A STANDARDS-BASED EVALUATIVE ANALYSIS OF THE EFFECTIVENESS OF THE ARABIC LANGUAGE LEARNING ASSESSMENT SYSTEM FOR EIGHTH-GRADE STUDENTS IN ARABIC SCHOOLS IN BRUNEI DARUSSALAM

Achmad Yani[1*]
[*1]Universiti Islam Sultan Sharif Ali (UNISSA), Brunei Darussalam
Siti Sara binti Haji Ahmad[2]
[2]Universiti Islam Sultan Sharif Ali (UNISSA), Brunei Darussalam
Hajah Rafidah binti Haji Abdullah[3]
[3]Universiti Islam Sultan Sharif Ali (UNISSA), Brunei Darussalam
Norhayati Binti Hj Abdul Karim[4]
[4]Universiti Islam Sultan Sharif Ali (UNISSA), Brunei Darussalam
Mohamed Fathy Mohamed Abdelgelil[5]
[5]Universiti Islam Sultan Sharif Ali (UNISSA), Brunei Darussalam
Faisal Mubarak Seff[6]
[6]Universitas Islam Negeri (UIN) Antasari Banjarmasin, Indonesia
Muhammad Sabri Bin Sahrir[7]
[7]Universiti Islam Antarabangsa Malaysia (IIUM), Malaysia
Hajah Rafizah binti Haji Abdullah[8]
[8]Universiti Islam Sultan Sharif Ali (UNISSA), Brunei Darussalam
Muhammad Zakir Bin Husain[9]
[9]Universiti Islam Sultan Sharif Ali (UNISSA), Brunei Darussalam
Abdelnaser Abdelgalil Mohamed Mousa[10]
[10]Universiti Islam Sultan Sharif Ali (UNISSA), Brunei Darussalam

Abstract
This study aimed to analyze the effectiveness of the Arabic language learning assessment system for eighth-grade students in Arabic schools in Brunei Darussalam in light of global standards for language assessment, with particular emphasis on validity, comprehensiveness, alignment, and balance across the four language skills. The study adopted a mixed-methods approach using an explanatory sequential design. Quantitative data were collected from 201 students through a five-point Likert-scale questionnaire, and the findings were subsequently enriched by qualitative evidence derived from semi-structured interviews with 15 teachers, document analysis, and classroom observations. Quantitative data were analyzed using percentages and relative weights, whereas qualitative data were examined through thematic analysis. The findings revealed that the assessment system demonstrates a generally good level of overall effectiveness, albeit with noticeable variation across its components. Writing emerged as the most strongly assessed skill (80.2%), followed by reading (78.9%), speaking (76.2%), and listening (74.5%), indicating a relative preference for written over oral skills. The results further showed a good level of test validity (81.2%) and alignment with curricular content (77.7%), alongside the continued dominance of grammar-based questions (82.1%), a relative imbalance between theoretical and practical dimensions (67.8%), and a limited presence of diagnostic assessment (63%). The study underscores the need to develop the assessment system toward a more communicative and integrated model that strengthens performance-based and formative assessment and measures learners' actual linguistic competence in a more comprehensive manner.

Keywords: Arabic language learning assessment; language assessment; communicative competence; four language skills; communicative assessment; teaching Arabic to non-native speakers; Arabic schools in Brunei Darussalam.

Received: November 20, 2025. Revised: January 21, 2025. Accepted: February 12,2026. Published: March 17, 2026.

1. Introduction

In contemporary educational scholarship, language learning assessment has become a determining component of educational quality, not merely because it serves as an instrument for measuring achievement, but because it defines what counts as valuable learning, shapes classroom practices, and restructures the relationship among curriculum, teaching, and learning outcomes. Accordingly, it is no longer acceptable for language tests to be reduced to the measurement of grammatical knowledge or formal memorization. Rather, they are now expected to be grounded in a functional conception of linguistic competence that integrates listening, reading, speaking, and writing, and that measures learners' ability to use language meaningfully in authentic communicative contexts (Canale & Swain, 1980, pp. 27–29; Bachman, 1990, pp. 84–87, 126–129; Bachman & Palmer, 2010, pp. 45–46, 121–123; Fulcher, 2015, pp. 88–96; Brown & Abeywickrama, 2019, pp. 28–31, 312–314). Moreover, modern reference frameworks, most notably the CEFR Companion Volume (2020), have expanded the notion of assessment to include interaction, mediation, and descriptive scales of language performance, thereby making balance across skills and standards a central requirement of any contemporary assessment system.

Within the field of teaching Arabic as a foreign language, the significance of this shift becomes even more pronounced. Arabic is characterized by a complex phonological, morphological, syntactic, and orthographic system, which renders test construction particularly sensitive in terms of content selection, authenticity, and the representation of learners' actual performance (Hughes, 2003, pp. 88–90; Weir, 2005, pp. 98–101, 143–145; McNamara, 2000, pp. 54–57, 77–79; Green, 2014, pp. 166–168; Luoma, 2004, pp. 15–18; Hyland, 2019, pp. 76–88; Ryding, 2014, pp. 112, 188, 215; Alosh, 2016, pp. 154, 173; Al-Batal, 2017, pp. 201, 214). Previous studies have shown that school-based assessment systems often privilege reading, writing, and grammar-based questions because they are easier to design and score, while listening, speaking, and performance-based tasks are marginalized, despite being more closely aligned with the logic of communicative competence (Savignon, 2018, pp. 31–34, 41–47; Fulcher & Davidson, 2007, pp. 112, 142; Messick, 1989, pp. 18–20; Biggs & Tang, 2011, pp. 95–104; Black & Wiliam, 2009, pp. 9–12; Brookhart, 2013, pp. 49–52; Coombe et al., 2020, pp. 5–9; Norris, 2016, pp. 230–233; Purpura, 2016, pp. 190–194).

Against this backdrop, the present study undertakes an evaluative analysis of the effectiveness of the Arabic language learning assessment system for eighth-grade students in Arabic schools in Brunei Darussalam in light of global language assessment standards. More specifically, it seeks to examine the extent to which the system represents the four language skills, achieves structural balance, produces valid outcomes, and moves from a traditional knowledge-based model toward a communicative, performance-oriented model. The significance of this study lies in the scarcity of field-based investigations focusing on this particular educational stage and context, as well as in its attempt to provide a standards-based reading informed simultaneously by quantitative and qualitative evidence (Alderson, 2000, pp. 23–25, 56–58; Creswell & Plano Clark, 2018, pp. 69–71, 215–219; Tashakkori & Teddlie, 2010, pp. 141–145).

2. Theoretical Framework and Previous Studies

The theoretical framework of this study is built upon the intersection of three major foundations. The first is the theory of communicative competence, which holds that the ultimate goal of language teaching is the ability to use language effectively in real-life contexts rather than merely to know its abstract rules (Canale & Swain, 1980, pp. 27–29; Savignon, 2018, pp. 31–34). The second is the theory of construct validity and assessment usefulness, which argues that the quality of a test should be judged by the extent to which it represents the intended language construct and by the fairness with which its results are used to inform educational decisions (Bachman, 1990, pp. 126–129; Bachman & Palmer, 2010, pp. 25–27, 45–46). The third is the principle of curriculum-assessment alignment, which links objectives, content, learning activities, and measurement tools within a coherent and unified system (Biggs & Tang, 2011, pp. 95–104).

The study also draws on the principles of formative and summative assessment, in addition to the CEFR's emphasis on balance among reception, production, interaction, and mediation in assessing

language competence (Black & Wiliam, 2009, pp. 9–12; Brookhart, 2013, pp. 49–52; Council of Europe, 2020, pp. 42–46). The literature on language assessment consistently indicates that effective testing should combine validity, comprehensiveness, authenticity, and positive washback, while avoiding the formal reduction of language to isolated grammar items (Messick, 1989, pp. 18–20; Fulcher, 2015, pp. 88–96; McNamara, 2000, pp. 54–57). Likewise, the literature on skill assessment demonstrates that listening and speaking require more complex performance-based tasks, whereas school systems often privilege reading and writing because they are easier to measure and score (Alderson, 2000, pp. 23–25; Hughes, 2003, pp. 88–90; Luoma, 2004, pp. 15–18; Weir, 2005, pp. 98–101; Green, 2014, pp. 166–168; Hyland, 2019, pp. 76–88). In the specific context of Arabic as a foreign language, studies have shown that the dominance of grammar, morphology, and spelling in school tests may constrain the assessment of genuine communicative performance, even though such structural elements remain necessary linguistic components (Ryding, 2014, pp. 112, 188, 215; Alosh, 2016, pp. 154, 173; Al-Batal, 2017, pp. 201, 214).

Previous studies generally converge on the conclusion that language assessment systems in Arab and Asian school settings continue to operate between two competing paradigms: a traditional model centered on written achievement and grammatical knowledge, and a modern model oriented toward performance tasks, communicative assessment, and standards-based alignment (Brown & Abeywickrama, 2019, pp. 312, 410; Fulcher & Davidson, 2007, pp. 112, 142; Norris, 2016, pp. 72–75; Purpura, 2016, pp. 190–194; Coombe et al., 2020, pp. 5–9). Accordingly, the present study seeks to address a gap in the literature by analyzing the effectiveness of the Arabic language learning assessment system for Grade Eight in Brunei in light of global standards and through a mixed-methods design integrating both quantitative and qualitative evidence (Creswell & Plano Clark, 2018, pp. 69–71, 215; Tashakkori & Teddlie, 2010, pp. 141–145).

3. Research Methodology

This study adopted a mixed-methods approach using an explanatory sequential design, as this is among the most suitable designs for evaluative educational studies that seek to combine quantitative measurement of trends with qualitative interpretation grounded in real classroom contexts. This design begins with the collection and analysis of quantitative data, followed by the collection of qualitative data to explain and deepen the interpretation of the statistical findings (Creswell & Plano Clark, 2018, pp. 69–71; Tashakkori & Teddlie, 2010, pp. 141–145). In line with the nature of evaluative inquiry in language assessment, this design was employed to measure the effectiveness of the Arabic language learning assessment system for eighth-grade students in Arabic schools in Brunei Darussalam against global standards of validity, comprehensiveness, alignment, and balance across language skills (Bachman & Palmer, 2010, pp. 25–27; Fulcher, 2015, pp. 93–96).

The quantitative sample consisted of 201 eighth-grade students, and data were collected through a closed-ended questionnaire based on a five-point Likert scale to measure students' perceptions of the effectiveness of tests in assessing listening, reading, speaking, and writing, as well as the structure and components of the tests. This procedure is considered appropriate in school-based evaluative research aimed at identifying general patterns in learners' attitudes and perceptions (Brown & Abeywickrama, 2019, pp. 28–31; Dörnyei, 2007, pp. 100–102). The qualitative data were collected through semi-structured interviews with 15 teachers, in addition to the analysis of relevant assessment documents and classroom observations, in order to achieve methodological triangulation and thus strengthen the credibility and interpretive richness of the findings (Patton, 2015, pp. 662–664; Cohen et al., 2018, pp. 643–645).

Quantitative data were analyzed using percentages and relative weights in order to determine the degree of effectiveness of the assessment system, whereas qualitative data were subjected to thematic analysis to identify interpretive patterns supporting the statistical results (Miles et al., 2014, pp. 73–76; Saldaña, 2021, pp. 9–12). In this way, the study achieved methodological alignment between the requirements of contemporary evaluative research and the principles of language assessment grounded in communicative competence (McNamara, 2000, pp. 4–6; Weir, 2005, pp. 11–14; Green, 2014, pp. 21–24).

4. Results and Discussion

First: Analysis of the Effectiveness of the Assessment System in Light of the Four Language Skills

Assessing linguistic competence in foreign language education constitutes a central pillar in ensuring the quality of the educational process. Assessment is no longer confined to measuring cognitive achievement alone; rather, it extends to evaluating learners' overall communicative ability across the four language skills: listening, reading, speaking, and writing. Contemporary literature in language assessment emphasizes that effective assessment systems should accurately reflect learners' linguistic competence and should be aligned with the principles of communicative language assessment, which focuses on language use in authentic and meaningful contexts (Bachman & Palmer, 2010, p. 45; Fulcher, 2015, p. 88).

In light of this premise, the present field study sought to analyze the effectiveness of the Arabic language learning assessment system for Grade Eight in Arabic schools in Brunei Darussalam through the analysis of responses from 201 students using a five-point Likert scale, followed by interpretation of the findings in light of contemporary educational literature and qualitative evidence from teacher interviews and classroom documents.

1. Assessment of Listening Skill

The quantitative analysis showed that the testing system assesses students' listening competence at 74.5%, indicating an acceptable level of effectiveness in this skill. Student responses were distributed as follows: 33 students strongly agreed that the tests assess listening skill, 93 agreed, 64 remained neutral, 9 disagreed, and only 2 strongly disagreed.

This result suggests that the majority of students perceive the presence of assessment elements targeting listening skill in the tests, which is consistent with modern trends in language assessment that underscore the necessity of incorporating listening into a comprehensive communicative competence framework (Brown & Abeywickrama, 2019, p. 312).

However, the 25.5% of students who did not affirm the effectiveness of assessment in this skill reveals a possible gap between instructional objectives and the assessment mechanisms employed. Bachman and Palmer argue that listening assessment is among the most complex areas in foreign language testing because it involves real-time auditory processing and comprehension of spoken discourse across diverse contexts (Bachman & Palmer, 2010, p. 121).

Language assessment studies likewise indicate that weakness in listening assessment is often associated with reliance on traditional tests that focus on written texts rather than authentic communicative situations (Green, 2014, p. 167). Thus, the percentage recorded in this study may be interpreted as evidence of attempts to include listening tasks in the tests, though these tasks may still be limited or insufficient to measure listening competence with full precision.

2. Assessment of Reading Skill

The analysis further showed that the tests assess students' reading competence at 78.9%, a higher percentage than that recorded for listening. A total of 40 students strongly agreed with this statement, 116 agreed, 40 were neutral, 4 disagreed, and only 1 strongly disagreed.

These findings indicate that reading enjoys a relatively higher status within the assessment system compared to the other language skills. This pattern is common in many foreign language education systems, where school tests tend to emphasize reading because reading questions are easier to design and score objectively (Alderson, 2000, p. 23).

Research on language assessment has consistently demonstrated that school-based reading examinations tend to rely predominantly on direct comprehension tasks, such as multiple-choice items and short-answer questions designed to elicit explicit understanding of the text. Consequently, the evaluation of reading comprehension has become far more prevalent within formal educational settings than the assessment of oral language skills, which often require more complex and resource-

intensive procedures for reliable measurement (Weir, 2005, p. 98). This tendency reflects a broader assessment practice in many educational systems, where written comprehension tasks are considered more administratively feasible and easier to standardize, thereby reinforcing the dominance of reading-based evaluation over the systematic assessment of learners' productive oral abilities.
Nevertheless, excessive reliance on reading assessment may lead to imbalance among the language skills, which runs counter to the principles of communicative assessment advocating integrated assessment of all language skills (Canale & Swain, 1980, p. 29). Accordingly, the high percentage for reading in this study should be understood within the context of school examinations, which tend to privilege written over oral skills.

3. Assessment of Speaking Skill
The study found that the tests assess speaking skill at 76.2%. Specifically, 40 students strongly agreed with this statement, 92 agreed, 61 were neutral, 7 disagreed, and 1 strongly disagreed.
This result indicates that speaking receives a moderate level of attention in the assessment system, although it remains less prominent than reading and writing. A number of scholars have emphasized that assessing speaking represents one of the greatest challenges in language education because of its interactive nature and the difficulty of designing objective measurement tools for it (Luoma, 2004, p. 15).

Studies in language assessment also show that oral testing requires greater human and temporal resources, including trained examiners and clearly defined rating criteria, which leads some educational institutions to reduce their reliance on it (Fulcher, 2015, p. 103).
Even so, the inclusion of speaking assessment at a rate exceeding 76% in this study constitutes a positive indication that Arabic schools in Brunei are gradually moving toward a more balanced assessment of communicative competence. This accords with Hughes's view that effective speaking assessment enhances learners' motivation to use language in real communicative situations (Hughes, 2003, p. 134).

4. Assessment of Writing Skill
The analysis showed that writing is the most strongly assessed skill in the testing system, reaching 80.2%. A total of 47 students strongly agreed that the tests assess writing skill, 116 agreed, 32 were neutral, 5 disagreed, and 1 strongly disagreed.
This finding suggests that writing occupies a central position in the assessment system used in Arabic schools in Brunei. Such a pattern is common in many Arabic language education systems, where tests rely heavily on written tasks such as paragraph writing or written responses to questions (Al-Batal, 2017, p. 201).

However, excessive emphasis on writing may create imbalance in the assessment of the language skills, as the CEFR stresses that comprehensive linguistic competence requires integrated assessment across all language skills (Council of Europe, 2020, p. 44).

At the same time, the high percentage recorded in this study may be explained by the fact that writing constitutes a convenient instrument for measuring multiple linguistic dimensions simultaneously, including vocabulary, grammar, and textual organization, thereby rendering it a multidimensional assessment tool (Hyland, 2019, p. 76).

General Discussion of the Four Language Skills
Overall, the results indicate that the Arabic language learning assessment system for Grade Eight in Arabic schools in Brunei Darussalam measures the four language skills with varying degrees of effectiveness. Writing was assessed at 80.2%, followed by reading at 78.9%, speaking at 76.2%, and listening at 74.5%.

These results reflect an assessment pattern that gives relative priority to written skills over oral skills, a pattern widely observed in many foreign language assessment systems (McNamara, 2000, p. 54).

Nevertheless, the relatively high percentages for listening and speaking suggest a gradual movement toward adopting a communicative assessment model that emphasizes the ability to use language in authentic communicative situations (Savignon, 2018, p. 33).

These findings are consistent with contemporary literature in language assessment, which emphasizes that effective assessment systems should ensure balance among the four language skills and should accurately reflect learners' communicative competence (Bachman & Palmer, 2010, p. 215; Fulcher & Davidson, 2007, p. 142).

Second: Analysis of the Structure of the Assessment System in Light of Language Test Components
Contemporary literature in language assessment affirms that the quality of an assessment system is determined not only by its ability to measure language skills, but also by the nature of the internal structure of the tests, the balance in mark distribution, the comprehensiveness of linguistic coverage, and the extent to which tests are linked to actual instructional content (Fulcher, 2015, p. 96; Bachman & Palmer, 2010, p. 214). In light of this, the present study examined the structure of the assessment system used in Grade Eight Arabic schools in Brunei Darussalam by investigating a set of indicators related to test design and content.

1. Balance of Mark Distribution Across the Four Language Skills
The quantitative analysis showed that the distribution of marks across the four language skills is balanced at 72.4%. A total of 42 students strongly agreed with this indicator, 60 agreed, 87 were neutral, 5 disagreed, and 7 strongly disagreed.

This result indicates that more than two-thirds of the students perceive a degree of balance in the distribution of marks across the language skills, which is consistent with the fundamental principles of communicative assessment emphasizing the need to balance the four language skills in test design (Brown & Abeywickrama, 2019, p. 410).

However, the high neutrality rate—87 students—is significant, suggesting that the balance in mark allocation may not be sufficiently transparent to students. Language assessment studies have shown that clarity in evaluation criteria is a key factor in strengthening learners' trust in the assessment system (Weir, 2005, p. 143). McNamara also notes that imbalance in mark distribution may inflate the importance of certain skills at the expense of others, negatively affecting the direction of the teaching process (McNamara, 2000, p. 77).

It may therefore be concluded that the findings indicate an acceptable level of balance, though this balance still requires greater transparency and reinforcement in test design.

2. Inclusion of Grammar Questions in the Tests
The study found that the testing system contains grammar questions at a notably high rate of 82.1%, the highest percentage recorded among the test content indicators in this study. Specifically, 67 students strongly agreed that grammar questions were present, 97 agreed, 29 were neutral, 7 disagreed, and 1 strongly disagreed.

This result indicates that the tests in Brunei's Arabic schools rely heavily on the assessment of grammatical knowledge. This tendency reflects long-standing instructional traditions in Arabic language education that assign major importance to grammar as the foundation for understanding language structure (Ryding, 2014, p. 112).

Yet modern literature in foreign language education indicates that excessive dependence on grammar-based questions may transform assessment into a form of formal cognitive testing rather than the assessment of actual communicative competence (Canale & Swain, 1980, p. 28). Bachman likewise argues that tests that focus excessively on grammar may fail to capture the true ability to use language in real communicative contexts (Bachman, 1990, p. 87).

Accordingly, the high proportion of grammar questions in this study may be interpreted both as an expression of the instructional traditions of Arabic teaching and as an indicator of the need to achieve greater balance between the assessment of linguistic knowledge and the assessment of communicative language performance.

3. Inclusion of Morphology Questions in the Tests

The analysis further revealed that the inclusion of morphology questions in the tests reached 68.1%, a lower rate than that recorded for grammar questions. In total, 35 students strongly agreed that morphology questions were present, 60 agreed, 67 were neutral, 25 disagreed, and 12 strongly disagreed.

These findings indicate that morphology questions constitute an important component of the assessment system, though they do not receive the same degree of emphasis as grammar questions. This reflects the nature of Arabic instruction, which often prioritizes syntactic structure over morphological structure (Alosh, 2016, p. 154).

The relatively high neutrality rate—67 students—also suggests that some students may not clearly perceive the presence of morphological elements in the tests. This accords with research in teaching Arabic to non-native speakers, which identifies Arabic morphology as one of the most complex linguistic domains for learners (Ryding, 2014, p. 215).

4. Inclusion of Questions Related to Writing Conventions

The study showed that the tests contain questions related to writing conventions at 76.4%. A total of 44 students strongly agreed with this indicator, 81 agreed, 67 were neutral, 5 disagreed, and 1 strongly disagreed.

This result suggests that the assessment system in Arabic schools in Brunei pays clear attention to writing conventions, which is expected in the context of Arabic language instruction, given that Arabic relies on a relatively complex writing system involving multiple orthographic and morphological conventions (Al-Batal, 2017, p. 214).

Educational literature indicates that assessing writing conventions is an important component of linguistic competence because it reflects the learner's ability to use the language accurately in writing (Hyland, 2019, p. 88). Even so, excessive focus on writing conventions may diminish attention to the communicative dimensions of language use (Savignon, 2018, p. 41).

5. Alignment of the Tests with Instructional Content

The study found that the tests align with the lessons and exercises taught during the semester at 77.7%. A total of 48 students strongly agreed with this indicator, 90 agreed, 55 were neutral, and 8 disagreed.

This result points to a good level of consistency between instructional content and the assessment system, which is one of the core principles of sound educational test design. In the literature, this principle is commonly referred to as instruction-assessment alignment (Biggs & Tang, 2011, p. 95).

Language assessment studies further affirm that consistency among educational objectives, instructional content, and assessment tools is essential for ensuring test validity (Bachman & Palmer, 2010, p. 222).

6. The Ability of Tests to Reflect Students' True Language Level

The analysis showed that test results reflect students' actual language level at 71.7%. A total of 33 students strongly agreed with this indicator, 78 agreed, 69 were neutral, 13 disagreed, and 7 strongly disagreed.

This result suggests a reasonable level of confidence in the ability of the tests to measure students' real language proficiency. However, the high neutrality rate indicates a degree of uncertainty among some students regarding the accuracy of test outcomes.

Bachman emphasizes that the ability of a test to measure genuine linguistic competence lies at the heart of construct validity in language assessment (Bachman, 1990, p. 126).

7. Validity of the Testing System
The study revealed that the testing system enjoys a high level of validity, reaching 81.2%. A total of 75 students strongly agreed with this indicator, 67 agreed, 52 were neutral, 4 disagreed, and 1 strongly disagreed.

This indicator suggests that the majority of students trust the fairness of the tests and their ability to measure language performance objectively. Validity is regarded as one of the most important criteria of test quality in educational literature, referring to the extent to which a test measures what it is intended to measure (Messick, 1989, p. 18).

8. Comprehensiveness of the Assessment System
The results showed that the assessment system possesses a level of comprehensiveness amounting to 73.7%, with 38 students strongly agreeing with this indicator, 65 agreeing, 91 remaining neutral, and

5 disagreeing.
This finding indicates that the assessment system covers multiple dimensions of linguistic competence. Yet the high neutrality rate suggests that some students may not clearly perceive the comprehensiveness of the system.

9. Balance Between Theoretical and Practical Dimensions
The study found that the tests combine theoretical and practical dimensions at 67.8%. A total of 20 students strongly agreed, 66 agreed, 87 were neutral, 26 disagreed, and 1 strongly disagreed.
This result indicates that the tests still lean relatively more toward theoretical dimensions than practical ones, thereby requiring further development in the direction of assessing actual language performance.

10. Diagnostic Tests
The findings showed that students undergo diagnostic tests before the beginning of instruction at only 63%, indicating a moderate level of use of this type of assessment.

11. Achievement Tests
The results further showed that students sit for achievement tests after the completion of instruction at 69.8%, indicating a clear reliance on summative assessment.
Third: Interpretation of the Findings in Light of Qualitative Data (Interviews, Document Analysis, and Classroom Observations)

Mixed-methods research confirms that integrating quantitative data with qualitative evidence provides a deeper understanding of educational phenomena, as it enables statistical patterns to be interpreted in light of actual instructional contexts (Creswell & Plano Clark, 2018, p. 215; Tashakkori & Teddlie, 2010, p. 137). Based on this principle, the questionnaire findings were interpreted through an analysis of semi-structured interviews conducted with 15 Arabic language teachers in Arabic schools in Brunei Darussalam, in addition to the analysis of classroom assessment documents and field observations during lessons.
The qualitative data revealed a number of interpretive patterns that help explain the quantitative findings.

1. Interpreting the Assessment of the Four Language Skills
The quantitative analysis showed that the tests assess the four language skills to varying degrees: writing (80.2%), reading (78.9%), speaking (76.2%), and listening (74.5%).
Teacher interviews indicated that this variation reflects the nature of test design in Arabic schools, as a number of teachers explained that written tests are easier to organize and score than oral tests. One teacher stated during the interview:

"We usually rely on written questions because they are easier to score and ensure fairness among students."

This interpretation is consistent with the literature on language assessment, which shows that school examinations tend to prioritize written skills because of the ease of measurement and objectivity of scoring (Alderson, 2000, p. 56; Hughes, 2003, p. 88).

Classroom observations also revealed that some teachers use listening and speaking activities during instruction, although these activities are not always formally assessed in official examinations. Bachman and Palmer note that this pattern is common in many foreign language education systems, where communicative activities form part of teaching but are not always reflected in formal assessment tools (Bachman & Palmer, 2010, p. 144).

2. Interpreting the Balance of Mark Distribution

The study found that the balance of mark distribution across language skills reached 72.4%. Teacher interviews suggested that mark allocation is often determined according to the traditional structure of school examinations, in which reading and writing receive a larger proportion of the final score than oral skills.

One teacher explained:
"Reading and writing tests make up the largest part of the final mark because they are easier to prepare for midterm and final examinations."
This interpretation corresponds with educational literature on the washback effect, according to which the nature of assessment influences teaching practices and determines the priority given to particular language skills (Fulcher & Davidson, 2007, p. 112).

3. Interpreting the Strong Focus on Grammar Questions

The analysis showed that the inclusion of grammar questions in the tests reached 82.1%, the highest percentage among the test content indicators. The qualitative data suggest that this emphasis reflects the nature of the curriculum in Arabic schools, where the teaching of grammar holds a prominent place.

Most teachers interviewed emphasized that understanding grammatical rules is essential for learning Arabic, especially for non-native speakers. This indicates the continued influence of the traditional approach to Arabic language teaching, which regards grammar as the foundation of language mastery (Ryding, 2014, p. 188).

At the same time, some teachers also expressed the need to redesign tests so that they assess students' ability to use grammar in communicative contexts rather than merely their theoretical knowledge of rules. This view is consistent with modern trends in language education, which stress the shift from assessing linguistic knowledge to assessing actual language performance (Savignon, 2018, p. 47).

4. Interpreting the Lower Percentage of Morphology Questions

The study found that the inclusion of morphology questions reached 68.1%, lower than the percentage for grammar questions. Interviews with teachers suggested that teaching morphology poses a relative challenge for students, especially in the earlier school stages.
One teacher remarked:
"Morphological concepts, such as verb patterns, can be difficult for students at this stage."

This observation aligns with research on teaching Arabic to non-native speakers, which identifies the Arabic morphological system as one of the most challenging linguistic dimensions for learners (Alosh, 2016, p. 173).

5. Interpreting Test Validity and Comprehensiveness
The analysis revealed that the validity of the testing system reached 81.2%, whereas the comprehensiveness of the system reached 73.7%. Teacher interviews indicated that test design is often based on models approved by the Ministry of Education in Brunei, which enhances consistency and objectivity in test construction.

Official test documents also showed that most questions are directly linked to the educational objectives specified in the curriculum, thereby reinforcing the concept of curriculum alignment (Biggs & Tang, 2011, p. 104).

However, some teachers observed that the tests still focus more strongly on linguistic knowledge than on communicative performance. This interpretation is consistent with recent literature in language assessment, which suggests that many educational systems remain in transition from traditional to communicative assessment (Fulcher, 2015, p. 201).

6. Interpreting the Limited Use of Diagnostic Assessment
The study found that the use of diagnostic tests before the beginning of instruction reached only 63%. Teacher interviews indicated that the main reason for this is time pressure at the beginning of the academic year.

Researchers in educational assessment emphasize that diagnostic testing is one of the most important tools of formative assessment because it helps identify students' proficiency levels before instruction begins (Black & Wiliam, 2009, p. 12).

7. Interpreting the Reliance on Achievement Testing
The results showed that the use of achievement tests reached 69.8%. The qualitative data suggest that this type of assessment constitutes the principal means of evaluating students at the end of the semester.

This pattern confirms what the educational literature has repeatedly shown: most educational systems continue to rely heavily on summative assessment rather than formative assessment (Brookhart, 2013, p. 52).

Summary of Results Analysis
Overall, the findings indicate that the Arabic language learning assessment system for Grade Eight in Arabic schools in Brunei Darussalam demonstrates a good degree of effectiveness, with the various indicators ranging from 63% to 82.1%.

The results reveal several major characteristics of the system, most notably:
- The clear presence of assessment for the four language skills, with greater emphasis on written skills;
- A marked reliance on grammar questions compared with morphology questions;
- A good level of validity and comprehensiveness in the testing system; and
- A greater dependence on achievement tests than on diagnostic tests.

These findings suggest that the assessment system used in Brunei's Arabic schools is moving in a positive direction toward meeting language assessment standards, but it still requires further development in order to strengthen communicative assessment and expand the use of formative assessment tools.

5. Conclusions and Recommendations
This field study concludes that the Arabic language learning assessment system for eighth-grade students in Arabic schools in Brunei Darussalam demonstrates a good overall level of effectiveness, although this effectiveness remains uneven across the various domains of assessment. The findings showed that the tests assess the four language skills at relatively comparable levels, with a clear advantage for written skills: writing ranked first at 80.2%, followed by reading at 78.9%, speaking at

76.2%, and listening at 74.5%. This ranking reveals that the assessment system tends to give relative priority to those skills that can be more easily tested through written formats, often at the expense of oral skills, which require more complex assessment procedures.

The results further showed that the structure of the tests demonstrates an acceptable degree of organization and consistency. The balance of mark distribution across the skills reached 72.4%, the alignment of tests with lessons and classroom exercises reached 77.7%, and test validity reached 81.2%, which is a strong indicator of students' confidence in the fairness of the assessment and its connection to instructional objectives. At the same time, however, the study revealed a clear dominance of grammar questions (82.1%) compared with the relatively lower proportion of morphology questions (68.1%). In addition, the balance between theoretical and practical dimensions did not exceed 67.8%, indicating the continued presence of a traditional tendency to assess linguistic knowledge more than actual communicative performance. The findings also showed greater reliance on achievement tests (69.8%) than on diagnostic tests (63%).

The qualitative evidence derived from teacher interviews, document analysis, and classroom observations strongly supported these results. It confirmed that written tests are the most common form of assessment because of the relative ease of preparing and scoring them, and that oral activities are sometimes practiced in the classroom without being clearly reflected in formal assessment tools. The documents also demonstrated a general alignment between the curriculum and the tests, albeit with a stronger emphasis on cognitive and rule-based components.

In light of these findings, the study proposes the reconstruction of the assessment system according to a broader communicative perspective, one capable of achieving genuine balance among listening, speaking, reading, and writing, while increasing the presence of performance-based and practical tasks. It also recommends expanding the use of diagnostic and formative assessment, developing clear rubrics for the assessment of oral and written skills, training teachers to design tests grounded in communicative competence, and aligning assessment with contemporary international standards such as the CEFR and modern principles of language assessment. Such reforms would help shift the assessment system from a predominant focus on abstract linguistic knowledge toward a more authentic measurement of learners' actual ability to use Arabic in educational and communicative contexts.

References

Alderson, J. C. (2000). Assessing reading. Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). Language test construction and evaluation. Cambridge University Press.

Al-Batal, M. (2017). Arabic as one language: Integrating dialect in the Arabic language curriculum. Georgetown University Press.

Alosh, M. (2016). Teaching Arabic as a foreign language: Issues and directions. Georgetown University Press.

Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). Language assessment in practice: Developing language assessments and justifying their use in the real world. Oxford University Press.

Biggs, J., & Tang, C. (2011). Teaching for quality learning at university (4th ed.). Open University Press.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability, 21(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Brindley, G. (2013). Assessment literacy in language education. Routledge.

Brookhart, S. M. (2013). How to create and use rubrics for formative assessment and grading. ASCD.

Brown, H. D., & Abeywickrama, P. (2019). Language assessment: Principles and classroom practices (3rd ed.). Pearson.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards & R. Schmidt (Eds.), Language and communication. Longman.
Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1(1), 1–47. https://doi.org/10.1093/applin/I.1.1
Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model. Issues in Applied Linguistics, 6(2), 5–35.
Chapelle, C. A. (2010). The spread of computer-assisted language learning. Language Teaching, 43(1), 66–74. https://doi.org/10.1017/S0261444809005850
Chapelle, C. A., & Douglas, D. (2006). Assessing language through computer technology. Cambridge University Press.
Cohen, L., Manion, L., & Morrison, K. (2018). Research methods in education (8th ed.). Routledge. https://doi.org/10.4324/9781315456539
Coombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn? Studies in Educational Evaluation, 64, 100822. https://doi.org/10.1016/j.stueduc.2019.100822
Council of Europe. (2020). Common European framework of reference for languages: Learning, teaching, assessment—Companion volume. Council of Europe Publishing.
Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). Sage.
Davies, A. (2008). Textbook trends in teaching language testing. Language Testing, 25(3), 327–347. https://doi.org/10.1177/0265532208090156
Dörnyei, Z. (2007). Research methods in applied linguistics. Oxford University Press.
Ellis, R. (2003). Task-based language learning and teaching. Oxford University Press.
Ellis, R. (2018). Reflections on task-based language teaching. Multilingual Matters.
Fulcher, G. (2012). Assessment literacy for the language classroom. Language Assessment Quarterly, 9(2), 113–132. https://doi.org/10.1080/15434303.2011.642041
Fulcher, G. (2015). Re-examining language testing: A philosophical and social inquiry. Routledge. https://doi.org/10.4324/9781315733050
Fulcher, G., & Davidson, F. (2007). Language testing and assessment: An advanced resource book. Routledge. https://doi.org/10.4324/9780203937761
González-Lloret, M., & Ortega, L. (2014). Technology-mediated task-based language teaching. John Benjamins.
Green, A. (2014). Exploring language assessment and testing: Language in action. Routledge. https://doi.org/10.4324/9781315889627
Hughes, A. (2003). Testing for language teachers (2nd ed.). Cambridge University Press.
Hyland, K. (2019). Second language writing (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108635547
Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), Sociolinguistics. Penguin.
Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base. Language Testing, 25(3), 385–402. https://doi.org/10.1177/0265532208090158
Levi, T., & Inbar-Lourie, O. (2020). Assessment literacy or language assessment literacy. Language Assessment Quarterly, 17(2), 168–182. https://doi.org/10.1080/15434303.2019.1692347
Long, M. (2015). Second language acquisition and task-based language teaching. Wiley-Blackwell.
Luoma, S. (2004). Assessing speaking. Cambridge University Press. https://doi.org/10.1017/CBO9780511733017
McNamara, T. (2000). Language testing. Oxford University Press.
McNamara, T., & Roever, C. (2006). Language testing: The social dimension. Blackwell.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). Macmillan.
Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). Qualitative data analysis: A methods sourcebook (3rd ed.). Sage.
Nation, I. S. P., & Macalister, J. (2010). Language curriculum design. Routledge.
Norris, J. M. (2016). Current uses for task-based language assessment. Annual Review of Applied Linguistics, 36, 230–244. https://doi.org/10.1017/S0267190516000027

North, B. (2014). The CEFR in practice. Language Teaching, 47(4), 459–465. https://doi.org/10.1017/S0261444814000201
Patton, M. Q. (2015). Qualitative research & evaluation methods (4th ed.). Sage.
Purpura, J. E. (2016). Second and foreign language assessment. The Modern Language Journal, 100(S1), 190–208. https://doi.org/10.1111/modl.12318
Richards, J. C., & Rodgers, T. S. (2014). Approaches and methods in language teaching (3rd ed.). Cambridge University Press.
Ryding, K. C. (2014). Teaching and learning Arabic as a foreign language: A guide for teachers. Georgetown University Press.
Saldaña, J. (2021). The coding manual for qualitative researchers (4th ed.). Routledge. https://doi.org/10.4324/9781003185536
Savignon, S. J. (2018). Communicative competence. Routledge. https://doi.org/10.4324/9781315228280
Skehan, P. (2018). Second language task-based performance. Routledge.
Stiggins, R. (2005). From formative assessment to assessment for learning. Phi Delta Kappan, 87(4), 324–328. https://doi.org/10.1177/003172170508700414
Tashakkori, A., & Teddlie, C. (2010). The SAGE handbook of mixed methods in social and behavioral research (2nd ed.). Sage. https://doi.org/10.4135/9781506335193
Taylor, L. (2013). Communicating the theory, practice and principles of language testing. Language Testing, 30(3), 403–412. https://doi.org/10.1177/0265532213480338
Wahba, K., Taha, Z., & England, L. (2013). Handbook for Arabic language teaching professionals. Routledge.
Weir, C. J. (2005). Language testing and validation: An evidence-based approach. Palgrave Macmillan. https://doi.org/10.1057/9780230514577