Exploring EFL Teachers' Perceptions of Artificial Intelligence Integration in E-Assessment Design: Implications for Efficiency and Objectivity

Shaje Ahmed Alhomami[1], Mohammed H. Albahiri[2], & Ali Albashir Mohammed Alhaj[3]

[1] King Khalid University, Abha, Saudi Arabia

[2] King Khalid University, Abha, Saudi Arabia

[3] Saudi Electronic University, Saudi Arabia

Correspondence: Dr. Ali Albashir Mohammed Alhaj, English Unit, Applied College, Dhahran Aljanoub, King Khalid University, SaudiArabia.

Received: November 3, 2025 Accepted: March 12, 2026 Online Published: April 28, 2026

doi:10.5430/wjel.v16n4p493 URL: https://doi.org/10.5430/wjel.v16n4p493

Abstract

The present study explores EFL educators' perceptions of AI-generated questions used in electronic examinations and examines whether these perceptions vary according to educators' educational stage, years of teaching experience, and academic specialization. Data were collected through a validated questionnaire administered to EFL teachers working in public education, and the study employed an empirical research design. The data were analyzed using descriptive statistics, including means and standard deviations, as well as inferential statistics, such as the Kruskal–Wallis test and the independent samples **t**-test. The findings indicate that teachers generally hold strong, positive perceptions of AI-generated e-examination tasks, particularly regarding objectivity, reduced bias, accurate feedback, time efficiency, and alignment with Bloom's cognitive taxonomy. Teachers also reported a high level of satisfaction with the integration of artificial intelligence in testing and assessment practices. Furthermore, the inferential analysis revealed no statistically significant differences in teachers' perceptions based on educational stage, years of teaching experience, or academic specialization, suggesting that educators share broadly similar professional views regarding the use of AI-assisted assessment. Overall, the study highlights the considerable potential of artificial intelligence to enhance fairness, efficiency, and pedagogical quality in EFL assessment. At the same time, it underscores the importance of maintaining human oversight, ethical guidance, and a culture of continual professional development to support the responsible and effective implementation of AI technologies in educational assessment.

Keywords: Artificial Intelligence, EFL Teachers' Perceptions, Electronic Assessment, Assessment Objectivity, Educational Technology

1. Introduction

The emergence of artificial intelligence (AI) has become a key driver of the contemporary integration of digital technologies in education, significantly influencing teaching, learning, and assessment practices. AI-powered technologies enable educators to analyze learner performance more effectively, personalize learning experiences, and automate various aspects of electronic assessment. These capabilities enhance the accuracy, efficiency, and consistency of evaluating students' academic performance. As a result, electronic assessment (e-assessment) has gained increasing importance in modern educational systems, as it enables timely feedback, continuous monitoring of student progress, and a substantial reduction in the administrative workload associated with traditional grading methods (Redecker & Johannessen, 2013).

Within EFL educational contexts, assessment plays a particularly critical role because language development involves multiple integrated skills, including reading, writing, listening, and speaking. Therefore, assessment practices must be accurate, objective, and pedagogically meaningful in order to effectively capture learners' communicative competence. Emerging research suggests that AI-supported e-assessment systems can improve test quality, enhance consistency in scoring, and reduce human bias in evaluation processes, thereby contributing to fairer and more reliable assessments (Baker & Inventado, 2014; Luckin et al., 2016). Nevertheless, the successful implementation of AI technologies in education requires institutional support. AI-based assessment tools largely rely on teachers' perceptions, attitudes, and willingness to adopt these technologies in their professional practices.

Despite the increasing number of studies on AI in education, significant research gaps remain, particularly regarding EFL teachers' perceptions of AI-generated electronic assessment. Moreover, scholars have raised important concerns about the reliability of automated assessment results, potential algorithmic bias, and the need for adequate professional training to promote the responsible use of AI in educational contexts (Selwyn, 2019).

In response to these research gaps, the present study investigates EFL teachers' perceptions of AI-generated electronic examination questions and examines whether such questions are perceived as efficient, objective, and unbiased. Furthermore, the study explores whether teachers' perceptions vary according to key demographic and professional variables, including educational stage, years of

teaching experience, and academic specialization. Accordingly, the study addresses two main research questions: (1) What are teachers' perceptions of AI-generated electronic examination questions? and (2) Are there statistically significant differences in teachers' perceptions at $\alpha = .05$ based on educational stage, years of teaching experience, and academic specialization? By addressing these questions, the study provides empirical insights into the fields of EFL assessment and applied linguistics. The findings may inform language educators, curriculum developers, and policymakers seeking to implement AI-supported assessment practices that are effective, equitable, and ethically grounded. In addition, the results may support targeted professional development initiatives for teachers working in AI-enhanced educational environments.

2. Literature Review

2.1 Artificial Intelligence (AI)

2.1.1 Concept of Artificial Intelligence (AI)

Artificial intelligence (AI) has emerged as a rapidly expanding field within modern computer science and is widely recognized as one of the most influential technological developments of the digital era. AI focuses on designing computational systems capable of emulating human cognitive processes, including learning, reasoning, decision-making, and problem-solving, and natural language understanding (Russell & Norvig, 2021). Through the analysis of large datasets and the use of advanced algorithms, AI enables machines to perform complex cognitive tasks with increasing levels of autonomy and precision. Core AI technologies include machine learning, which improves system performance through experience and data, and deep learning, which employs multilayered artificial neural networks to detect complex patterns in data (Goodfellow et al., 2016). Additional components such as computer vision and natural language processing further enhance AI's capacity to interpret visual and linguistic information. In education, AI supports personalized learning, intelligent assessment, and data-driven feedback, thereby enhancing the effectiveness and efficiency of teaching and learning.

2.1.2 The Emergence of Artificial Intelligence

Artificial intelligence developed as a formal academic discipline in the mid-twentieth century, influenced by advances in computer science, mathematical logic, and early investigations into human cognition. A pivotal milestone occurred in 1956 during the Dartmouth Conference, where the term artificial intelligence was introduced by John McCarthy to describe the development of machines capable of simulating aspects of human intelligence, such as learning, reasoning, and problem-solving. Early AI research was largely based on symbolic and rule-based systems and was characterized by strong scientific optimism. However, limited computational capacity and insufficient data led to periods of stagnation known as the "AI winter" during the 1970s and 1980s (Russell & Norvig, 2021). Renewed progress emerged with the development of statistical algorithms, machine learning methods, and deep learning models supported by large datasets and high-performance computing, advancements that transformed AI into a practical technology applied across multiple sectors. In education, AI now supports personalized learning, learning analytics, and intelligent assessment systems (Goodfellow et al., 2016).

2.1.3 The Importance of Artificial Intelligence in Education

Artificial intelligence has become a transformative force in contemporary education, significantly improving teaching, learning, and assessment practices. Through intelligent systems and data-driven technologies, it enables more efficient and adaptive educational processes. One of its most important contributions is the development of personalized learning environments. By analyzing learners' behavioral patterns and academic performance, AI-powered systems can adapt instructional materials and teaching strategies to meet individual learning needs, thereby enhancing students' engagement, motivation, and academic achievement. In addition, AI technologies provide advanced learning analytics that allow educators to monitor student progress more accurately and make evidence-based instructional decisions, supporting flexible teaching approaches suited to diverse learning styles.

Furthermore, AI has improved electronic assessment through automated grading, intelligent test design, and real-time feedback mechanisms, innovations which have increased the efficiency, transparency, and fairness of evaluation while reducing teachers' administrative workload (Redecker & Johannessen, 2013). Moreover, AI supports inclusive and lifelong learning by offering accessible digital resources and assistive technologies for diverse learners. International organizations highlight that responsible AI integration can foster educational innovation and sustainable development in modern knowledge-based societies (UNESCO, 2023).

2.1.4 Classification of Artificial Intelligence (In Brief)

Artificial intelligence is a multifaceted field commonly classified according to its functional capabilities and cognitive complexity, offering a clearer understanding of both its current applications and future potential. It is also one of the most widely discussed categories of AI, referring to systems designed to perform specific tasks without possessing general cognitive abilities, examples of which include recommendation systems, speech recognition, and machine translation technologies that support various digital services and educational tools (Russell & Norvig, 2021). In contrast, artificial general intelligence (AGI) represents a theoretical form of intelligence capable of performing a wide range of cognitive tasks comparable to human reasoning and learning, although such systems have not yet been realized. A more speculative stage is artificial superintelligence (ASI), where machines may surpass human intelligence in reasoning, creativity, and decision-making, raising complex ethical and societal concerns. From a functional perspective, AI systems are commonly classified into reactive machines, limited-memory systems, theory-of-mind systems, and self-aware systems. However, most existing applications remain within the lower levels of this hierarchy.

2.1.5 Applications of Artificial Intelligence in Education

Artificial intelligence has become a major driver of digital transformation in contemporary education, supporting innovative teaching practices and improving instructional effectiveness at both pedagogical and institutional levels. By analyzing learners' data, interaction patterns, and learning behaviors, AI technologies enable the creation of personalized learning environments that adapt instructional materials to students' individual needs and learning preferences, thereby enhancing engagement and academic performance.

AI systems also provide flexible instructional support through intelligent tutoring tools that deliver guided explanations, instant feedback, and continuous assistance, helping learners overcome difficulties and encouraging self-directed learning and problem-solving abilities (Luckin et al., 2016). Furthermore, AI-driven learning analytics allow educators to analyze large datasets related to students' performance, attendance, and engagement. These insights support data-informed pedagogical decisions, early identification of academic challenges, and the design of more effective instructional strategies.

Furthermore, AI technologies have enhanced assessment practices through automated grading systems and intelligent feedback mechanisms that improve the efficiency, transparency, and fairness of evaluation while reducing teachers' administrative workload (Redecker & Johannessen, 2013). AI-based assistants and chatbots also provide immediate academic support, contributing to more flexible and inclusive learning environments.

2.2. Electronic Assessments

2.2.1 Concept of Electronic Assessments

Electronic tests (e-assessment) are digital assessments that use computer and internet technologies for the design, administration, scoring, and analysis of tests, either fully or partially automated. That means assessment of performance and cognitive skills, based on test scores, for learners can be done more quickly and objectively than on paper. These tests vary by time and location and incorporate various question formats, such as multiple-choice, true/false, and short-answer questions, interactive items, and digital simulations. Variety is a means of assessing higher-order cognitive skills, such as analysis, critical thinking, and problem-solving. Electronic tests also provide rapid feedback, can be used for formative and continuous assessment, and can further reduce the time and effort required for grading and result recording (Redecker & Johannessen, 2013). They improve the precision and fairness of evaluation since they reduce human biases, furnish accurate analytical data on the performance of learners, and help educators and decision-makers in enhancing educational practices and making data-driven teaching decisions in light of the necessity for digital transformation and quality

2.2.2 The Emergence of Electronic Tests

Electronic testing emerged in the 1960s when early computers were introduced in educational institutions. During this period, centralized computing systems were used to administer basic computerized assessments and automatically score students' responses, a practice later known as computer-based testing (CBT). With the advancement of computing technologies and the widespread adoption of personal computers in the 1980s and 1990s, electronic testing became more accessible and sophisticated. Educational institutions began to employ diverse question formats and interactive interfaces that enhanced learner engagement and improved the quality of assessment processes. The expansion of the internet in the late 1990s and early 2000s further transformed assessment practices by enabling online testing through learning management systems, providing real-time feedback and continuous evaluation (Redecker & Johannessen, 2013). More recently, rapid developments in web technologies, artificial intelligence, and learning analytics have expanded e-assessment beyond static testing to include adaptive assessments, interactive simulations, and performance-based evaluation. Consequently, e-assessment has become a key tool for improving efficiency, transparency, and objectivity in modern digital education (Bennett, 2015).

2.2.3 Key Objectives of Electronic Assessment in Modern Educational Systems

Artificial intelligence has become a key driver of digital transformation in contemporary education, supporting innovative teaching practices and improving instructional effectiveness at both pedagogical and institutional levels. By analyzing learners' data, interaction patterns, and learning behaviors, AI technologies enable the creation of personalized learning environments that adapt instructional materials to individual needs and learning styles, thereby enhancing engagement and academic achievement. AI systems also provide intelligent tutoring tools that deliver guided explanations, instant feedback, and continuous academic support, helping students develop self-directed learning and problem-solving skills (Luckin et al., 2016). Moreover, AI-driven learning analytics allow educators to monitor performance and engagement more effectively, supporting data-informed decisions and improving assessment efficiency and transparency in educational contexts (Redecker & Johannessen, 2013).

2.2.4 Types of AI-Powered Electronic English Language Assessments

Artificial intelligence has significantly transformed automated English language assessment, particularly within digital learning environments. AI-based assessment systems employ machine learning and natural language processing technologies and speech recognition technologies to improve the validity, reliability, and efficiency of language testing. Unlike traditional computer-based tests, AI-driven assessments can adapt the difficulty of test items in real time according to learners' performance across reading, listening, writing, and speaking skills, providing more precise evaluations of language proficiency (Chapelle & Voss, 2016). In writing assessment, automated scoring systems analyze linguistic complexity, grammatical accuracy, cohesion, and lexical diversity while delivering immediate feedback for both formative and summative purposes (Dikli, 2006; Shermis & Burstein, 2013). Similarly, AI-enabled speech recognition technologies support standardized speaking assessments and reduce rater bias. Furthermore, adaptive listening and reading assessments generate diagnostic insights into learners' comprehension strategies and difficulties. Overall, AI-driven assessment tools promote continuous

evaluation, personalized feedback, and data-informed instructional decisions in contemporary language education.

2.2.5 Advantages of Electronic English Language Tests

Electronic English language testing is a contemporary assessment method, and its implementation on e-learning platforms will be significantly expedited with effective systems. The four fundamental language skills(listening, speaking, reading, and writing) are further improved by the use of multimedia elements (e.g., audio fragments or video clips) and interactive language tasks, thereby increasing the validity of the measure by approximating authentic language use (Fulcher, 2015). Additionally, electronic testing can ensure greater objectivity and accuracy in scoring tasks, particularly those involving grammar, vocabulary, and reading comprehension, and reduce human bias and types of errors (Dikli, 2006; Dörnyei & Taguchi, 2010). Real-time feedback enables formative assessment of student performance (Redecker & Johannessen, 2013), increasing students' awareness of their strengths and weaknesses and enabling them to improve their language performance. Digital assessment systems acknowledge students' varied needs by providing flexibility and adaptive learning strategies, such as adjusting the difficulty of assessment questions to match learners' proficiency levels, thereby better capturing performance as a function of language ability (Chapelle & Voss, 2016). In addition, these tests enable analysis of language performance data and longitudinal tracking of learners' progress; they have been shown to be a useful platform for data-driven educational decision-making and effectiveness in EFL and ESL settings in the postmodern world.

2.2.6 Disadvantages of Electronic English Language Tests

The growing integration of electronic English language testing within modern assessment systems offers many advantages; however, it also introduces several challenges that may affect the validity, reliability, and fairness of evaluation processes if not carefully addressed. Technical challenges such as system failures, unstable internet connections, and software incompatibility can disrupt test administration and potentially influence learners' performance, particularly in high-stakes examinations (Dikli, 2006). In addition, the persistent digital divide—reflected in unequal access to devices, reliable connectivity, and adequate digital literacy—may create structural disadvantages for some learners, thereby introducing construct-irrelevant variance and limiting the comparability of assessment results across different groups (Redecker & Johannessen, 2013). From a pedagogical perspective, many electronic language tests still rely heavily on objective item formats such as multiple-choice or short-answer questions, which may not fully capture complex language competencies, including authentic communication, extended writing, and speaking performance (Fulcher, 2015). Although automated scoring technologies have improved significantly, they often struggle to evaluate discourse coherence, creativity, and contextual appropriateness, highlighting the continued importance of human judgment in language assessment. Therefore, effective electronic EFL assessment frameworks should incorporate technological innovation in a manner that preserves pedagogical validity, upholds ethical responsibility, and ensures equitable access to learning opportunities.

2.3 AI Integration in EFL E-Assessment: A Critical Review of Previous Studies

The integration of artificial intelligence into English as a Foreign Language (EFL) assessment has received increasing scholarly attention, particularly in the areas of writing evaluation and electronic test design. Recent studies highlight the potential of AI-driven technologies to improve assessment efficiency, objectivity, and scalability. For instance, AI-based grading systems such as CoGrader enable automated essay evaluation and rapid feedback, supporting teachers in managing large volumes of student work (Alsalem, 2024). Nevertheless, researchers consistently emphasize that AI-generated scores should not replace professional human judgment: teacher mediation remains essential to interpret automated feedback and ensure pedagogically meaningful evaluation (Barrot, 2026).

Despite their potential benefits, the widespread adoption of AI-based assessment tools remains challenged by a range of pedagogical, technical, and ethical considerations. Studies indicate that excessive reliance on algorithmic systems may weaken teachers' evaluative expertise and raise concerns regarding algorithmic bias, transparency, and fairness (Benek, 2025). Similarly, Bessadat and Korichi (2025) reported that although teachers generally perceive AI-assisted writing assessment positively, they strongly support the need for continuous human supervision to maintain assessment credibility. From a technological perspective, AI systems have demonstrated promising capabilities in automated item generation, adaptive testing, and exam design; however, technical limitations, linguistic inaccuracies, and ethical considerations remain unresolved (Luc Ha & Nguyen, 2025).

Concerns about assessment validity are also widely documented. Kaldaras et al. (2024) found that while AI systems may track certain cognitive patterns, questions remain regarding their ability to measure higher-order thinking skills accurately. Likewise, Benek (2025) observed that EFL teachers viewed generative AI tools such as ChatGPT as useful for supporting lower-level reading tasks, yet emphasized the need for more sophisticated models capable of evaluating complex language performance. Teacher readiness and digital competence also appear to be crucial determinants of successful AI integration. Research indicates that many instructors possess limited AI literacy and express concerns about system accuracy, bias, and academic integrity. To respond to these challenges, Arslan (2025) highlighted the importance of professional development programs grounded in the Technological Pedagogical Content Knowledge (TPACK) framework to strengthen teachers' competencies in effectively integrating AI within assessment practices.

Furthermore, contextual challenges may influence AI adoption. Uddin et al. (2024) reported resistance to AI-mediated e-assessment in resource-constrained environments, where inadequate infrastructure, limited training, and concerns about academic misconduct hinder implementation. Overall, existing research suggests that while AI technologies offer significant potential to enhance assessment efficiency and personalization, their effective use depends on robust institutional infrastructure, ethical governance, and teachers' professional competence. Consequently, the responsible integration of AI in EFL assessment demands a balanced approach that aligns technological

innovation with pedagogical expertise and robust ethical oversight.

3. Methodology

3.1 Research Design

This study employed a descriptive survey design, which was appropriate for examining EFL teachers' perceptions of and attitudes toward emerging assessment practices in English language education. (Creswell & Creswell, 2018; Fraenkel et al., 2019). In order to assess EFL teachers' perceptions of AI-generated electronic test questions, a structured questionnaire was developed. This design allowed for the systematic analysis of participants' responses and helped identify trends related to the effectiveness, objectivity, and pedagogical relevance of AI-based e-assessment in EFL contexts.

3.1.1 Research Participants

The study population consisted of 150 EFL teachers working in the English language departments of public schools in Sharurah Governorate, Saudi Arabia, during the 2024 academic year. Using simple random sampling, 90 teachers were selected to participate in the study, a procedure that enhances representativeness and minimizes potential sampling bias. (Creswell & Creswell, 2018; Fraenkel et al., 2019). Participants included teachers from elementary, intermediate, and secondary educational levels. Data were collected using an online questionnaire, a widely used instrument in educational research due to its efficiency, accessibility, and ability to reach participants across different locations (Dörnyei & Taguchi, 2010). Ethical considerations were carefully observed, and informed consent was obtained from all participants prior to data collection. Participants were informed about the purpose of the study, the voluntary nature of their participation, and their right to withdraw at any stage without penalty. The study also ensured confidentiality and anonymity by coding participants' identities and adhering to established ethical standards for research involving human participants. (American Educational Research Association, 2011).

3.1.2 Statistical Description of the Study Sample

The following section describes the statistical characteristics of the sample and presents summary findings. Table 1 presents the proportions of study participants by the study's main variables, along with their demographic and contextual characteristics in the sample. This distribution was used to assess the representativeness of the sample and served as the basis for interpreting subsequent statistical analyses.

Table 1. Distribution of the Study Sample According to the Study Variables

Variable	Levels	Frequency (n)	Percentage (%)
Educational Stage	Primary	32	35.6%
	Secondary	35	38.9%
	Intermediate	23	25.6%
Years of Experience	Less than 5 years	43	47.8%
	5-10 years	32	35.6%
	More than 10 years	15	16.7%
Specialization	Scientific	52	57.8%
	Literary	38	42.2%
Total Sample		90	100%

The key figures and demographic characteristics presented in Table 1 of the study sample were highlighted to provide important contextual information for drawing inferences. The distribution of educational stages indicates that the primary, intermediate, and secondary levels are represented, with the largest proportion from the secondary level. This diversity enhances the generalizability of the findings across different educational settings. (Creswell & Creswell, 2018). The experience profile indicates that approximately half of the sample has less than five years of teaching experience, suggesting that many early-career teachers may have influenced the sample's attitudes toward innovation and the use of new technology, which may, in turn, affect attitudes and approaches (Fraenkel et al., 2019). The predominance of participants with scientific specializations also reflects the composition of the sample also reflects the sample's composition and aligns with calls to report background variables that show how this diversity may be affecting participants' pedagogical values and assessment processes. This balance indicates that Table 1 presents a distribution of the sample generally consistent with the study's research objectives.

3.2 Instruments

Based on a review of the relevant literature and prior empirical research, and in line with the research questions, a questionnaire was designed to investigate teachers' perceptions of AI-generated electronic test questions. The instrument consisted of 26 items organized around one primary dimension and used a five-point Likert scale, with response options ranging from Strongly Agree to Strongly Disagree, rated 5 to 1. A standardized five-point Likert scale was used because it is suitable for measuring participants' attitudes and response tendencies along a continuum of agreement and disagreement (Dörnyei & Csizér, 2012). The questionnaire items were developed through a critical literature review and a content validity analysis conducted by an independent linguist and educational technology expert. An electronic survey was administered via Google Forms to facilitate efficient data collection by minimizing time, effort, and logistical constraints.

3.2.1 The Criterion Adopted in the Study Instrument

The interpretation criterion for the five-point Likert scale was established using a standard statistical procedure commonly used in educational research. The interval length of 0.80 was determined by dividing the range of the highest and lowest scale values ($5 - 1 = 4$) by the total number of scale categories (5). This value was then added to the lowest scale value (1) to determine the upper limit of each category. Consequently, participants' mean scores were classified into levels, representing the degree of teachers' attitudes toward AI-generated electronic test questions (see the table for details). This method is one of the most common and valid techniques for interpreting Likert-scale data. (Creswell & Creswell, 2018; Pallant, 2020).

Table 2. The Criterion Adopted in the Study (Scale)

No	Category	Category Range (From)	Category Range (To)
1	Strongly Agree (Very High Degree)	More than 4.20	5.00
2	Agree (High Degree)	More than 3.40	4.20
3	Neutral (Moderate Degree)	More than 2.60	3.40
4	Disagree (Low Degree)	More than 1.80	2.60
5	Strongly Disagree (Very Low Degree)	1.00	1.80

The selected criterion for the interpretation of the mean scores derived from the five-point Likert scale employed in the present research. For the 5 equal intervals of the scales by cell length ($4 \div 5 = 0.80$), the standard values of the scales were 1.00 to 5.00, a common procedure used in educational and social sciences research. It is based on a mean score greater than 4.20, indicating a very high degree of agreement; in this case, it was 3.40–4.20, a high degree of agreement. Moderate (neutral) scoring indicates that average scores range from 2.60 to 3.40, and low agreement scores range from 1.80 to 2.60. Finally, mean scores of approximately 1.00 to 1.80 indicate a very low agreement. Additionally, this categorization procedure improves the interpretability and reliability of the analysis and confirms the interpretation of participants' perceptions, as suggested in the methodological literature, using a Likert-type scale (Alhaj & Albahiri, 2020; Creswell & Creswell, 2018; Pallant, 2020).

3.2.2 Validity of the Study Instrument

Validity refers to the extent to which an instrument accurately measures the constructs it is intended to assess the constructs for which it was designed, that it comprehensively encompasses all elements necessary for meaningful analysis of the results, and has items are clear and aligned with the questionnaire, so that they are easily understood by all respondents. The researchers established the validity of the study instrument through the following procedures:

A. Face Validity of the Study Instrument (Expert Judgment Validity):

A panel of subject-matter experts reviewed the preliminary version of the questionnaire to ensure face validity and to confirm that the instrument assesses the constructs it was designed to measure. The judges were instructed to determine the relevance of the items to the study objectives, clarity of wording, appropriateness of the items, linguistic accuracy, and the extent to which the items matched the overall instrument. They were also invited to suggest adjustments, deletions, or additions whenever needed. Items that attained an agreement rate of $\geq 80\%$ among the judges after review were retained or revised. The final questionnaire incorporated revisions approved by the majority of the expert panel. (Creswell & Creswell, 2018; DeVellis, 2017; Lynn, 1986).

B. Internal Consistency Validity of the Instrument

The internal consistency validity of the research instrument was assessed using a pilot sample of 35 public-school teachers from the Sharurah Governorate who were not included in the main study sample. Based on the pilot data, Pearson's product–moment correlation coefficient was calculated to examine the relationship between each questionnaire item and the overall questionnaire score. This procedure helped determine the degree to which individual items were consistent with the total scale. Such analysis is commonly used to evaluate the internal consistency and construct validity of survey instruments. The item–total correlation analysis is an established method of measuring internal consistency, as it ensures that all individual items are relevant to the whole construct being measured and promotes the uniformity of the instrument (Creswell & Creswell, 2018; Pallant, 2020)

Table 3. Pearson's Correlation Coefficients Between the Questionnaire Items and the Total Questionnaire Score

Item Number	Correlation Coefficient	Item Number	Correlation Coefficient
1	0.968**	14	0.807**
2	0.984**	15	0.807**
3	0.966**	16	0.890**
4	0.963**	17	0.810**
5	0.834**	18	0.774**
6	0.868**	19	0.750**
7	0.939**	20	0.775**
8	0.864**	21	0.859**
9	0.917**	22	0.772**
10	0.886**	23	0.750**
11	0.889**	24	0.720**
12	0.838**	25	0.761**
13	0.826**	26	0.703**

Note: Significant at the 0.01 level ($p \leq 0.01$)

Table 3 presents Pearson's correlation coefficients between each questionnaire item and the total score, providing evidence of acceptable internal consistency and supporting the construct validity of the instrument. The results indicate that each item is significantly and positively correlated with the total questionnaire score. ($r = .703$-$.984$, $p < .01$)—above the commonly cited minimum of .30 for item–total correlations in the educational and behavioral science literature. These coefficient estimates confirm that each item meaningfully contributes to the construct measured by the questionnaire within its respective scale dimension. Thus, none of the items were excluded, as low item–total correlations were absent, attesting to the instrument's consistency, homogeneity, and internal structure (Creswell & Creswell, 2018; DeVellis, 2017; Pallant, 2020).

3.2.3 Reliability of the Study Instrument

The reliability of the study instrument was examined using Cronbach's alpha coefficient and the split-half method, a statistical value used to determine the degree of internal consistency among items of the instrument. Cronbach's alpha is one of the most frequently used indicators of internal consistency. (Cronbach, 1951; Pallant, 2020). Table 4 presents the Cronbach's alpha coefficients for each dimension of the questionnaire, as well as the overall reliability coefficient of the instrument. High data values obtained here indicate high instrument reliability, making it most appropriate for field use and data analysis with high confidence (Creswell & Creswell, 2018).

Table 4. Cronbach's Alpha Coefficient for Measuring the Reliability of the Study Instrument

Dimension	Split-Half Reliability	Cronbach's Alpha (α)
Overall Reliability	0.836	0.891

The data shown in Table 4 demonstrate a high reliability characteristic for the study instrument. Overall, the average Cronbach's alpha coefficient ($\alpha = 0.891$) exceeds the widely reported threshold of 0.70, indicating excellent internal consistency across all questionnaire items. Additionally, the split-half reliability coefficient (0.836) supports the instrument's stability. The findings indicate that the items are homogeneous and measure the intended construct reliably, making the instrument applicable for field and future data analysis. Such consistency levels align with published benchmarks in educational and social science research (Creswell & Creswell, 2018; Cronbach, 1951; Pallant, 2020).

3.2.4 The Final Version of the Study Instrument

The completed questionnaire comprises 26 items across a single main dimension. The items are measured on a five-point Likert scale, an established instrument for assessing attitudes and perceptions in education. This scale was utilized to identify teacher stances and perceptions of electronic test questions developed by artificial intelligence, with participants being able to represent their answers in varying degrees of agreement in a rational and objective manner. Because the response options are Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree, the instrument provides a more sensitive measure for quantitative analysis (Creswell & Creswell, 2018).

3.2.5 Procedures for Implementing the Study Instrument

In this study, the instrument was developed based on a review of the relevant literature and previous research on the topic as well as sufficiently covering the dimensions for measurement (Creswell & Creswell, 2018; Fraenkel et al., 2019). After being developed, the questionnaire was submitted to a panel of subject-matter experts for validation and feedback. The proposed changes were implemented, and finally, the instrument was created in its officially approved form (DeVellis, 2017). The instrument was then pilot-tested on an appropriate sample, and statistical reliability coefficients were calculated, which were significant for internal consistency and reliability in the study's primary sample (Pallant, 2020). The questionnaire was administered electronically to the study sample of public-school teachers in Sharurah Governorate via a dedicated online link; this approach enabled the investigators to collect the survey data, and the response rate subsequently increased. A total of 90 completed questionnaires were received and used in the statistical analysis. Before statistical analyses, data were checked for completeness and for outliers and invalid values (Tabachnick & Fidell, 2019). Statistical analysis was performed using the Statistical Package for the Social Sciences (SPSS) and statistical techniques (Pallant, 2020) appropriate to the nature of the study questions. The findings were interpreted in the context of the study's aims and in relation to relevant theories and data from previous studies (Creswell & Creswell, 2018). Finally, the study presented conclusions and recommendations, as well as directions for future research.

3.3 Statistical Analyses

The generated data were statistically analyzed using the Statistical Package for the Social Sciences (SPSS), widely recognized as a leading statistical software package for educational and social science research, to address the research questions and test study hypotheses (Pallant, 2020). Participants' overall response levels were assessed using descriptive statistics, including means and standard deviations. (Creswell & Creswell, 2018). Standard deviations were computed to determine the extent of variation and spread around their means (Field, 2018). Inferential analyses were also conducted using the independent-samples t-test to test differences between groups, with two categories (i.e., academic specialization group) and the Kruskal–Wallis test was used to compare groups with more than two categories when the data were not normally distributed. (Pallant, 2020; Tabachnick & Fidell, 2019).

4. Results and Discussion

4.1 Data Analysis and Presentation of Results

This section presents the analysis of the data and the results of the study, detailing a clear and systematic procedure that ensures the instrument's reliability and validity, confirms its suitability for the field test, and presents it to a representative sample of public education

teachers. The data were analyzed statistically, based on the study variables and research purpose, to address the research questions and test the main hypothesis systematically. Here, quantitative scores related to teachers' perceptions concerning artificial intelligence and constructed electronic test questions were reported by means, standard deviations, and response levels. These measures help identify response patterns among participants and the extent to which they hold favorable or unfavorable perceptions of new electronic assessment practices. Moreover, this report is also helpful for presenting the general tendencies of data and evaluating teacher perceptions in light of the data, thereby facilitating the discussion of the findings in relation to relevant theories and previous empirical studies. In order to facilitate discussion of research findings and their relationship to theories and previous empirical studies presented in the next section of the chapter (as outlined by the research methods used in quantitative education research), Creswell and Creswell (2018) and Pallant (2020).

4.2 Discussion of the Study Results

This section focuses on the interpretation of the findings and their educational implications, and the comparison of the results to the previous studies on the application of artificial intelligence in electronic assessment and education in general (Fraenkel et al., 2019). Furthermore, we examine our findings from the educational perspective of the research context, considering teachers' professional characteristics, the requirements of digital transformation in education, and the increasing role of AI in making assessment tools more efficient and objective. The discussion illustrates the alignment of these findings with prevailing perceptions in educational research that advocate the use of intelligent technologies in assessment, provided that pedagogical and ethical sensibilities are maintained (UNESCO, 2021). This section further interprets the findings to provide insights into teachers' attitudes and their implications for the development of electronic tests, educational decision-making, and AI-based assessment practices.

4.3 Results Related to the Study Questions and Hypothesis and Their Interpretation:

4.3.1 Results Related to the First Research Question

To address the first research question, EFL teachers' attitudes toward AI-generated electronic examination questions were examined. Descriptive statistics, including means, standard deviations, and ranks, were calculated for the study sample to examine responses to the questionnaire items and address this question. These statistical analyses identified the overarching trends, variation, and relative significance of teachers' perceptions of artificial intelligence-based electronically generated test questions. The results are reported as follows.

Table 5. Responses of the Study Sample Participants Regarding Teachers' Perceptions of Electronic Examination Questions Generated by Artificial Intelligence

No	Item	Mean (Average)		Standard Deviation	Ranks
		Mean Value	Level		
1	The AI-generated questions are high quality	3.87	High	1.153	24
2	Questions with AI outputs are appropriate for students at various levels	3.81	High	1.131	25
3	Questions that were generated by AI are diverse and span the entire curriculum	4.12	High	0.872	11
4	AI-generated questions acknowledge individual differences among students	3.73	High	1.149	26
5	Cheating among students is curbed with the use of AI-generated questions	3.88	High	1.069	23
6	Questions generated by AI will reduce errors of different types	4.22	Very High	0.909	3
7	AI-generated assessments meet national and international standards	4.01	High	1.076	21
8	AI-generated questions balance theoretical and practical aspects	4.01	High	0.954	19
9	The AI-generated questions can bring a mix of theory and practice together	3.98	High	1.07	22
10	AI saves EFL teachers time and effort in preparing assessment tasks.	4.06	High	0.94	17
11	AI-generated questions help reduce bias in examination design.	4.33	Very High	0.912	1
12	AI provides accurate and constructive feedback to students.	4.26	Very High	0.943	2
13	AI-generated questions are diverse and cover the full curriculum	4.18	High	0.943	9
14	AI-generated queries improve critical thinking skills	4.09	High	0.944	16
15	It is possible to get varied learning environments adapted to AI-generated question types.	4.01	High	1.022	20

No	Item	Mean (Average)		Standard Deviation	Ranks
		Mean Value	Level		
16	Simple software is required for AI-generated questions which are easy to use	4.21	Very High	Very High	6
17	AI can provide questions that are culturally and socially unbiased	4.11	High	1.033	13
18	Such assessments of AI-generated data are error-free	4.20	High	0.974	7
19	Objectivity in evaluation processes is increased through AI	4.11	High	0.988	12
20	Artificial intelligence reduces EFL teachers' control of the content in the preparation of the AI assessments	4.10	High	1.017	14
21	AI-created questions lead to greater student readiness for tests	4.04	High	1.131	18
22	AI presents opportunities for developing interactive assessments	4.17	High	0.951	10
23	AI-generated questions are suitable for use across most academic subjects.	4.22	Very High	0.98	4
24	AI-generated questions increase student preparedness for assessments	4.19	High	0.982	8
25	AI offers opportunities for developing interactive assessments	4.21	Very High	0.896	5
26	AI-generated questions are suitable for most academic subjects	4.10	High	1.05	15
	Overall Mean	4.085	High	0.765	---

Table 5 indicates that teachers hold a generally positive attitude toward the use of AI-based electronic examination questions, with a high mean score (M = 4.09). Respondents strongly agreed that AI can reduce bias and errors while providing more accurate feedback in assessment design, suggesting that AI-supported tools contribute to fairer, more objective, and reliable evaluation processes. Teachers also rated the role of AI highly in supporting curriculum alignment, including the incorporation of Bloom's cognitive taxonomy and the promotion of higher-order thinking skills. These findings suggest that educators perceive AI-based assessment as offering meaningful pedagogical benefits. In addition, teachers acknowledged that AI technologies could reduce the time and effort required for preparing tests, enhance consistency in scoring, and increase the interactive nature of electronic examinations. Although responses were consistently high, slight variations in standard deviations may reflect differences in teachers' technical competence or institutional readiness. Overall, the findings support previous research highlighting AI's potential to enhance assessment efficiency while emphasizing the continued importance of human supervision to ensure ethical and pedagogically sound evaluation practices.

4.3.2 Results Related to the Second Research Question

The second research question asked whether there were statistically significant differences in teachers' perceptions based on educational stage, years of experience, and academic specialization at the 0.05 significance level in teachers' perceptions based on educational stage, years of experience, and academic specialization at the significance level of 0.05? To examine this question, the study formulated the null hypothesis stating that there are no statistically significant differences at the significance level ($\alpha \leq 0.05$) in teachers' perceptions of AI-generated electronic examination questions with respect to the variables of educational stage, years of experience, and academic specialization. This hypothesis was tested to determine whether demographic and professional characteristics influence teachers' perceptions of AI-generated assessment tools.

First: Educational Stage Variable

Normality was assessed using the Shapiro–Wilk test for small sample groups, compared with the overall score of the variable. This strategy was employed because most parametric tests require normally distributed data, and the intermediate-stage category had a relatively small sample size. As the data were not normally distributed, non-parametric tests were used to assess differences in participants' responses.

Table 6. Kruskal–Wallis Test Results for Differences Based on Educational Stage

Comment	Significance (p-value)	Test Value	Rank	Number	Educational Stage	
Not Statistically Significant	0.351	2.096	48.44 38.67 47.24	32 23 35	Primary Intermediate Secondary	EFL teachers' perceptions

The Kruskal–Wallis test (Table 6) indicates differences in EFL teachers' perceptions of AI-generated electronic examination questions across educational stages (primary, intermediate, and secondary). The test statistic ($\chi^2 = 2.096$) is not statistically significant ($p = 0.351 > \alpha \leq 0.05$), and teachers' perceptions do not differ by educational stage. This finding suggests that EFL teachers at all levels share similar views

regarding the effectiveness and importance of artificial intelligence in electronic assessment. One possible explanation is that teachers across all educational stages face similar assessment-related challenges, including time pressure, workload, and concerns about objectivity. As artificial intelligence is used to address such issues by automating test preparation, grading, and creation (thereby increasing precision and mitigating human bias), it is considered a useful and effective means of assessment (UNESCO, 2021). Furthermore, the ongoing digital transformation of education has reshaped teaching, learning, and assessment practices, encouraging the development of integrated professional development programs that may foster consensus on AI in education and strengthen positive attitudes toward its use in assessment. (Fraenkel et al., 2019). The lack of significant differences may be interpreted as evidence of a cohesive approach to the use of artificial intelligence, supporting its claimed capacity to enhance both the throughput and quality of electronic tests across all levels of education (Creswell & Creswell, 2018).

Second: Years of Experience Variable

The normality of the years of teaching experience variable was assessed using the Shapiro–Wilk test because the groups within this variable were relatively small compared with the overall sample size. From a distributionality perspective, normality of the data is among the underlying assumptions of parametric statistical testing (Field, 2018; Pallant, 2020). For the over-10-year group, because the sample size was small, the data did not follow a normal distribution, as indicated by the Shapiro–Wilk test. Therefore, the author used the non-parametric Kruskal–Wallis test to show if differences in the responses of the study sample were significant because of the difference in the years of experience variable, appropriate when the normality assumption is violated, with small samples or ordinal data (Creswell & Creswell, 2018; Field, 2018). Statistical analysis was carried out, and the results of the analysis are presented in the table below.

Table 7. Kruskal-Wallis Test Results for Differences Based on Years of Experience

Comment	Significance (p-value)	Test Value	Mean Rank	Number	Years of Experience	
Not Statistically Significant	0.505	1.365	43.23	43	Less than 5 years	EFL Teachers' Perceptions
			45.33	32	5-10 years	
			52.37	15	More than 10 years	

The findings suggested that the significance value (Sig.) for teachers' general perceptions of AI-generated electronic exam questions with respect to years of teaching experience was 0.505, exceeding the accepted significance level ($\alpha = 0.05$), which is in favor of the non-statistically significant differences by teaching experience. It suggests that all teachers perceive artificial intelligence in assessment in a remarkably similar manner, regardless of experience. This convergence suggests a general endorsement of AI as a pedagogically useful and relatively unbiased tool for assessment and is considered pedagogically valid at professional levels (Creswell & Creswell, 2018; Fraenkel et al., 2019). This similarity could be attributed to convergences across educational settings and teacher training programs, where the development of AI capabilities is widely acknowledged as collaborative practice that promotes shared understanding of AI's teaching and evaluation processes (Pallant, 2020). In addition, teachers of all ages and levels of experience face similar issues in assessment, such as time pressure, workload demands, and the need for objectivity, which AI-enabled software can be expected to address efficiently and to consistently improve assessment practices (Albahiri & Alhaj, 2020; Field, 2018). Overall, this finding corresponds to the overall pattern of an increase in institutional acceptance of AI in educational assessments (where the value perceived is bigger than the individual's own experience in teaching) and its contribution to educational assessment quality and teaching practice (UNESCO, 2021)

Third: Academic Specialization Variable

The independent-samples t-test was used to assess whether teachers' responses differed significantly according to academic specialization. Using this check, we examined whether the average scores for these sample respondents differed significantly by this variable. The statistical analysis yielded the results shown in the following table, indicating statistical significance and differences between the two groups.

Table 8. Results of the Independent Samples t-Test for Differences in the Responses of the Study Sample According to the Academic Specialization Variable

Comment	Test Value	Test Value	Standard Deviation	Mean	No	Academic Specialization	
Not Statistically Significant	0.111	-1.232	0.73520	4.0007	52	Literary	تصورات المعلمين
			0.79993	4.2014	38	Scientific	

Statistical results confirmed that the significance value (Sig. = 0.111) exceeded the acceptable level ($\alpha = 0.05$), indicating that perceptions of AI-generated electronic examination questions were not associated with significant differences among teachers in their academic specialization. Teachers from literary studies, linguistics, and English language teaching backgrounds demonstrate similar attitudes toward the usefulness and effectiveness of artificial intelligence in assessment design. This result indicates that academic specialization might not be a decisive factor in teachers' perceptions of AI-based assessment technology. The convergence of views is likely due to the flexibility and cross-disciplinary nature of artificial intelligence, which supports common educational goals, for instance, objectivity, accuracy, and bias reduction across areas (Alhaj & Albahiri, 2022; Creswell & Creswell, 2018; Fraenkel et al., 2019) .Professional development (PD) on AI-enabled pedagogy and assessment is an additional component of the shared understanding among teachers., regardless of discipline (Pallant, 2020). The adoption of AI in modern assessment further strengthens shared professional attitudes toward

fairness, efficiency, adaptability, and timely feedback (UNESCO, 2023). However, although AI-based assessment tools are generally considered helpful in EFL education, there remain enduring pedagogical, technical, and ethical issues. Human judgment, therefore, remains vital for scoring reliability, linguistic accuracy, and validity (Alhaj & Albahiri, 2020; Alsalem, 2024; Barrot, 2026). Ultimately, effective integration of AI into assessment relies on teacher preparedness, institutional support, and continued, contextualized professional training (Alhaj, 2024; Arslan, 2025).

5. Conclusion

We investigated the attitudes of EFL teachers toward artificial intelligence-based electronic examination questions and whether these attitudes differed among educational level, teaching experience, and academic specialization. Results showed a strong, positive attitude toward AI-based assessment across all groups, supporting teachers' belief that a computerized approach can generate fair, accurate, efficient, and objective assessments. The descriptive data indicated strong support for AI's ability to reduce bias, enhance feedback quality, promote higher-order thinking, and reduce preparation time and effort. Inferential analyses further supported these findings and revealed no significant differences based on teachers' educational level, professional experience, or academic specialization, thus suggesting a shared professional perspective among teachers. This convergence reflects, for example, the greater perceived pedagogical value of AI and the alignment of contextual and professional factors, indicating the rise of digital transformation in education. However, results also highlight the need for human intervention, validity issues, and pedagogical fit. Based on the research findings, artificial intelligence was identified as a valuable and widely accepted tool for evaluating EFL learners when responsibly deployed, embedded in appropriate pedagogical and ethical considerations, and reinforced through ongoing professional development.

5.1 Recommendations

Educational institutions should adopt AI-based assessment incrementally— by implementing AI-based assessment within clear ethical and regulatory frameworks. Continual professional development is crucial for improving teachers' AI literacy, ethical awareness, and pedagogical competence. AI-generated assessment items must align with course objectives and Bloom's taxonomy to ensure cognitive validity and fairness. Policymakers must ensure sustainable technological infrastructure, ongoing monitoring, and quality assurance for the responsible implementation of AI. AI-based assessment practices supported by teacher training programs should be embedded in annual teacher development plans and should also foster collaborative research partnerships.

5.2 Limitations and Future Research

Despite the valuable insights generated by this study, several limitations should be acknowledged. First, the research was conducted with a relatively limited sample of EFL teachers from public schools in a single Saudi governorate, which may limit the generalizability of the findings to other educational contexts. Second, the study relied on a questionnaire-based survey that captured teachers' perceptions but did not examine their actual classroom practices or direct experiences with AI-generated assessment tools. Future research should therefore adopt mixed-methods approaches, including interviews, classroom observations, and experimental designs, to provide deeper insights into how AI-based assessment is implemented in real educational settings. Moreover, longitudinal studies are needed to explore how teachers' perceptions and practices evolve as AI technologies become increasingly integrated into educational systems. Comparative research across different countries, educational levels, and academic disciplines may also contribute to a more comprehensive understanding of the contextual factors influencing the adoption of AI-supported assessment in language education.

Acknowledgments

The authors extend their sincere appreciation to the Deanship of Scientific Research at King Khalid University for funding this study through the Large Research Project under Grant Number (G.R.P/1/180/1446).

Authors' Contributions

The authors made significant contributions to the conception and design of the study. Shaje Ahmed Alhomami performed the textual analysis. Mohammed H. Albahiri and Ali Albashir Mohammed Alhaj contributed to the interpretation and analysis of the collected data. All authors were involved in editing, proofreading, and critically revising the manuscript in response to the editor's and reviewers' comments. All authors reviewed and approved the final version of the manuscript and accepted responsibility for the integrity and accuracy of all aspects of the work.

Funding

The authors gratefully acknowledge funding support from the Deanship of Scientific Research at King Khalid University.

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Informed Consent

Obtained.

Ethics Approval

Approved by the Publication Ethics Committee of Sciedu Press. The journal adheres to the core practices established by the Committee

on Publication Ethics (COPE).

Provenance and Peer Review

Not commissioned; externally double-blind peer reviewed.

Data Availability Statement

The data that support the findings of this study are available upon request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data Sharing Statement

No additional data are available

Open access

References

Albahiri, M. H., & Alhaj, A. A. M. (2020). Role of the visual element in spoken English discourse: Implications for YouTube technology in EFL classrooms. *The Electronic Library, 38*(3), 531-544. https://doi.org/10.1108/EL-07-2019-0172

Alhaj, A. A. M., & Albahiri, M. H. (2020). Using pedagogic video to enhance English for specific purposes teaching programs for Saudi university students: A new prospective approach. *Arab World English Journal (AWEJ), Special Issue on CALL, 6*. https://doi.org/10.31235/osf.io/4yb5j

Alhaj, A. A. M., & Albahiri, M. H. (2022). Exploring the impact of utilizing weblog platform technology to enhance female translation students' written translation performance at King Khalid University. *Arab World English Journal, 13*(4), 43-60. https://doi.org/10.24093/awej/vol13no4.4

Ali, A. (2016). Exploring the problems of machine translation from Arabic into English faced by Saudi university students of translation at the Faculty of Arts, Jazan University. *IOSR Journal of Humanities and Social Science, 21*(4), 55-66. https://doi.org/10.9790/0837-2104025566

Alsalem, M. S. (2024). EFL teachers' perceptions of the use of an AI grading tool (CoGrader) in English writing assessment at Saudi universities: An activity theory perspective. *Cogent Education, 11*(1). https://doi.org/10.1080/2331186X.2024.2430865

American Educational Research Association. (2011). *Code of ethics*. Retrieved from https://www.aera.net/About-AERA/AERA-Rules-Policies/Professional-Ethics

Arslan, S. (2025). English-as-a-foreign-language university instructors' perceptions of integrating artificial intelligence: A Turkish perspective. *System, 131*, 103680. https://doi.org/10.1016/j.system.2025.103680

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61-75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4

Barrot, J. S. (2026). Generative artificial intelligence for automated essay scoring: Exploring teacher agency through an ecological perspective. *Assessing Writing, 67*, 100990. https://doi.org/10.1016/j.asw.2025.100990

Benek, K. (2025). EFL learners' and teachers' perceptions of AI-powered language learning technologies: Benefits and challenges. *International Journal of Instruction, 18*(2), 103-120. https://doi.org/10.29333/iji.2025.1827a

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39*(1), 370-407. https://doi.org/10.3102/0091732X14554179

Bessadat, A. I., & Korichi, S. (2025). Perceptions of AI-based assessment for formative writing in EFL: Evidence from Algerian teachers. *Aleph, 12*(3), 59-71.

Chapelle, C. A., & Voss, E. (2016). Evaluation of language assessment technology. In F. Farr & L. Murray (Eds.), *The Routledge handbook of language learning and technology* (pp. 321-336). Routledge. https://doi.org/10.4324/9781315674639

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/BF02310555

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE Publications.

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

Dörnyei, Z., & Csizér, K. (2012). How to design and analyze surveys in second language acquisition research. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 74-94). Wiley-Blackwell.

https://doi.org/10.1002/9781444347340

Dörnyei, Z., & Taguchi, T. (2010). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). Routledge. https://doi.org/10.4324/9780203864739

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). SAGE Publications.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2019). *How to design and evaluate research in education* (10th ed.). McGraw-Hill Education.

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Routledge. https://doi.org/10.4324/9781315694385

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Kaldaras, L., Akaeze, H. O., & Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education, 9*, 1399377. https://doi.org/10.3389/feduc.2024.1399377

Luc Ha, D. N., & Nguyen, A. T. (2025). Artificial intelligence-based assessment in ELT exam creation: A case study of Van Lang University lecturers. *Saudi Journal of Language Studies, 5*(1), 34-49. https://doi.org/10.1108/SJLS-06-2024-0030

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382-385. https://doi.org/10.1097/00006199-198611000-00017

Pallant, J. (2020). *SPSS survival manual* (7th ed.). Open University Press. https://doi.org/10.4324/9781003117452

Redecker, C., & Johannessen, Ø. (2013). Changing assessment: Towards a new assessment paradigm using ICT. *European Journal of Education, 48*(1), 79-96. https://doi.org/10.1111/ejed.12018

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. Polity Press.

Shermis, M. D., & Burstein, J. (2013). *Handbook of Automated Essay Evaluation*. Routledge. https://doi.org/10.4324/9780203122761

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Uddin, M. S., Islam, M. N., Haque Nirjon, M. I., Hilaly, M. R., Hosain Mazed, M. F., & Hasan, M. M. (2024). University EFL teachers' perceptions about the effectiveness of AI-enhanced e-assessments in Bangladesh: A phenomenological study. *Bulletin of Advanced English Studies, 9*(2). https://doi.org/10.31559/BAES2024.9.2.4

UNESCO. (2021). *Artificial intelligence in education: Guidance for policy-makers*. UNESCO Publishing.

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing.