Academic Journal of Research and Scientific Publishing

International, peer-reviewed scientific journal

The issue eighty-four

Publication date: 05 April 2026

ISSN: 2706-6495

Doi: doi.org/10.52132/Ajrsp.e.2026.84

Email: editor@ajrsp.com

Dedication

It is our pleasure and great privilege to present the issue eighty-four of the Academic Journal of Research and Scientific Publishing to all researchers and professor who published their research in the issue, and we thank and appreciate to all contributors and supporters of the academic journal and those involved in the production of this scientific knowledge edifice.

Academic Journal of Research and Scientific Publishing

Editorial Board

Chief Editor:

Prof. Khetam Ahmed Al-Nagdi

Advisory Members:

Prof. Abdul Hakim Ahmed Sirr Al-Khatim Jinni

Prof. Riad Said Ali Al-Mutairi

Editorial Members:

Prof. Khaled Mohamed Abdel-Fattah Abu Shaira

Prof. Azab Alaziz Alhashemi

Prof. Khaled Ibrahem Khalil Hijazi Abu Alqumsan

Dr. Abdel Razek Wahba Sayed Ahmed

Prof. Abdel Fattah Hussein

Table of Content:

Academic Journal of Research and Scientific Publishing | Vol 7 | Issue 84
Publication Date: 5 April 2026 ISSN: 2706-6495

The Role of EFL Teachers' Productive Vocabulary Size in Explaining Public School Students' Vocabulary Knowledge

Abdullah Ibrahim Alfairouz

English language lecturer at Technical College of Environmental Sciences in Buraidah and postgraduate student in the department of Educational Technology, Qassim University, Saudi Arabia

Email: a.alfairouz@tvtc.gov.sa

Received:
13 March 2026
First Decision:
15 March 2026
Accepted:
25 March 2026
Published:
5 April 2026

Copyright © 2026 by Abdullah Ibrahim Alfairouz and AJRSP. This is an open-access article distributed under the terms of the Creative Commons Attribution license (CC BY NC).

Abstract:

This study aimed to measure the productive vocabulary size of English language teachers in public schools in Saudi Arabia and to examine its role in explaining students' low vocabulary knowledge in light of previous research findings. The study was motivated by the noticeable gap between the educational goals for English vocabulary and the actual level achieved by students by the end of secondary school. Given learners' limited exposure to English outside the classroom, there is a need to investigate the sources of linguistic input available within the classroom. The study adopted a quantitative descriptive approach and involved a sample of 42 male and female English language teachers working in public schools in Qassim region. The X-LexP test (Al-Falah, 2010) was used to measure the participants' productive vocabulary size within the 5,000 most frequent words in English. The results showed that the teachers' mean productive vocabulary size was 4,058 words, indicating that most participants possessed a relatively high productive vocabulary knowledge. A comparison with the findings of previous studies also revealed a clear gap between teachers' vocabulary size and students' vocabulary knowledge at the level of the most common words in English. This finding suggests that students' low vocabulary knowledge may not be primarily attributable to teachers' productive vocabulary size, but may instead be related to other classroom factors, such as how teachers use their productive vocabulary in classroom interaction and the actual amount of English exposure provided during the lesson. The study recommends greater attention to the pedagogical use of teachers' vocabulary knowledge in classroom practices.

Keywords: Vocabulary acquisition, vocabulary size, vocabulary knowledge, teacher talk, linguistic input

1. Introduction:

Over the past few decades, English language education in Saudi public schools has received increasing attention due to the growing importance of English in the fields of economics, technology, and science. English was officially introduced into school curricula in the 1960s (Al-Seghayer, 2005), and since then, several developmental reforms have been implemented. One of the most notable reforms was the introduction of English at earlier stages of schooling, particularly at the primary level, after it had previously been limited to the intermediate and secondary levels (Mitchell & Alfuraih, 2017). This reform increased instructional time and created greater opportunities for enhancing learners' language proficiency.

Despite these efforts, empirical research continues to point to a gap between the learning outcomes targeted by the Ministry of Education and the actual outcomes of English instruction (Mitchell & Alfuraih, 2017). In terms of vocabulary development, the Ministry set a target of approximately 3,000 English words to be acquired by students by the end of secondary school. (Alsaif, 2011). However, several previous studies have shown that the actual average vocabulary size of secondary school graduates is only around 1,000 words (Al-Hazemi, 1993; Alsaif, 2011; Alhaj et al., 2019). Alsuhaibani (2025) likewise found that secondary school graduates' performance on vocabulary levels tests remains below the expected level. This concern is further reflected in the researcher's teaching experience in English courses at the Technical College of Environmental Sciences in Buraidah, where newly enrolled trainees graduating from secondary school often appear to have very limited English vocabulary knowledge. Research on second language acquisition suggests that vocabulary size is an important indicator of overall language proficiency and that a minimum knowledge of 2,000 of the most frequent words in English is essential to achieve a basic level of comprehension and communication (Nation, 2001). Given the limited opportunities for exposure to English outside formal education in the Saudi Arabia, classroom input becomes the main source of English and often the only meaningful opportunity for vocabulary acquisition.

Since the classroom constitutes the primary source of English exposure for many learners in Saudi Arabia, examining the nature of the linguistic input provided by the teacher becomes essential for understanding students' limited vocabulary development. The teacher serves as one of the main sources of linguistic input in the classroom. This role extends beyond presenting the curriculum content to include enriching the classroom environment through teacher talk, repeated exposure to vocabulary, and vocabulary use across multiple contexts. Previous research has shown that teacher talk may provide learners with vocabulary exposure equal to, or even greater than, that offered by textbook

content (Donzelli, 2007; Vassiliu, 2001, as cited in Milton, 2009). In addition, vocabulary that appears in teacher talk, particularly when repeated across different contexts, may support incidental vocabulary acquisition (Milton, 2009). This raises an important question regarding the extent to which teachers' vocabulary knowledge may affect students' vocabulary development.

1.1. Research Problem:

Despite ongoing efforts to improve English language education in Saudi Arabia, public school students continue to demonstrate low levels of English vocabulary knowledge in relation to the expected outcomes. Given that the teacher is one of the major sources of linguistic input in the classroom, there is a need to examine the productive vocabulary size of English language teachers and to explore its possible relevance to students' vocabulary knowledge in Saudi public education.

1.2. Research Questions:

- What is the average productive vocabulary size of English language teachers in public schools in Saudi Arabia?
- To what extent can English language teachers' productive vocabulary size help explain students' low vocabulary knowledge in light of previous studies?

1.3. Significance of the Study:

Many previous studies have focused on analyzing curriculum and textbook content in order to explain their effect on learners' vocabulary acquisition (Alsaif & Milton, 2012; Alhudithi, 2017; Alshumrani & Al-Ahmadi, 2022). However, the role of the teacher within the classroom still requires further investigation, given that the teacher is one of the main sources of the linguistic input to which learners are exposed. In this respect, the present study contributes to literature by offering a deeper understanding of one factor that may influence learners' vocabulary acquisition through measuring teachers' productive vocabulary size and examining its relevance to students' vocabulary knowledge.

The findings of this study may also contribute to ongoing efforts to improve English language education in Saudi public schools by drawing attention to the importance of teachers' productive vocabulary knowledge as a factor affecting the quantity and quality of the linguistic input students receive in the classroom. The findings may also help guide English teacher preparation and professional development programs toward greater attention to productive vocabulary knowledge, thereby enhancing opportunities for vocabulary acquisition and improving learners' overall language proficiency.

1.4. Definition of Terms:

Vocabulary Knowledge: Vocabulary knowledge refers to what a learner knows about words,

including their form, meaning, and use in different linguistic contexts (Nation, 2001; Milton, 2009).

Vocabulary Size: Vocabulary size refers to the number of words a learner knows in a language (Nation, 2001; Milton, 2009).

Productive Vocabulary: Productive vocabulary refers to the words that a learner can retrieve and use appropriately in productive language skills such as speaking and writing (Nation, 2001).

1.5. Limitations of the Study:

This study is limited to a purposively selected sample of English language teachers working in public schools in Qassim region in Saudi Arabia. Therefore, its findings cannot be generalized to all English language teachers in the Kingdom of Saudi Arabia. It should also be noted that the interpretation of the relationship between teachers' productive vocabulary size and students' vocabulary knowledge was based on an analytical comparison with previous studies rather than on direct measurement of students taught by the participating teachers.

2. Theoretical Framework and Previous Studies:

2.1. The Effect of Vocabulary Size on Language Skills:

In recent years, vocabulary size has attracted increasing attention as an important indicator of second language learners' proficiency. Nation (2001) argued that vocabulary forms the foundation of language skills, while Milton (2009) emphasized the close relationship between learners' overall language proficiency and the size of their vocabulary. Accordingly, several studies have examined the relationship between vocabulary size and learners' performance across different language skills.

With regard to listening, Stæhr (2009) examined the relationship between learners' vocabulary size and their listening comprehension ability. The results revealed a strong correlation ($\rho = 0.70$), highlighting the important role of vocabulary size in listening comprehension. In terms of reading and writing, Karakoç and Durmuşoğlu-Köse (2017) investigated the impact of receptive and productive vocabulary knowledge on learners' performance in reading and writing during an intensive English program. The findings showed a statistically significant positive relationship between vocabulary knowledge and performance in both skills and also indicated a relationship between learners' vocabulary knowledge and their overall language proficiency.

Similarly, Kılıç (2019) examined whether learners' performance in writing and speaking could be predicted by their vocabulary knowledge. The findings revealed a positive correlation between different dimensions of vocabulary knowledge; receptive, productive, and depth of knowledge, and learners' performance in writing and speaking. Li et al. (2024) also found that learners' vocabulary test results

may help predict both their general language proficiency and academic success in educational settings where English is used as the medium of instruction.

Overall, the findings of previous studies indicate that broader vocabulary knowledge enhances learners' ability to understand and use language in communicative contexts and improves their chances of academic success. Accordingly, it is important to examine the factors that may contribute to vocabulary development among learners, including the linguistic input to which they are exposed in the classroom, with the teacher serving as one of its most important sources.

2.2. The Role of Teacher Talk in Learners' Vocabulary Acquisition

In many countries where English is taught as a foreign language, the classroom represents the primary source of linguistic input, as learners have limited exposure to English outside the classroom. In such settings, the teacher becomes one of the main sources of classroom linguistic input. Research in second language acquisition has shown that repeated exposure to language is one of the most important factors supporting vocabulary learning (Ellis, 1994). Milton (2009) noted that the spoken input provided by teachers in class has not received sufficient attention in vocabulary acquisition research, despite its potential importance as a source of vocabulary exposure.

Donzelli (2007) was among the earliest studies to address this issue. The study compared the vocabulary learners were exposed to through teacher talk with the vocabulary available in textbooks. It concluded that teacher talk can provide an amount of vocabulary exposure equal to, or even greater than, that provided by the textbook, which highlights the important role of teacher discourse in enriching the classroom linguistic environment and enhancing learners' opportunities for vocabulary acquisition.

Recognizing the need to better understand the role of teacher discourse in vocabulary growth, some recent studies have explored this issue more closely. Bastidas (2023), for example, compared recordings of three teachers delivering beginner level Spanish as a second language lessons. The analysis focused on the amount of linguistic input provided by the teacher, the degree of student interaction, and the repetition and distribution of vocabulary in teacher talk. The study found that teacher talk constitutes an important source of classroom input and that repetition and lexical diversity in teacher discourse create opportunities for incidental vocabulary acquisition.

Grøver et al. (2022) investigated the role of teacher talk during shared reading activities in supporting second language learners' vocabulary development. The findings showed that teachers' use of varied vocabulary and their explanations of words during classroom interaction contribute to vocabulary learning. The study concluded that the quality of teacher discourse plays an important role in

supporting vocabulary development. A similar conclusion was reported by Wanzek et al. (2023), who found that the range of vocabulary used by the teacher and the way it is presented in classroom interaction are positively associated with improvements in students' vocabulary and broader language outcomes.

Farrow et al. (2025) investigated the relationship between teacher talk and learners' vocabulary development. The study focused on three features of teacher's talk; the use of sophisticated vocabulary, complex syntax, and decontextualized language that goes beyond the immediate classroom context. The findings showed that the quality of teacher discourse was related to learners' vocabulary growth, and that complex syntax was the feature most strongly associated with vocabulary improvement. These findings indicate that the language used by the teacher during classroom interaction may play an important role in supporting vocabulary acquisition.

Overall, previous studies indicate that teacher talk constitutes an important source of classroom linguistic input and that its features, such as lexical diversity, repetition, and the quality of linguistic structures, may enhance opportunities for vocabulary acquisition. These studies also show that the teacher's role extends beyond delivering textbook content, as teacher-student interaction enriches the classroom linguistic environment and broadens the range of vocabulary to which learners are exposed.

Although the present study does not directly measure teachers' actual classroom discourse, it assumes that teachers' productive vocabulary size represents one of the linguistic resources that may shape the nature of the linguistic input they provide. Accordingly, greater attention should be paid to English language teachers' productive vocabulary size as one of the factors that may influence the linguistic input available to learners.

3. Methodology:

3.1. Research Design:

The study adopted a quantitative descriptive design to measure the productive vocabulary size of English language teachers in Saudi Arabia and to provide a comparative and interpretive analysis in light of previous studies.

3.2. Population and Sample:

The population of the study consisted of English language teachers working in public schools in Saudi Arabia. The sample included 42 male and female English language teachers employed in public schools in Qassim region. All participants held a bachelor's degree in English language. Their teaching experience ranged from one year to twenty years, and they taught at different educational levels, including the primary, intermediate, and secondary levels.

3.3. Instrument:

The study used the X-LexP test (Al-Falah, 2010) to measure teachers' productive vocabulary knowledge. The test is based on the widely used X-Lex test for receptive vocabulary knowledge (Meara & Milton, 2003), which measures vocabulary knowledge within the 5,000 most frequent words in English. Previous studies have shown that these tests have acceptable levels of validity and reliability in measuring vocabulary knowledge among English language learners.

Al-Falah's (2010) test consists of 100 Arabic words distributed across five columns, with each column representing one frequency level in English and covering 1,000 of the most frequent words. Participants are required to write the English equivalent of each Arabic word. To reduce possible translation errors, the first letter of the target English word is provided. Each correct response is awarded 50 points, making the final score an approximate estimate of the participant's productive vocabulary size.

3.4. Procedures:

The teachers were contacted directly, and the purpose of the study was explained to them. They were then provided with an informed consent form to indicate their voluntary agreement to participate. After completing a brief demographic questionnaire, each participant was asked to complete the vocabulary test individually in paper form. The test was administered in person, with some participants completing it at their schools and others in locations arranged outside the school setting. The researcher monitored the testing process to prevent practices that might compromise the accuracy of the results, such as the use of dictionaries or translation tools during the test.

3.5. Statistical Methods:

Descriptive statistics, including means and standard deviations, were used to estimate the productive vocabulary size of the teachers.

4. Results:

4.1. Results for the First Research Question:

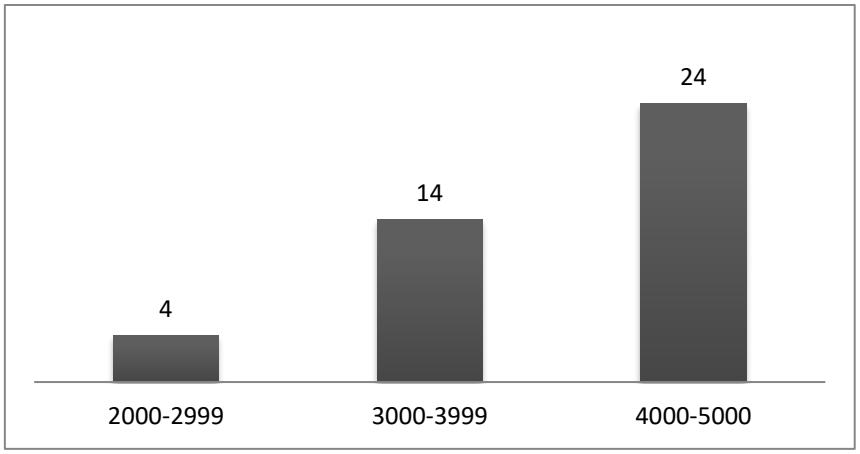
To answer the first research question, which stated "What is the average productive vocabulary size of English language teachers in public schools in Saudi Arabia?", the analysis of the test results showed that the mean productive vocabulary size of the participating teachers was approximately 4,058 words out of 5,000, with a standard deviation of 732. This indicates noticeable, though not substantial, variation in vocabulary knowledge across the sample. Table 1 presents a summary of the findings.

Table 1: *Mean and Standard Deviation of the X-LexP Test Results*

Sample	42
Mean Score	4058.33
St. Deviation	732.346

As shown in Figure 1, most teachers demonstrated a relatively high productive vocabulary size. 24 out of the 42 participants scored within the 4,000–5,000 word range, representing the largest proportion of the sample. By contrast, 14 teachers fell within the 3,000–3,999 range, while only four teachers scored within the 2,000–2,999 range.

Figure 1: *Distribution of the sample according to their scores on the X-LexP*



The results also showed that the lowest score was 2,500 words, whereas the highest score was equivalent to 5,000 words; only one participant achieved the maximum score. Overall, these findings indicate that the participating English language teachers were able, on average, to produce around 4,000 of the 5,000 most frequent English words. However, this result should be interpreted in light of the nature of the instrument used. Since the X-LexP test provided the first letter of the target English word, this may have facilitated the words retrieval and contributed to a higher estimate of productive vocabulary size. Accordingly, the present findings should be viewed as an approximate indicator of teachers' productive vocabulary size rather than a direct reflection of their actual spontaneous vocabulary use during classroom talk.

Despite this limitation, the findings suggest that most teachers possess a relatively high productive vocabulary size, which may enable them to use a sufficient range of common vocabulary in the classroom. This finding is important considering the literature suggesting that the classroom environment is a major source of linguistic input and that teacher talk can broaden the range of vocabulary to which learners are exposed (Donzelli, 2007; Vassiliu, 2001, as cited in Milton, 2009).

4.2. Results for the Second Research Question:

To answer the second research question, which stated "To what extent can English language teachers' productive vocabulary size help explain students' low vocabulary knowledge in light of previous studies?", the comparison with previous studies on the vocabulary knowledge of public school students in Saudi Arabia revealed a noticeable gap between teachers' productive vocabulary size and learners' vocabulary knowledge at the end of secondary school. Several studies have reported that students' average vocabulary size at secondary school graduation is approximately 1,000 words (Al-Hazemi, 1993; Alsaif, 2011; Alhaj et al., 2019). This comparison suggests that learners' low vocabulary knowledge may not be primarily attributable to teachers' productive vocabulary size, since the present findings indicate that teachers possess a substantially larger productive vocabulary than learners.

This gap may therefore be better explained by other classroom related factors, such as how teachers make use of their vocabulary in class, the range of words they provide to learners, and the extent to which vocabulary is repeated and recycled during classroom interaction. This interpretation may also be supported by the findings of Mitchell and Alfuraih (2017), who reported that more than 60% of English language teachers in Saudi Arabia use Arabic for about 30% of lesson time, which may reduce both the quantity and the quality of English input available to students in the classroom. These findings highlight the need for further investigation into classroom related factors, particularly the nature of the linguistic input provided by the teacher and the actual amount of English used during instruction.

5. Summary of Findings:

- The average productive vocabulary size of the English language teachers who participated in the study was 4,058 words out of the 5,000 most frequent English words, indicating that most teachers possessed a relatively high productive vocabulary size.
- A comparison of the present findings with previous studies on the vocabulary knowledge of students in public school in Saudi Arabia suggests that learners' low vocabulary knowledge is not primarily associated with teachers' productive vocabulary size.
- The findings indicate that the issue may lie not in teachers' productive vocabulary size itself, but rather in the extent to which this vocabulary is used in classroom practice in ways that provide learners with meaningful opportunities for vocabulary exposure and acquisition.

6. Recommendations:

- English teacher preparation programs should give greater attention to productive vocabulary size and its relationship to actual classroom practice.

- Educational supervision should incorporate indicators for evaluating the quality of English teachers' classroom discourse, such as lexical variety, clarity, repetition, and the extent to which vocabulary is used in interaction with students.
- Blended learning approaches should be integrated into English language teaching in ways that increase learners' exposure to vocabulary both inside and outside the classroom and support repetition and recycling through a range of digital activities.

7. Suggestions for Future Research:

- Future studies should directly examine the relationship between teachers' productive vocabulary size and the vocabulary size of their students within the same classroom to provide a more accurate statistical test of the relationship.
- Further research should measure teachers' productive vocabulary size through analysis of recordings of their actual classroom talk rather than relying solely on vocabulary tests.
- Additional studies should investigate other classroom related factors affecting vocabulary acquisition, such as the amount of language exposure, patterns of classroom interaction, types of language activities, and the use of multimedia resources.

8. References:

Al-Hazemi, H. (1993). *Low level EFL vocabulary tests for Arabic speakers* [Unpublished doctoral dissertation]. University of Wales, Swansea.

Al-Falah, K. (2010). *The EFL vocabulary knowledge of university students in the Kingdom of Saudi Arabia* [Unpublished master's thesis]. Swansea University.

Alfairouz, A. (2015). *Measuring receptive and productive vocabulary sizes of EFL English teachers in public schools in Saudi Arabia* [Unpublished master's thesis]. Swansea University.

Alhaj, A. A. M., Alwadai, M. A. M., & Albuhairi, M. H. (2019). Evaluating Saudi EFL secondary school students' performance on Paul Nation's standardized vocabulary level tests. *LLT Journal: A Journal on Language and Language Teaching, 22*(1), 126–136. https://doi.org/10.24071/llt.v22i1.1687

Alhudithi, E. (2017). *A corpus-based analysis of English vocabulary input provided in K–12 textbooks used in Saudi Arabia* (Doctoral dissertation, Colorado State University). Colorado State University Libraries. https://mountainscholar.org/handle/10217/183959

Alsaif, A. (2011). *Investigating vocabulary input and explaining vocabulary uptake among EFL learners in Saudi Arabia* [Unpublished doctoral dissertation]. Swansea University.

Alsaif, A., & Milton, J. (2012). Vocabulary input from school textbooks as a potential contributor to the small vocabulary uptake gained by English as a foreign language learners in Saudi Arabia. *The Language Learning Journal, 40*(1), 21–34. https://doi-org.sdl.idm.oclc.org/10.1080/09571736.2012.658221

Al-Seghayer, K. (2005). Teaching English in the Kingdom of Saudi Arabia: Slowly but steadily changing. In G. Braine (Ed.), *Teaching English to the world* (pp. 125–134). Lawrence Erlbaum Associates.

Donzelli, G. (2007). Foreign language learners: Words they hear and words they learn, a case study. *Estudios de Lingüística Inglesa Aplicada (ELIA), 7*, 103–126.

Ellis, R. (1994). Factors in the incidental acquisition of second language vocabulary from oral input: A review essay. *Applied Language Learning, 5*, 1–32.

Farrow, J. M., Wasik, B. A., & Hindman, A. H. (2025). Exploring the relations between teachers' high-quality language features and preschoolers' and kindergarteners' vocabulary learning. *Journal of Child Language, 52*(6), 1338–1366. https://doi.org/10.1017/S0305000924000485

Grøver, V., Rydland, V., Gustafsson, J.-E., & Snow, C. E. (2022). Do teacher talk features mediate the effects of shared reading on preschool children's second-language development? *Early Childhood Research Quarterly, 61*, 193–206. https://doi.org/10.1016/j.ecresq.2022.06.002

Kalovski-Ravenhorst, E., & Laufer, B. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30. https://doi.org/10.64152/10125/66648

Karakoç, D., & Köse, G. D. (2017). The impact of vocabulary knowledge on reading, writing and proficiency scores of EFL learners. *Journal of Language and Linguistic Studies, 13*(1), 352–378.

Kılıç, M. (2019). Vocabulary knowledge as a predictor of performance in writing and speaking: A case of Turkish EFL learners. *PASAA: Journal of Language Teaching and Learning in Thailand, 57*, 133–164. https://doi.org/10.58837/chula.pasaa.57.1.6

Li, Z., Li, J. Z., Zhang, X., & Reynolds, B. L. (2024). Mastery of listening and reading vocabulary levels in relation to CEFR: Insights into student admissions and English as a medium of instruction. *Languages, 9*(7), 239. https://doi.org/10.3390/languages9070239

López Bastidas, L. G. (2023). *Vocabulary use and classroom practices through teacher talk: A comparative and longitudinal study* (Doctoral dissertation, University of California, Davis). eScholarship, University of California. https://escholarship.org/uc/item/72q8w9nf

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition, 31*(4), 577–607. https://doi.org/10.1017/S0272263109990039

Wanzek, J., Wood, C., & Schatschneider, C. (2023). Teacher vocabulary use and student language and literacy achievement. *Journal of Speech, Language, and Hearing Research, 66*(9), 3574–3587. https://doi.org/10.1044/2023_JSLHR-22-00605

Copyright © 2026 by Abdullah Ibrahim Alfairouz, and AJRSP. This is an Open-Access Article Distributed under the Terms of the Creative Commons Attribution License (CC BY NC)

Doi: https://doi.org/10.52132/Ajrsp.e.2026.84.1

Academic Journal of Research and Scientific Publishing | Vol 7 | Issue 84
Publication Date: 5 April 2026 ISSN: 2706-6495

Validating AI-Enhanced Assessment in Higher Education (A Qualitative Multi-Phase Study on Fairness, Trust, and Cultural Adaptation in the Middle East)

By: Noura F. Assaf

PhD in Education, The British University in Dubai, United Arab Emirates

Email: Noura_assaf@hotmail.com

Received:
9 March 2026
First Decision:
15 March 2026
Revised:
22 March 2026
Accepted:
28 March 2026
Published:
5 April 2026

Copyright © 2026 by Abdullah Ibrahim Alfairouz and AJRSP. This is an open-access article distributed under the terms of the Creative Commons Attribution license (CC BY NC).

Abstract:

Artificial intelligence is reshaping assessment in higher education by enabling scalable, adaptive, and data-driven feedback. Yet, some concerns about reliability, fairness, and contextual validity still persist, especially in multilingual and culturally diverse systems, which research has generally overlooked. Hence, this study investigates how AI-based assessment tools can be validated rigorously and transparently to build trust among educators, developers, and policymakers. To do so, a qualitative multi-phase study was followed with a PRISMA-guided systematic review and semi-structured interviews involving. The review identified persistent methodological gaps: few replication studies, overreliance on accuracy metrics, and limited attention to fairness or cultural adaptation, while the interviews revealed convergent priorities, namely the need for replicability, linguistic and cultural sensitivity, algorithmic transparency, stakeholder co-design, and continuous monitoring with ethical oversight. Synthesizing both strands, the study proposes a Four-Stage Validation Framework incorporating algorithmic validation, contextual adaptation, stakeholder engagement, and continuous monitoring. This framework aims to reframe validation as an iterative form of institutional governance that is essential for equitable and trustworthy AI assessment. Although anchored in Middle Eastern contexts, the findings offer transferable guidance for educational settings worldwide seeking to align technological innovation with human-centered and culturally responsive assessment practices.

Keywords: AI assessment; validation; higher education; fairness; transparency; stakeholder engagement; Middle East; educational technology

1. Introduction:

Artificial intelligence (AI) has become a defining feature of educational transformation in higher education; automated scoring, adaptive testing, and data-driven feedback systems are now integral to institutional strategies for improving teaching and learning towards efficiency as well as effectiveness. These tools often promise efficiency and personalization that are generally difficult to achieve through conventional assessment practices. Prior work shows that AI-based assessment can enhance formative feedback and support evidence-based instruction (Boulhrir & Hamash, 2025; Holmes et al., 2019; Luckin et al., 2016). Across the Middle East, governments have invested heavily in digital education and international benchmarking in an attempt to modernize higher education institutions and to expand access to high-quality learning opportunities (Alghamdi & Li, 2022; World Economic Forum, 2021).

Despite this momentum, there is a growing body of research that warns about the adoption having outpaced methodological scrutiny. This is because many AI systems are introduced without thorough validation of their reliability, fairness, or cultural suitability (Zawacki-Richter et al., 2019; Williamson & Piattoeva, 2022). For clarity, validation refers here to the systematic process of verifying that assessment outcomes are consistent, interpretable, and equitable across diverse populations; it requires testing both algorithmic accuracy and contextual fit, including linguistic and curricular alignment. In practice, only few institutions report such procedures, and even fewer make them replicable. This gap presents ethical and practical risks: when scoring rules or data sources remain opaque, trust in digital assessment erodes among faculty, students, and policymakers.

Available scholarship illustrates this tension between innovation and methodological integrity, as many studies frequently optimize for predictive accuracy but seldom address replicability or fairness, especially in multilingual environments (Al-Zahrani & Alasmari, 2025). Classrooms in the Middle East are linguistically and culturally diverse; they often include students following different curricular traditions and communicating in both Arabic and English (Mazawi, 2020). Under such conditions, models trained primarily on Western data may misrepresent student ability. Without region-specific validation; hence, AI systems risk amplifying inequities rather than reducing them.

Higher education institutions stand at the center of this challenge; they increasingly rely on AI-supported assessments to manage large enrollments and provide rapid feedback, and even admission decisions, yet they operate within social contexts that demand transparency and fairness. A natural question arises here: how can AI assessment systems be validated in ways that preserve both technical rigor and cultural responsiveness? Addressing this question requires empirical evidence from two fronts: existing research on validation practices and the lived experiences of those implementing AI tools in the classrooms.

To examine these issues, this study adopts a mixed-methods design that integrates a review of 25 systematically selected empirical studies with interviews of (a matching number by mere coincidence) 25 stakeholders from universities and educational authorities in the UAE, SA, and Qatar. The review aims to identify methodological patterns in literature, whereas the interviews reveal practitioners' and policymakers' perspectives on fairness, transparency, and trust. This will provide a foundation for constructing a validation framework that is tailored to multilingual and multicultural higher education systems.

Hence, the research is guided by three Research Questions (RQ):

RQ1. How have existing studies addressed the methodological validation of AI-based assessment systems in higher education, particularly regarding replicability, fairness, and cultural adaptability?

RQ2. What are the perceptions and expectations of key stakeholders concerning the reliability, fairness, and transparency of AI assessment systems in the Middle Eastern context?

RQ3. How can insights from the literature and stakeholder perspectives be synthesized to develop a culturally responsive validation framework for AI-driven assessment in higher education?

2. Methods:

2.1. Research Design:

A sequential qualitative multi-phase design integrating systematic review and semi-structured interviews was adopted; this design allowed evidence from prior research to inform data collection and interpretation of stakeholder perspectives. The systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines to ensure transparency and replicability. Following the analysis of the selected studies, the interview phase built on the findings to explore how policymakers, developers, and educators in higher education perceive and implement AI-based assessment validation (see key interview questions in the appendix). This two-stage structure aligns with the three research questions to be answered and supports triangulation between empirical and experiential evidence.

2.2. Systematic Review:

To strengthen coverage and replicability, we ran verification searches in Web of Science (Core Collection), ERIC, and IEEE Xplore, in addition to Scopus, arXiv, ResearchGate, and Google Scholar, using broadened Boolean strings spanning reliability, fairness/bias, transparency/explainability, trust, and cultural adaptation. No additional eligible studies meeting our inclusion criteria were identified;

the retained set remains n = 25. Representative search strings included, for example: ("artificial intelligence" or AI) and (assessment or testing or grading) and (validation or validity or reliability or fairness or bias or transparency or explainability or trust or "cultural adaptation").

Screening was conducted independently by two reviewers with a third adjudicator for conflicts. Title/abstract screening and full-text eligibility were assessed using predefined criteria (Table 1). We also conducted backward and forward citation chasing. Because the verification searches did not change the eligible set (n = 25), the PRISMA stages reported earlier remain valid.

Coding Scheme Transparency and Analytic Traceability:

A codebook was developed a priori and refined iteratively. Top-level codes included: (1) methodological rigor (replicability, disclosure), (2) fairness and cultural adaptation (subgroup analyses, multilingual data), (3) transparency and explainability (model cards, user-facing rationales), and (4) stakeholder participation (co-design, feedback loops). Each code was defined with inclusion/exclusion rules and example indicators. A study × code matrix was created to trace how each included article contributed to each theme; representative studies are cited under each theme in the Results.

Coder training involved joint calibration on five pilot articles followed by independent coding with periodic consensus meetings. Disagreements were resolved through discussion; decision trails were logged to ensure auditability.

Search Strategy and Data Sources

The review focuses on empirical studies published between 2015 and 2024 that examine the validation, fairness, or reliability of AI-driven assessment in higher education. Searches were conducted in Scopus, arXiv, ResearchGate, and Google Scholar using predefined keywords: *"AI in assessment," "validation," "replicability," "fairness," "educational technology,"* and *"cultural adaptation."* The search targeted peer-reviewed journal articles, conference proceedings, and authoritative reports in English and Arabic. Additional sources were identified through backward and forward citation tracking.

Duplicates were removed before screening, and titles and abstracts were reviewed for relevance and the remaining papers underwent full-text evaluation according to explicit inclusion and exclusion criteria (Table 1).

Each article meeting these criteria was coded using NVivo to identify methodological trends, validation approaches, and reported challenges.

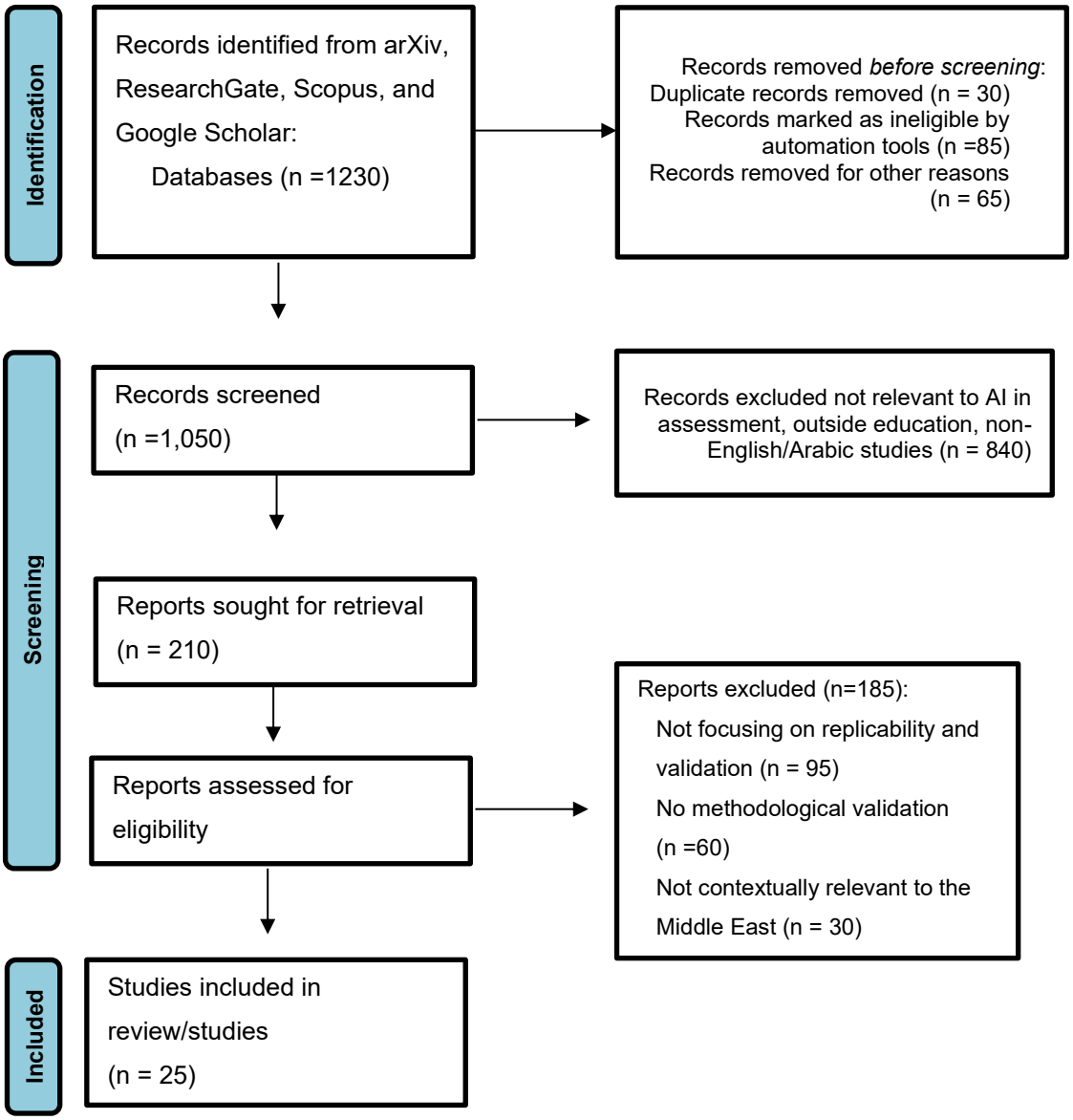
Table 1: Selection Criteria for the Systematic Review

Category	Inclusion Criteria	Exclusion Criteria
Publication Type	Peer-reviewed journal articles, conference papers, or official research reports	Non-scholarly sources (blogs, editorials, news articles, non–peer-reviewed essays)
Publication Years	2015–2024	Published before 2015 or after 2024
Language	English or Arabic	Other languages
Topic Focus	Research examining AI-based or algorithmic assessment systems used for grading, feedback, or evaluation in education	Studies on AI for general learning analytics, tutoring systems, or non-assessment purposes
Validation Dimension	Reports evidence or discussion of validation, reliability, fairness, replicability, or transparency in AI assessment	Studies mentioning AI assessment tools but without reference to validation, reliability, or fairness
Data and Methodology	Provides sufficient methodological details (sample, data source, validation method, or evaluation metrics)	Conceptual papers or commentaries lacking empirical or methodological description
Relevance to Middle Eastern or comparable multilingual contexts	Studies conducted in, or providing transferable insights to, multilingual or culturally diverse educational settings	Studies limited to homogeneous, monolingual, or non-comparable contexts

Review Process and Synthesis:

The screening process began with 1,230 initial records. After removing duplicates and non-relevant studies, 210 full texts were assessed, resulting in 25 retained for analysis (see details in Figure 1 below). Each study was coded according to publication year, research design, AI application type, data context, and validation dimension. Thematic synthesis grouped findings into three overarching categories: (1) methodological rigor and replicability, (2) fairness and cultural adaptability, and (3) stakeholder or user engagement in validation. These categories form the analytical foundation for the interview phase and the development of the proposed validation framework

Figure 1: PRISMA Flow Diagram illustrating the selection process



2.3. Stakeholder Interviews:

Participants and Setting

The interview phase included 25 stakeholders drawn from higher-education institutions and related agencies across the UAE, SA, and Qatar. these three countries were purposefully chosen because they represent the most active hubs of AI integration and educational reform in the Arab world, each with distinct policy environments, institutional structures, and levels of digital infrastructure development (Alghamdi & Li, 2022; Kayan Fadlelmula & Qadhi, 2024). This diversity enabled a comparative perspective on how AI assessment validation is perceived and enacted across varied governance and curricular contexts within the broader Middle Eastern higher education landscape. Other countries in

the region were not included at this stage due to differences in the availability of AI-enhanced assessment systems, institutional access, and the maturity of educational technology governance frameworks; however, future research should expand sampling to include additional MENA countries for broader representativeness.

Participating interviewees comprised 10 policymakers, 8 AI developers, and 7 university instructors who had direct experience with AI-driven assessment tools. Purposeful sampling ensured diversity in institutional type, discipline, and level of digital integration. Each participant had at least five years of professional experience in technology-enhanced assessment or educational governance.

Interview Protocol

Interviews followed a semi-structured guide organized around five thematic areas: experience with AI assessments, reliability and fairness, cultural and linguistic adaptation, stakeholder collaboration, and policy directions for ethical implementation. The interview questions were developed based on the thematic gaps and methodological patterns identified in the systematic review phase of this study, supplemented by validated frameworks in the educational AI and assessment validation literature (Zawacki-Richter et al., 2019; Bond et al., 2020; Burstein & LaFlair, 2024). This evidence-based grounding ensured that the questions were theoretically anchored and empirically relevant. Interviews were conducted either face-to-face or online, depending on participants' location and accessibility. Face-to-face interviews were primarily carried out with locally available participants, while online interviews were conducted with participants based in other countries using secure videoconferencing platforms such as Microsoft Teams and Google Meet. Each session lasted between 45 and 60 minutes and was conducted in either English or Arabic, according to participants' preferences. All interviews were audio-recorded with participants' consent and transcribed verbatim for analysis.

2.4. Data Management:

Transcripts and anonymized notes were stored on encrypted drives accessible only to the researcher, and each participant was assigned a code to maintain confidentiality. After transcription, researcher conducted independent readings to ensure accuracy and contextual fidelity before analysis.

2.5. Data Analysis:

The systematic review data and interview transcripts were analyzed using NVivo software. For the review analysis, a descriptive coding framework identified validation methods, datasets used, and indicators of fairness or transparency. The data collected from the interviews were analyzed thematically following Braun and Clarke's (2006) six-phase procedure. The initial codes captured recurring issues such as replicability, bias, and transparency; they were then grouped into higher-order

themes: validation and replicability, fairness and cultural sensitivity, transparency and explainability, stakeholder engagement, and continuous monitoring. The identified patterns were compared across stakeholder groups to identify convergent and divergent perspectives. This process culminated in synthesizing insights from both datasets to construct a four-stage validation framework that emphasizes algorithmic validation, contextual adaptation, stakeholder engagement, and continuous monitoring.

2.6. Ethical Considerations:

All participants provided written informed consent prior to data collection, and institutional ethical approval was obtained from the lead author's university to ensure compliance with regional research governance standards. Participants were informed of their right to withdraw at any time and assured of anonymity in all publications. Responsible AI use is also taken into consideration; OpenAI's GPT-5, Microsoft Copilot, and Grammarly were used under the researcher' supervision for language refinement, grammar, and editorial consistency; no analytical content was generated by the model. The author assumes full responsibility for interpretation and conclusions.

3. Results:

This section presents findings from the systematic review and the stakeholder interviews. The two strands of evidence are reported separately and then integrated to inform the proposed validation framework. All results, hence, are interpreted in light of the research questions in the higher education context as the anchor, though the implications extend to other levels of education.

3.1. Systematic Review Findings:

To deepen synthesis beyond descriptive percentages, we link themes to exemplars. For instance, Arabic AES work (e.g., Lotfy et al., 2023; Ghazawi & Simpson, 2024) demonstrates reliability gains yet underscores the need for subgroup fairness checks; educator-focused studies in the region (e.g., Khlaif et al., 2024; Al-Abdullatif, 2024) show how trust and AI literacy mediate adoption; and policy-oriented analyses (e.g., Traidi, 2024) highlight governance gaps. These cases ground the themes in concrete empirical contexts and clarify how methodological choices intersect with context.

Overview

As stated before, the systematic review identified 25 eligible studies published between 2015 and 2024. Collectively, these studies examined AI-based assessment tools across higher-education contexts in Asia, Europe, North America, and the Middle East. Figure 1 (PRISMA Flow Chart) summarizes the screening process, and Table 2 presents the analytical matrix of included studies.

Table 2: Matrix of the included studies

Study	Main Topic / Idea	Key Findings / Contributions
Aboalela (2023)	Saudi Arabia context; generative AI used for assessment items; addresses validity of question generation	Explores the use of ChatGPT for creating valid assessment items; highlights potential and challenges for reliability and fairness in Middle Eastern accreditation contexts
Alghamdi and Li (2022)	Review of AI applications and challenges in Middle Eastern education	Summarizes AI adoption trends, highlights challenges including fairness, validity, cultural adaptation, and provides recommendations for region-specific implementation
Al-Khalidi and Al-Shehri (2023)	Fairness and interpretability of AI assessment models	Found that AI models can show bias if not carefully designed; interpretability improves trust and fairness in assessment outcomes in Saudi Arabia
Al-Mutawa and Javid (2023)	Validity of automated essay scoring in Gulf universities	Demonstrated that AES systems produce scores aligned with human raters, but cultural and linguistic context affects validity and reliability
Mahmoud and Ameen (2021)	AI for equitable assessment in Egypt	AI systems can reduce human bias; equitable access and careful design needed to ensure fairness and reliability
Lotfy et al., (2023)	Arabic automated essay scoring	Proposed ML-based AES system; found strong reliability and validity compared to human scoring, supporting automated evaluation in Arabic contexts
Ghazawi and Simpson (2024)	Arabic AES using BERT	Introduced Arabic AES dataset; BERT-based AES achieved high agreement with human raters, demonstrating effective scoring in linguistic context

Alzahrani (2022)	Systematic review of AI in Arab education	Highlights trends, gaps, and challenges, including validation, fairness, and adaptation to local cultures; useful for guiding AI assessment research
Bashendy et al., (2024)	Arabic AES corpus with trait-specific annotations	Created dataset supporting cultural and linguistic adaptation; provides benchmark for scoring validity, reliability, and fairness in AES
Khlaif et al., (2024)	Teacher perspectives on AI assessment tools	Revealed teachers' concerns on reliability, fairness, and validity; adoption requires training and context-specific adaptation
Hobeika et al., (2024)	Validation of AI literacy scale in Arabic	Successfully validated AILS across multiple Arab countries; emphasizes cross-cultural adaptation and psychometric reliability
Aboalela (2023)	AI for generating assessment items	Explores generative AI for creating valid assessment items; highlights potential and challenges for reliability and fairness in Middle Eastern accreditation
Traidi (2024)	Policy implications of AI in education	Discusses fairness and equity implications; stresses need for culturally and linguistically adapted AI assessment systems
Al-Sharoufi (2022)	English writing curriculum & technology integration	Focus on alignment of assessment with AI tools; highlights validity and cultural adaptation issues in Gulf region
Alazemi (2024)	AI-supported formative assessment	Demonstrates that AI-based formative tools improve learning outcomes; highlights reliability and fairness considerations in Arabic instruction
Alobed et al., (2021)	Arabic AES using hybrid ML	Hybrid AES system shows strong validity and reliability; emphasizes linguistic and cultural adaptation for scoring accuracy

Al-Zahrani and Alasmari (2024)	Saudi higher-ed context; touches on assessment/ grading, fairness, bias, & AI's educational impact	Discusses the ethical and social implications of AI in higher education; highlights risks and considerations for fairness and bias in AI-based assessment practices
Calderwood (2024)	UAE context; student perceptions of generative AI use in assessment	Reveals complex interactions between student agency, tool affordances, and academic ethics; highlights concerns regarding fairness, reliability and integrity when generative AI is used for assignments
Kayan Fadlelmula and Qadhi (2024)	Systematic review of AI in GCC higher education	Identifies methodological gaps in assessment validity, fairness, reliability; provides recommendations for AI adoption in Gulf context
Khan et al., (2024)	UAE/Iraq/UK mixed context: develops a framework for AI-based assessment; policy, validity, fairness issues featured	Proposes a conceptual framework for AI-driven assessment; emphasizes policy alignment, validity, and fairness considerations, particularly in Middle Eastern higher education contexts
Alenezi and Alenezi (2024)	Saudi Arabia: deployment of AI-powered formative assessment in higher education; addresses cultural/linguistic localization and feedback via AI	Investigates how AI formative assessment tools are implemented across Saudi universities; highlights challenges in fairness, reliability, and validity in local contexts
Al-Abdullatif (2024)	Saudi Arabia: explores educators' acceptance of GenAI for assessment and teaching; touches on trust, fairness, reliability through human-tool interface	Explores teachers' perceptions of generative AI, revealing that trust and AI literacy influence adoption; emphasizes fairness and reliability considerations in classroom assessment

Khan et al., (2024)	Afghanistan/region, AI-assisted assessment of young learners; addresses validity and reliability of AI assessments	Evaluates AI-assisted vocabulary tests for young learners; highlights reliability challenges and the need for culturally adapted assessment approaches
Al-Kaabi (2024)	UAE: effects of AI in assessment and learning; implications for valid assessment via AI	Investigates AI's impact on critical thinking and assessment outcomes; highlights validity and reliability concerns when AI supports higher education evaluation
Khlaif et al., (2024)	Middle East: surveys teachers' perceptions of GenAI in assessment contexts; explores fairness, validity, adoption	Provides insights on educators' acceptance and challenges in using generative AI for assessment; emphasizes importance of fairness, reliability, and contextual adaptation in Arab higher education

Across the analyzed studies, four dominant trends emerged. To begin with, most research emphasized algorithmic performance metrics such as predictive accuracy and internal reliability while providing limited discussion of cross-context replication. Only a few studies tested fairness or bias across multilingual or multicultural groups. More interestingly, transparency and interpretability of AI scoring remained underreported, particularly in commercial systems. Last but not least, relatively few publications involved educators or policymakers in validation design, which suggests that stakeholder engagement remains underdeveloped in empirical practice. Below are the key patterns identified by the analysis.

Key Pattern:

1. Methodological Rigor and Replicability

Quantitatively, 68% of studies employed experimental or quasi-experimental designs, yet only three explicitly replicated findings across independent samples. Studies relying on proprietary algorithms rarely disclose model parameters, which makes replication nearly impossible. Researchers often reported internal consistency coefficients above .80, but few provided inter-rater reliability or test–retest data. These patterns indicate that reproducibility in AI validation remains the exception rather than the rule.

2. Fairness and Cultural Adaptability

Only five studies analyzed differential performance by language or cultural group. Those who did found measurable disparities in AI scoring for bilingual learners. For instance, one Middle Eastern study noted a 12 percent lower agreement rate between human and AI raters for essays containing Arabic-English code-switching. This gap reflects the broader challenge of training language models on data dominated by Western linguistic norms.

3. Transparency and Explainability

Most of the papers (72%) mentioned transparency as desirable but did not operationalize it. Only two offered visual or textual explanations of algorithmic decisions accessible to instructors. The lack of standardized reporting protocols is a limitation of both peer verification and practitioner confidence.

4. Stakeholder Involvement

Engagement of educators, students, and policymakers in validation was limited, as only 6 studies incorporated user feedback loops or participatory design elements. This omission can be viewed as a suggestion that validation remains primarily a technical process, disconnected from classroom realities and institutional governance.

These findings collectively indicate a fragmented methodological landscape; validation research in AI-based assessment has advanced technically but lags in contextual sensitivity and human participation. These patterns provided the foundation for the interview phase, which explored how stakeholders perceive and address these shortcomings in practice.

3.2. Stakeholder Interview Findings:

Policymakers emphasized accountability instruments (mandatory model documentation, periodic bias audits); educators prioritized score explainability and alignment with local rubrics; developers focused on data access, licensing constraints, and the feasibility of interpretable architectures. This differentiation clarifies how validation requirements vary by role and informs the staged framework that follows.

Overview

Insights from twenty-five stakeholder interviews complement the literature by revealing how methodological challenges identified in prior studies play out in practice. Participants included ten policymakers, eight AI developers, and seven higher-education instructors across the 3countries (UAE, SA, and Qatar). Discussions were primarily centered on five thematic areas, with follow-up questions when needed: validation and replicability, fairness and cultural sensitivity, transparency and

explainability, stakeholder engagement, and continuous monitoring with ethical oversight. These perspectives clarify, in particular, how those in charge of designing, governing, and deployment/use of AI assessment systems interpret trust, reliability, and fairness in higher-education settings.

Theme 1. Validation and Replicability

Participants steadily described inconsistent AI scoring as the main threat to confidence in digital assessment; instructors reported that identical essays sometimes produced different scores across submissions, a pattern echoed by policymakers who questioned the absence of regional benchmarks. Developers traced the problem to limited access to local training data and sparse model documentation. One educator from the UAE noted, *"The same essay scored differently when submitted a day apart which is inconsistent and undermines trust."* These accounts point to a fundamental issue: reproducibility is rarely verified in real teaching contexts. Validation procedures must therefore include replication across institutions and learner populations, not only laboratory testing.

Theme 2. Fairness and Cultural Sensitivity

Concerns about fairness dominated the interviews; educators observed that bilingual or culturally hybrid expression often triggered lower scores, and policymakers framed fairness as an ethical duty within educational reform. Developers acknowledged that many algorithms rely on Western-based contexts datasets lacking regional dialects. A Saudi developer remarked, *"We need to expand our datasets to include regional dialects; only then can we talk about fairness."* These reflections point out that cultural adaptation is not peripheral to validation because it defines whether AI assessments measure ability or linguistic conformity. Fairness requires bilingual corpora, context-appropriate prompts, and regionally informed evaluation metrics.

Theme 3. Transparency and Explainability

Transparency emerged as a cornerstone of trust when educators requested visible rationales for automated scores, so they could justify grades to students, who were reported to be often skeptical of the ability of AI to score accurately and fairly. Policymakers argued in support of the educators' concerns that explainability should be a regulatory requirement rather than a technical enhancement, and developers conceded that interpretable models remain uncommon. One of the policymakers explained, *"Transparency isn't about open code; it's about accessible reporting educators can interpret."* These perspectives clearly suggest that transparency in higher education depends on human-readable explanations of scoring logic that students understand and accept as fair. Validation protocols must therefore document how results are generated and communicated to users.

Theme 4. Stakeholder Engagement

Collaboration among educators, developers, and policymakers surfaced as both a value and a validation mechanism in the sense that teaching faculty members who participated in pilot testing reported greater confidence in AI tools and quicker identification of contextually irrelevant items. According to the developers, faculty engagement as co-designers reduced debugging time and improved cultural alignment. This was also emphasized by policymakers who regarded such engagement as a safeguard against misuse. *"Validation should include everyone who uses the system,"* as stated by one of the interviewees. The findings show and stress the fact that participatory validation strengthens both technical performance and social legitimacy, transforming assessment from a vendor-driven product into a shared institutional practice.

Theme 5. Continuous Monitoring and Ethical Oversight

One of the policymakers summarized the general sentiment towards monitoring and ethical oversight saying, *"Validation isn't a checkbox; it's a cycle that keeps AI accountable over time."* As a matter of fact, all participating policymakers and most developers described validation as an ongoing governance process and that bias detection and curriculum alignment were continuous responsibilities rather than final checks. Teachers also called for continuous periodic reviews to ensure that models evolve with pedagogical changes. This reinforces that ethical oversight and adaptive monitoring are essential for sustaining fairness and reliability after deployment; hence, regular audits and independent evaluations should become standard components of institutional quality assurance.

Summary of Themes

The five themes above (described in more detail in Table 3) form a coherent narrative linking technical rigor to human-centered accountability; participants generally envision trustworthy AI assessment as a reproducible system, fair across languages and cultures, transparent in operation, collaboratively governed, and continuously monitored. The close alignment of expectations with the methodological gaps identified in the systematic review indicates a clear convergence between scholarly and practitioner perspectives.

Table 3: Summary of Stakeholder Interview Themes and Implications for Validation

Theme	Representative Perspective	Illustrative Quote	Implication for Validation
Validation and Replicability	Scoring results vary across contexts; lack of	"We can't rely on results that vary from	Develop region-specific validation standards and

	regional benchmarks undermines trust.	one institution to another." Policymaker	replication studies across sites.
Fairness and Cultural Sensitivity	AI models penalize bilingual expression and ignore local conventions.	"Language differences are mistaken for lack of understanding." Instructor	Use bilingual data and culturally adapted test items to ensure equitable outcomes.
Transparency and Explainability	Educators need clear rationales for AI-generated scores.	"If teachers can't explain the score, they won't trust it." Instructor	Require transparent reporting and user-friendly explainable-AI interfaces.
Stakeholder Engagement	Co-design among teachers, developers, and policymakers improves contextual fit.	"Validation should include everyone who uses the system." Policymaker	Embed feedback loops and participatory validation in each implementation phase.
Continuous Monitoring and Ethical Oversight	Validation must remain dynamic and accountable over time.	"Validation should evolve with the system." Developer	Implement periodic bias audits and adaptive oversight within institutional policy.

In short, the interviews shed light on the fact that stakeholders and prior research studies diagnose the same weaknesses (e.g., replicability, fairness, transparency, and contextual adaptation), but from complementary standpoints. In other words, stakeholders add the pragmatic dimension of collaboration and continuous governance, which were largely absent from published studies. The following sub-section (3.3) integrates both data strands to articulate a four-stage framework for validating AI-enhanced assessment in higher education.

3.3. Integrated Interpretation:

Table 4: Evidence-to-Framework Mapping (Review ↔ Interviews ↔ Framework Stages)

Framework Stage	Evidence from Systematic Review	Evidence from Stakeholder Interviews
Algorithmic Validation	Sparse replication; limited disclosure in proprietary systems; accuracy emphasized over reproducibility.	Inconsistent scores reported; call for regional benchmarks and cross-institution checks.

Contextual Adaptation	Few subgroup analyses; bilingual/code-switching issues underexplored in many studies.	Bilingual expression penalized; demand for Arabic–English data and rubric localization.
Stakeholder Engagement	User participation rare; validation treated as technical task.	Co-design improves fit and trust; pilots reduce debugging time and resistance.
Continuous Monitoring & Ethical Oversight	Post-deployment auditing rarely reported in the literature.	Monitoring framed as governance duty; periodic bias audits and public reporting requested.

The integration of the systematic-review and interview findings reveals a consistent pattern: the core challenges of validating AI-enhanced assessment in higher education are not isolated technical problems but interdependent social and methodological processes. Across both datasets, four recurrent dimensions, replicability, fairness, transparency, and stakeholder participation, define what trustworthy validation entails. A fifth dimension, continuous monitoring and ethical oversight, emerged uniquely from practitioner experience and extended the empirical literature toward post-deployment governance.

Converging Evidence:

The systematic review showed that published studies focus heavily on algorithmic accuracy yet seldom verify results across contexts. Stakeholders echoed this concern, describing inconsistent scoring as a daily obstacle to adoption. Similarly, both sources highlighted fairness as a chronic weakness: the literature notes the absence of multilingual testing, and practitioners confirmed that bilingual students are often disadvantaged by models trained on monolingual data. Transparency occupies the same dual position. Scholars report a lack of open documentation, and educators explained that they cannot defend AI-generated grades without clear rationales. Both data-sources/strands are indicative of the idea that limited stakeholder participation constrains credibility; when validation remains a purely technical exercise, users perceive AI systems as opaque and externally imposed.

The additional practice-based theme (e.g. continuous monitoring) advances this conversation. Stakeholders insisted that validation is cyclical and must evolve as curricula, datasets, and technologies change. This insight reframes validation from a pre-implementation test to an enduring governance function. Together, the five dimensions form the conceptual scaffolding for a human-centered validation model.

The Four-Stage Validation Framework (FSVF)

Synthesizing the convergent and emergent insights from the literature and interviews yields a Four-Stage Validation Framework (FSVF)that redefines validation as an iterative partnership between technology and institutional practice. Each stage corresponds to a distinct but interlinked domain of activity:

1- Algorithmic Validation

This stage focuses on verifying the technical reliability of AI assessment models through replicable statistical procedures; it includes cross-sample testing, inter-rater comparisons, and transparent documentation of algorithms and training data. Its purpose is to establish reproducible evidence that the system performs consistently across student cohorts.

2- Contextual Adaptation

Validation must extend beyond technical accuracy to ensure cultural and linguistic fit, which means that contextual adaptation involves integrating bilingual data, aligning scoring rubrics with local curricula, and adjusting for sociolinguistic variation common in Middle Eastern higher-education settings. This stage operationalizes fairness as contextual relevance rather than universal standardization.

3- Stakeholder Engagement

Validation gains legitimacy only, and only, when those affected participate in it. Engagement of stakeholder participation is part of a participatory design cycles that include educators, students, developers, and policymakers. Co-design workshops, pilot testing, and structured feedback loops translate expert validation into socially recognized credibility. Engagement also serves as a professional-learning mechanism, increasing digital assessment literacy among instructors.

4- Continuous Monitoring and Ethical Oversight

This is the final stage which institutionalizes validation as an ongoing governance process; it involves regular audits of algorithmic bias, monitoring for model drift, and alignment with evolving ethical and curricular standards. Independent oversight bodies within universities can perform these reviews, ensuring accountability and public transparency.

It is worth noting that these stages are cyclical rather than sequential; algorithmic validation initiates the process, contextual adaptation situates it, stakeholder engagement sustains it, and continuous monitoring renews it. Hand in hand, they create a dynamic feedback system that balances innovation with accountability. The framework places validation as an evolving ecosystem embedded in educational practice, more so than as a static verification checklist.

The synthesis demonstrates that methodological rigor and human-centered governance are mutually reinforcing. Technical improvements without contextual and participatory validation will not build trust; conversely, stakeholder dialogue without algorithmic evidence cannot guarantee fairness. The FSVF offers a structured path to reconcile these dimensions. Although derived from Middle Eastern higher-education contexts, its principles (e.g., replicability, cultural adaptation, participatory design, and sustained oversight) are transferable to other multilingual and data-rich environments seeking equitable digital assessment.

4. Discussion:

For the purpose of interpreting the results and answering the research questions, this study focused on examining how AI-enhanced assessment can be validated in higher education so that it is both methodologically rigorous and socially credible. Evidence from the 25 reviewed studies and the stakeholder interviews points to a persistent tension between the drive for technical innovation and the need for contextual trust. Across both datasets, reproducibility, fairness, and transparency remain fragile, and stakeholders' calls for continuous monitoring suggest that validation is not a one-off procedure but a sustained institutional responsibility. These findings extend prior syntheses of AI in higher education that emphasized efficiency and scalability (Holmes et al., 2019; Zawacki-Richter et al., 2019) by demonstrating that credibility depends equally on cultural adaptation and participatory governance.

4.1. Methodological validation in higher education (RQ1)

The literature and stakeholder accounts converge on the conclusion that reproducibility is the weakest link in AI assessment research. Studies report high accuracy yet seldom provide the documentation necessary for replication or cross-context testing. This captured pattern likely reflects structural incentives within educational technology research, where proprietary datasets and competitive innovation limit transparency (Williamson & Piattoeva, 2022). Teachers' reports of inconsistent scoring and developers' admission of incomplete documentation confirm that these systemic constraints translate directly into classroom uncertainty, where students internalize the idea that their academic achievement is assessed by AI, instead of a human.

The mechanism appears circular: opacity limits replication, replication deficits erode trust, and diminished trust discourages data sharing, perpetuating opacity. Thus, reproducibility emerges not only as a methodological concern but as an ethical commitment to learners and instructors who rely on consistent evidence of competence.

Fairness is similarly underdeveloped as only a few studies examined multilingual or multicultural biases, and stakeholders described visible inequities for bilingual students, which challenge the assumption that universal rubrics guarantee validity. In contrast, they support arguments that fairness must be interpreted through cultural and linguistic context (Bond et al., 2020). As a rule of thumb, validation should therefore test performance differentials by language and discourse type, not merely by accuracy thresholds. Transparency completes this triad; both data sources confirm that reporting and interpretability are preconditions for legitimacy. With (over)reliance on AI, educators cannot justify grades, and policymakers cannot defend procurement, when system logic remains inaccessible. Consistent with Burstein and LaFlair (2024), transparency must be treated as evidence, not as publicity.

4.2. Stakeholder perceptions and expectations in the Middle East (RQ2)

Stakeholders' expectations shed light on how validation practices can regain credibility in regions of rapid digital expansion. Participants framed trust as a composite of technical stability, cultural fairness, and interpretive clarity. Their insistence on bilingual data and culturally adapted prompts contests the "one-size-fits-all" orientation prevalent in global EdTech discourse. In this sense, the Middle Eastern perspective refines earlier models that focused mainly on algorithmic optimization (Zawacki-Richter et al., 2019) by grounding validation in linguistic pluralism and curricular alignment. This further supports and illustrates what Bond et al. (2020) describe as *digital equity*: systems gain legitimacy when they accommodate local diversity rather than erase it.

Stakeholders also redefined transparency as pedagogical usability in the sense that teachers want explanations they can communicate with/to students; policymakers seek documentation that satisfies public accountability; developers need standards that balance disclosure with intellectual-property constraints. These positions expand the concept of explainability beyond code visibility to *interpretive translation*, a process through which technical reasoning becomes educationally meaningful. Such translation aligns with a human-centered approach and vision of technology in higher education. Moreover, participants framed collaboration itself as a validation mechanism. Their accounts support Tlili et al. (2023), who argue that co-design transforms AI ethics from compliance into shared stewardship. By engaging educators in pilot cycles and iterative feedback, validation becomes an act of collective inquiry rather than unilateral certification.

More importantly, stakeholders introduced the idea of *validation as governance*. For instance, continuous monitoring, bias auditing, and curriculum alignment were described as institutional obligations. This stance adds a temporal dimension that is absent from most empirical studies and parallels calls for lifelong quality assurance in digital learning ecosystems (Sun & Chen, 2022).

The upshot is that practitioners see validation not as a finish line but as an evolving accountability framework embedded in academic life.

4.3. Toward a culturally responsive validation framework (RQ3)

The FSVF synthesizes these insights into a pragmatic model. Algorithmic Validation institutionalizes reproducible testing, including inter-rater checks and open documentation. Contextual Adaptation embeds cultural and linguistic relevance by aligning data, rubrics, and curricula. Stakeholder Engagement ensures participatory co-design and shared interpretation of evidence. Continuous monitoring and ethical oversight close the cycle through periodic audits and public reporting. Each stage addresses a distinct deficit identified in the literature and enacts the stakeholder priorities identified here. Together, they convert validation from a pre-deployment technical task into an ongoing relational practice.

This framework extends previous work on AI validation by reframing it as human-centered governance, especially since earlier studies treated validation as methodological hygiene; the present synthesis conceptualizes it as institutional learning, where evidence, interpretation, and ethics intersect. The framework also clarifies external validity; although derived from Middle Eastern contexts, its logic applies to other multilingual or data-constrained systems, provided local teams adapt procedures to their languages and norms.

4.4. Practical implications for assessment in education

From a governance perspective, several actionable mechanisms can be proposed for reinforcement:

1. Education institutions can require validation records for AI tools that include cross-sample reliability, subgroup fairness analyses, and plain-language explanations.
2. Quality-assurance units can schedule bias audits each semester and publish results internally to build a culture of transparency.
3. Curriculum committees can integrate bilingual materials into scoring rubrics that consider cultural particularities, ensuring that cultural expression is assessed as knowledge, not deviation.
4. Developer contracts can mandate data-sharing clauses that enable independent replication.
5. Faculty training can emphasize interpretation of AI feedback as a pedagogical resource rather than a grading shortcut.

4.5. Limitations

This qualitative study employs a multi-phase design that provides triangulated insights, but it also has inherent limitations. While the interview sample is diverse in terms of roles and includes participants from multiple Middle Eastern countries, it is still regionally constrained.

As a result, the findings may not fully apply to higher education systems in different linguistic, cultural, or policy contexts. The systematic review consists of twenty-five published studies, reflecting only accessible English and Arabic scholarly work, which means it may exclude potentially relevant proprietary or non-indexed research.

Furthermore, because this study is qualitative and interpretive, it does not measure the causal impact of specific validation practices on learning outcomes. Patterns related to fairness are described rather than quantitatively analyzed, due to inconsistent reporting in the existing literature. These limitations are common in exploratory qualitative research and underscore the need for future longitudinal, cross-regional, and experimental studies to empirically test and refine the proposed validation framework.

4.6 Future directions:

Future studies could implement the Four-Stage Framework in multi-institutional pilots using bilingual corpora and pre-registered protocols. Comparative research across regions would reveal how contextual adaptation varies by language and policy regime. It would also be valuable to examine how explainability features influence teacher trust and student engagement, perhaps through controlled rollouts. Cost and capacity analyses could inform policy on sustainable auditing and governance structures. Expanding the evidence base to include grey literature and practitioner reports would further balance technical and experiential perspectives. Through these extensions, validation research can evolve from identifying gaps to building globally adaptable, culturally responsive systems.

5. Conclusion:

The evidence from the qualitative multi-phase design (systematic review + interviews) demonstrates that validation of AI-enhanced assessment in higher education is as much a social and ethical process as it is a technical one. First off, across published studies and practitioner accounts, reproducibility, fairness, and transparency emerged as recurring vulnerabilities that erode trust in digital assessment. Stakeholders added a vital dimension by framing validation as an ongoing institutional obligation requiring collaboration and continuous monitoring. This synthesis clarifies that credible AI assessment depends on a cycle of technical verification, contextual adaptation, participatory governance, and ethical oversight.

The FSVF proposed here organizes these interdependent dimensions into a practical guide for researchers, developers, and higher-education institutions. Algorithmic validation secures methodological rigor, contextual adaptation ensures cultural and linguistic fit, stakeholder engagement embeds collaboration, and continuous monitoring sustains accountability. Together, these stages translate the idea of trustworthy AI into routines that can be tested, replicated, and improved across

contexts. In this way, the framework extends previous models that treated validation as a pre-deployment quality check, positioning it instead as a cornerstone of human-centered digital transformation.

From a theoretical standpoint, the study advances understanding of validation as governance: a process through which evidence and ethics co-produce legitimacy in technology-enhanced learning. Practically, it offers institutions a structure for linking policy and pedagogy in AI assessment, aligning with the goal of connecting technological innovation to equitable educational practice. Although anchored in Middle Eastern higher education, the framework is adaptable to other multilingual or data-rich environments seeking to balance efficiency with justice in assessment. Future research that applies and tests this model will refine its metrics and evaluate its influence on trust, learning outcomes, and institutional accountability.

6. Recommendations:

Building on the findings and conclusions of this study, the following recommendations are directed for researchers, practitioners, developers, and policymakers working to strengthen the validation of AI-enhanced assessment in higher education, particularly in multilingual and culturally diverse settings.

6.1. For Researchers and Academics:

Future studies should adopt pre-registered protocols and publish validation procedures in sufficient detail to enable cross-institutional replication. Researchers are encouraged to move beyond accuracy-only metrics and incorporate subgroup fairness analyses, inter-rater reliability, and test-retest procedures as standard components of AI assessment validation. Expanding research samples beyond the UAE, Saudi Arabia, and Qatar to include other MENA countries and additional multilingual contexts would strengthen the generalizability of findings.

6.2. For Higher Education Institutions and Policymakers:

Higher education institutions should adopt the Four-Stage Validation Framework (FSVF) as a governance model, embedding algorithmic validation, contextual adaptation, stakeholder engagement, and continuous monitoring into their quality-assurance cycles. Institutions are encouraged to require AI tool vendors to provide transparent documentation of training data, model parameters, and fairness testing prior to procurement. Regulatory bodies and ministries of education should formulate national standards for AI assessment validation that mandate culturally and linguistically adapted rubrics, independent bias audits conducted each academic cycle, and publicly accessible transparency reports.

Funding should be directed toward the development of Arabic-English bilingual assessment corpora and the establishment of regional centers of excellence for AI validation in education.

6.3. For AI Developers:

Developers of AI assessment systems should integrate explainability features as a core design principle rather than an afterthought, enabling educators to communicate scoring rationale to students in plain, pedagogically meaningful language. Systems intended for multilingual or multicultural contexts should be trained on regionally representative data and tested for differential performance across language groups before deployment. Participatory co-design models that involve educators and students in iterative pilot testing should become standard practice, as they demonstrably improve cultural alignment, reduce debugging cycles, and build institutional trust. Data-sharing agreements with educational partners should be structured to permit independent replication and third-party auditing.

6.4. For Educators:

Faculty members and academic practitioners are encouraged to actively engage in AI tool evaluation processes at their institutions rather than treating AI-generated scores as fixed outputs. Professional development programs should be designed to build digital assessment literacy, equipping educators to critically interpret AI feedback, identify scoring anomalies, and advocate for student fairness. Educators should document and report inconsistencies in AI scoring through institutional channels, contributing to the evidence base needed to trigger meaningful audit and review cycles. Ultimately, educators serve as the most proximate safeguard for equitable assessment practice, and their sustained engagement with AI validation processes is essential to the credibility of AI-enhanced education.

7. Acknowledgments:

I would like to thank Dr. Taoufik Boulhrir for his valuable feedback during the initial stage of this work.

8. Funding statement:

The author received no financial support for the research.

9. References:

Aboalela, R. A. (2023). *ChatGPT for generating questions and assessments based on accreditations. arXiv preprint.* https://arxiv.org/abs/2312.00047

Al-Abdullatif, A. M. (2024). *Modeling teachers' acceptance of generative artificial intelligence use in higher education: The role of AI literacy, intelligent TPACK, and perceived trust. Education Sciences, 14*(11), 1209. https://doi.org/10.3390/educsci14111209

Al-Kaabi, S. (2024). *Reliance on AI and its effects on critical thinking and graduate readiness: Evidence from UAE higher education. International Journal of Learning, Teaching and Educational Research.* https://doi.org/10.26803/ijlter.24.8.17

Al-Khalidi, R., & Al-Shehri, T. (2023). *Fairness and interpretability of AI models in Saudi educational assessment. Heliyon, 9*(5), e15490. https://doi.org/10.1016/j.heliyon.2023.e15490

Al-Mutawa, H., & Javid, C. Z. (2023). *Automated essay scoring validity in Gulf higher education institutions. Language Testing in Asia, 13*(1), 16. https://doi.org/10.1186/s40468-023-00203-4

Al-Zahrani, A. M., & Alasmari, T. (2024). *Exploring the impact of artificial intelligence on higher education: The dynamics of ethical, social, and educational implications. Humanities and Social Sciences Communications, 11,* Article 912. https://doi.org/10.1038/s41599-024-03432-4

Al-Zahrani, A. M., & Alasmari, T. M. (2025). *A comprehensive analysis of AI adoption, implementation strategies, and challenges in higher education across the Middle East and North Africa (MENA) region. Education and Information Technologies.* https://doi.org/10.1007/s10639-024-13300-y

Alazemi, T. F. (2024). *Formative assessment in artificial integrated instruction: Delving into the effects on reading comprehension progress, online academic enjoyment, personal best goals, and academic mindfulness. Language Testing in Asia, 14,* 44. https://doi.org/10.1186/s40468-024-00319-8

Alenezi, A., & Alenezi, A. R. (2024). *AI formative assessment in Saudi education: A study across universities. Journal of Teaching and Learning.* https://doi.org/10.22329/jtl.v19i4.10012

Alghamdi, A., & Li, L. (2022). *Artificial intelligence in education: A review of applications and challenges in the Middle East. Education and Information Technologies, 27*(5), 6023–6045. https://doi.org/10.1007/s10639-022-11089-3

Alobed, M., Altrad, A., & Abu Bakar, Z. B. (2021). *Automated Arabic essay scoring based on hybrid machine learning techniques. Malaysian Journal of Computer Science.*

Al-Sharoufi, H. (2022). *Towards a unified English technology-based writing curriculum in the Arabian Gulf countries: The case of Oman. Language Testing in Asia, 12,* 33. https://doi.org/10.1186/s40468-022-00178-1

Alzahrani, A. M. (2022). *A systematic review of artificial intelligence in education in the Arab world. Amazonia Investiga, 11*(54), 293–305. https://doi.org/10.34069/AI/2022.54.06.28

Bashendy, M., Albatarni, S. M., Eltanbouly, S., Zahran, E., Elhuseyin, H., Elsayed, T., Massoud, W., & Bouamor, H. (2024). *QAES: First publicly available trait-specific annotations for Arabic automated essay scoring (AES).* In *Proceedings of The Second Arabic Natural Language Processing Conference* (pp. 337–351).

Braun, V., & Clarke, V. (2006). *Using thematic analysis in psychology. Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., & Wongvorachan, T. (2024). *The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. arXiv.* https://doi.org/10.48550/arXiv.2406.18900

Burstein, J., & LaFlair, G. T. (2024). *Where assessment validation and responsible AI meet. arXiv.* https://doi.org/10.48550/arXiv.2411.0257

Calderwood, S. J. (2024). *Evaluation of higher education students' views of the use of generative AI in a Middle Eastern university* (Doctoral thesis). Zayed University, UAE. https://doi.org/10.21203/rs.3.rs-3869266/v1

Ghazawi, R., & Simpson, E. (2024). *Automated essay scoring in Arabic: A dataset and analysis of a BERT-based system. arXiv preprint.* https://arxiv.org/abs/2407.11212

Hamash, M., Ghreir, H., Tiernan, P., & Boulhrir, T. (2026). From NPCs to AI Assistants: A Scoping Review of AI-Driven Agents in Immersive STEM Learning. *Generators, Bots, and Tutors: Creative Approaches to Human-AI Synergy in Classroom Instruction*, 211-244. https://doi.org/10.4018/979-8-3373-0847-0.ch008

Hobeika, E., et al. (2024). *Multinational validation of the Arabic version of the Artificial Intelligence Literacy Scale (AILS) in university students. Cogent Social Sciences.* https://doi.org/10.1080/23311908.2024.2395637

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning.* Center for Curriculum Redesign.

Kayan Fadlelmula, F., & Qadhi, S. M. (2024). *A systematic review of research on artificial intelligence in higher education: Practice, gaps, and future directions in the GCC. Journal of University Teaching & Learning Practice.* https://doi.org/10.53761/pswgbw82

Khan, M. A., Kurbonova, O., Abdullaev, D., Basim, N., & Others. (2024). *Is AI-assisted assessment liable to evaluate young learners? Parents' support, teacher support, immunity, and resilience in testing vocabulary learning. Language Testing in Asia, 14,* 48. https://doi.org/10.1186/s40468-024-00324-x

Khan, W., Topham, L. K., Al-Shabandar, R., Kolivand, H., Hussain, A., & Khan, I. (2024). *Auto-assessment of assessment: A conceptual framework towards fulfilling the policy gaps in academic assessment practices. arXiv preprint.* https://arxiv.org/abs/2411.08892

Khlaif, Z. N., Ayyoub, A., Hamamra, B., Bensalem, E., Mitwally, M. A. A., Ayyoub, A., Hattab, M. K., & Shadid, F. (2024). *University teachers' views on the adoption and integration of generative AI tools for student assessment in higher education. Education Sciences, 14*(10), 1090. https://doi.org/10.3390/educsci14101090

Lotfy, N., Shehab, A., Elhoseny, M., & Abu-Elfetouh, A. (2023). *An enhanced automatic Arabic essay scoring system based on machine learning algorithms. Computers, Materials & Continua, 77*(1), 1227–1249. https://doi.org/10.32604/cmc.2023.039185

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

Mahmoud, S., & Ameen, K. (2021). *AI and equitable assessment in Egyptian schools. Education Sciences, 11*(7), 325. https://doi.org/10.3390/educsci11070325

Mazawi, A. (2020). *Education and cultural diversity in the Arab States: Towards inclusive and equitable systems. International Review of Education, 66*(5–6), 713–737. https://doi.org/10.1007/s11159-020-09873-1

Traidi, A. (2024). *AI integration in education in the MENA region: Will it be a driver of social inequality? Global Campus of Human Rights Policy Brief.*

Williamson, B., & Piattoeva, N. (2022). *Objectivity as standardization in data-scientific educational governance: Grasping the global through the local. Research in Education, 106*(1), 3–22. https://doi.org/10.1177/00345237221104756

World Economic Forum. (2021). *The Future of Jobs Report 2021*. World Economic Forum. https://www.weforum.org/reports/the-future-of-jobs-report-2021

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., & Gašević, D. (2023). *Practical and ethical challenges of large language models in education: A systematic scoping review. arXiv.* https://doi.org/10.48550/arXiv.2303.13379

Zawacki-Richter, O., Marín, V., Bond, M., & Gouverneur, F. (2019). *Systematic review of research on artificial intelligence applications in higher education. International Journal of Educational Technology in Higher Education, 16,* 39. https://doi.org/10.1186/s41239-019-0171-0

10. List of abbreviations

- AI: Artificial Intelligence
- UAE: United Arab Emirates
- SA: Saudi Arabia
- RQ: Research Questions
- FSVF: Four-Stage Validation Framework

11. Appendix

Interview Questions

1. How would you describe your experience with AI-based assessment systems in your educational or professional setting?
2. In your opinion, how reliable and consistent are the results generated by AI assessment tools?
3. What measures do you believe are essential to ensure fairness and transparency in AI-generated scores?
4. How well do you think current AI assessment tools adapt to the linguistic and cultural diversity of Middle Eastern classrooms?
5. What challenges have you encountered regarding the implementation or validation of AI assessments?
6. How do you think stakeholders (teachers, policymakers, developers) should collaborate to ensure trustworthy AI assessment practices?
7. What policies or frameworks do you think are needed to enhance the validation and ethical use of AI in education?
8. Can you describe a situation where AI assessment outputs did not align with human judgment or expected outcomes?
9. How important is it to include teachers and students in the validation and feedback processes of AI systems?
10. What recommendations would you offer to improve the credibility and acceptance of AI assessments in your context?

Copyright © 2026 by Noura F. Assaf, and AJRSP. This is an Open-Access Article Distributed under the Terms of the Creative Commons Attribution License (CC BY NC)

Doi: https://doi.org/10.52132/Ajrsp.e.2026.84.2