Insyirah: Jurnal Ilmu Bahasa Arab dan Studi Islam
Vol. 7, No. 2, December 2024, p. 219-231
DOI: 10.26555/insyirah.v7i2.11863

Test Evaluation Techniques in Arabic Language Learning: Theory and Practice at SDIT Insan Mulia Pekalongan

[1]Abd Rofi Fathoni Nidhomillah *, [2]Aidah Fithtriyah, [3]Amrina Rosyada, [4]Nur Qomari

[1]230104220046@student.uin-malang.ac.id, [2]230104220052@student.uin-malang.ac.id,
[3]230104220041@student.uin-malang.ac.id [4]qomari@uin-malang.ac.id
[1234]Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia

ARTICLE INFO

Article history
Received: 22 October 2024
Revised: 9 November 2024
Accepted: 23 December 2024

Keywords
Evaluation,
Test,
Arabic Learning,

*Corresponding Author

ABSTRACT

Arabic language learning at the elementary level is crucial for understanding Islamic values and developing general language skills. Evaluation plays a key role in ensuring effective teaching and learning, especially with the increasing diversity of assessment methods. This study aims to analyze the implementation of Arabic language evaluation at SDIT Insan Mulia Pekalongan, focusing primarily on test evaluation techniques. The research employs a qualitative approach with a case study design, involving interviews, observations, and documentation. The results indicate that evaluations are systematically conducted through various types of tests, including Daily Summative Tests (SH) and End-of-Semester Summative Assessments (ASAS), which assess students' abilities in reading, writing, listening, and speaking. Additionally, oral assessments and memorization practices using songs are implemented to provide constructive feedback. This study also identifies that the use of Item Response Theory (IRT) and Classical Test Theory (CTT) in evaluations can offer a more comprehensive understanding of students' abilities. The findings are expected to contribute to the development of more effective teaching strategies and improve the quality of Arabic language learning at the primary school level.

1. Introduction

The importance of learning Arabic at the elementary school level is highly significant, especially in the context of religious education, as this language is the language of the Qur'an and

Hadith (Andriani, 2015). By learning Arabic, students can gain a deeper understanding of Islamic teachings directly. Additionally, mastering Arabic from an early age supports the development of general language skills, including listening, speaking, reading, and writing. Arabic also serves as a gateway to understanding the culture and traditions of Arabic-speaking countries, thereby broadening students' insights into cultural diversity (Ridwan, 2023).

In the context of careers, proficiency in Arabic can open opportunities in various fields, given the increasing demand for a workforce skilled in the language. Additionally, learning a foreign language such as Arabic can enhance students' cognitive abilities, including problem-solving and critical thinking (Mauludiyah & Murdiono, 2023). Therefore, mastering Arabic is not only beneficial for religious education but also supports the development of social, cultural, and career-related skills for the future (Zaidar, 2023).

The development of evaluation methods in education has undergone significant changes, driven by shifts in learning paradigms. Initially, evaluation focused more on quantitative aspects, emphasizing learning outcomes in the form of numbers and grades through multiple-choice tests and final exams. However, with the emergence of competency-based approaches, evaluation has transitioned toward measuring students' skills and abilities, placing greater importance on the learning process rather than just final results.

The term "evaluation" is derived from the English word "evaluation." According to Mehrens and Lehmann, as quoted by Ngalim Purwanto, it refers to a process that involves planning, gathering, and providing essential information to support decision-making among various alternatives (Purwanto, 2004). An educator needs to evaluate the success of teaching to improve and direct the learning process effectively. Learning evaluation is a systematic assessment of this process, covering several components: students' initial behavior (raw input), teachers' professional competence (instrumental input), curriculum (study programs, methods, and media), administrative resources (tools, time, and funding), teaching implementation procedures (process), and learning outcomes (output), which reflect the achievement of objectives (Hamalik, 1995).

Evaluation is a systematic process of collecting, analyzing, and interpreting information to determine the extent to which specific objectives or criteria have been achieved. In the context of education, evaluation aims to assess students' learning progress, teaching effectiveness, and curriculum quality (Magdalena et al., 2020). Evaluation can be conducted through various methods, including tests, observations, and project-based assessments. The results of evaluation are used to provide feedback, improve the learning process, and make informed decisions regarding education (Eka Saputri et al., 2024).

In educational settings, evaluation helps in decision-making, such as determining whether a specific approach, method, or technique should be implemented (Sofyan & dkk, 2006). There are two primary types of evaluation in schools: test-based evaluation and non-test evaluation. Test-based evaluation involves measuring students' abilities through tasks or questions (Sudijono, 2006). Tests are divided into three types: diagnostic tests, which identify students' weaknesses for targeted intervention (Suharsimi Arikunto, 2002), formative tests, which assess students' progress toward learning objectives and are often referred to as "daily quizzes"; and summative tests, which are conducted at the end of a teaching program, known as "general exams," and are used for report cards or certification purposes. Test-based evaluation can also be categorized into written and oral tests, depending on how questions are presented. Meanwhile, non-test evaluation involves methods such as rating scales, questionnaires, checklists, interviews, observations, and biographies (Arikunto, 2002).

Formative assessment conducted during the learning process is essential for providing constructive feedback, while summative assessment evaluates students' achievements at the end of a learning period (Pramita, 2023). Moreover, information technology has transformed evaluation practices, facilitating data collection and enabling more in-depth analysis. Project-based and authentic assessments emphasize the application of knowledge in real-world contexts, helping measure collaboration, creativity, and problem-solving skills (Warsah & Habibullah, 2022).

Continuous feedback is increasingly recognized as a crucial part of the learning process, helping students identify their strengths and areas for improvement. Thus, the evolution of evaluation methods reflects a shift from traditional approaches to more holistic assessments aimed at enhancing the quality of learning and students' overall development.

Item Response Theory (IRT) dan Classical Test Theory (CTT) are two important approaches in educational evaluation, including in Arabic language testing. IRT focuses on the relationship between students' abilities and their responses to test items, allowing for a detailed analysis of question characteristics and the use of adaptive tests that align with students' ability levels. n contrast, CTT emphasizes total scores, assuming that scores consist of true ability and measurement error. Although CTT is simpler and easier to implement, IRT offers greater precision in measurement. Combining these two theories provides a more comprehensive understanding of students' abilities, enhances test quality, and offers more specific feedback for developing Arabic language learning (Mustafidah & Harjono, 2019). In contrast, CTT emphasizes total scores, assuming that scores consist of true ability and measurement error. Although CTT is simpler and easier to implement, IRT offers greater precision in measurement. Combining these two theories provides a more comprehensive understanding of students' abilities, enhances test quality, and

offers more specific feedback for developing Arabic language learning (Mustafidah et al., 2018). By leveraging the strengths of both approaches, educators can design more effective evaluations to support continuous learning.

Previous studies relevant to this research include: 1) A study by Khoirotun Ni'mah and Durrotun Nafisah titled *"Pelaksanaan Evaluasi Pembelajaran Bahasa Arab di SD Negeri Tlogorejo Sukodadi Lamongan"*. The research found that Arabic teachers use three types of assessment: first, oral questioning after class sessions; second, periodic quizzes at the end of basic competency (KD) instruction; and third, mid-semester or final exams, which combine material from several KDs over a specific period. Various types of tests are used according to different language skills, including listening, speaking, reading, and writing (Ni'mah & Nafisah, 2020). 2) Another study by Uswatun Hasanah and Andi Prastowo titled *"Evaluasi Kurikulum Pembelajaran Bahasa Arab Untuk Perbaikan Mutu Akademik Di MI Mambaul Ma'arif"* found that four types of curriculum evaluation were implemented: formative, summative, input, and product evaluation. The instruments used included tests (divided into ten types) and non-tests (divided into twelve types). The selection of instruments had to align with students' needs and characteristics while being easy for teachers to design and implement. If evaluation results did not meet minimum standards, remedial activities were required; if the standards were met, enrichment activities could be provided (Chasanah & Prastowo, 2021).

Arabic language learning at the elementary school level plays a crucial role in building foundational language skills and understanding Islamic teachings. At SDIT Insan Mulia Pekalongan, Arabic learning is integrated into religious and general curricula, requiring effective evaluation methods to assess students' abilities in reading, writing, listening, and speaking comprehensively. Previous studies have highlighted the importance of evaluation in Arabic learning.

Khoirotun Ni'mah and Durrotun Nafisah (2020) examined Arabic evaluation methods in SD Negeri Tlogorejo, focusing on oral questioning, periodic quizzes, and mid-term exams, aligned with basic competencies. Meanwhile, Uswatun Hasanah and Andi Prastowo (2021) explored curriculum evaluation in MI Mambaul Ma'arif, identifying a combination of formative, summative, input, and product evaluations using diverse instruments tailored to students' needs.

However, these studies primarily focus on evaluation types and their general application. This research distinguishes itself by integrating Item Response Theory (IRT) and Classical Test Theory (CTT) in analyzing test-based evaluations. By bridging theoretical approaches and practical implementation, this study provides deeper insights into how comprehensive evaluation strategies can enhance Arabic language learning outcomes at SDIT Insan Mulia Pekalongan.

Arabic language learning at the elementary school level holds significant importance, particularly in Islamic educational contexts, as it helps students understand Islamic teachings and develop language skills such as reading, writing, listening, and speaking. At SDIT Insan Mulia Pekalongan, Arabic learning is integrated into religious and general subjects, reflecting its essential role in the curriculum. However, ensuring effective Arabic learning requires systematic evaluations to assess and enhance students' language skills comprehensively.

Existing evaluation methods often lack a balance between traditional and modern approaches, resulting in incomplete assessments of students' abilities. Moreover, the integration of advanced theories, such as Item Response Theory (IRT) and Classical Test Theory (CTT), in evaluations is rarely explored in primary education settings. This gap highlights the necessity of this research to analyze how test-based evaluation methods, incorporating IRT and CTT principles, can improve the quality of Arabic language assessments at SDIT Insan Mulia Pekalongan. By addressing these challenges, the study aims to contribute to more effective evaluation strategies and better educational outcomes.

2. Method

This study employs a qualitative approach with a case study design. This approach is chosen to gain an in-depth understanding of the implementation of test-based evaluation in Arabic language learning at SDIT Insan Mulia Pekalongan. The data collection techniques include interviews, observations, and documentation (Lexy J. Moleong, 2006). The collected data consists of narratives reflecting the views and experiences of participants, particularly teachers responsible for teaching Arabic. The interviews in this study are in-depth interviews, allowing respondents to freely and comprehensively provide answers and explanations. The chosen interview technique is unstructured interviews, enabling participants to describe themselves and their environment using their own words, which reflect their culture and traditions (Mulyana, 2020). The second data collection technique involves observation, which focuses on monitoring the evaluation processes carried out in the classroom, including the types of evaluations applied. The third technique is documentation, which consists of test results from the evaluation of Arabic language learning at SDIT Insan Mulia Pekalongan.

Data analysis follows the model proposed by Miles and Huberman, which includes data reduction, data display, and conclusion drawing (Hermawan Wasito, 1995). In the data reduction phase, the researcher summarizes and selects important information from the study regarding the implementation of test-based evaluations in Arabic language learning. Next, in the data display phase, the findings are presented to facilitate the final data analysis, allowing the researcher to draw conclusions related to the implementation of the Arabic language evaluation. To ensure data

validity, the researcher performs data triangulation by comparing information from various sources, including interviews, observations, and documentation (Sugiyono, 2017).

3. Results and Discussion

3.1. Practice of Test Evaluation Techniques in Arabic Learning

SDIT Insan Mulia is an elementary-level educational institution located in Tanjungkulon Village, Kajen District, Pekalongan Regency, Central Java. Its activities operate under the supervision of the Ministry of Education and Culture. Education under the Ministry of Education and Culture (Kemendikbud) includes Arabic language learning, which is an essential part of the curriculum. In every learning process, evaluation is necessary to assess students' abilities and their progress in mastering the Arabic language.

This evaluation aims not only to measure academic achievement but also to provide constructive feedback to students, helping them identify areas that need improvement. Through various evaluation methods, including written and oral tests, educators can design more effective learning strategies tailored to students' needs. Consequently, Arabic language learning is expected to run optimally, facilitating students in developing their language skills.

"I teach Arabic for grades 1 through 4. For evaluation, we use test-based evaluation techniques, which include written tests such as SH (Daily Summative) and ASAS (End-of-Semester Summative Assessment). SH is conducted at the end of each chapter, before moving on to a new one, and during the tests, students are not allowed to open their books. SH consists of 15 questions (10 multiple-choice and 5 fill-in-the-blank). The grading system for multiple-choice questions is 1 point for each correct answer and 0 for incorrect answers. For fill-in-the-blank questions, correct answers are multiplied by 2, partially correct answers by 1, and incorrect answers by 0.5." (Ustadzah Durrotul Khikmah, S.Pd, Guru Bahasa Arab di SDIT Insan Mulia Pekalongan)

Table 1. Evaluation of Class 3 C for the Odd Semester of the 2024/2025 Academic Year

No.	Name	PH1	PH2	PTS	PH3	PH4	PAS	B.3	B.4	B.2	B.1
1	Adhityahayu	100	100	100	100	100	96	90	90	83	90
2	Akhlam	68/75	96	70/75	96	100	96	90	90	83	83
3	Akhza	100	80	55/75	60/75	52/75	72/75	85	90	75	75
4	Aisyah	88	72/75	55/75	50/75	36/75	44/75	80	85	79	80
5	Kauren	100	96	90	84	98	90	90	90	83	89
6	Arsyil Ady	96	84	100	76	98	98	85	90	81	81
7	Syafiq I.	88	84	80	74/75	94	98	85	90	83	89
8	Safa H.	100	100	100	100	100	100	90	85	85	90
9	Raka Jati	100	96	100	96	100	100	85	90	83	85
10	F.N Rafie Zain	64/75	56/75	50/75	44/75	44/75	50/75	85	90	78	80
11	Belva	96	96	100	98	100	100	90	90	80	85
12	Gian	90	90	90	90	92	92	85	90	75	78

13	Jihan	92	68/75	75	54/75	40/75	46/75	75	75	75	75
14	Bintang Y	92	96	95	96	96	96	85	90	75	75
15	Lintang Aska	36/75	36/75	45/75	52/75	44/75	62/75	90	90	78	80
16	Najwan Al.	100	96	100	100	100	98	90	90	75	75
17	Mahira	96	100	100	98	100	100	90	90	85	90
18	Nada	96	84	95	74/75	100	98	80	85	83	90
19	Fatih	68/75	64/75	50/75	50/75	78	67/75	90	90	78	81
20	Alfin	92	80	95	62/75	68/75	86	85	85	78	80
21	Ibas Z.	80	76	95	88	86	94	85	90	79	83
22	Umar	96	80	85	84	94	96	90	90	83	85
23	Zain Al.G	92	68/75	85	86	92	84	85	90	78	80
24	Rama	96	100	95	92	84	98	90	90	85	87
25	Rasya	75/75	80	95	72/75	88	80	75	75	75	75
26	Sakeeza (Ckeke)	92	92	95	100	96	100	90	90	82	83
27	Siti Aisyah H.	96	100	100	96	100	98	90	90	82	90

At SDIT Insan Mulia Pekalongan, the implementation of Arabic language learning evaluation for grades 1 to 4, taught by Ustadzah Durrotul Khikmah, S.Pd, employs test-based evaluation techniques. This evaluation includes written and oral tests. For the written tests, there is a Daily Summative (SH) activity conducted each time there is a change of chapter. Students are required to answer 15 questions, consisting of 10 multiple-choice questions and 5 fill-in-the-blank questions. In this grading system, students earn 1 point for each correct answer on the multiple-choice questions, while incorrect answers receive no points. For fill-in-the-blank questions, correct answers earn 2 points, partially correct answers earn 1 point, and incorrect answers receive 0.5 points. During the test, students are prohibited from opening their books.

"The Final Summative Assessment (ASAS) at SDIT Insan Mulia Pekalongan comprises 25 questions in multiple-choice, short answer, and essay formats. Scoring varies: multiple-choice questions award 1 point for correct answers and 0 for incorrect ones, short answers earn up to 2 points for complete responses and 0 for unanswered ones, and essays receive up to 3 points for thorough answers with 0 for entirely incorrect ones. Additionally, daily task assessments evaluate students' understanding of the material taught. (Ustadzah Durrotul Khikmah, S.Pd, Guru Bahasa Arab di SDIT Insan Mulia Pekalongan)

Table 2. Evaluation of Class 2 A for the Odd Semester of the 2024/2025 Academic Year

No.	**Name**	**Tugas Bab 5**	**PH Bab 5**	**Tugas Bab 6**	**PH Bab 6**	**Bab 7 L1**	**Bab 7 L2**	**PH Bab 7**	**PAT**
1	Abbad	100	100	90/100	95	80	100	100	98
2	Adira	100	88	75/100	45/75	100	75	40/75	82
3	Aya	100	100	100	90	100	85	92	98
4	Aira	100	76	100	100	100	85	92	93
5	Aisyah	100/90	100	100/75	90	100	75	76	89
6	Reifan	100	72/75	100	70/75	100	92	36/88	67/75
7	Adela	100	100	100	95	100	100	100	94

8	Bilqis	100	100	100	100	100	100	88	98
9	Bondan	100	100	100	95	100	90	96	98
10	Clara	100	100	100	100	100	85	80	80
11	Dipa	100	64/75	100	73/75	80	85	88	79
12	Fathan	100	60/75	90/100	92	100	90	96	89
13	Hafza	100	100	100	100	100	90	92	89
14	Cinta	100	52/75	100	73/75	100	80	88	82
15	Ivan	100	68/75	80	68/75	100	89	89	84
16	Keinara	100	92	80	68/75	100	85	92	90
17	Alif	100	80	100	88	100	89	88	95
18	Abi	100	100	100	100	100	85	100	100
19	Sira	100	100	100	100	100	100	100	100
20	Abim	100	100	80/100	98	80	100	100	87
21	Rakha	100/80	92	100	78	100	100	72/75	73/75
22	Niken	100	100	100	100	75	100	96	98
23	Ratu	100	100	100	98	100	90	76	93
24	Javas	100	92	100	95	100	90	92	95
25	Rayyan	100/80	80	80	68/75	100		44/75	95
26	Salsa	100	100	100	100	100	90	92	95
27	Shakila	100	100	90/100	93	100	100	96	98
28	Alma	100	96	100	98	100	90	100	88
29	Viola	100	92	100/80	93	100	100	96	68/75
30	Zaim	100	64/75	80/100	75	100	90	80	95

In the Final Summative Assessment (ASAS), the evaluation is conducted through three types of questions: multiple-choice, short answer, and essay questions. For multiple-choice questions, correct answers receive 1 point, while incorrect answers receive no points. This method simplifies the assessment process and allows students to demonstrate their basic understanding of the material while reducing pressure, as there is no penalty for wrong answers, encouraging students to respond with more confidence. Next, for short answer questions, correct answers receive 2 points, unanswered questions receive no points, and partially correct answers receive 1 point. This system provides students the opportunity to show a deeper understanding and motivates them to make an effort, even if they are not completely certain. Finally, for essay questions, correct answers earn 3 points, unanswered questions earn no points, partially correct responses earn 2 points, and incorrect answers earn 1 point. The assessment for essay questions requires students to think critically and articulate their arguments clearly, with diverse criteria helping educators understand the extent of students' grasp of the material and their ability to express ideas. With this varied assessment approach, it is hoped to provide a comprehensive picture of students' abilities in Arabic language learning.

“In addition to written tests, we also conduct oral tests that include the practice of memorizing vocabulary using songs (for example, the melody of "Sayonara" incorporated with vocabulary) and conversation practice. In grades 1 to 4, conversation practice is typically conducted in front of the class, while in grades 5 and 6, students are required to create conversation videos. There are four

assessment categories for this practice: not fluent, somewhat fluent, fluent, and very fluent." (Ustadzah Durrotul Khikmah, S.Pd, Guru Bahasa Arab di SDIT Insan Mulia Pekalongan).

Table 3. Evaluation of Class 4 A for the Odd Semester of the 2024/2025 Academic Year

No.	Name	T1	P	T2	SH	T3	SH2	T4	SH3	T5	SH4	Total
1	Adzkiya	100	100	92								
2	Fakhira		100	89								
3	Abyan	100	100	86								
4	Alfa	100	70	85								
5	Aleena	100	100	90								
6	Alghazali	100	80	90								
7	Alfaro Nizam	60	80	85								
8	Azka	100	80	87								
9	Saskara	85	100	89								
10	Naima	100	100	92								
11	Evano	100	100	90								
12	Syasya	100	100	89								
13	Guntur	100	60	89								
14	Hasna	100	100	89								
15	Ibanes	100		85								
16	Rafasya	100	90	80								
17	Khansa	100	80	90								
18	Syamil	92	80	90								
19	Alovie	100	80	88								
20	Alvaro	100	90	89								
21	Fatih	100	70	87								
22	Izzul	85	50	85								
23	Miska	60	80	88								
24	Wafa	85		85								
25	Najma	100	100	92								
26	Nizam	85	90	87								
27	Salsa	100	100	89								
28	Rumaisya	100	80	92								
29	Thafana	100	90	85								

Oral tests are conducted through the practice of memorizing vocabulary using songs, which helps students remember vocabulary in a fun way. The assessment for this oral test is categorized as: not fluent, somewhat fluent, fluent, and very fluent. Additionally, students in grades 5 and 6 are assigned a project to create conversation videos, while students in grades 1 to 4 simply practice conversation in front of the class, with the teacher's assessment based on how well the practice aligns with the conversation book used.

3.2. Analysis of Learning Evaluation Theory

Based on the findings of the study on the evaluation of Arabic language learning at SDIT Insan Mulia Pekalongan, the implementation of diverse evaluation methods reflects the principles of Item Response Theory (IRT) and Classical Test Theory (CTT). In this context, IRT focuses on a detailed analysis of students' ability to memorize Arabic vocabulary (mufradat). Meanwhile, CTT emphasizes an in-depth evaluation of students through tests, such as written assessments, by measuring their abilities based on the scores obtained.

In this context, the use of *mufradat* memorization within the framework of Item Response Theory (IRT) aims to analyze the relationship between students' ability to memorize vocabulary and their responses to the given assessments. When students are evaluated through memorization practices, IRT enables educators to assess the characteristics of the vocabulary used. For example, teachers can identify the difficulty level of specific vocabulary based on the number of students who successfully memorize it. Thus, IRT provides deeper insights into students' abilities and helps design adaptive evaluations where difficulty levels are adjusted according to individual student capabilities.

This approach also facilitates the development of adaptive tests, where question difficulty is aligned with students' abilities, providing more accurate feedback on their understanding of the material (Hambleton, R. K., Swaminathan & Rogers, 1991). Using IRT, educators can create more effective assessment tools that not only focus on final outcomes but also on students' learning processes, offering clearer insights into areas that require improvement. Hence, IRT enables the development of more effective instruments and provides more precise feedback on areas needing improvement, both in vocabulary memorization and in overall Arabic language learning.

On the other hand, the evaluation of Arabic language learning through Classical Test Theory (CTT) in this study is reflected in the use of written tests such as *Sumatif Harian* (SH) and *Asesmen Sumatif Akhir Semester* (ASAS). These tests align with CTT principles, which emphasize the total scores achieved by students (McDonald, 1999). Assessment is carried out through various question types, including multiple-choice, short-answer, and essay questions, providing a comprehensive picture of students' abilities. For instance, in SH, students are required to answer 15 questions consisting of 10 multiple-choice questions and 5 short-answer questions, with scoring based on the number of correct answers.

CTT assumes that the scores obtained reflect students' true abilities, although there may be elements of measurement error. In this context, the grading system, where correct answers receive specific points and incorrect answers receive none, reduces pressure on students, encouraging them to attempt questions without fear of penalty. This approach enables teachers to gain clear insights into students' understanding of the material and makes it easier to decide on

remedial or enrichment actions. Thus, the application of CTT in the evaluation of Arabic language learning at SDIT Insan Mulia not only focuses on final outcomes but also provides useful feedback for developing a more effective learning process.

Additionally, CTT, which emphasizes total scores, is reflected in both summative and formative evaluations. The study reveals that assessments are designed with a variety of question types, allowing teachers to collect comprehensive data on students' achievements. By employing both approaches simultaneously, the school can design assessments that not only meet measurement standards but also provide valuable information about students' learning processes.

Therefore, the use of these theories in the application of diverse evaluation methods at SDIT Insan Mulia Pekalongan whether through the principles of Item Response Theory (IRT) or Classical Test Theory (CTT)—significantly supports the measurement of students' abilities in learning Arabic. The combination of these two approaches offers a more comprehensive view of students' academic achievements and provides constructive feedback to enhance the learning process (Haladyna, T. M. and Rodriguez, 2013). In this way, effective evaluation can contribute to the development of teaching strategies that are better aligned with students' needs, ensuring they achieve proficiency in Arabic.

The use of both Item Response Theory (IRT) and Classical Test Theory (CTT) in the evaluation methods at SDIT Insan Mulia Pekalongan significantly enhances the accuracy and comprehensiveness of measuring students' abilities in Arabic learning. IRT, which focuses on the relationship between students' abilities and their responses to test items, allows for a more detailed analysis of each student's performance. This theory helps to determine the difficulty level of each test item based on how students respond, providing more precise insights into their strengths and weaknesses (Purwa Antara, 2023). For instance, when analyzing vocabulary memorization through oral assessments, IRT enables teachers to assess the difficulty of specific words by observing how many students can recall them correctly.

On the other hand, CTT provides a broader view of students' overall ability by focusing on total scores. While it is easier to implement and understand, it has limitations in its ability to account for varying levels of student ability. CTT assumes that the test score reflects both the true ability of the student and potential measurement errors (Setiawan & Susanah, 2023). Thus, while CTT is practical for general assessments like the Daily Summative (SH) and End-of-Semester Summative Assessments (ASAS), it may not provide as detailed a picture of individual students' strengths.

By combining IRT and CTT, this study offers a more holistic view of students' academic achievements in Arabic learning. The integration of both approaches not only enhances the reliability of the evaluations but also provides constructive feedback that helps educators identify specific areas for improvement. This comprehensive evaluation system, rooted in both IRT and

CTT, allows for more tailored teaching strategies that align with students' unique needs, ensuring they can achieve proficiency in Arabic more effectively.

4. Conclusion

This study highlights the implementation of Arabic language learning evaluation at SDIT Insan Mulia Pekalongan, with a focus on test-based evaluation techniques. The findings indicate that the evaluation is conducted systematically through various methods, including written and oral tests, designed to assess students' abilities in reading, writing, listening, and speaking in Arabic. The diverse evaluation methods, such as *Sumatif Harian* (SH) and *Asesmen Sumatif Akhir Semester* (ASAS), reflect the application of principles from Item Response Theory (IRT) and Classical Test Theory (CTT). These approaches enable educators to obtain more comprehensive data regarding students' abilities.

The implications of this research include improving the quality of Arabic language learning through constructive feedback for students, helping them identify their strengths and areas for improvement. These findings can also be used to design a curriculum that is more responsive to students' needs by incorporating appropriate evaluation methods for varying skill levels. Additionally, the use of IRT and CTT allows teachers to develop more adaptive teaching strategies by aligning assessment instruments with students' abilities. The study also emphasizes the importance of teacher training in the effective implementation of evaluation methods and encourages further research on the effectiveness of other evaluation techniques and their impact on student learning outcomes in Arabic language education. Thus, this research makes a significant contribution to the development of evaluation practices in Arabic language education while supporting the improvement of educational quality at the primary school level.

References

Andriani, A. (2015). Urgensi Pembelajaran Bahasa Arab dalam Pendidikan Islam. *Ta'allum: Jurnal Pendidikan Islam*, *3*(1), 39–56. https://doi.org/10.21274/taalum.2015.3.1.39-56

Chasanah, U., & Prastowo, A. (2021). Evaluasi Kurikulum Pembelajaran Bahasa Arab Untuk Perbaikan Mutu Akademik Di Mi Mambaul Ma'Arif. *Ta'allum: Jurnal Pendidikan Islam*, *9*(2), 272–299. https://doi.org/10.21274/taalum.2021.9.2.272-299

Eka Saputri, R., Firmansyah, R., & Silfiya, S. (2024). Pentingnya Evaluasi Pembelajaran Untuk Meningkatkan Kompetensi Peserta Didik di Sekolah Dasar. *Sindoro Cendekia Pendidikan*, *3*(8), 10–20.

Haladyna, T. M. and Rodriguez, M. C. (2013). *Developing And Validating Test Items.* Taylor and Prancis.

Hamalik, O. (1995). *Kurikulum dan Pembelajar*. Bumi Aksara.

Hambleton, R. K., Swaminathan, H. and, & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* SAGE Publications.

Hermawan Wasito. (1995). *Pengantar Metodologi Penelitian*. gramedia pustaka.

Lexy J. Moleong. (2006). *Metodologi Penelitian Kualitatif*. PT. Remaja Rosdakarya Offset.

Magdalena, I., Dea, K. Y., & Puspitasari. (2020). Rendahnya mutu hasil belajar siswa Sekolah Dasar

dengan adanya pembelajaran online. *Jurnal Edukasi Dan Sains*, *2*(2), 292–305. https://ejournal.stitpn.ac.id/index.php/edisi

Mauludiyah, L., & Murdiono. (2023). Pendampingan Pembelajaran Bahasa Arab Berbasis Pjbl-Steam Pada Guru Bahasa Arab Di Kota Malang. *Journal of Research on Community Engagement (JRCE)*, *5*(1), 21–26.

McDonald, R. . (1999). *Test Theory: A Unified Treatment, Larvrence Erbaum Associates*. New Jersey.

Mulyana, D. (2020). *Metodologi Penelitian Kualitatif: Paradigma Baru Ilmu Komunikasi dN Ilmu Sosial Lainnya* (Pipih Latifah (ed.)). PT Remaja Rosdakarya.

Mustafidah, H., & Harjono, H. (2019). Implementation of QUEST Program to analyze test items for SMP Muhammadiyah 2 Karanglewas Teachers. *JPPM (Jurnal Pengabdian Dan Pemberdayaan Masyarakat)*, *3*(2), 321.

Mustafidah, H., Harjono, H., & Purwo Wicaksono, A. (2018). Peningkatan Kemampuan Menganalisis Butir Soal Tes bagi Guru-guru MGMP IPS Menggunakan Program QUEST. *JPPM (Jurnal Pengabdian Dan Pemberdayaan Masyarakat)*, *2*(1), 47. https://doi.org/10.30595/jppm.v2i1.1430

Ni'mah, K., & Nafisah, D. (2020). Pelaksanaan Evaluasi Pembelajaran Bahasa Arab Di Sd Negeri Tlogorejo Sukodadi Lamongan. *Al-Fakkaar: Jurnal Ilmiah Pendidikan Bahasa Arab*, *1*(1), 23–39.

Pramita, K. N. (2023). Evaluasi Pembelajaran Dalam Ranah Aspek Kognitif Pada Jenjang Pendidikan Dasarpada MI Assalafiyah Timbangreja. *Jurnal Review Pendidikan Dan Pengajaran*, *6*(2), 403–411.

Purwa Antara, A. A. (2023). *Karakteristik Tes Prestasi Belajar Berdasarkan Pendekatan Klasik Dan Item Response Theory* (Issue April).

Purwanto, N. (2004). *Prinsip-Prinsip dan Teknik Evaluasi Pengajaran*. PT. Remaja Rosdakarya.

Ridwan, M. (2023). Membuka Wawasan Keislaman: Kebermaknaan Bahasa Arab Dalam Pemahaman Islam. *Jazirah: Jurnal Peradaban Dan Kebudayaan*, *4*(2), 102–115. https://doi.org/10.51190/jazirah.v4i2.100

Setiawan, D. B., & Susanah, S. (2023). Penerapan Goal-Free Problems dalam Pembelajaran Matematika secara Kolaboratif untuk Melatih Kemampuan Siswa dalam Memecahkan Masalah. *MATHEdunesa*, *12*(1), 275–288. https://doi.org/10.26740/mathedunesa.v12n1.p275-288

Sofyan, A., & dkk. (2006). *Evaluasi Pembelajaran IPA Berbasis Kompetensi*. UIN Jakarta Press.

Sudijono, A. (2006). *Pengantar Evaluasi Pendidikan*. PT. Raja Grafindo Persada.

Sugiyono. (2017). *Metode Penelitian Kuantitatif Kualitatif dan R&D*. Alfabeta.

Suharsimi Arikunto. (2002). *Dasar-Dasar Evaluasi Pendidikan*. Bumi Aksara.

Warsah, I., & Habibullah. (2022). Implementasi Evaluasi Hasil Belajar Pendidikan Agama Islam di Madrasah. *JOEAI (Journal of Education and Instruction)*, *5*(1), 213–225. https://dataindonesia.id/sektor-riil/detail/angka-konsumsi-ikan-ri-naik-jadi-5648-kgkapita-pada-2022

Zaidar, M. (2023). Pembelajaran Bahasa Arab dalam Pengembangan Karakter Anak di Era Modern: Kajian Konseptual. *Islamic Insights Journal*, *5*(1), 42–55. https://islamicinsights.ub.ac.id/index.php/insights/article/view/89