CURRENT ISSUES IN ENGLISH LANGUAGE ASSESSMENT: HOW TO DEAL WITH THE CHALLENGES

Yusti Arini[1]
[1] *English Language Education, Cultures and Languages Faculty, UIN Raden Mas Said Surakarta, Indonesia*
Corresponding author: yusti.arini@staff.uinsaid.ac.id

Abstract
Language assessment is undergoing significant transformation due to rapid technological advancements, evolving educational paradigms, and increasing demands for equitable and authentic evaluation practices. This paper reviews current issues in language assessment, especially in English learning, by examining key challenges and emerging trends in educational contexts. The discussion focuses on five major issues: the integration of artificial intelligence in assessment, the implementation of formative and performance-based assessment, concerns regarding validity and reliability in digital testing, fairness and inclusivity for diverse learners, and the ethical implications of automated scoring systems. The paper also explores the impact of online learning environments on assessment design and the growing emphasis on assessing communicative competence and higher-order thinking skills rather than isolated linguistic knowledge. Through a review of recent literatures, the study highlights the need for assessment practices that balance technological innovation with pedagogical principles and ethical responsibility. Findings suggest that while digital technologies enhance efficiency, accessibility, and personalized feedback, they also introduce challenges related to data privacy, algorithmic bias, academic integrity, and unequal access to technological resources. The paper concludes that future language assessment should adopt a learner-centered, evidence-based, and context-sensitive approach that promotes fairness, authenticity, and continuous improvement while preparing learners for communication in increasingly diverse and digital global environments.
Keywords: language performance-based assessment, artificial intelligence-based assessment, validity, algorithmic bias, academic integrity

INTRODUCTION

Language proficiency assessment is a crucial component of the language learning process, especially in learning English, because it serves to determine students' level of mastery of the language competencies they have learned. Assessment is not only used to assign grades, but also to obtain information about the development of listening, speaking, reading, and writing skills and to support decision-making in the learning process. Therefore, the quality of the assessment significantly determines the quality of information obtained by teachers, lecturers, educational institutions, and other stakeholders. A good language proficiency assessment must adhere to the principles of validity, reliability, fairness, practicality, authenticity, and washback so that the assessment results truly represent students' language abilities.

The development of digital technology has brought about significant changes in language proficiency assessment practices. Previously largely conducted conventionally, assessments are now increasingly evolving toward online assessment, computer-based

assessment, mobile assessment, and various forms of technology-based assessment. These changes offer a number of advantages, such as flexibility in test administration, time efficiency, ease of data management, rapid feedback, and the possibility of using more interactive and authentic tasks. However, digital transformation also presents new challenges, particularly related to infrastructure readiness, the digital competence of students and teachers, data security, access to technology, and equality of opportunity in assessment.

One of the most prominent issues in language proficiency assessment today is the increasing use of Artificial Intelligence (AI) and Generative Artificial Intelligence (GenAI). AI technology has been utilized in various stages of assessment, from item creation and test administration to feedback delivery, to automated assessment of writing and speaking skills. Recent studies indicate that research on AI in language assessment is growing rapidly, with the use of AI focusing heavily on automated scoring of speaking and writing skills. However, the study also revealed gaps in the validation of AI use, particularly in ensuring that AI systems actually measure the language ability constructs they are intended to measure (O'Grady et al., 2026).

The use of AI in assessment also raises issues regarding the validity and reliability of scores. Automated assessment systems do have the potential to improve the efficiency and consistency of assessments, especially in large-scale assessments. However, the agreement between AI-generated scores and human assessments is not always consistent across contexts. A recent meta-analysis of English language proficiency assessments showed that, on average, there was no significant difference between AI and human scores, but there was significant heterogeneity across studies. This suggests that the effectiveness of AI in assessment is strongly influenced by the type of AI system, the proficiency level of the learners, the number of human assessors, the assessment instrument, and the context in which it is used. Therefore, AI is more appropriately viewed as part of an assessment quality assurance system rather than as a complete replacement for human assessors (Li et al., 2026).

In addition to validity and reliability, fairness and equity are important issues in AI-based assessments. Algorithms can introduce bias if they are developed using data that does not adequately represent the diversity of languages, cultures, accents, or

backgrounds of learners. This situation has the potential to disadvantage learners who speak certain language varieties or come from different linguistic backgrounds. Research on automated scoring also shows that bias based on language background can emerge in automated assessment systems. Therefore, the development and use of AI in language assessment must consider fairness, transparency, explainability of assessment mechanisms, and the possibility of algorithmic bias (Andersen et al., 2025).

Another increasingly important issue is academic integrity and the authenticity of language skills. The advent of GenAI allows learners to generate text, correct grammar, translate, summarize, and even receive assistance in writing the answers. This situation challenges traditional concepts of authorship, originality, and authentic language proficiency. In this context, the output of written assignments cannot always be viewed as a direct representation of a learner's language ability, as the production process may involve assistance from AI technology. Consequently, educators need to redesign assessment methods by prioritizing the learning process, higher-order thinking skills, reflection, interaction, and authentic performance, rather than focusing solely on the final product. UNESCO also emphasizes that the rise of Generative AI (GenAI) compels educational institutions to re-examine what is learned, how learning takes place, and how that learning is assessed and validated (Miao, Fengchun, UNESCO, Holmes, Wayne, 2023).

Furthermore, technological advancements pose challenges to the roles of teachers and assessors. Over-reliance on automated systems can diminish the scope for human professional judgment, particularly when assessments involve complex aspects such as creativity, pragmatic competence, communication skills in specific contexts, and authentic language use. Therefore, a human-centered assessment approach is becoming increasingly relevant. Recent research highlights the importance of balancing technological capabilities with human professional expertise to uphold validity, fairness, interpretability, and ethical standards in AI-based language assessment (Jang & Sawaki, n.d.).

Moreover, issues regarding privacy and data security require attention. Digital-based assessments can generate vast amounts of learner data, including written responses, audio recordings, videos, personal identities, and information on learning patterns. The

use of such data by digital platforms and AI systems raises questions regarding data ownership, storage, usage, and the extent to which learners are informed about how their data is utilized. UNESCO underscores the importance of data protection, a human-centered approach, inclusivity, equity, and cultural and linguistic diversity in the implementation of GenAI within education (Miao, Fengchun, UNESCO, Holmes, Wayne, 2023).

Given these developments, it is evident that the issue of language proficiency assessment in the digital era is no longer solely about creating valid and reliable tests; it also entails ensuring that the assessment process and outcomes remain fair, authentic, transparent, secure, inclusive, and relevant to technological advancements. Systematic reviews of GenAI in language assessment indicate that current key challenges include validity, reliability, fairness, practicality, ethics, and the sustainability of the technology's implementation across various assessment contexts. Researchers also caution that the integration of GenAI should not be driven merely by technological trends but must be grounded in pedagogical needs and evidence regarding its effectiveness (Chuang & Yan).

Consequently, examining current issues in language proficiency assessment is crucial for understanding the shifting assessment paradigm in the digital era. Educators, test developers, researchers, and educational institutions need the ability to critically evaluate assessment technologies—not solely based on efficiency and ease of use, but also on principles of quality measurement and assessment. This paper discusses various contemporary issues in language proficiency assessment, particularly those concerning the evolution of online assessment and AI, the validity and reliability of automated assessment, fairness and bias, academic integrity, assessment authenticity, data privacy, the digital divide, and the professional and ethical challenges educators face when designing language assessments in the technology era. In detail, this paper tries to answer four questions: 1) What are the current issues in language proficiency assessment? 2) How do advancements in AI and digital technology influence language assessment practices? 3) What challenges are encountered in the implementation of technology-based language assessment? 4) What strategies can be used to develop valid, reliable, fair, and authentic language assessments in the digital era?

By exploring current issues in language assessment and its challenges, this paper can contribute to the enhancement of readers' knowledge concerning the topic. Practically, the results of the literature review can become references for those having authority in making policies. Ultimately, this paper's contribution lies in proposing the perspective that the future of language proficiency assessment is not merely a transformation from conventional to digital or automated testing, but rather a paradigm shift toward assessment that is more authentic, adaptive, inclusive, transparent, and learning-oriented. The paper reinforces the argument that technological innovation must be balanced with principles of validity, reliability, fairness, authenticity, and ethics. Therefore, future research needs to develop validation models for AI-based assessments, examine the fairness of automated systems across diverse linguistic and cultural groups, explore assessments of multimodal skills and low-resource languages, and formulate the new competencies learners require to use AI critically and responsibly in language-based communication.

LITERATURE REVIEW

2.1 The Concept of Language Proficiency Assessment

Language proficiency assessment is a systematic process of gathering, analyzing, and interpreting information regarding an individual's ability to use language. In the context of language learning, assessment encompasses various competencies—such as listening, speaking, reading, and writing—as well as linguistic aspects like vocabulary, grammar, pronunciation, and pragmatic competence. Assessment serves not only to determine learners' levels of achievement but also to provide information that can be used to improve the learning process and offer feedback (Ozoda, et al., 2024).

High-quality language proficiency assessment must be grounded in robust measurement principles. Validity is a primary principle, as it concerns the extent to which the interpretation and use of scores can be justified. Reliability relates to the consistency of assessment results, while fairness concerns providing learners with equal opportunities to demonstrate their abilities, free from the influence of factors irrelevant to the construct being measured. Furthermore, language assessment must also consider authenticity—the degree to which assessment tasks reflect language use in real-world situations (Andersen et al., 2025).

Technological advancements have transformed the concept of language assessment. Assessment is no longer confined to conventional written tests but has evolved to include computer-based assessment, online assessment, mobile assessment, computer-adaptive testing, automated assessment, and various forms of digital assessment. These changes offer opportunities for more flexible and efficient assessment but simultaneously introduce new challenges regarding quality, fairness, security, and the interpretation of results (Grab, 2025).

2.2 Online Assessment in Language Ability Assessment

Online assessment is an important development in assessing language skills. This system allows language tests and assignments to be carried out via digital platforms so that students can carry out assessments more flexibly. In language assessment, digital technology allows the use of various forms of stimulus such as audio, video, images, interactive text, as well as communication simulations that are closer to authentic language use (Dang & Habók, 2026).

The advantages of online assessment lie in administrative efficiency, flexibility in time and place, automatic data collection, and the ability to provide feedback quickly. Technology also allows test developers to use student response data to analyze error patterns and skill development in more detail. In certain contexts, digital systems even enable adaptive assessments that adjust the difficulty level of questions based on student responses.

However, online assessment also faces various challenges. The quality of the assessment can be influenced by the availability of devices, the quality of the internet connection, digital literacy, and the environment in which students take the test. Differences in access to technology can create a digital divide so that assessment results do not fully reflect students' language abilities. Thus, technology in assessment needs to be seen as a tool that must fulfill the principles of fairness and accessibility, not just as an instrument to increase efficiency (Zhou, 2022).

2.3 Artificial Intelligence and Generative AI in Language Assessment

One of the most significant issues in the current development of language assessment is the use of Artificial Intelligence (AI), especially Generative AI (GenAI) and Large Language Models (LLMs). AI has been used in various aspects of assessment,

such as creating questions, administering tests, analyzing responses, automated scoring, providing feedback, and interpreting assessment results. A scoping review of 212 studies published between 2019 and 2024 shows that the use of AI in language assessment is growing rapidly, with dominant attention being paid to automated scoring of speaking and writing skills. However, research on validating the use of AI and its application to underrepresented receptive skills and languages remains relatively limited (Yang & Wang, 2025).

GenAI also changes the relationship between learners and assessment. Previously, students' writing abilities could be assessed based on written products produced independently. With the emergence of applications like ChatGPT and various other LLMs, students can use AI to generate ideas, correct grammar, translate, organize text, or even generate entire answers. This condition causes the concepts of language proficiency and authentic performance to need to be reviewed.

A systematic review of several articles regarding GenAI in language testing and assessment shows that the main issues that arise are related to reliability, validity, fairness and practicality. The study also emphasizes that the use of GenAI should not be done solely because it follows technological developments, but must be based on pedagogical needs and empirical evidence regarding its benefits and impact on assessment quality (Hussein et al., 2019).

2.4 Automated Scoring and Validity Challenges

Automated scoring is one of the most widely developed AI applications in language ability assessment. This technology allows written and oral responses to be analyzed and scored automatically. In writing assessments, for example, AI can analyze aspects such as grammar, vocabulary, text organization, coherence, and other linguistic characteristics. In speaking assessments, the system can utilize automatic speech recognition and natural language processing to analyze students' language production (Chikhalikar, n.d.).

Although automated scoring has the potential to improve efficiency and consistency, its validity remains an important issue. Validity is not only concerned with how close the AI scores are to human scores, but also whether the system actually measures the language skills that are the assessment construct. A system that produces consistent scores will not necessarily produce valid interpretations if it gives too much

weight to the surface characteristics of the text rather than the actual quality of the communication (Chikhalikar, n.d.).

Research on the use of LLM for automated essay scoring shows that the performance of various models is not always the same. Some models demonstrate fairly good reliability and validity, but performance can change based on the model used, response characteristics, and assessment conditions. (ScienceDirect) Other research on LLMs as raters in large-scale writing assessments also found that LLMs can show good consistency in certain tasks, but still differ from human raters in absolute scoring and in the application of complex scoring criteria (Asrifan et al., 2026a).

Therefore, the use of automated scoring needs to be placed within the framework of validity arguments. An AI system should not be immediately deemed valid just because it is able to generate scores quickly or has a certain level of agreement with human raters. Evidence is required to demonstrate that the scores can be used for the intended decision-making purposes.

2.5 Reliability and Consistency of AI-Based Assessments

Reliability is another important aspect in assessing language skills. Human assessments of productive abilities such as speaking and writing often face issues of subjectivity and differences between assessors. AI offers the possibility to increase the consistency of assessment by applying the same scoring procedure to all student responses.

However, the use of AI does not automatically guarantee perfect reliability. A 2026 meta-analysis of 21 studies with a total of 401,698 participants found that on average there was no significant difference between AI scores and human scores in English language assessments. However, the study also found very high heterogeneity. Differences in scores are influenced by the type of AI system, the number of human raters, the agreement index used, the learner's ability level, and other factors. These findings show that AI performance is highly dependent on the context in which it is used (Zhou, 2022).

Thus, AI is more appropriately positioned as part of an assessment system involving a quality assurance process rather than as an absolute replacement for human assessors.

In the context of high-stakes assessment, using a combination of automated scoring and human judgment can be a more defensive approach.

2.6 Fairness, Bias, and Equity in AI-Based Assessments

Fairness is one of the main challenges in using AI for language assessment. AI systems are developed based on certain data and algorithms. If training data does not adequately represent the diversity of learners, languages, accents, cultures, and linguistic backgrounds, the system can produce bias against certain groups. This issue is very relevant in language assessment because students have very wide linguistic variations. In speaking assessments, for example, automatic speech recognition systems may perform differently when encountering accents or language varieties that are underrepresented in the training data. Recent research on automated language systems shows that limited representation of language varieties in training data can give rise to issues of fairness and validity (Bailey et al., 2026).

In automated writing assessment, 2026 research also shows that the level of fairness can differ based on student groups and their ability levels. This shows that evaluating AI systems is not enough to just use general accuracy measures, but also needs to carry out fairness audits of various groups of students (Younes, 2025). Therefore, fairness must be an integral part of the development and implementation process of AI-based assessments. System developers and users need to check whether there are systematic differences in assessment results based on linguistic background or learner characteristics that are not relevant to the construct of language ability.

2.7 Academic Integrity and Authenticity of Assessment

The presence of GenAI poses a major challenge to academic integrity. Learners can use AI to generate answers or improve their work so that the final product does not fully represent their individual abilities. This issue is very complex in assessing writing ability because writing produced by AI can have relatively good structure, grammar and vocabulary. These conditions encourage a paradigm shift from assessments that only focus on the final product to assessments that also pay attention to the process. Teachers or lecturers can use process-oriented assessment, portfolio, drafting, reflection, oral defense, presentation, peer assessment, and authentic assignments to obtain evidence of students' abilities from various sources (Asrifan et al., 2026b).

UNESCO emphasizes that the development of AI has disrupted traditional concepts of assessment because machines can now produce essays, solve problems and carry out various activities that were previously considered evidence of a learner's individual abilities. Therefore, assessments need to be increasingly directed at higher-order thinking skills, creativity, problem solving and ethical reasoning (UNESCO, 2025). Thus, the main challenge is not just how to detect whether students are using AI, but how to design an assessment system that is still able to provide authentic evidence regarding students' abilities and learning processes.

2.8 Privacy, Data Security, and Ethics

The digitalization of assessments has also increased attention to privacy and data security issues. AI-based assessment systems can collect various types of data, including written answers, voice recordings, videos, student identities, and interaction data during the assessment process. This data can be used for various purposes, so clear policies regarding data collection, storage, processing and use are needed (Bailey et al., 2026).

Ethical issues also arise regarding algorithm transparency. Students and educators need to know how decisions or scores are generated, especially when those scores are used for important decisions. If the AI system provides an inappropriate score, there must be a mechanism for human review, appeal, or re-scoring. In this context, the use of AI in assessment requires the principle of human oversight. Humans still have an important role in monitoring systems, interpreting results, handling unusual cases, and making final decisions when assessment results have major consequences for students (Li et al., 2026).

2.10 Challenges and Future Directions for Language Proficiency Assessment

Based on these various studies, it can be concluded that the future of language ability assessment will be increasingly influenced by digital technology and AI. However, these developments do not mean that all conventional forms of assessment must be abandoned. The main challenge lies in how to integrate technology with the basic principles of language assessment. A promising approach is hybrid assessment, namely a combination of technology-based assessment and human judgment. AI can be used for tasks of an administrative nature, providing initial feedback, assisting with response analysis, or providing diagnostic information. Meanwhile, humans still play a role in interpretation, decision making, moderation, and handling complex responses (Grab, 2025).

Apart from that, future assessment development needs to lead to learning-oriented assessment and assessment for learning. Assessment is not only used to measure final results, but also to help students understand their strengths and weaknesses and determine their next learning steps (Yang & Wang, 2025). Overall, current literature shows that the biggest challenge of language assessment in the AI era lies not in the technology's ability to produce scores, but in the ability of the education system to ensure that the scores are valid, reliable, fair, transparent, authentic, safe, and accountable. Therefore, the integration of AI in language assessment must be carried out critically, evidence-based, and oriented towards pedagogical needs. Recent studies even confirm that the use of GenAI that is not selective and only follows technological trends can cause new problems if it is not accompanied by an evaluation of its validity, fairness, reliability, practicality and impact on learning (Khasawneh, 2024).

RESEARCH METHOD

This study employs a qualitative approach with a literature review design to identify, examine, and synthesize current issues in language proficiency assessment and the associated challenges. The research utilizes secondary data sourced from international journal articles relevant to topics such as language assessment, language testing, language proficiency assessment, digital language assessment, artificial intelligence in language assessment, generative AI, validity, reliability, fairness, authenticity, and assessment ethics. Articles were obtained through searches in academic databases—including Scopus, Web of Science, ScienceDirect, ERIC, and Google Scholar—prioritizing those published between 2020 and 2026, particularly recent publications addressing developments in language assessment within the digital and AI era. Source selection was based on inclusion criteria: international journal articles directly addressing developments, issues, challenges, or innovations in language proficiency assessment that were available in full text; articles lacking direct relevance to the study's focus were excluded.

The selected literature was analyzed using thematic content analysis, involving stages of identification, categorization, comparison, and synthesis of findings. Based on the literature search and categorization, the key issues analyzed include the use of AI and GenAI in assessment; the validity and reliability of technology-based assessment; fairness

and potential algorithmic bias; academic integrity and the authenticity of learner responses; authentic and multimodal assessment; and ethical concerns and educator readiness regarding the use of assessment technology.

RESULTS AND DISCUSSION

4.1 Paradigm Shift in Language Proficiency Assessment in the Digital Era

Research findings indicate that language proficiency assessment is undergoing a fundamental paradigm shift. Assessments that were previously oriented primarily toward measuring learning outcomes via written tests are now evolving into approaches that are digital, multimodal, adaptive, and process-oriented. This shift is driven not only by technological advancements but also by changes in the nature of linguistic communication within society. In the digital era, language proficiency is no longer limited to the ability to use language in text-based forms; it also encompasses the ability to communicate across various media, interpret multimodal information, and utilize technology appropriately and critically.

4.2 AI and Generative AI: Opportunities and Challenges

One of the most notable findings is the increasing use of AI and Generative AI (GenAI) across various stages of language assessment. AI can assist in creating test items, generating stimuli, analyzing responses, providing feedback, and scoring writing and speaking proficiency. On the positive side, the use of AI can enhance assessment efficiency. In large classes, for instance, teachers or lecturers can utilize AI to provide preliminary feedback on students' writing, thereby reducing the time required for feedback delivery. AI also holds the potential to facilitate more personalized assessments and provide diagnostic information regarding students' language errors.

However, a systematic review of articles on GenAI in language testing and assessment indicates that its application still faces challenges regarding reliability, validity, fairness, and practicality. The review also cautions that GenAI integration should not be driven merely by the technology's rapid development; rather, it must be grounded in pedagogical needs and empirical evidence concerning its impact on assessment quality. (ScienceDirect). Consequently, the central issue is no longer whether AI can be used in language assessment, but rather for what purpose and in what manner it is employed, and what evidence supports such usage. Adopting this perspective is crucial to ensure that

technology does not become an end in itself, but rather serves as a tool to enhance the quality of learning and assessment.

4.3 Automated Scoring and Validity Issues

Automated scoring represents one of the most significant developments in language assessment. This technology is particularly compelling for the assessment of writing and speaking skills, as evaluating these areas entirely by human means requires substantial time and effort. A meta-analysis of studies involving a great amount participants revealed that, on average, there is no significant difference between AI-generated scores and human-assigned scores. However, the research identified a high degree of heterogeneity. Discrepancies between AI and human scores are influenced by factors such as the type of AI system, the learners' proficiency levels, the number of human raters, the agreement indices employed, and specific study characteristics.

These findings have significant implications. It is incorrect to interpret the similarity between average AI and human scores as an indication that AI can directly replace human assessors. Instead, the high degree of heterogeneity suggests that the quality of automated scoring depends heavily on the context of its application. Other research on Large Language Models (LLMs) in automated essay scoring also indicates that generative models may offer superior validity in capturing specific aspects of writing quality, yet their reliability can differ from that of systems based on linguistic representations. (ScienceDirect). Therefore, the validity of AI-based assessment needs to be comprehensively examined. Validity cannot be determined solely by the level of correlation between AI scores and human scores. Researchers and practitioners must ensure that the system genuinely measures the intended construct, rather than merely surface-level linguistic characteristics such as text length, vocabulary complexity, or specific grammatical patterns.

4.4 Fairness and Bias in AI-Based Assessment

The issue of fairness represents one of the most serious challenges in using AI for language proficiency assessment. AI systems learn from the data used during their development. If this data does not adequately represent the diversity of languages, accents, cultures, or learner groups, bias may arise in the assessment results. This issue is particularly critical in speaking assessments. Automatic Speech Recognition (ASR)

systems, for instance, may perform better with language varieties or accents that are prevalent in their training data compared to those that are underrepresented. Research on automated language assessment indicates that limited training data for specific language varieties can lead to issues regarding fairness and validity.

Similar issues arise in automated essay scoring. Recent research indicates that the fairness of AI-based writing assessment systems can vary depending on learner groups and their proficiency levels. No single model is entirely free from between-group disparities. (ScienceDirect). This demonstrates that overall accuracy is not the sole indicator of an AI system's quality. A system with high aggregate accuracy can still produce unfair outcomes if it systematically assigns lower scores to specific groups.

In the context of language education, this issue is particularly significant given the diverse linguistic backgrounds of learners. Consequently, the development of AI for assessment must incorporate fairness testing, bias analysis, and evaluations across various user groups before the system is deployed widely.

4.5 Academic Integrity and the Changing Meaning of "Language Proficiency"

The advent of GenAI has also altered issues surrounding academic integrity in language assessment. Prior to the rise of GenAI, teachers and lecturers could use essays as evidence of a learner's writing proficiency. However, now that learners can use AI to generate or refine their writing, the final product cannot always be viewed as a direct reflection of the individual learner's own abilities. This situation raises a fundamental question: Can AI-assisted writing still serve as evidence of an individual's writing proficiency? This question cannot be answered simply by banning the use of AI. In practice, AI has become an integral part of the digital communication and literacy landscape. Therefore, a more constructive approach is to redesign assessments to distinguish between the skills learners must possess independently and the ability to use AI responsibly.

Assessment can focus on the process through portfolios, staged drafts, reflections, presentations, discussions, oral defenses, peer assessments, and authentic tasks. Such approaches enable teachers or lecturers to gather more diverse evidence regarding learners' abilities. Consequently, the advent of GenAI is not merely an issue of academic integrity; it also creates a need to redefine the construct of language proficiency. Recent

studies indicate that AI is even prompting a potential reconceptualization of communicative competence to encompass the ability to communicate within digital and multimodal environments that involve AI technology.

4.6 Challenges of Online Assessment: Access, Infrastructure, and the Digital Divide

Beyond AI, online assessment also presents issues regarding equitable access. Digital-based assessment presupposes that learners possess adequate devices, internet connections, learning environments, and digital literacy. In reality, however, these conditions vary. Learners with reliable devices and internet connections hold an advantage over those facing technological limitations. Consequently, assessment scores may be influenced by factors unrelated to actual language proficiency.

This issue can be categorized as "construct-irrelevant variance"—that is, score variation stemming from factors unrelated to the construct intended for measurement. For instance, if the test aims to measure English speaking proficiency, factors such as audio equipment quality, internet connectivity, or the ability to operate the application should not be the dominant determinants of the score. Consequently, the implementation of online assessments must take into account the principles of accessibility and equity. Test developers need to ensure that the technology employed does not create additional barriers for learners with limited digital access.

4.7 The Role of Human Evaluators in AI-Based Assessment Systems

Study findings indicate that the advancement of AI has not eliminated the need for human evaluators. Instead, the human role has shifted from merely assigning scores to acting as supervisors, validators, interpreters, and decision-makers. A recent meta-analysis comparing AI and human scoring concludes that AI cannot yet fully replace human judgment; rather, it is best utilized as a diagnostic tool or reference within a broader quality assurance framework. Significant discrepancies between AI and human scores require further investigation. (Nature)

This view is reinforced by research on automated writing scoring, which indicates that educators tend to support the use of automated systems when combined with human scoring. Educators also express concerns regarding potential bias against multilingual learners, as well as accessibility issues for students with specific limitations or low levels of digital literacy. Therefore, the most realistic model in an educational context is not AI

versus human, but rather AI-human collaboration. AI can assist with repetitive and analytical tasks, while humans remain essential for providing professional judgment and understanding the social, cultural, and pedagogical contexts of assessment results.

4.8 Data Privacy and Assessment Ethics

The use of digital technology and AI in assessment also gives rise to ethical and privacy concerns. Language assessment data may include written work, audio recordings, video footage, learner identities, and interaction data. When such data is processed by third-party platforms, questions arise regarding data ownership, storage, usage, and protection. Another issue is transparency. Learners may receive scores generated by algorithms without knowing how those scores were derived. This situation is particularly problematic when assessment results are used for critical decisions such as graduation, class placement, or selection. Therefore, the use of AI in assessment must adhere to principles of transparency, data security, accountability, and human oversight. Learners also need to be adequately informed about the use of AI in the assessment process.

4.9 Implications for Teachers, Lecturers, and Test Developers

These issues indicate a need for educators to enhance their assessment-related competencies. It is insufficient for teachers and lecturers merely to understand concepts such as validity, reliability, and test construction techniques; they also require AI literacy and digital assessment literacy. Educators must be able to address several critical questions before employing AI technology in assessment, namely, a) is the technology truly necessary? b) what language ability construct is being measured? c) are the AI-generated scores valid for the intended purpose? d) is there a potential for bias against specific groups? e) how are AI results verified by humans? f) how is student data protected? g) does the use of the technology provide equitable access for all students? The ability to answer these questions is a crucial aspect of educator professionalism in the digital era. Recent studies indicate a current research gap regarding the lack of sufficient evidence-based guidelines for stakeholders on the use of AI for language assessment.

4.10 Synthesis of Study Findings and Key Challenges

Based on the study findings and discussion above, several key issues regarding language proficiency assessment in the digital era can be identified in Table 4.1.

Tabel 4.1 Key Issues Regarding Language Proficiency Assessment

Current Issue	Opportunities	Challenges
Online assessment	Flexible, efficient, rapid	Digital divide, connectivity, security
AI in assessment	Automation, feedback, efficiency	Validity, transparency, technological dependency
Automated scoring	Consistency and scalability	Bias, construct representation, human oversight
Generative AI	Personalization and learning support	Academic integrity and authorship
AI-based speaking assessment	More efficient speaking assessment	Accent bias and ASR limitations
AI-based writing assessment	Rapid scoring and feedback	Fairness and construct validity
Multimodal assessment	More authentic and contextual	Design and assessment complexity
Digital assessment	Faster data collection	Data privacy and security
Human-AI collaboration	Combining AI efficiency and human judgment	Educators' AI literacy competence

Overall, the study results reveal a paradox in the development of language assessment technology. On one hand, AI and digital technologies offer opportunities to enhance efficiency, flexibility, scalability, and the quality of feedback. On the other hand, these technologies can give rise to new sources of inequity, bias, breaches of academic integrity, and validity issues. Recent findings even indicate that technically high-performing AI systems are not necessarily entirely fair. A 2026 study on automated essay

scoring found that while certain models could achieve a high degree of alignment with human scores, inter-group disparities could still emerge.

Thus, the primary challenge in language proficiency assessment in the AI era lies not merely in developing increasingly sophisticated technology, but in ensuring that such technology is employed within a framework of validity, reliability, fairness, authenticity, practicality, transparency, and ethics. High-quality assessment must remain grounded in the language proficiency construct being measured, without allowing technology to displace the pedagogical objectives of the assessment.

4.11 Discussion: Towards a Human-Centered Language Assessment System

Based on a synthesis of the literature, the most relevant direction for the development of language proficiency assessment is towards a human-centered system. In this approach, AI technology serves as a tool to enhance the assessment process rather than as a replacement for the full scope of human functions. This approach can be implemented through several strategies. First, AI can be utilized to provide preliminary scoring or feedback, while the final decision remains with the teacher or lecturer. Second, AI assessment results should be periodically compared with human assessments to identify any discrepancies. Third, the system must be tested across diverse groups of learners to detect potential biases. Fourth, assessments should be designed to measure not only the final product but also the process and the learners' ability to use language authentically.

This approach aligns with recent research findings indicating that while AI can offer significant benefits in language assessment, its implementation must be measured and selective. Systematic reviews of Generative AI (GenAI) underscore the importance of considering reliability, validity, fairness, practicality, impact, and sustainability before integrating the technology into assessment systems. Consequently, the future of language assessment should not be viewed as a shift from human assessment to AI assessment, but rather as a transformation toward assessment that is intelligent, evidence-based, equitable, and human-centered. While technology can enhance the capabilities of assessment systems, decisions regarding what is assessed, the rationale for assessment, the interpretation of results, and the application of those results still require professional human judgment.

Ultimately, the success of language assessment innovation is determined not by the sophistication of the technology employed, but by the extent to which that technology yields information that is valid, reliable, fair, authentic, and beneficial to the learning process. Therefore, future assessment development must prioritize measurement principles and learner needs as the primary foundation, while positioning AI and digital technologies as supporting instruments to achieve those goals.

CONCLUSIONS

Current issues in language ability assessment show that language assessment continues to change along with the development of digital technology, artificial intelligence (AI), online learning, and increasingly complex and authentic communication demands. Assessment is no longer only oriented towards mastery of grammar, vocabulary, reading, writing, listening and speaking, but also needs to measure communicative abilities, critical thinking, digital literacy and multimodal competence. Although technology provides opportunities to provide more flexible, adaptive, efficient, and personalized assessments, its implementation also presents challenges in the form of instrument validity and reliability, academic integrity, inappropriate use of AI, data security, gaps in technology access, and fairness and inclusivity in assessments.

Therefore, the development of language ability assessment needs to be carried out in an innovative and responsive manner to current developments while adhering to the principles of validity, reliability, fairness, authenticity and transparency. Educators also need to have adequate assessment literacy and digital literacy to be able to use technology and AI ethically and responsibly. Thus, language assessment today is expected to not only function as a tool for measuring learning outcomes, but also as an integral part of the learning process which is able to provide meaningful feedback and support the development of students' competencies in facing communication needs in the digital era.

REFERENCES

Andersen, N., Mang, J., Goldhammer, F., & Zehner, F. (2025). Algorithmic Fairness in Automatic Short Answer Scoring. *International Journal of Artificial Intelligence in Education*, *35*(5), 3128–3165. https://doi.org/10.1007/s40593-025-00495-5

Asrifan, A., Oliveira De Barros Cardoso, L. M., & Vargheese, K. J. (2026b). Automated feedback for speaking and writing skills: Deep learning in English language assessment. *EduLite: Journal of English Education, Literature and Culture*, *11*(1), 67. https://doi.org/10.30659/e.11.1.67-85

Bailey, A. L., Johnson, A., Shankar, N. B., Veeramani, H., Washington, J. A., & Alwan, A. (2026). Addressing Bias in Spoken Language Systems Used in the Development and Implementation of Automated Child Language-Based Assessment. *Journal of Educational Measurement*, *63*(1), e12435. https://doi.org/10.1111/jedm.12435

Chikhalikar, D. R. S. (n.d.). *English language speaking assessment tests and evaluations, incorporating AI innovative methods*.

Chuang & Yan. (2025). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions. *System 134: 103846*. https://doi.org/10.1016/j.system.2025.103846

Dang, H. T., & Habók, A. (2026). Digital reading comprehension assessment in English language education: A systematic review. *Cogent Education*, *13*(1), 2633011. https://doi.org/10.1080/2331186X.2026.2633011

Grab, M. Ö. (2025). Assessing language proficiency in the digital age: Exploring innovative methods for assessment and feedback in ESL classrooms. *Journal of Digital Educational Technology*, *5*(2), ep2510. https://doi.org/10.30935/jdet/16286

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208. https://doi.org/10.7717/peerj-cs.208

Jang, E. E., & Sawaki, Y. (n.d.). *Advancing language assessment for teaching and learning in the era of the artificial intelligence (AI) revolution: Promises and challenges.*

Khasawneh, M. A. S. (2024). Investigating the impact of automated instruments used for assessing the writing skill: Perspectives of language e-learners. *Research Journal in Advanced Humanities*, *5*(2). https://doi.org/10.58256/4fd2qt78

Li, B., Qunhan, X., & Mao, C. (2026). Differences between human and AI scoring: A meta-analysis of english language assessments. *Scientific Reports*, *16*(1), 17412. https://doi.org/10.1038/s41598-026-48053-w

O'Grady, S., Gokturk, N., & Rossi, O. (2026). Artificial Intelligence in Language Assessment Research: A Scoping Review. *Research Synthesis in Applied Linguistics*, *2*(1), 103–139. https://doi.org/10.1080/29984475.2026.2633389

Ozoda et al. (2024). Principles of Language Assessment. *Modern Education and Development, 14: 45 – 48.*

Yang, Z., & Wang, P. (2025). Current status and research trend of English language assessment: A bibliometric analysis. *Language Testing in Asia*, *15*(1), 11. https://doi.org/10.1186/s40468-024-00317-w

Younes, Z. B. B. (2025). Investigate How AI Algorithms Can Be Used to Automate English Language Proficiency Assessments. *World Journal of English Language*, *15*(5), 285. https://doi.org/10.5430/wjel.v15n5p285

Zhou, L. (2022). Construction of English Writing Hybrid Teaching Model Based on Machine Learning Automatic Composition Scoring System. *Procedia Computer Science*, *208*, 384–390. https://doi.org/10.1016/j.procs.2022.10.054