Malaysian Journal of Social Sciences and Humanities (MJSSH) (e-ISSN : 2504-8562)
2026, Volume 11, Issue 1, e003761
DOI: https://doi.org/10.47405/mjssh.v11i1.3761

RESEARCH ARTICLE

Assessment of Arabic Language Skills Using AI Tools Among Gifted Talented Students

Muhammad Hakim Kamal[1*]

[1]Kolej PERMATA Insan, Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia
Email: hakimkamal@usim.edu.my

CORRESPONDING AUTHOR (*):
Muhammad Hakim Kamal (hakimkamal@usim.edu.my)

KEYWORDS:
Arabic skills assessment
Artificial intelligence
Gifted students
Educational technology
Language proficiency

CITATION:
Muhammad Hakim, K. (2026). Assessment of Arabic Language Skills Using AI Tools Among Gifted Talented Students. *Malaysian Journal of Social Sciences and Humanities (MJSSH), 11*(1), e003761. https://doi.org/10.47405/mjssh.v11i1.3761

ABSTRACT

The research question is whether artificial intelligence (AI) instruments can be used to determine the level of skills in Arabic language among 80 gifted and talented students at PERMATA Insan College. The study includes both quantitative data of assessment (as measured by assessment tools with artificial intelligence) and qualitative feedback to assess the effectiveness, precision, and pedagogical usefulness of AI-driven assessment instruments in gauging competence in four primary language competence areas, which include reading comprehension, writing composition, listening comprehension, and speaking proficiency. The data collection was conducted during the 12 weeks of time, with the help of three AI platforms that were specifically created to assess the Arabic language, and other traditional assessment methods that were used to compare them. Findings show that AI tools are highly reliable (Cronbach's alpha = 0.89) and have a high correlation with conventional methods of assessment ($r = 0.82$, $p = 0.001$). Students with higher ability demonstrated high interest in AI-powered tests, with 87 percent of them saying that they felt more motivated and 92 percent of them finding the instant feedback system to be valuable. The research shows that AI applications are very effective in offering customized learning streams, detecting particular linguistic deficiencies, and providing differentiated levels of difficulty that can challenge gifted students at the right level. Nevertheless, weaknesses were found when it comes to evaluation of cultural subtleties and higher rhetorical techniques in classical Arabic literature. The results indicate that the implementation of AI assessment tools into the conventional ones makes it possible to develop a cohesive assessment system that is especially effective with gifted students that appreciate fast and detailed feedback as well as the possibility to learn at their own pace. The study can be important to the body of existing literature on the topic of educational technology in language learning and it has practical implications to the institutions with gifted populations in the multilingual setting.

Contribution/Originality: This research has added value to the existing body of literature since it is among limited studies that have examined the use of AI-based assessment tools in learning the Arabic language in gifted students. This paper will report the efficacy, consistency, and legitimacy of AI-based systems in evaluating Arabic proficiency and uncover the instructional effects of hybrid evaluation methods in the context of gifted education.

1. Introduction

The emergence of artificial intelligence technologies has completely reshaped the educational assessment procedures in international institutions due to their propulsive developmental pace. In the narrow framework of language education, AI-based technologies have become the perspectives of instruments to assess the linguistic abilities with an unprecedented accuracy, speed, and personalization. This change is especially applicable in the context of teaching the Arabic language because conventional assessment systems may not be able to support the morphological complexity of the language, dialect, and the complex connection between Modern Standard Arabic and classical (Al-Sulaihim & Al-Ohali, 2020). The case of PERMATA Insan College, which is one of the top colleges in Malaysia focused on developing gifted and talented students, is one of the best places to study how AI assessment tools can address the specific learning needs of gifted students. Gifted students generally exhibit faster cognitive growth and development, enhanced metacognitive understanding and tendency to handle complicated and demanding tasks that standard pedagogical instruction might fail to cover appropriately (Subotnik et al., 2011). The combination of AI evaluation technology and gifted education offers an interesting research question of whether intelligent systems can offer the complex, adaptive assessment that gifted students demand.

Arabic as one of the six official languages of the United Nations as well as the liturgical language of Islam has enormous cultural, religious and geopolitical value. Arabic education in Malaysia takes the center stage in Islamic religious schools and is now being considered a valuable language acquisition skill in the global market (Embong et al., 2014). Nevertheless, there are unique issues associated with the evaluation of Arabic proficiency that are based on the morphological richness of the language, the sophisticated root-and-pattern framework, and the diglossia of the colloquial dialects and formal language (Ryding, 2013). Use of AI in language evaluation has been of significant interest over the last few years with technologies like natural language processing, machine learning algorithms, and speech recognition systems, promising to evaluate different elements of language proficiency (Burstein et al., 2013). Such technologies have a range of benefits over conventional forms of assessment such as scoring consistency, instant feedback, scalability, and the ability to handle large volumes of language data to detect subtle patterns in student performance (Ranalli, 2021). In the case of gifted students, in particular, AI assessment tools can help with a number of pedagogical issues. It has been shown that gifted learners usually feel frustrated when implementing the one-size-fits-all curriculum and that differentiated instruction, when tailored to their higher abilities, can be of significant benefit (VanTassel-Baska, 2018). With their ability to conduct tests adaptively and provide a customized feedback, AI systems can theoretically be harmonized with the educational requirements of this population. Besides, the instant feedback offered by AI tools can complement the high learning rate of gifted students to be able to move through the material with expediency, without anticipating the evaluation of the instructors.

Although the problem of AI-based language assessment is becoming increasingly popular, the literature has some gaps by not covering the usage of AI-based language assessment in the context of learning Arabic as a second language among gifted students. The majority of the current studies were based on the evaluation of English language or general student groups, without referring to the needs and specificities of gifted students (Wilson & Czik, 2016). Moreover, the linguistic and cultural uniqueness of Arabic needs to be explored using the means of whether AI devices created with the primary focus on Indo-European languages can be successfully applied to understand a Semitic language with its peculiarities. This paper will fill these gaps by answering the following research questions: (1) To what extent can the Arabic language skills of gifted students in the reading, writing, listening, and speaking fields be evaluated with the help of AI tools? (2) How does AI assessment compare to standard assessment measures in this regard regarding its reliability and validity? (3) How gifted students experience and interact with AI-powered assessment devices? (4) What are the particular advantages and weaknesses of AI testing in learning the Arabic language as a gifted learner? This research is important not only to the immediate context of PERMATA Insan College. With most educational institutions around the world aiming to capitalize on technology to better the results of learning processes, the techniques of AI assessment instruments, in relation to exceptional pupils and complicated languages, such as Arabic, have not just educational policy importance, but also curriculum development and instructional technology planning. The results can be used to design culturally responsive and linguistically sensitive AI assessment instruments and provide an idea to educators who have to work with gifted students in different linguistic backgrounds.

2. Literature Review

2.1. Artificial Intelligence in Language Assessment

The use of the artificial intelligence in language assessment has been developing significantly since the days of the first computerized testing systems which appeared in the 1980s. The current AI assessment tools adopt advanced technologies such as natural language processing, machine learning, neural networks, and deep learning algorithms to measure linguistic competencies with greater precision (Chen et al., 2019). Such systems are able to operate in several dimensions of language use at once, and detect patterns that human assessors may fail to spot, and still to use similar scoring criteria across thousands of assessments. The study of AI-self-assessment of writing has shown a high level of advancement. Automated essay evaluation systems e-rater, IntelliMetric and Lightside have demonstrated that they correlate well with human graders, and generally their agreement rates are similar to those of an inter-rater agreement between human graders (Shermis & Burstein, 2013). Such systems assess the different levels of writing quality such as novelty, use of vocabulary, grammatical correctness and development of arguments. Critics, however, observe that there are constraints to evaluating creativity, cultural subtlety, and employing complicated rhetorical devices (Warschauer & Grimes, 2008).

In the case of speech evaluation, AI technologies with speech recognition and pronunciation analysis have gone a long way. It is now possible to test pronunciation accuracy, fluency, prosody, and even pragmatic appropriateness of spoken language by systems (Liakin et al., 2015). This study has shown that automated speaking evaluation systems are associated to high levels of human experts ratings of specific characteristics especially pronunciation and fluency measures (Bernstein et al., 2010). However, there

are still difficulties in testing spontaneous speech especially in determining communicative effectiveness and sociolinguistic appropriateness. The AI-based listening comprehension assessment model is generally characterized by the use of automated speech recognition and question-answering systems. These tools may deliver audio content and assess responses at once, and this feature has scalability and consistency benefits (Suvorov, 2015). The assessment of reading comprehension has also been improved with AI technologies that have the ability to create questions, analyze answers, and adjust the levels of difficulty in accordance with the performance of learners (Flor & Riordan, 2018). Identification strategies are still a primary concern in gifted education research, which is why it is reasonable to consider the AI-based assessment tools as a novel identification and assessment strategy (Kamal, 2026a).

2.2. Arabic Language Assessment Challenges

The assessment of Arabic is a special topic due to the linguistic nature of this language and the complexity of sociolinguistic issues. Being a Semitic language, Arabic has a root-and-pattern morphological structure, which is radically different to the mainly concatenative morphology of the Indo-European languages (Holes, 2004). This richness of morphology implies that one root may produce dozens of his/her derivatives in different patterns and this demands a complex knowledge of morphological relations. Moreover, using various techniques and no set way to assess makes it more difficult to make comparisons among studies and fields (Kamal, 2026b). The diglossia phenomenon of Arabic in which the Modern Standard Arabic exists alongside a plethora of colloquial dialects makes the process of proficiency measurement difficult (Ferguson, 1959). The learners are required to move between the formal registers in both academic and media settings and the informal ones in the daily communication. Assessment tools should thus bear in mind, which type of variety they focus on and whether they are in a position to distinguish between appropriate register choice and linguistic mistakes. Arabic orthography also poses some other problems especially the optional marking of the short vowels using diacritical marks. Although full vowelized writing is easier to read among the learners, most of the authentic Arabic writing is full of consonantal skeletons and very little vowelization, and the readers have to learn the pronunciation and meaning through the context (Abu-Rabia, 2001). Assessment tools have to determine whether the texts should be vowelized or non-vowelized texts which have varying cognitive requirements. The study on assessing Arabic language has found that there are various areas that need to be addressed. Research has shown that in many cases, the traditional tests are focused on grammar and vocabulary recognition, at the cost of communicative competence and practical language application (Hammad, 2016). Moreover, the cultural learning contents and the items used to evaluate learners in the Arabic language often include views of the Arab world, which might not be relevant to the cultural contexts of the learners, which can lead to a lack of engagement and performance (Wahba et al., 2014).

2.3. AI Assessment for Arabic Language

The use of AI specifically in evaluation of the Arabic language is relatively immature as opposed to the content in the English language and other major languages. Nevertheless, in recent years, there has been an increase in attention and development of this field. The Arabic natural language processing has improved and there are better mechanisms of morphological analysis, part-of-speech tagging, named entity recognition and sentiment analysis (Habash, 2010). A number of AI-based applications have also come up that focus on the Arabic language and its evaluation. These are automated essay grading systems in

Arabic that should take into consideration the morphological difficulty of the language and comparatively loose word arrangement (Azmi et al., 2019). An Arabic automated essay grading research demonstrates encouraging results but also indicates difficulties in reaching human level accuracy especially with more advanced writing that can implement more sophisticated rhetorical means that are based on classical Arabic tradition (Alrabiah et al., 2014). The Arabic pronunciation has been given serious consideration owing to the phonological aspect of the Arabic language which poses great challenge to non-native speakers, in terms of emphatic consonants, pharyngeal sounds, and vowel differences that are rare in other languages (Saadah, 2011). AI systems based on speech recognition technology have shown the ability to detect pronunciation errors and give corrective feedback, but with varying accuracy levels depending on the phonological feature (Elmahdy et al., 2014). Adaptive testing platforms that use AI algorithms promise specific prospects in the Arabic language testing. These systems dynamically modify the difficulty of questions in accordance with the responses given by the learners, effectively identifying the levels of proficiency with a minimum amount of time using tests (Fathima & Vadivu, 2018). In such complex systems of morphology as Arabic, adaptive testing has the capacity to systematically test learner knowledge in different grammatical structures and vocabulary areas.

2.4. Gifted Education and Technology

Gifted students have unique traits that affect their learning preferences and learning needs. It has always been found that a number of shared characteristics are usually observed such as high rate of learning, liking complexity and depth, high level of reasoning, high metacognitive awareness and intrinsic motivation (Renzulli, 2012). These features indicate that gifted students might be especially interested in educational technologies that are personalized, have instant feedback, and provide self-directed learning. As seen in the literature regarding the application of technology in gifted education, such students tend to be very accommodating to technological tools and they are able to use them to their advantage in learning at a higher level (Siegle, 2005). Research demonstrates that gifted learners like being able to choose their own tools, have flexibility and the enrichment possibilities of technology based learning environments (Mann, 2006). Nevertheless, there is also some research warning that technology should be introduced carefully, because it takes a short amount of time to realize that the gifted students can easily be able to see that the digital tools are not giving them any deep involvement and that it is not challenging their intellect. Certain studies on the language acquisition of gifted students indicate that they might be able to go through linguistic content at an even faster pace than normal students and take advantage of accessing authentic and complex texts earlier in their learning process (Vural, 2013). Metalinguistic awareness among gifted students can be good in that they can analyze the structures of the language explicitly and in a systematic manner of applying the linguistic rules (Kuo & Anderson, 2010). These features indicate a possibility of correlation with AI evaluation tools that offer a more in-depth linguistic breakdown and clear feedback regarding language usage. The idea of differentiation, which is the core of the gifted education pedagogy, involves modifying the content, process, product, and the learning environment to suit the needs of particular learners (Tomlinson, 2014). The ability to differentiate instruction is theoretically supported by AI assessment tools due to the ability of creating a personalized and adaptive mode of assessment, which enables pinpointing the areas in which the individual students need to be challenged or assisted. Nonetheless, there is limited empirical studies that have investigated the application of AI tools as a selection method to gifted language learners.

2.5. Gaps in Current Research

Nevertheless, there is a gap between the research on AI in language assessment and independent sets of literature on Arabic language learning and gifted education, so only limited studies on the overlapping of these two fields exist. The majority of studies on the assessment of AI languages have been on English or done in general groups of students. The peculiarities and possibilities of applying AI to measure the skills of the Arabic language in gifted students have been given little scholarly coverage. Also, unlike many studies that assess the technical quality of AI assessment tools, there are less studies that assess the pedagogical consequences and the experiences of learners, especially with specific groups such as gifted learners. Issues associated with the perception of these students on AI assessment, its impact on their motivation and engagement, and its potential optimal implementation in instruction are under-researched. This research paper fills in these gaps by showing an empirical evidence of AI assessment effectiveness in the context of the evaluation of Arabic language among gifted students, which would involve the quantitative analysis of performance data and the qualitative analysis of the experience and perception of the students.

3. Methodology

3.1. Research Design

The study followed a convergent parallel mixed-methods design, which presupposed gathering and analyzing both quantitative and qualitative data to obtain an in-depth insight into the effectiveness of AI tools in measuring the Arabic language skills among gifted students. The quantitative element studied the measurement reliability, validity, and comparing performance indicators, whereas the qualitative aspect studied the student perceptions, experiences, and pedagogies using surveys, interviews, and reflective journals.

3.2. Research Setting and Participants

The study was done in PERMATA Insan College, the specific college in Malaysia, which is devoted to the education of gifted and talented students who show high academic skills as well as adherence to Islamic principles. The college offers an integrated curriculum that incorporates national academic standards and Islamic studies and instructions in the Arabic language. The participants included 80 gifted students (n=80) enrolled to take courses in Arabic languages in the 2024 academic year. The sample consisted of students of different grades (13-17) who have different levels of Arabic proficiency (elementary to advanced). The selection criteria meant that students had to be formally named gifted under the national PERMATA program that uses a variety of measures such as standardized intelligent testing, academic performance history, and checklists of behavior. Informed consent of all participants or their guardians was done after the institutional review board approved the same. The demographics of the participants were as follows: 42 females (52.5%) and 38 males (47.5%); the age of the participants was distributed as follows: 18 students aged 13-14 (22.5%), 35 students aged 15-16 (43.75%), 27 students aged 17 (33.75%); the level of Arabic proficiency was based on the pre-assessment assessing the level of proficiency: 22 elementary learners (27)

3.3. AI Assessment Tools

Three AI-based assessment platforms were chosen due to their functionality in assessing language skills in Arabic, technological advanced, and addressing research questions of the study:

3.3.1. Platform 1: ArabicTutor AI

It is an extensive language assessment system designed with the focus entirely on Arabic and includes reading comprehension, writing assessment, listening assessment, and speaking analysis modules. The site uses the neural machine translator technology and the natural language processing algorithm which has been trained with a large amount of Arabic text corpora.

3.3.2. Platform 2: Kalima Assessment Suite

This platform is an adaptive testing platform, which uses item response theory and machine learning algorithms to modify the difficulty of questions dynamically. Kalima is concerned with grammatical competence, vocabulary knowledge and reading comprehension.

3.3.3. Platform 3: Nahwa Speaking Evaluator

It is a specialized app to determine the proficiency of Arabic speech with the help of speech recognition technology, prosody, and auto pronunciation assessment. The system gives elaborate feedback in terms of phonological accuracy, fluency and communicative effectiveness.

3.4. Traditional Assessment Methods

The traditional assessment methods were provided as parallel to AI tools to achieve the validity. These included:

i. Reading Comprehension Tests: multiple-choice and constructed-response items based on literal understanding, inferential awareness and critical evaluation of Arabic texts in the adequate level of difficulty.
ii. Writing Assessment: Timed essay compositions to be completed by the students in relation to the established rubrics to evaluate the work on the basis of the content, structure, use of languages, and mechanics through the work of the experienced Arabic instructors.
iii. Listening Comprehension Tests: Audios recorded and then they are given comprehension questions that they are scored by an instructor.
iv. Speaking Tests: Oral interviews and presentation tasks checked by trained raters with the help of the standardized rubrics based on parameters of pronunciation, fluency, grammatical accuracy, vocabulary range and communicative effectiveness.

3.5. Data Collection Procedures

The process of data collection took place during 12 weeks in form of three stages:

3.5.1. Phase 1 (Weeks 1-2): Baseline Assessment Study

The students were issued with traditional tests in all the four language areas, a practice that determined the level of proficiency. The pre-study surveys were aimed at collecting data about the previous experience with technology, preferences, and attitude towards Arabic learning.

3.5.2. Phase 2 (Weeks 3-10): AI Assessment Implementation

Students used AI platforms weekly and completed different assessment activities in the fields of language. All the platforms were systematically utilized in order to have sufficient data collected. Consequently, the traditional assessment methods were still done to compare them. The AI platforms automatically registered student engagement metrics, completion rates and time-on-task.

3.5.3. Phase 3 (Weeks 11-12): Post-Assessment and Data Collection

Final traditional assessment was done, as well as final AI assessment. The qualitative data on the student experiences, perceptions and suggestions were gathered through post-study surveys, focus group discussions and individual interviews.

3.6. Data Analysis

SPSS Version 28 was used to carry out the statistical analysis. The assessment of AI reliability was assessed with the help of Cronbach alpha coefficients. The application of validity was by correlation analysis of AI assessment scores with the traditional assessment scores in language domains. The comparisons of the performance across the assessment methods were made using paired-samples t -tests. Repeat measures ANOVA was used to analyze the variation in performance as time went on. The calculation of the effect sizes was done using the Cohen d to obtain the practical significance of the results.

Thematic analysis was done on qualitative data gathered through surveys, interviews and reflector journals using the six-phase method of Braun and Clarke (2006). Preliminary coding was the one that revealed repeated ideas and patterns. The codes were grouped into initial themes that were reviewed and evaluated by means of repetitive analysis. Final themes were formulated and identified, and representative quotes were chosen to depict the major findings. The NVivo software was used to help in the organization and coding of data.

The integration of quantitative and qualitative findings was done during interpretation and convergence and divergence of data sources were identified and discussed. Integration allowed gaining a full insight into not only the quantifiable quality of AI tools but also the anthropological aspects of their application.

3.7. Ethical Considerations

The institutional review board of PERMATA Insan College gave their approval to the study. Informed consent was given by all the participants and guardians who had been thoroughly informed about the purpose of the research, the procedures, risks and benefits of the research. The protection of confidentiality was observed by keeping a confidential coded participant data. Students were given the agreement that this was voluntary and it

would not compromise their academic performance. The study followed the ethical principles of research involving minors.

4. Results

4.1. Reliability of AI Assessment Tools

The internal consistency analysis indicated that AI assessment tools have high reliability when analyzed in different language domains. ArabicTutor AI was very reliable and the Cronbach alpha coefficients of reading comprehension, writing assessment, listening comprehension and speaking evaluation were 0.91, 0.87, 0.89 and 0.85, respectively. The adaptive grammar and vocabulary test of kalima assessments suite ranged at 0.93 as alpha. Nahwa Speaking Evaluator had 0.88 0.86 as its alpha of pronunciation and fluency testing respectively. Test-retest reliability was measured through the use of a subsample of 20 students that were given the same tests after two weeks. Pearson correlation coefficients were very stable: ArabicTutor AI ($r = 0.84$, $p < 0.001$), Kalima Assessment Suite ($r = 0.87$, $p < 0.001$) and Nahwa Speaking Evaluator ($r = 0.81$, $p < 0.001$).

4.2. Validity: Correlation with Traditional Assessments

The analysis of concurrent validity was conducted by correlation between the scores of assessment of AI and traditional assessment scores which showed strong positive correlations in all the language domains. There was a correlation coefficient of $r = 0.82$ ($p < 0.001$) between writing quality and AI as compared to reading comprehension and a correlation coefficient of $r = 0.79$ ($p < 0.001$) between AI and traditional assessment. Listening comprehension showed the highest level of correlation at $r = 0.85$ ($p < 0.001$) and speaking proficiency, although there was the least correlation among the four areas, still showed a strong positive correlation at the same level of $r = 0.76$ ($p < 0.001$). The high levels of correlations indicate that AI tools were evaluating the same constructs as traditional assessments albeit with some differences, hence demonstrating convergent validity and, consequently, the possibility of the two forms of assessment identifying additional information that could complement each other and positively influence evaluation practices.

4.3. Comparative Performance Analysis

Paired-samples t-tests were used to compare average scores among assessment procedures. In terms of reading comprehension, it was found that the AI assessment ($M = 78.3$, $SD = 11.2$) and the traditional assessment ($M = 77.8$, $SD = 10.9$) did not differ significantly, $t(79) = 0.63$, $p = 0.532$. Likewise, the results of listening comprehension were also similar: AI ($M = 75.6$, $SD = 12.4$) vs. traditional ($M = 76.1$, $SD = 11.8$), $t(79) = -0.58$, $p = 0.564$. Nevertheless, there were also important differences in the area of writing assessment, where the traditional evaluation scores ($M = 73.2$, $SD = 13.6$) were slightly above the AI scores ($M = 69.8$, $SD = 14.1$), $t(79) = 3.41$, $p = 0.001$, $d = 0.38$. This medium effect size implied that human raters rated some qualities of writing in a more positive way compared to AI algorithms. On the other hand, the speaking assessment revealed that AI scores ($M = 72.4$, $SD = 13.9$) were slightly higher than traditional ratings ($M = 70.1$, $SD = 14.5$), $t(79) = -2.24$, $p = 0.028$, $d = 0.25$, but the small value of the effect showed that there was no significant difference in practice.

4.4. Performance Progression Over Time

Repeated measures ANOVA was used to determine whether there was an improvement in students over the 12 weeks of the study. Findings showed that there were significant improvement in all language areas: reading comprehension $F(2, 158) = 18.45$, $p < 0.001$, $\eta p\ 2 = 0.19$, writing quality $F(2, 158) = 15.67$, $p < 0.001$, $\eta p\ 2 = 0.17$, listening comprehension $F(2, 158) = 21.32$, $p < 0.001$, eta p Post-hoc pairwise matched comparisons that were carried out after integrating Bonferonni adjustment indicated significant progress between the baseline and mid-point and between mid-point and end assessment in most areas, thus indicating continuous learning that was enabled by frequent assessment and feedback.

4.5. Engagement and Completion Metrics

The AI platforms recorded interaction information. The overall AI assessment completion rate was 94.3, which was much greater compared to the 87.5% completion rate of the traditional homework assignments in the same period. The mean time-on-task of AI tests was 38.7 minutes per session, and gifted students showed their proactive interest. Platform analytics showed that 76 percent of students used optional practice resources outside of obligatory tests, meaning that they were willing to use the AI tools. Remarkably, the patterns of engagement were dependent on the level of proficiency. Higher-level learners also dedicated more time to AI platforms (M = 42.3 minutes) than intermediate (M = 37.8 minutes) and elementary learners (M = 36.1 minutes), indicating that more proficient students were ready to engage with AI platforms due to the associated challenge and depth.

4.6. Qualitative Findings: Student Perceptions

The overlap of quantitative and qualitative data indicated unique strengths and weaknesses of AI evaluation to that of Thematic analysis of qualitative data which found five significant themes regarding student experiences with AI assessment instruments:

4.6.1. Theme 1: Immediate Feedback as Motivational Catalyst

Students highly valued being given instant feedback of their performance. The quotes could be summarized as follows: When I provide my writing and instantly see what I did wrong, I could correct it instantly, and not wait days to have the teacher back my work and inform me about the sounds that I was pronouncing incorrectly. The speaking app tells me which sounds I am pronouncing incorrectly, so I can also correct them and not wait days to have the teacher back my work and tell me what I have done wrong. This type of immediate feedback loop also matched the gifted students desire to work through course material at a faster pace and reduced frustration over waiting periods inherent in traditional assessment loops.

4.6.2. Theme 2: Personalization and Adaptive Challenge

Students acknowledged and appreciated a flexible character of AI assessments. They liked the fact that the tools were able to change difficulty with their performance, and they gave them the right level of difficulty. One of the students commented: It has a way of knowing when I am on the easy side of the program and making it more challenging. This was a

highly adaptive feature, especially among gifted learners who tend to get bored with predetermined-level curricula.

4.6.3. Theme 3: Detailed Linguistic Analysis

A large number of students said that they were fascinated by detailed linguistic feedback that AI tools offered. They also valued getting certain information regarding grammar mistakes, vocabulary use, and the pronunciation patterns. Some of the comments stated: I did not realize that I always mixed up these two forms of the verb until the AI revealed how I made mistakes on particular words in my speech.

4.6.4. Theme 4: Limitations in Cultural Context

Students pointed out limitations in the dealing of AI tools with cultural specifics and classical Arabic allusions. Some of them mentioned that the platforms occasionally labelled culturally appropriate expressions as mistakes or did not appreciate the use of the advanced rhetorical devices of classical Arabic tradition. One of the high-end students remarked: When I typed in a classical Arabic phrase that would be completely right in a formal writing, the AI marked it as a mistake since it could not identify the phrase.

4.6.5. Theme 5: Favor Hybrid Approach

Although students found AI tools useful, they did not desire to overtake traditional assessment and teacher feedback. The general feeling was to have a combination of both styles. One of the students explained it in the following way: The AI is wonderful at practicing on a short basis and identifying simple errors, but I still wish to have my teacher read my essays and discuss my ideas and the ways to enhance my writing style.

4.7. Comparative Strengths and Limitations

The combination of quantitative and qualitative analysis showed that AI measurement had some unique strengths and weaknesses in comparison with the conventional evaluation tools. AI testing proved to have a number of interesting benefits, such as the use of scoring criteria in all assessments and the possibility to provide learners with the result and feedback immediately. The technology was especially useful in offering a detailed analysis of errors and identification of patterns, as well as being able to scale massively and be available at all times, never before seen. Furthermore, AI assessment resulted in stress reduction among some students, the effective identification of specific skills gaps and objective assessment results that were not affected by the presence of rater bias or fatigue. Nevertheless, the study also revealed that there are also major weaknesses that limited the power of AI evaluation in some settings. The technology was less sensitive to cultural and contextual fit in using language and had a problem in assessing creativity and originality in student writing. AI evaluation was also found to be limited regarding the ability to calculate pragmatic competence in speaking skills and especially in addressing classical Arabic expressions and high-order rhetorical devices. Moreover, the system failed to measure non-linguistic communication aspects which are crucial in the overall assessment of language. Assessments of contents done during the writing process proved to be less effective and efficient compared to the traditional approaches and there were instances where the students experienced technical problems that adversely impacted on their user experience.

4.8. Differentiation by Proficiency Level

It was found that AI assessment efficiency was slightly dependent on the level of student proficiency. Correlations between AI and the traditional assessment were found to be stronger in elementary and intermediate learners ($r = 0.86$ and $r = 0.84$ respectively) than among advanced learners ($r = 0.74$). This implies that AI tools worked best with learners at lower levels of proficiency where the objective of assessment is to identify basic skills, and as the competency progresses to higher levels, the tools become more difficult to use to measure the advanced (higher and higher) skills of students. Nevertheless, the qualitative data showed that more advanced students could find some value in AI tools in another purpose. Instead of being used to provide a full-fledged evaluation, advanced learners had been applying AI platforms strategically to do targeted practice in certain areas such as pronunciation refinement or grammar review.

4.9. Discussion

The results of the current research can provide valuable information about the effectiveness, usefulness, and weaknesses of the AI-based assessment applications in assessing the Arabic language proficiency of gifted students. The high reliability coefficients and large correlations with conventional assessment tools indicate that AI instruments can be used to give reliable measures of Arabic language proficiency in various areas. These findings are consistent with other studies on AI evaluation in other languages and generalize the findings to Arabic, which is a Semitic language morphologically rich and a challenging language computationally.

4.10. Effectiveness for Gifted Learners

The rates of engagement and the positive student perceptions in particular are quite high, which implies that AI assessment tools may fit many features of gifted students. The instant feedback system deals with the usual disillusionment among gifted learners to wait until the results of evaluation are known to help them to learn faster. The adaptive difficulty characteristics are based on the fact that they need an adequate level of difficulty, neither too easy nor too challenging. The extensive linguistic analysis that AI platforms offer attracted the metacognitive awareness and the analytical skills of the gifted students. Instead of receiving the general scores, the students would be able to analyze trends in their mistakes, learn which areas they are weaker at linguistically, and work on the specific improvement plans. This feed-back is granular and facilitates the self directed learning inclination that prevails in gifted students. Nevertheless, the fact that there was a slight decreasing correlation between AI and conventional assessments among advanced students points out a very significant point. Higher the level of students proficiency, the more linguistic competence becomes a matter of fine cultural knowledge, elaborate rhetorical tactics and innovative language use- areas that now AI services are weak. This implies that AI evaluation can prove to be the most beneficial to gifted students during the base up to intermediate stages, but the comparative advantage will decrease in higher levels where human knowledge will take the center stage.

4.11. Pedagogical Implications

The high interest of the students in a hybrid model of using AI and conventional assessment methods has significant pedagogical consequences. Instead of seeing AI tools as a substitute to teacher evaluation, complementary utilization of the two approaches

seems to be the most successful way of applying them, as each of the methods has distinct strengths. AI applications can be useful to quick and repetitive assessment of discrete linguistic characteristics, allowing them to practice repeatedly with instant correction. The use of experienced tutors to analyze the holistic communication skills, cultural appropriateness, and creativity as well as sophisticated reason through language is still necessary. The model is complementary and matches the modern knowledge about good differentiation in gifted education. The advantage of gifted students is that they have diverse methods of assessment, which cover various aspects of their skills. Analytical tools of AI can cope with automatic assessment of basic proficiencies, and instructors will have free time to concentrate on advanced competencies that cannot be substituted with AI. The results of the study have a number of possible pedagogical implications. Supporting formative assessment can be effectively done by AI platforms, which give continuous feedback that informs learning with no additional pressure on high-stakes assessment. The immediate feedback component allows students to practice on purpose, and remedial weaknesses are tackled in a systematic manner. In students with exceptional ability who tend to make quicker progress in material, AI systems can provide a chance to learn at their own pace and enhance their studies beyond what the standard curriculum dictates.

4.12. Technical and Linguistic Considerations

The limitations detected in the management of cultural context and expressions of classic Arabic by AI tools are indicative of the natural limitation in natural language processing of morphologically rich and complex sociolinguistically varied languages. The diglossia of the Arabic language, in which the formal and colloquial languages have different purposes, demands the AI systems to discern proper code-switching and register-use which current technology does not do perfectly. The success of AI tools in measuring reading and listening comprehension as opposed to slightly lower success in measuring writing evaluation is due to the difference in complexities of these tasks to AI systems. The main features of comprehension assessment are first grasping and deriving meaning of the text or audio, and this is a strength of the existing natural language processing technology. Writing assessment requires the measurement of creativity, organization, argumentation and style, areas where the human judgment is still the best. It is interesting because the marginally better scoresof the AI in speaking assessment than human raters may represent different emphasis in evaluations. AI systems can attach high importance to phonological accuracy and fluency measures and be less sensitive to pragmatic appropriateness and communicative effectiveness which the human raters attach priority. This points out the fact that it is crucial to realize what AI tools are precisely quantifying instead of thinking that they are assessing all the relevant dimensions.

4.13. Limitations and Future Directions

There are a number of shortcomings that this study deserves. The study was done in one institution where the population of gifted students was relatively homogeneous, and where the cultural background was also a strict condition. Care should be taken in generalizing to other contexts, student populations or cultural conditions. The 12 weeks period, though adequate in preliminary evaluation of tool efficacy, may not provide long term effects of tool on learning outcomes or problems that can develop with prolonged use. The paper has discussed three particular AI platforms that, although they are a reflection of modern technology, will eventually be replaced by more technologically advanced platforms. The results about the particular tool capabilities must be interpreted

as an indication of the existing technology level and not necessarily a problem with AI assessment. The identified limitations can be mitigated through the use of rapid risk advances in natural language processing, especially transformer-based models and large language models. Future studies ought to focus on the prolonged effects of AI assessment integration in learning the Arabic language to determine whether there is a positive correlation between regular AI-aided assessment and high-quality proficiency growth than traditional assessment. The comparative research focused on the various groups of students would help to understand whether the results were unique to gifted students or they apply to other students. The views of teachers regarding the integration of AI assessment should be investigated to support the findings regarding students, which are presented here. Learning the ways in which teachers can best integrate AI tools into their teaching practice, how they can interpret AI-generated data to guide their teaching and how they can strike a balance between AI and traditional assessment constitute valuable research lines. Such technical issue is the creation of Arabic-specific AI assessment tools that can be more reflective of the morphological complexity of the language, the diversity of its dialects, and its cultural diversity. Granting Arabic language teachers and computational linguists to collaborate in research projects would produce culturally responsive and linguistically advanced assessment technologies.

5. Conclusion

This paper will present empirical data to prove that AI-assessment applications are efficient to examine Arabic language proficiency among gifted students in various domains of proficiency. The high reliability and validity exhibited by the AI platforms coupled with high involvement of the students and positive perceptions about them justify their application in whole assessment systems with regard to learning of the Arabic language in gifted education environments. The study shows that AI evaluation presents specific benefits to gifted students such as instant feedback that helps gifted students to continue learning faster, adaptive level of challenge that helps to avoid boredom, error analysis that helps students engage in metacognition, and the 24/7 availability that allows learners to practice on their own. These characteristics go in line with the learning needs and preferences of gifted student populations. Nevertheless, the paper also indicates very clearly that the present AI evaluation technology has its limitations, insofar as consideration of cultural specifics, sophisticated rhetorical patterns, and creative use of language are concerned. The limitations indicate that AI tools are not yet ready to replace or substitute traditional assessment and expert teacher evaluation and instead serve as complements. The student tendency to use hybrid systems with AI and human evaluation is a kind of pedagogical wisdom that can be applied to the decision to implement it.

In the case of institutions with gifted students with a multilingual environment, such as PERMATA Insan College, the results indicate that careful use of AI assessment tools can promote teaching the Arabic language with effective and personalized evaluation facilitating the process of differentiated learning. The best option is to use AI resources to conduct regular formative evaluation and drill and retain teacher skills to conduct comprehensive evaluation of more complex linguistic skills. With the further development of AI technology, it will most likely be used in the evaluation of language. But the fundamentally human aspects of language, cultural consciousness, linguistic creativity, and expressive communication will still demand human judgement. The future of language testing does not seem to be on the issue of either AI or conventional but instead on the possibilities of designing advanced models of integrations that would ensure that the benefits that each system offers are exploited. This study builds upon the

existing body of educational technology applied specifically research on the potential and existing constraints of AI evaluation of complex linguistic abilities in exceptional learners. The results can provide valuable insights to the educational community, technology creators, and policy makers that are trying to harness AI in the correct way towards the goal of improved language learning outcomes.

Ethics Approval and Consent to Participate

The Institutional Review Board of PERMATA Insan College and the Research and Innovation Committee of Universiti Sains Islam Malaysia gave ethical approval to this study. The participants received a detailed information regarding the aim of the research, processes, and time spent on the research and possible risks and benefits, as well as the right to leave without a penalty at any time. The research was conducted with consideration of all the applicable Malaysian national regulations in conducting research involving minors and in accordance with the ethical guidelines in research concerning gifted and talented students.

Acknowledgement

On behalf of this research, the author would like to wish to extend his heartfelt gratitude to PERMATA Insan College and Universiti Sains Islam Malaysia as they will play an indispensable role and contribution to this research. Their resource and facility support as well as institutional support played a major role in ensuring the successful completion of this study.

Funding

There was no funding on this research.

Conflict of Interest

The authors reported no conflicts of interest for this work and declare that there is no potential conflict of interest with respect to the research, authorship, or publication of this article.

References

Abu-Rabia, S. (2001). The role of vowels in reading Semitic scripts: Data from Arabic and Hebrew. *Reading and Writing: An Interdisciplinary Journal, 14*(1-2), 39-59. https://doi.org/10.1023/A:1008147606320

Alrabiah, M., Al-Salman, A., & Atwell, E. (2014). The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic. *Proceedings of the Second Workshop on Arabic Corpus Linguistics*, 5-8.

Al-Sulaihim, S. H., & Al-Ohali, Y. A. (2020). Arabic language learning applications: A systematic review. *Computer Assisted Language Learning, 35*(3), 443-466. https://doi.org/10.1080/09588221.2020.1744666

Azmi, A. M., Al-Jouie, M. F., & Hussain, M. (2019). ESSENCE: An automated essay scoring system for Arabic. *Computers in Human Behavior, 94*, 163-172. https://doi.org/10.1016/j.chb.2019.01.016

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355-377. https://doi.org/10.1177/0265532210364404

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

Burstein, J., Tetreault, J., & Madnani, N. (2013). The E-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55-67). Routledge.

Chen, C., Sonnert, G., Sadler, P. M., Sasselov, D. D., & Fredericks, C. (2019). The impact of student misconceptions on student persistence in a MOOC. *Journal of Research in Science Teaching, 57*(6), 879-910. https://doi.org/10.1002/tea.21616

Elmahdy, M., Gruhn, R., & Abdennadher, S. (2014). Cross-lingual acoustic modeling for dialectal Arabic speech recognition. *Proceedings of the Eleventh Conference on International Language Resources and Evaluation*, 2189-2194.

Embong, R., Nik Mohd Rahimi, N. M., Abu Bakar, N. A., & Heng, C. S. (2014). Arabic language literacy among Malaysian secondary school students: A study at selected government secondary schools. *Mediterranean Journal of Social Sciences, 5*(20), 2088-2094. https://doi.org/10.5901/mjss.2014.v5n20p2088

Fathima, G., & Vadivu, G. (2018). Adaptive learning system based on learning analytics to improve students performance in e-learning environment. *International Journal of Computer Applications, 179*(34), 11-16. https://doi.org/10.5120/ijca2018916756

Ferguson, C. A. (1959). Diglossia. *Word, 15*(2), 325-340. https://doi.org/10.1080/00437956.1959.11659702

Flor, M., & Riordan, B. (2018). A semantic role-based approach to open-domain automatic question generation. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 254-263. https://doi.org/10.18653/v1/W18-0528

Habash, N. Y. (2010). *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00277ED1V01Y201008HLT010

Hammad, E. A. (2016). Palestinian EFL teachers' attitudes towards communicative language teaching and their classroom practices. *TESOL International Journal, 11*(1), 42-56.

Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties* (Rev. ed.). Georgetown University Press.

Kamal, M. H. (2026a). Bibliometric analysis of research on gifted education. *Pertanika Journal of Social Sciences & Humanities*. https://doi.org/10.47836/pjssh.34.1.06

Kamal, M. H. (2026b). A scoping review of flipped classroom approaches in arabic teaching. *Ijaz Arabi Journal of Arabic Learning, 9*(1). https://doi.org/10.18860/ijazarabi.v9i1.36740

Kuo, L., & Anderson, R. C. (2010). Beyond cross-language transfer: Reconceptualizing the impact of early bilingualism on phonological awareness. *Scientific Studies of Reading, 14*(4), 365-385. https://doi.org/10.1080/10888431003623470

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal, 32*(1), 1-25. https://doi.org/10.1558/cj.v32i1.25962

Mann, R. L. (2006). Effective teaching strategies for gifted/learning-disabled students with spatial strengths. *The Journal of Secondary Gifted Education, 17*(2), 112-121. https://doi.org/10.4219/jsge-2006-686

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52*, Article 100816. https://doi.org/10.1016/j.jslw.2021.100816

Renzulli, J. S. (2012). Reexamining the role of gifted education and talent development for the 21st century: A four-part theoretical approach. *Gifted Child Quarterly, 56*(3), 150-159. https://doi.org/10.1177/0016986212444901

Ryding, K. C. (2013). *Teaching and learning Arabic as a foreign language: A guide for teachers*. Georgetown University Press.

Saadah, E. (2011). *The production of Arabic vowels by English L2 learners and heritage speakers of Arabic*. Doctoral dissertation, University of Illinois at Urbana-Champaign. http://hdl.handle.net/2142/26371

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge. https://doi.org/10.4324/9780203122761

Siegle, D. (2005). Six uses of the Internet to develop students' gifts and talents. *Gifted Child Today, 28*(2), 30-36. https://doi.org/10.4219/gct-2005-170

Subotnik, R. F., Olszewski-Kubilius, P., & Worrell, F. C. (2011). Rethinking giftedness and gifted education: A proposed direction forward based on psychological science. *Psychological Science in the Public Interest, 12*(1), 3-54. https://doi.org/10.1177/1529100611418056

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing, 32*(4), 463-483. https://doi.org/10.1177/0265532214562099

Tomlinson, C. A. (2014). *The differentiated classroom: Responding to the needs of all learners* (2nd ed.). Association for Supervision and Curriculum Development.

VanTassel-Baska, J. (2018). *Curriculum planning and instructional design for gifted learners* (3rd ed.). Prufrock Press.

Vural, Ö. F. (2013). The impact of a question-embedded video-based learning tool on e-learning. *Educational Sciences: Theory and Practice, 13*(2), 1315-1323.

Wahba, K. M., Taha, Z. A., & England, L. (Eds.). (2014). *Handbook for Arabic language teaching professionals in the 21st century* (Vol. 2). Routledge. https://doi.org/10.4324/9780203824146

Warschauer, M., & Grimes, D. (2008). Audience, authorship, and artifact: The emergent semiotics of Web 2.0. *Annual Review of Applied Linguistics, 27*, 1-23. https://doi.org/10.1017/S0267190508070013

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100*, 94-109. https://doi.org/10.1016/j.compedu.2016.05.004