and scoring, self-assessment, adaptive testing, giving immediate feedback, and speech recognition) contribute to our understanding of this critical area of inquiry. This finding broadens the literature on the role of AI in assessment, which previous review studies have identified as two major categories of the role of AI in assessment: automatic scoring and predicting student performance (Chiu, et al., 2023), Martínez-Comesaña, et al., 2023) and automatic grading and giving immediate feedback (González-Calatayud, et al., 2021). Our work offers a more explicit typology of AI used in English language assessment. However, the use of AI in English language assessment should be further theorised and underpinned by educational and assessment theories to ensure its learning and teaching functions. The intersections of AI, assessment, and educational theories will optimise the impact of using AI in English language assessment, ensuring effective learning and teaching, and may increase the rigour and acceptability of assessment results for other functions, including compliance, accountability, and higher-level decision-making. This intersection may also facilitate the scalability (Bezirhan & von Davier, 2023; Zhang, 2023) and sustainability (Lee, et al., 2023; Peng, et al., 2023; Rodriguez-Barrios, et al., 2021; Srinivasan, 2021) of AI because teachers and school leaders will see the critical roles of AI for improving learning and teaching.

Second, there is a strong consensus among studies that utilising AI technologies in English language assessment yields high accuracy and efficiency. The quality of prompts generated is comparable to that of teacher-generated prompts in terms of coherence, appropriateness, and readability (Bulut & Yildirim-Erbasli, 2022; Kohnke, et al., 2023). Also, AI technologies can generate a variety of high-quality assessment items (Chompooyod, et al., 2023; Kumar & Boulanger, 2020; Rodriguez-Barrios, et al., 2021). However, the uptake of AI technologies for English language assessment is relatively low due to their perceived low trustworthiness. The human-technology interface was highlighted as a critical factor for ensuring

its effectiveness and trustworthiness. However, this interface currently needs improvement due to teachers' limited knowledge, skills, and dispositions regarding AI. An, et al. (2022) emphasise the role of teachers' beliefs as influencing factors that can enhance the use of AI. They suggest that promoting the understanding of AI's usefulness among teachers can improve their behavioural intention to integrate AI in learning, teaching, and assessment. They noted that if teachers have a strong belief in the usefulness of AI in facilitating meaningful interactions among students and teachers, regardless of time and place, and can also help reduce student anxiety, they are most likely to use it. The features and capabilities of AI that meet teachers' expectations of how AI can help improve their teaching efficiency (Performance Expectancy) influence teachers' behavioural intention and willingness to adopt AI.

Third, the outcomes of using AI technologies in English language assessment provide evidence for its capabilities in supporting or leveraging teachers' assessment practices. As such, AI technologies can enhance adaptive practice (e.g., Derakhshan & Ghiasvand, 2024; Ericsson & Johansson, 2023; Lee, et al., 2022; Moghadam, et al., 2023; Pandorava, et al., 2019; Srinivasan, 2021), meet compliance and accountability requirements (e.g., Ng, et al., 2022; Rahimi, et al., 2017; Safi, et al., 2023; Wei, et al., 2023; Wilson, et al., 2021), ensure efficient administration of assessment (e.g., An, et al., 2022; Bezirhan & von Davier, 2023), enhance motivation and engagement (e.g., Lee, at al., 2022; Liu & Wang, 2024; Moghadam, et al., 2023; Ng, et al., 2022; Pandorava, et al., 2019; Rahimi, et al., 2017; Rodriguez-Barrios, et al., 2021), and enhance pedagogical approaches (e.g., Ericsson & Johansson, 2023; Jeon, 2021; Kumar & Boulanger, 2020; Lee, at al., 2022; Liu, et al., 2021; Moghadam, et al., 2023; Shruti, et al., 2025). The importance of these outcomes in learning and teaching highlights the potential of AI technologies to support teachers, students, and school leaders in ensuring effective learning and teaching.

Fourth, we have synthesised the recommendations offered by the reviewed articles to further expand the theoretical and practical knowledge of using AI in English language assessment. There is a strong recommendation for utilising existing AI technologies for other macro-skills ($n = 12$). Currently, most AI technologies reported are used in either one or two macro skills: reading ($n = 7$), writing ($n = 12$), grammar ($n = 3$), vocabulary ($n = 2$), and speaking ($n = 4$). The application of AI technologies in assessing various macroskills provides evidence of their wide-ranging application and potential for scalability (Bezirhan & von Davier, 2023; Wei, et al., 2023; Zhang, 2023). However, no study reported the use of AI in listening skills. This apparent gap in the application of AI in English assessment presents an opportunity for further study. The limited application of one AI technology to a single macro skill implies the need to subscribe to multiple AI technologies for a comprehensive English language assessment. There is also a recommendation relating to improving the functionalities of existing AI technologies. In addition, most studies are either pilot studies or single-stage studies, and hence, the sustainability of using AI in English assessment needs further exploration. Additionally, there are no longitudinal studies, and most reported studies are quantitative. Moreover, most studies are done using one or a few classes, and hence, the scalability of using AI technologies for assessment is inconclusive. Furthermore, there is a strong emphasis on enhancing teachers' AI knowledge and skills to ensure

their acceptability and effective integration into learning, teaching, and assessment activities. The issues relating to sustainability and scalability hinge on teachers' capability to use AI.

Fifth, the ethical implications and risks associated with using AI in English assessment require rigorous investigation. The use of AI raises a wide range of ethical issues, including concerns about data privacy and the dissemination of inappropriate content. Policy development and enactment regarding the use of AI in English assessment are needed to ensure that AI technologies are used ethically and legally. Therefore, identifying all stakeholders and fostering their collaborative engagement are crucial to ensuring the ethical deployment of AI in education (Klimova, et al., 2023).

Conclusion

Using a scoping review, we synthesised findings from 34 peer-reviewed studies (2011 – 2025) on the use of AI in English language assessment. These studies reveal diverse applications, including automated test item generation, conversational agents, marking and scoring, self-assessment, adaptive testing, real-time feedback, and speech recognition. We have also identified the reported outcomes of using AI technologies, including fostering adaptive practices, complying with accountability standards, increasing efficiency, enhancing learning motivation, supporting pedagogy, meeting compliance and accountability requirements, ensuring the efficient administration of the assessment, supporting pedagogical practices, and enhancing motivation and engagement during assessments. We also synthesised the recommendations made by extant studies, including their application to other micro-skills, sustainability, scalability, teachers' AI knowledge and skills, functionalities, and geographical applications. Taking these results as a body of knowledge, it is evident that AI technologies are reconfiguring the landscape of English assessment.

While our findings present critical insights into the current use of AI in English language assessment, several important areas warrant further exploration. First, future studies should investigate how the integration of AI is reshaping educational authority and pedagogical relationships, particularly as evaluative control shifts from teachers to algorithmic systems. Research is needed to examine how this transition affects both teacher agency and student experience, especially for students whose linguistic identities, learning styles, or cultural backgrounds may not align with data-driven models. Second, the evolving teacher–student–assessor dynamic, increasingly mediated by automated systems, raises questions about the role of teachers in designing versus delivering assessment. Future research should explore the impact of AI on pedagogical values, such as dialogic feedback, formative assessment, and holistic learning, and whether these are sustained or undermined in AI-supported contexts. Third, to ensure that AI promotes educational equity and justice, studies should focus on ethical design principles, including transparency, inclusivity, and explainability. Further systematic reviews should also broaden the methodological approach by including grey literature, policy documents, and practitioner perspectives to fully capture the spectrum of AI's impact on assessment practices. Finally, the long-term sustainability of AI in assessment should be


evaluated not only in terms of technical functionality but also in terms of its potential to enhance, rather than replace, equitable and human-centred teaching and learning. Longitudinal studies and system-wide evaluations are needed to assess how AI can support inclusive pedagogy over time and at scale. Finally, another study that warrants investigation is the theoretical framing of using AI for assessment purposes, as the paper tends to focus more on practical applications than on underlying conceptual models. Exploring robust theoretical frameworks such as assessment theory, learning analytics, or sociocultural perspectives could provide deeper insights into how AI reshapes notions of validity, reliability, and fairness in assessment. Such theoretical grounding would also help clarify the role of teachers, ethical considerations, and the pedagogical implications of integrating AI tools in diverse educational contexts.


In terms of limitations of our study, although careful attention was paid to searching the databases for relevant studies, one limitation of our process is the deliberate exclusion of review papers, opinion papers, reports, theses, and grey literature. In our effort to include only peer-reviewed papers, which are perceived as more rigorous, we may have missed other insights reported in the excluded categories. Also, our keywords might have limited our search. Papers that do not use "artificial intelligence" might have been excluded from the databases. Hence, for future review papers, the keywords should be expanded to include AI-related terminologies and specific AI technologies.


Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, we utilised ChatGPT to ensure the coherence of our arguments. After using this tool, we reviewed and edited the content as needed. We take full responsibility for the content of the publication.

Author Biographies

Dennis Alonzo researches the intersections of curriculum, assessment, equity, evaluation of educational programs, and teacher education and development. He works with educational systems and schools nationally and internationally to lead their assessment reforms focused on articulating policies, developing assessment resources, implementing professional development, and transforming teachers' beliefs and practices. 

Jan Michael Vincent N. Abril is an MRes Education candidate at UNSW Sydney, supported by the University International Postgraduate Scholarship. A former UK Chevening scholar with a Master's from UCL IOE, his research models teachers' digital and AI competencies using large-scale datasets (e.g., PISA), framed by TPACK and Digital Capital Theory. 

Cherry Zin Oo is a lecturer at Myanmar Imperial College and a PhD graduate from the University of New South Wales (UNSW), Australia. Her research focuses on teacher education, assessment literacy, professional development, and large-scale assessment. 

References

- Alam, A. 2022. Employing adaptive learning and intelligent tutoring robots for virtual classrooms and smart campuses: Reforming education in the age of artificial intelligence. In Shaw, R.N., Das, S., Piuri, V. & Bianchini, M. *Advanced Computing and Intelligent Technologies*. Springer: Singapore, 395-406.
- Alghamdy, R.Z. 2023. Pedagogical and ethical implications of artificial intelligence in EFL context: A review study. *English Language Teaching*, 16(10): 87.
- *Almegren, A., Almegren, R.M., Hazaea, A.N., Mahdi, H.S. & Ali, J.K.M. 2025. AI powered ELT: Instructors' transformative roles and opportunities. *PLOS ONE*, 20(5): e0324910.
- Alonzo, D. 2020. Teacher education and professional development in Industry 4.0: The case for building a strong assessment literacy. In Ashadi, Priyana, J., Basikin, Triastuti, A. & Putro, N.H.P.S. (Eds.). *Teacher Education and Professional Development in Industry 4.0*. Taylor & Francis Group.
- Alonzo, D., Baker, S., Knipe, S. & Bottrell, C. 2023. A scoping study relating Australian secondary schooling, educational disadvantage and assessment for learning'. *Issues in Educational Research*, 33: 874-896.
- Alonzo, D. & Loughland, T. 2022. Variability of students' responses to assessment activities: The influence of achievement levels. *International Journal of Instruction*, 15(4): 1071-1090.
- *An, X., Chai, C.S., Li, Y., Zhou, Y., Shen, X., Zheng, C. & Chen, M. 2022. Modeling English teachers' behavioral intention to use artificial intelligence in middle schools. *Education and Information Technologies*, 28(5): 5187-5208.
- Baroudi, Z. 2007. Formative assessment: definition, elements and role in instructional practice. *Post-Script: Postgraduate Journal of Education Research*, 8(1): 37-48.
- Bennett, R.E. 2011. Formative assessment: a critical review. *Assessment in Education: Principles Policy and Practice*, 18(1): 5-25.
- *Bezirhan, U. & von Davier, M. 2023. Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, 5.
- Black, P. 2017. Assessment in science education. In Taber, K.S. & Akpan, B. (Eds.). *Science Education: An International Course Companion*. SensePublishers, 295-309.
- *Bulut, O. & Yildirim-Erbasli, S.N. 2022. Automatic story and item generation for reading comprehension assessments with transformers. *International Journal of Assessment Technology in Education*, 9(Special Issue): 72-87.
- *Chen, C.-M. & Lee, T.-H. 2011. Emotion recognition and communication for reducing second-language speaking anxiety in a web-based one-to-one synchronous learning environment. *British Journal of Educational Technology*, 42(3): 417-440.
- *Chomphooyod, P., Suchato, A., Tuaycharoen, N. & Punyabukkana, P. 2023. English grammar multiple-choice question generation using Text-to-Text Transfer Transformer. *Computers and Education: Artificial Intelligence*, 5.

- Chiu, T.K.F., Xia, Q., Zhou, X., Chai, C.S. & Cheng, M. 2023. Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 4: 100118.
- Cui, K. 2022. Artificial intelligence and creativity: Piano teaching with augmented reality applications. *Interactive Learning Environments*, 1-12.
<https://doi.org/10.1080/10494820.2022.2059520>
- Davison, C. & Michell, M. 2014. EAL assessment: What do Australian teachers want? *TESOL in Context*, 24(2): 51-72.
- *Derakhshan, A. & Ghiasvand, F. 2024. Is ChatGPT an evil or an angel for second language education and research? A phenomenographic study of research-active EFL teachers' perceptions. *Computer Assisted Language Learning*.
- Ellis, N. J., Alonzo, D. & Nguyen, H.T.M. 2020. Elements of a quality pre-service teacher mentor: A literature review. *Teaching and Teacher Education*, 92: 103072.
- *Ericsson, E. & Johansson, S. 2023. English speaking practice with conversational AI: Lower secondary students' educational experiences over time. *Computers and Education: Artificial Intelligence*, 5.
- *Gayed, J.M., Carlon, M.K.J., Oriola, A.M. & Cross, J.S. 2022. Exploring an AI-based writing assistant's impact on English language learners. *Computers and Education: Artificial Intelligence*, 3: 100055.
- *Ghafouri, F., Sahragard, R. & Meihami, H. 2024. From virtual assistant to writing mentor: Exploring the impact of a ChatGPT-based writing instruction protocol on EFL teachers' self-efficacy and learners' writing skill. *System*, 118: 103203.
- González-Calatayud, V., Prendes-Espinosa, P. & Roig-Vila, R. 2021. Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, 11(5467).
- *Hannah, L., Kim, H. & Jang, E.E. 2022. Investigating the effects of task type and linguistic background on accuracy in automated speech recognition systems: Implications for use in language assessment of young learners. *Language Assessment Quarterly*, 19(3): 289-313.
- Hattie, J. 2008. *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement*. Hoboken: Routledge.
- Hwang, G.-J., Xie, H., Wah, B. W. & Gašević, D. 2020. Vision, challenges, roles and research issues of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1: 100001.
- *Jamshed, A., Rehman, M. & Younas, M. 2024. The impact of ChatGPT on English language learners' writing skills: An assessment of AI feedback on mobile. *Education and Information Technologies*.
- *Jeon, J. 2021. Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, 36(7): 1338-1364.
- Klimova, B., Pikhart, M. & Kacetl, J. 2023. Ethical issues of the use of AI-driven mobile apps for education. *Frontiers in Public Health*, 10.

- *Kumar, V. & Boulanger, D. 2020. Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education*, 5.
- Lampropoulos, G. 2023. Augmented reality and artificial intelligence in education: Toward immersive intelligent tutoring systems. In Geroimenko, V. (Ed.). *Augmented Reality and Artificial Intelligence: The Fusion of Advanced Technologies*. Springer Nature Switzerland, 137-146.
- *Lee, D., Kim, H.H. & Sung, S.H. 2022. Development research on an AI English learning support system to facilitate learner-generated-context-based learning. *Educational Technology Research and Development*, 71(2), 629-666.
- *Liu, C.-C., Liu, S.-J., Hwang, G.-J., Tu, Y.-F., Wang, Y. & Wang, N. 2023. Engaging EFL students' critical thinking tendency and in-depth reflection in technology-based writing contexts: A peer assessment-incorporated automatic evaluation approach. *Education and Information Technologies*, 28(10): 13027-13052.
- * Liu, W. & Wang, Y. 2024. The effects of using AI tools on critical thinking in English literature classes among EFL learners: An intervention study. *European Journal of Education*, 59(4): e12804.
- Loughland, T. & Alonzo, D. 2019. Teacher adaptive practices: A key factor in teachers' implementation of assessment for learning. *Australian Journal of Teacher Education*, 44(7).
- Martínez-Comesaña, M., Rigueira-Díaz, X., Larrañaga-Janeiro, A., Martínez-Torres, J., Ocaranza-Prado, I. & Kreibel, D. 2023. Impact of artificial intelligence on assessment methods in primary and secondary education: Systematic literature review. *Revista de Psicodidáctica (English Ed.)*, 28(2): 93-103.
- Minn, S. 2022. AI-assisted knowledge assessment techniques for adaptive learning environments. *Computers and Education: Artificial Intelligence*, 3: 100050.
- *Moghadam, T.S., Darejeh, A., Delaramifar, M. & Mashayekh, S. 2024. Toward an artificial intelligence-assisted language learning ecosystem: Affordances, challenges, and future directions. *Education and Information Technologies*.
- *Moorhouse, B.L. & Kohnke, L. 2024. The effects of generative AI on initial language teacher education: Opportunities and challenges. *System*, 121: 103239.
- *Nazaretsky, T., Ariely, M., Cukurova, M. & Alexandron, G. 2022. Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, 53(4): 914-931.
- Ng, D.T.K., Luo, W., Chan, H.M.Y. & Chu, S.K.W. 2022. Using digital story writing as a pedagogy to develop AI literacy among primary students. *Computers and Education: Artificial Intelligence*, 3.
- Oo, C.Z., Alonzo, D. & Asih, R. 2022. Acquisition of teacher assessment literacy by pre-service teachers: A review of practices and program designs. *Issues in Educational Research*, 32(1): 352-373.
- Oo, C.Z., Alonzo, D., Asih, R., Pelobillo, G., Lim, R., San, N.M.H. & O'Neill, S. 2023. Implementing school-based assessment reforms to enhance student learning: A systematic review.

- Educational Assessment, Evaluation and Accountability*. <https://doi.org/10.1007/s11092-023-09420-7>
- Oo, C.Z., Alonzo, D. & Davison, C. 2023. Using a needs-based professional development program to enhance pre-service teacher assessment for learning literacy. *International Journal of Instruction*, 16(1): 781-800.
- *Ormerod, C., Lottridge, S., Harris, A.E., Patel, M., van Wamelen, P., Kodeswaran, B., Woolf, S. & Young, M. 2022. Automated short answer scoring using an ensemble of Neural Networks and Latent Semantic Analysis Classifiers. *International Journal of Artificial Intelligence in Education*, 33(3): 467-496.
- Pandarova, I., Schmidt, T., Hartig, J., Boubekki, A., Jones, R.D. & Brefeld, U. 2019. Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *International Journal of Artificial Intelligence in Education*, 29(3): 342-367.
- *Peng, Y., Wang, Y. & Hu, J. 2023. Examining ICT attitudes, use and support in blended learning settings for students' reading performance: Approaches of artificial intelligence and multilevel model. *Computers & Education*, 203.
- *Rahimi, Z., Litman, D., Correnti, R., Wang, E. & Matsumura, L.C. 2017. Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education*, 27(4): 694-728.
- *Rodriguez-Barrios, E.U., Melendez-Armenta, R.A., Garcia-Aburto, S.G., Lavoignet-Ruiz, M., Sandoval-Herazo, L.C., Molina-Navarro, A. & Morales-Rosales, L.A. 2021. Bayesian approach to analyse reading comprehension: A case study in elementary school children in Mexico. *Sustainability*, 13(8).
- Sadler, D. 1989. Formative assessment and the design of instructional systems. *Instructional Science*, 18(2): 119-144.
- Safi, M. F., Al Sadrani, B. & Mustafa, A. 2023. Virtual voice assistant applications improved expressive verbal abilities and social interactions in children with autism spectrum disorder: a Single-Subject experimental study. *Int J Dev Disabil*, 69(4): 555-567.
- *Sargazi Moghadam, T., Darejeh, A., Delaramifar, M. & Mashayekh, S. 2023. Toward an artificial intelligence-based decision framework for developing adaptive e-learning systems to impact learners' emotions. *Interactive Learning Environments*, 1-21. <https://doi.org/10.1080/10494820.2023.2188398>
- *Srinivasan, V., & Murthy, H. (2021). Improving reading and comprehension in K-12: Evidence from a large-scale AI technology intervention in India. *Computers and Education: Artificial Intelligence*, 2. <https://doi.org/10.1016/j.caeai.2021.100019>
- Stacey, M., Wilson, R. & McGrath-Champ, S. 2022. Triage in teaching: The nature and impact of workload in schools. *Asia Pacific Journal of Education*, 42(4): 772-785.
- Su, J., Ng, D.T.K. & Chu, S.K.W. 2023. Artificial intelligence (AI) literacy in early childhood education: The challenges and opportunities. *Computers and Education: Artificial Intelligence*, 4: 100124.

- Sun, Z., Anbarasan, M. & Praveen Kumar, D. 2021. Design of online intelligent English teaching platform based on artificial intelligence techniques. *Computational Intelligence*, 37(3): 1166-1180.
- *Tafazoli, D. 2024. Exploring the potential of generative AI in democratizing English language education. *Computers and Education: Artificial Intelligence*, 7: 100275.
- *Wei, P., Wang, X. & Dong, H. 2023. The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. *Frontiers in Psychology*, 14: 1249991.
- William, D. 2011. *Formative Assessment: Definitions and Relationships*. Institute of Education, University of London.
- *Wilson, J., Huang, Y., Palermo, C., Beard, G. & MacArthur, C.A. 2021. Automated feedback and automated scoring in the elementary grades: Usage, attitudes, and associations with writing outcomes in a districtwide implementation of MI write. *International Journal of Artificial Intelligence in Education*, 31(2): 234-276.
- *Xia, Q., Chiu, T., Chai, C.S., & Xie, K. 2023a. The mediating effects of needs satisfaction on the relationships between prior knowledge and self-regulated learning through artificial intelligence chatbot. *British Journal of Educational Technology*, 54(4): 967-986.
- *Xia, Q., Chiu, T.K.F. & Chai, C.S. 2023b. The moderating effects of gender and need satisfaction on self-regulated learning through Artificial Intelligence (AI). *Education and Information Technologies*, 28(7): 8691-8713.
- Zawacki-Richter, O., Marín, V.I., Bond, M. & Gouverneur, F. 2019. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1).
- *Zhang, F. 2023. Design and application of an automatic scoring system for English composition based on artificial intelligence technology. *International Journal of Advanced Computer Science and Applications*, 14(8): 195-205.
- *Zhang, H. & Han, X. 2021. Influence of vocalized reading practice on English learning and psychological problems of middle school students. *Front Psychol*, 12: 709023.
- *Zhao, R., Zhuang, Y., Zou, D., Xie, Q. & Yu, P.L.H. 2023. AI-assisted automated scoring of picture-cued writing tasks for language assessment. *Education and Information Technologies*, 28(6): 7031-7063.