



Item Analysis of Arabic Daily Test for Fifth-Grade Students Using the Rasch Model at MIN 1 Pasaman Barat

Ulfi Amelia

Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia

Corresponding author e-mail: ameliaulfi1503@gmail.com

Keywords:	Abstract
Rasch Model; Item Analysis; Arabic Language Assessment	<i>Assessment in Arabic language learning plays an essential role in measuring students' cognitive achievement; however, the quality of test instruments is often not empirically validated. This study aims to analyze the quality of Arabic daily test items for fifth-grade students at MIN 1 Pasaman Barat using the Rasch model, focusing on item validity, reliability, and difficulty level. This research employed a quantitative approach with a descriptive-exploratory design. Data were collected from students' responses to multiple-choice items and analyzed using Rasch measurement through fit statistics, reliability indices, and logit-based item difficulty. The results showed that only a small number of items met the validity criteria based on Rasch fit statistics, while the majority were identified as misfit. The item reliability was categorized as moderate, indicating that the internal consistency of the instrument has not yet reached an optimal level. In addition, the distribution of item difficulty was unbalanced, with most items classified as difficult, suggesting a mismatch between item difficulty and students' ability levels. These findings indicate that the test instrument requires improvement in terms of item construction and difficulty distribution. The study implies that Rasch analysis can serve as an effective diagnostic tool to enhance the quality of assessment instruments, contributing to more valid, reliable, and standardized evaluation practices in Arabic language learning at the primary school level.</i>

INTRODUCTION

Background of the Study

Assessment in Arabic language learning plays a strategic role as an instrument for measuring the achievement of students' competencies, particularly in the cognitive domain (Fauzi & Rahmawati, 2025). One of the most commonly used forms of evaluation in elementary schools is the daily test, which functions as a formative assessment to continuously monitor students' mastery of the material (Latifah & Suwandi, 2024). Through daily tests, teachers not only obtain an overview of students' learning outcomes but can also use the results as a basis for making pedagogical decisions to improve the learning process (A. M. Ibrahim, 2024).



However, a number of studies in the past five years indicate that (Utama & Nawas, 2026). Ideally, evaluation instruments should meet the principles of validity, reliability, and a proportionally calibrated level of difficulty (Saputra, 2025). In the context of Arabic language learning, the complexity of materials such as *mufradat* (vocabulary), *qawa'id* (grammar), and language skills requires instruments that are not only content-valid but also psychometrically stable (Rahmadon & Oktarina, 2025).

The Problem of The Study

Preliminary findings at MIN 1 Pasaman Barat indicate that Arabic daily test items have not been developed based on empirical item analysis. The test items used have not undergone psychometric testing, so their validity and reliability cannot yet be objectively ensured. This condition reflects a gap between evaluation practices in the field and modern data-driven assessment standards.

Moreover, item characteristics such as difficulty level, discrimination power, and the possibility of item misfit have never been systematically analyzed. This may lead to instruments that are unable to accurately and consistently identify students' abilities (Fitraynsyah & Hilmiyati, 2024). Furthermore, the absence of empirical analysis can result in an imbalanced distribution of item difficulty, which ultimately affects the quality of interpretation of learning outcomes (Elmananda et al., 2024).

Research's State of the Art

Studies on item analysis in educational evaluation in recent years have developed through two main approaches, namely Classical Test Theory (CTT) and Item Response Theory (IRT) (Pratama, 2020). CTT-based studies are still predominantly used in various educational research due to the ease of calculating difficulty and discrimination indices (Stemler & Naples, 2021). However, recent studies indicate that CTT has fundamental limitations, as item parameters are highly dependent on sample characteristics, making them less stable for cross-population generalization (Maria et al., 2025).

In more recent developments, the IRT approach particularly the Rasch Model has been increasingly applied in educational evaluation research. International studies show that the Rasch Model is capable of providing sample-invariant estimates (N. L. Ibrahim, 2025), allowing both person ability and item difficulty to be placed objectively on the same logit scale (Riskawati et al., 2025). In addition, the Rasch Model enables the analysis of item misfit and person fit, which cannot be optimally achieved using CTT (Novriyanti & Arthur, 2024).

However, based on recent literature mapping, the application of the Rasch Model in the context of Arabic language learning at the elementary school level remains very limited. Existing studies generally focus on numerical subjects such as mathematics or science, while the domain of Arabic language learning is still at an exploratory stage and has not widely adopted modern psychometric approaches in a comprehensive manner. Thus, there is a research gap in integrating Arabic language assessment with diagnostic Rasch-based analysis.

Novelty, Research Gap, & Objective

Based on the synthesis of previous studies, three major gaps can be identified. First, in terms of context, Rasch-based item analysis in Arabic language learning at the elementary level is still scarce. Second, methodologically, most studies still rely on the Classical Test Theory approach, which has not been able to provide in-depth diagnostic information. Third, practically, there is no study that specifically evaluates the quality of Arabic daily test items with a focus on detecting item misfit and calibrating difficulty levels using a modern measurement model. Based on these gaps, the novelty of this study lies in the application of diagnostic Rasch analysis for Arabic assessment quality control in the primary school context, specifically used to analyze the quality of Arabic daily test items. This approach not only focuses on measuring validity and reliability but also on identifying misfit items and mapping difficulty levels based on a logit scale that is more objective and standardized.

This study contributes to three main aspects. Theoretically, it strengthens the development of educational evaluation literature based on Item Response Theory in the context of Arabic language learning. Methodologically, it introduces the application of the Rasch Model as a diagnostic tool for analyzing evaluation instruments at the elementary school level. Practically, this study is expected to serve as a basis for improving the design of Arabic language assessment instruments to make them more valid, reliable, and aligned with modern measurement standards. The objective of this study is to analyze item validity, examine the reliability of the instrument, and identify item difficulty levels as well as potential misfit items based on the Rasch Model in the Arabic daily test of fifth-grade students at MIN 1 Pasaman Barat.

METHOD

Type and Design

This study employs a quantitative approach with a descriptive-exploratory design based on the Rasch Model (Patonah et al., 2023). This design is selected because the study aims to empirically evaluate the quality of Arabic daily test items, including item validity, instrument reliability, difficulty level, and the identification of misfitting items. The Rasch Model, as part of Item Response Theory (IRT), is used due to its advantage in producing measurements that are invariant across samples and items, thereby providing more objective and stable results in educational measurement (Riskawati et al., 2025). Therefore, this study falls into the category of evaluative research on test instruments that have been used in the learning process

Data and Data Sources

This study was conducted at MIN 1 Pasaman Barat during the current academic year, with fifth-grade students who had taken the Arabic daily test as the research subjects. All students who participated in the test were included as respondents through a total sampling technique. The selection of the entire population as the sample aims to obtain more accurate parameter estimates in Rasch analysis, particularly in mapping students' abilities and item characteristics in a more representative and stable manner.

Data Collection Technique

Data were collected through documentation and testing techniques. The documentation technique was used to obtain the daily test items and students' answer sheets, while students' response data served as the primary data for quantitative analysis. All collected data were then tabulated into an item-person response matrix format, which was subsequently used as input for the Rasch model analysis.

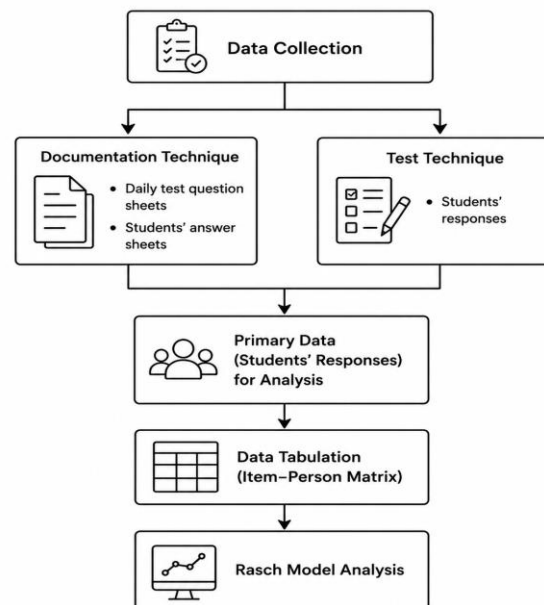


Figure 1. Data Collection and Analysis Procedure Using the Rasch Model

Data Analysis

Data analysis was conducted using the Rasch model approach with the assistance of psychometric analysis software such as Winsteps or similar programs. The analysis focused on examining item fit and person fit to determine the consistency between empirical data and the Rasch model through infit and outfit mean square indicators. In addition, item difficulty levels were estimated on a logit scale to observe the distribution of item difficulty. Instrument reliability was analyzed using person reliability and item reliability to assess measurement consistency (Maghfirotnunnisa, 2023). Furthermore, the separation index was used to measure the instrument's ability to distinguish between different levels of student ability as well as item quality. Visualization using the Wright Map was also employed to map the distribution of students' abilities and item difficulty levels on the same logit scale, thereby providing a comprehensive overview of instrument quality.

Instrument validity in this study is not only based on content validity but also on construct validity grounded in the Rasch model (Ramadhan, 2024). An item is considered valid if its infit and outfit mean square values fall within the acceptable range according to Rasch standards. Items that show misfit with the model will be identified as part of the item quality evaluation and considered for revision or elimination in subsequent instrument development.

RESULTS AND DISCUSSIONS

The results of data analysis using the Rasch Model indicate that the Arabic daily test instrument exhibits varied psychometric characteristics in terms of validity, reliability, and item difficulty levels. In general, the findings show that the instrument has a fairly sound measurement foundation; however, several weaknesses remain, particularly in aspects of construct validity and the distribution of item difficulty.

Validity Test Results

The analysis of test quality in distinguishing students' abilities was examined from the perspectives of validity, reliability, and item difficulty. Item validity was initially tested using the product-moment correlation method. An item is considered valid if the correlation coefficient $r_{xy} > r_{\text{tabel}}$ whereas $r_{xy} \leq r_{\text{tabel}}$ item is considered invalid (Ida & Musyarofah, 2021). Within the Rasch Model framework, item validity is determined based on several criteria (Suciainsani, Okky, Euis Eti Roeshaeti, 2021).

- a) The acceptable Outfit Mean Square (MNSQ) range is: $0,5 < \text{Outfit -MNSQ} < 1,5$
 - b) The acceptable Outfit Z-Standard (ZSTD) range is: $-2,0 < \text{ZSTD} < +2,0$
 - c) The acceptable Point Measure Correlation (Pt Measure Corr) range is: $0,4 < \text{Point Measure Corr} < 0,85$
- After conducting validity testing for each item, the results can be seen in Table 1:

Table 1. Results of Item Validity Analysis Using the Rasch Model Approach

Results	Question number
Valid	4,6,7,9,13,16,18
Invalid	1,2,3,5,8,10,11,12,14,15,17,19,20,21,22,23,24,25

Based on the results of item quality analysis in terms of validity using the Rasch Model, it was found that 7 items were categorized as valid, while 23 items were considered invalid because they did not meet the criteria for Outfit MNSQ, Outfit ZSTD, and Point Measure Correlation (Pt Measure Corr). The Rasch Model is considered more accurate in determining item validity because an item must satisfy all three criteria to be classified as valid.

Furthermore, based on the Rasch-based validity testing using the criteria of Outfit MNSQ (0.5–1.5), Outfit ZSTD (–2.0 to +2.0), and Point Measure Correlation (0.4–0.85) (Fahmi, 2022), the analysis of 25 items revealed that only 7 items namely items 4, 6, 7, 9, 13, 16, and 18 were valid. Meanwhile, the remaining 18 items did not meet one or more fit statistics criteria and were therefore categorized as invalid. These findings indicate that most items are not fully aligned with the Rasch measurement model, suggesting that the construct validity of the instrument still needs improvement (Maria et al., 2025). These results also imply that although the items were developed in accordance with the instructional content, empirically there are still inconsistencies between students' response patterns and the model expectations. This condition highlights that instrument validity is not solely

determined by content alignment but also by the statistical consistency of items in measuring the same construct.

Item Reliability

The results of the reliability analysis show that the item reliability value is 0.67, which falls into the moderate category based on established criteria (Endah Pratiwi, 2025). This indicates that, overall, the instrument demonstrates a moderate level of internal consistency in measuring students' Arabic language ability, although it has not yet reached a high level of reliability. Therefore, the instrument still requires improvement to produce more stable and consistent measurements in future studies. In the Rasch Model framework, item and person reliability are interpreted based on specific criteria used to determine the quality of measurement (Hadiyanti & Susongko, 2024):

Table 2. Reliability Criteria in the Rasch Model

Reliability value (person/item)	Interpretasi
>0,94	Special
0,91-0,94	Very good
0,81-0,90	Good
0,67-0,80	Enough
<0,67	Weak

Table 3. Results of Item Reliability Analysis

Model rasch	
Item Reliability	Category
0,67	Enough

Based on Table 3, the item reliability value is 0.67, which falls into the moderate category. Therefore, it can be concluded that the quality of the test items in the instrument demonstrates an adequate level of reliability.

Item Difficulty Level

In terms of item difficulty, the analysis using the Rasch Model shows that the distribution of items is uneven. Based on the logit values, there are 2 items categorized as very easy, 3 items as easy, 17 items as difficult, and 3 items as very difficult. The dominance of items in the difficult category indicates that the instrument tends to be more challenging for students and is therefore not fully able to accommodate the variation in students' abilities in a balanced manner. In the Rasch Model, item difficulty is categorized based on logit measures and the standard deviation (SD) of item logits, which are generally divided into four categories as follows (Sri Astuti Iriyani, Elyakim N.S Patty, 2025):

Table 4. Criteria for Item Difficulty Levels

Measure Value (logit)	Interpretation of the difficulty of the question item
Measure logit > SD logit	It's very easy
-SD logit $\leq$ measure logit $\leq$ 0	Easy
0 $\leq$ measure logit $\leq$ SD logit	Difficult
Measure logit > SD logit	very difficult

Table 5. Distribution of Item Difficulty Levels Using the Rasch Model

Measure Value (logit)	Interpretation of the difficulty of the question item	A number of questions
Measure logit > -1,02	It's very easy	2
-1,02 $\leq$ measure logit $\leq$ 0,0	Easy	3
0,0 $\leq$ measure logit $\leq$ 1,02	Difficult	17
Measure logit > 1,02	very difficult	3

Based on Table 5, the results of item analysis in terms of difficulty level using the Rasch Model based on logit measure values show that the items are grouped into four categories: 2 items are classified as very easy, 3 items as easy, 17 items as difficult, and 3 items as very difficult. This imbalance in the distribution of item difficulty affects the instrument's ability to optimally differentiate students' abilities, particularly at the lower to intermediate levels. In the context of educational evaluation, an ideal instrument should have a proportional distribution of difficulty levels to measure the full range of students' abilities more accurately.

Furthermore, the results of item fit analysis using the Rasch Model indicate that most items fall within the acceptable range of infit and outfit mean square (Zuhri, 2024), although several items show indications of misfit. These misfitting items reflect inconsistencies between students' response patterns and the model's expectations, suggesting the need for revision or elimination of certain items. This finding reinforces the previous validity test results, which showed that not all items possess good psychometric quality (Danni & Wahyuni, 2021).

The Wright Map analysis also reveals a mismatch between the distribution of students' abilities and item difficulty levels. Some students are positioned below the average item difficulty level, indicating that the instrument is relatively too difficult

for a portion of the participants (Laili, 2020). Conversely, several items are too easy and do not provide significant measurement information. This condition suggests that the instrument has not yet optimally mapped students' abilities in a comprehensive manner. These findings are consistent with previous studies indicating that the Rasch Model provides more stable estimates of reliability and validity compared to Classical Test Theory, as it is invariant across samples (Anastya Nida Alhana, 2024). However, the presence of misfitting items indicates that the quality of item construction in instructional practice still needs improvement to align with modern measurement principles based on Item Response Theory.

From a pedagogical standpoint, classroom assessment should be understood as a fundamental element of the instructional process rather than simply a means of assigning grades. High-quality assessments provide empirical evidence of students' achievement, reveal learning strengths and weaknesses, and enable teachers to evaluate the effectiveness of instructional strategies. Daily tests, therefore, should function as diagnostic and formative tools that support continuous improvement in teaching and learning. When assessment instruments are developed systematically, they not only measure learning outcomes but also generate meaningful information that can guide curriculum implementation, instructional planning, and targeted academic interventions.

The development of effective assessment instruments requires a comprehensive evaluation framework that extends beyond content accuracy and alignment with curriculum standards. Although content validity remains essential, the psychometric properties of an instrument must also be examined to determine whether it measures students' abilities with sufficient accuracy, consistency, and fairness. Reliable and valid instruments minimize measurement error and ensure that students' scores genuinely reflect their level of competence rather than weaknesses in test construction. Consequently, psychometric evaluation should become a routine component of classroom assessment development to strengthen the credibility and educational value of assessment outcomes.

Within this context, the Rasch Model offers a rigorous analytical approach for evaluating the quality of educational assessment instruments. Unlike conventional methods that primarily emphasize total test scores, the Rasch Model investigates the interaction between item characteristics and student ability on a common measurement scale. This approach enables educators to determine whether individual items perform according to measurement expectations, identify inconsistencies in item functioning, and estimate the relative difficulty of each question with greater precision. Such information provides an objective basis for refining assessment instruments and improving the overall quality of educational measurement.

The application of the Rasch Model is particularly relevant in Arabic language education because language proficiency encompasses multiple interconnected competencies that differ in both complexity and cognitive demand. Students are expected to demonstrate mastery of vocabulary, grammatical structures, reading comprehension, and the practical use of language in meaningful contexts. These diverse competencies cannot be adequately assessed through poorly constructed or

unbalanced test items. Instead, they require carefully designed instruments capable of distinguishing varying levels of student proficiency while maintaining consistency across different dimensions of language learning. Psychometric analysis therefore plays a crucial role in ensuring that assessment results accurately represent students' actual language abilities.

Integrating Rasch-based evaluation into classroom assessment practices also contributes to the establishment of a more evidence-based assessment culture in elementary education. Teachers are provided with objective information that supports informed decisions regarding item revision, instrument improvement, and instructional enhancement. This continuous evaluation process promotes greater transparency, fairness, and accountability in student assessment while encouraging the development of instruments that remain responsive to evolving educational standards. Ultimately, combining pedagogical considerations with rigorous psychometric evaluation enables daily tests to function as reliable indicators of student learning, thereby enhancing both the quality of assessment practices and the effectiveness of Arabic language instruction.

The Rasch Model analysis demonstrates that the Arabic daily test instrument possesses a moderate level of reliability, indicating that it is capable of producing reasonably consistent measurements of students' abilities. This finding suggests that the instrument has the potential to serve as an effective classroom assessment tool, although its measurement precision has not yet reached an optimal level. In addition, the analysis confirms that several test items satisfy the validity requirements established by the Rasch measurement framework. These items exhibit acceptable fit statistics, indicating that they function consistently with the expectations of the measurement model and contribute meaningfully to the assessment of students' competencies. The presence of valid items provides empirical evidence that a substantial portion of the instrument has been appropriately designed and can accurately represent the targeted learning outcomes.

Nevertheless, the analysis also identifies several psychometric weaknesses that limit the overall effectiveness of the instrument. One of the primary concerns is the uneven distribution of item difficulty, which indicates that the test does not adequately cover the full spectrum of student abilities. An assessment instrument should ideally include a balanced combination of easy, moderate, and difficult items to distinguish effectively among learners with different proficiency levels. The current imbalance reduces the instrument's discriminative power and may result in less accurate estimates of student achievement. Furthermore, the identification of several misfitting items suggests that some questions do not function consistently within the Rasch measurement model. These items may contain ambiguous wording, inappropriate distractors, inconsistent cognitive demands, or content that does not align closely with the intended construct, thereby reducing the overall validity and precision of the assessment.

The findings further demonstrate the practical value of Rasch analysis as a comprehensive framework for improving educational assessment. Unlike conventional item analysis, which primarily focuses on item difficulty and discrimination indices, the Rasch Model provides a more sophisticated evaluation of

the interaction between item characteristics and student ability. This approach enables educators to identify problematic items, detect unexpected response patterns, evaluate item functioning across different ability levels, and calibrate the instrument based on empirical evidence. Consequently, Rasch analysis should not be viewed solely as a statistical procedure but as a valuable decision-support tool for teachers, curriculum developers, and educational researchers seeking to enhance the quality of classroom assessment. Regular psychometric evaluation can promote more accurate, objective, and equitable measurement practices while strengthening the credibility of assessment results used to monitor students' academic progress.

Overall, this study concludes that the Arabic daily test instrument demonstrates satisfactory psychometric characteristics and provides a reasonable foundation for assessing students' learning achievement at the elementary school level. However, the identified limitations indicate that the instrument has not yet achieved the level of quality required for optimal educational measurement. Continuous validation, systematic revision, and periodic recalibration based on Rasch analysis are therefore essential to improve the instrument's reliability, validity, and measurement accuracy. Future development should prioritize the refinement of misfitting items, the establishment of a more balanced distribution of item difficulty, and the expansion of item coverage to better represent the targeted competencies. Through sustained evidence-based improvement, Arabic language assessment instruments can provide more dependable information about students' learning outcomes, support fairer educational evaluation, and contribute to the enhancement of instructional quality and student achievement in elementary education..

CONCLUSION

This study aims to analyze the quality of Arabic daily test items for fifth-grade students at MIN 1 Pasaman Barat using the Rasch Model approach, particularly in terms of validity, reliability, and item difficulty. The results indicate that the instrument used has not fully met optimal measurement standards. This is reflected in the limited number of items that meet the validity criteria based on Rasch fit statistics, as well as the reliability value, which falls into the moderate category. In addition, the unbalanced distribution of item difficulty dominated by difficult items suggests that the instrument has not been able to proportionally accommodate the variation in students' abilities. These findings confirm that the use of the Rasch Model as a diagnostic analysis approach is capable of revealing instrument weaknesses more comprehensively compared to conventional approaches. Thus, this study highlights that evaluating Arabic language learning requires not only content alignment but also psychometrically sound instruments that have been empirically tested. This also reinforces the importance of integrating Item Response Theory into evaluation practices at the elementary school level.

Practically, the results imply the need to enhance teachers' competencies in designing and evaluating test instruments based on modern analytical approaches such as the Rasch Model. Assessment instruments should be continuously developed by considering a balanced distribution of difficulty levels and the alignment of items with measurement models. In this way, the evaluation process in Arabic language

learning can be conducted more objectively, accurately, and in a manner that truly reflects students' abilities.

For future research, this study is still limited by the sample size and its focus on a single school context. Therefore, further studies are recommended to involve larger samples and to develop more indepth analyses, such as the use of polytomous Rasch models or Differential Item Functioning. This is expected to enrich the study of Arabic language assessment and strengthen the development of higher-quality, internationally standardized assessment instruments

REFERENCES

- Anastya Nida Alhana, S. R. (2024). ANALISIS KUALITAS BUTIR SOAL ULANGAN HARIAN SEKOLAH DASARMATA PELAJARAN BAHASA INDONESIA MENGGUNAKAN RASCH MODEL. *Jurnal Tawadhu*, 8(2), 194–204. <https://doi.org/https://doi.org/10.52802/twd.v8i2.1316>
- Danni, R., & Wahyuni, A. (2021). ITEM RESPONSE THEORY APPROACH : KALIBRASI BUTIR SOAL. *Arabi: Journal of Arabic Studies*, 6(1), 93–104. <https://doi.org/https://doi.org/10.24865/ajas.v6i1.320>
- Elmananda, Y. R., Baedhowi, S., & Arisyanto, P. (2024). Analisis Kompetensi Pedagogik dan Kompetensi Profesional Guru dalam Meningkatkan Hasil Belajar Mata Pelajaran IPS Siswa Kelas VI SDN Gaplokan Kecamatan Japah Kabupaten Blora. *INNOVATIVE: Journal Of Social Science Research*, 4(4), 2726–2740. <https://doi.org/https://doi.org/10.31004/innovative.v4i4.13390>
- Endah Pratiwi, B. S. (2025). ANALISIS VALIDITAS BUTIR DAN KETEPATAN TARGET UKUR INSTRUMEN BIOLOGI KELAS XII: APLIKASI PEMODELAN RASCH. *Jurnal Pendidikan Biologi*, 16, 155–167. <https://doi.org/http://dx.doi.org/10.17977/um052v16i2p155-168>
- Fahmi, B. dkk. (2022). Analisis Butir Soal Bahasa Arab MAS Pondok Pesantren Assalam Kampar Riau. *Ta'lim al-'Arabiyyah: Jurnal Pendidikan Bahasa Arab & Kebahasaaraban*, 6(1), 95–105. <https://doi.org/10.15575/jpba.v6i1.16193>
- Fauzi, A., & Rahmawati, S. (2025). Evaluation of Speaking Competence in Arabic Language Learning: A Performance Assessment Approach. *INTERNATIONAL JOURNAL OF ARABIC STUDIES*, 2(1), 12–22.
- Fitraynsyah, A., & Hilmiyati, F. (2024). Peran Tingkat Kesukaran dan Daya Pembeda dalam Analisis Butir Tes: Kajian Literatur untuk Pendidikan Menengah. *Jurnal Riset Evaluasi Pendidikan*, 1(4), 252–262. <https://doi.org/https://doi.org/10.51574/jrep.v1i4.2250>
- Hadiyanti, H., & Susongko, P. (2024). Pengembangan Instrumen Higher Order Thinking Skill Mata Pelajaran Matematika Dengan Rasch Model. *Journal of Education Research*, 5, 399–407. <https://doi.org/https://doi.org/10.37985/jer.v5i1.765>
- Ibrahim, A. M. (2024). Studi tentang Pengaruh Kompetensi Pedagogik dan Motivasi Guru terhadap Hasil Belajar Siswa di lingkungan pendidikan Journal of Holistic Education. *Journal of Holistic Education*, 1(1), 19–38.
- Ibrahim, N. L. (2025). Analisis Literatur Penggunaan Model Rasch untuk Validasi Instrumen Asesmen Pendidikan: Tren , Temuan , dan Implikasi. *Lencana: Jurnal Inovasi Ilmu Pendidikan*, 3(2), 46–60. <https://doi.org/https://doi.org/10.55606/lencana.v3i2.5012>
- Ida, F. F., & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. *Al-Mu'Arrib: Journal of Arabic Education*, 1(1), 34–44. <https://doi.org/10.32923/al-muarrib.v1i1.2100>

-
- Laili, M. (2020). Ketepatan Kontruksi Butir Pilihan Ganda Bahasa Arab. *ALSUNIYAT: Jurnal Penelitian Bahasa, Sastra, Dan Budaya Arab*, 3(2), 111–124. <https://doi.org/https://doi.org/10.17509/alsuniyat.v3i2.25272>
- Latifah, H., & Suwandi, M. A. (2024). Innovative Assessment Methods: Enhancing Student Learning through Formative Assessment and Peer Review. *Innovative Journal of Educational Research and Insights*, 1(1), 40–50.
- Maghfirotnunnisa, A. dkk. (2023). Validitas Isi dan Reliabilitas Instrumen Penilaian Harian Mata Pelajaran Bahasa Arab MI Manarul Islam Malang. *Shaut al Arabiyyah*, 11(1), 19–28. <https://doi.org/10.24252/saa.v11i1.34368>
- Maria, A., Larantukan, F., & Yudiarto, A. (2025). Validitas Alat Ukur Zarit Burden Interview (ZBI) Menggunakan Model Rasch pada Populasi Indonesia. *Humanitas (Jurnal Psikologi)*, 9(1), 18–40. <https://doi.org/https://doi.org/10.28932/humanitas.v9i1.11007>
- Novriyanti, E., & Arthur, R. (2024). Analisis Kualitas Butir Soal Ujian Tengah Semester Biologi Umum Menggunakan Model Rasch. *JagoMIPA: Jurnal Pendidikan Matematika Dan IPA*, 4(4), 718–733. <https://doi.org/https://doi.org/10.53299/jagomipa.v4i4.927>
- Patonah, I., Sambella, M., & Az-Zahra, S. M. (2023). Pendekatan Penelitian Pendidikan: Penelitian Kualitatif, Kuantitatif Dan Kombinasi (Mix Method). *Pendas : Jurnal Ilmiah Pendidikan Dasar*, 08(1989), 5378–5392. <https://doi.org/https://doi.org/10.23969/jp.v8i3.11671>
- Pratama, D. (2020). Analisis Kualitas Tes Buatan Guru Melalui Pendekatan Item Response Theory (IRT) Model Rasch. *Tarbawy : Jurnal Pendidikan Islam*, 7(1), 61–70. <https://doi.org/https://doi.org/10.32923/tarbawy.v7i1.1187>
- Rahmadon, & Oktarina, M. (2025). PENGEMBANGAN INSTRUMEN PENILAIAN PEMBELAJARAN BAHASA ARAB. *ISLAMICA: Journal Of Islamic Education Research*, 1(2), 1–8.
- Ramadhan, M. F. dkk. (2024). Validitas and Reliabilitas. *Journal on Education*, 6(2), 10967–10975. <https://doi.org/10.31004/joe.v6i2.4885>
- Riskawati, Diana, N., Rahman, M., Tonang, A. P., & Barumbun, M. (2025). CJPE : Cokroaminoto Juornal of Primary Education Analisis Rasch Terhadap Pemahaman Guru Sekolah Dasar terhadap Model Pembelajaran Kooperatif Pendahuluan Pembelajaran abad ke-21 menuntut peserta didik memiliki keterampilan berpikir. *CJPE: Cokroaminoto Juornal of Primary Education*, 8, 1954–1970. <https://doi.org/https://doi.org/10.30605/cjpe.8.4.2025.7391>
- Saputra, A. (2025). Pengembangan Instrumen Evaluasi (Analisis Pengembangan Penilaian Satuan Pendidikan MTs / SMP). *Ar-Raudah: Jurnal Pendidikan Dan Keagamaan*, 2(4), 1–14. <https://doi.org/https://doi.org/10.61891/ar-raudah.v2i4.651>
- Sri Astuti Iriyani, Elyakim N.S Patty, V. L. L. (2025). Jurnal Ilmiah Pendidikan Citra Bakti MAHASISWA BERBASIS PEMODELAN RASCH. *Jurnal Ilmiah Pendidikan Citra Bakti*, 12, 940–952. <https://doi.org/https://doi.org/10.38048/jipcb.v12i3.5469>
- Stemler, S. E., & Naples, A. (2021). Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line. *Practical Assessment, Research, and Evaluation*, 26(11). <https://doi.org/https://doi.org/10.7275/v2gd-4441>
- Suciainsani, Okky, Euis Eti Roeshaeti, W. Y. (2021). VALIDITAS DAN RELIABILITAS ANGKET KEDISIPLINAN SISWA. *Fokus*, 4(5), 401–406. <https://doi.org/10.22460/fokus.v7i4.7410>
- Utama, A., & Nawas, A. (2026). Measuring Critical Thinking in Physics: A Rasch Analysis of Instrument Quality and Gender Equivalence. *Jurnal Pendidikan MIPA*, 27(1). <https://doi.org/http://doi.org/10.23960/jpmpipa>
-

Zuhri, N. Z. (2024). Analisis Validitas , Reliabilitas , dan Tingkat Kesukaran Soal Bahasa Arab Tingkat SMP Berbasis Artificial Intelligence (AI) melalui Platform QuestionWell. *Jurnal Pendidikan Dan Pembelajaran Indonesia (JPPI)*, 4, 693–704. <https://doi.org/https://doi.org/10.53299/jppi.v4i2.576>

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest.

