Dwija Cendekia: Jurnal Riset Pedagogik
Volume 9 (3), 2025, pp. 433-448

Transformative learning based on numeracy assessment in Islamic boarding schools: Item response theory

Rosid Bahar [a] *, Syafrudin Syafrudin [b]

Sekolah Tinggi Agama Islam Idrisiyyah. Cisayong, Tasikmalaya, Jawa Barat 46153, Indonesia
[a] rosidbahar@idrisiyyah.ac.id; [b] syafrudin@idrisiyyah.ac.id
* Corresponding Author

Receipt: 14 November 2025; Revision: 4 December 2025; Accepted: 16 December 2025

Abstract: Numeracy skills are a key competency in addressing 21st-century challenges, including in religious-based educational environments such as Islamic boarding schools (pesantren). This study aims to measure students' numeracy skills through a transformative learning approach based on numeracy assessment. The research employs a descriptive quantitative method involving 383 students from four Islamic boarding schools located in different regencies/cities in West Java Province: Purwakarta, Tasikmalaya, and Sukabumi. The assessment instrument was developed based on the Minimum Competency Assessment (AKM) framework and underwent content validity testing involving experts, construct validity testing using Exploratory Factor Analysis (EFA), and reliability testing with the Item Response Theory (IRT) approach. The analysis results indicate that the Rasch (1PL) model is the most suitable for measuring students' numeracy skills, with item difficulty as the primary indicator. The ability distribution shows that 21.67% of students are in the advanced category, 27.68% proficient, 27.42% basic, and 23.24% require special intervention. These findings suggest that while some students already possess strong numeracy skills, learning reinforcement remains necessary. Overall, this study confirms that numeracy assessment-based transformative learning is highly feasible to implement in pesantren environments, as it offers a contextual, equitable, and data-driven learning approach.

Keywords: Transformative Learning; Islamic Boarding School; Numeracy; Minimum Competency Assessment; Item Response Theory

INTRODUCTION

Islamic boarding schools (pesantren) have long served as the foundation of Islamic education in Indonesia, functioning not only as centers for religious instruction but also as catalysts for community development and cultural preservation. The enactment of Law No. 18 of 2019 concerning Pesantren marked a significant milestone in recognizing and formalizing the role of pesantren in the national education system (Dewi & Wajdi, 2023; Khoirurrijal et al., 2023). This law mandates that pesantren also teach general subjects such as Mathematics, Indonesian Language, and Natural Sciences, highlighting the need for adaptation and enhanced teaching quality (Nuha et al., 2024).

A strategic step to address this challenge is strengthening mathematics education, particularly in numeracy. Numeracy plays a crucial role in preparing individuals for the digital era of Society 5.0 while enabling the integration of scientific knowledge and religious values within the pesantren context (Estrada-Mejia et al., 2016). Numeracy skills support cross-curricular learning and contribute to lifelong learning and quality

DOI:10.20961/jdc.v9i3.111029

of life (Goos et al., 2024). For students (santri), developing numeracy skills must continue throughout adulthood, especially for those with diverse learning needs, ensuring they can navigate the complexities of modern life effectively (Tout, 2020). This approach requires careful attention to factors such as learning duration, literacy and numeracy skills, and teacher competencies (Zaqiah et al., 2024)

However, numeracy learning in pesantren faces significant challenges, including limited integration with transformative learning and pedagogical constraints. Transformative education, which emphasizes contextual and collaborative understanding, is rarely implemented in pesantren learning systems (Santos & Pechliye, 2023; V. X. Wang, 2018). Numeracy learning teaches mathematical knowledge in real-life contexts, fosters critical thinking, and employs diverse tools for problem-solving. Cross-curricular and innovative approaches to numeracy instruction can enhance student skills and prepare them for future challenges (Díez-Palomar et al., 2023). Thus, this study integrates mathematics with religious knowledge, such as inheritance laws in fiqh, measurements of religious symbols like the Ka'bah, and other relevant contexts.

Furthermore, the pesantren environment, which often limits parental involvement, poses additional barriers to numeracy skill development. Research indicates that home-based activities with parents significantly shape children's mathematical skills in school (Gashaj et al., 2023). Other challenges include epistemological obstacles, such as difficulties in understanding mathematical notation and communication, as well as limited access to educational technology, despite its critical role in 21st-century learning (Us et al., 2023). Nevertheless, many pesantren have begun integrating technology into their curricula, producing students with high digital competencies. Therefore, a specialized assessment of students' mathematical abilities, particularly in numeracy and its analysis, is essential. This assessment, for example, comes from a mathematics test instrument whose questions are in the form of an integration of Islamic boarding school knowledge that is included in mathematics education, such as inheritance, zakat, or daily life in Islamic boarding schools.

Currently, many assessment reports are limited in scope (e.g., covering only one class) and rely on local needs or classical test theory (CTT). CTT has limitations in validity, reliability, and generalizability (Meguellati et al., 2024), and does not account for item characteristics, making it suboptimal for measuring individual student abilities (Amusa et al., 2022). To address these limitations, this study proposes a more comprehensive approach using Item Response Theory (IRT). IRT provides more accurate measurements by considering item parameters such as difficulty, discrimination, and guessing probability, estimated separately from test-taker abilities (Lim & Wells, 2022). IRT has been effectively applied in educational assessment, enabling precise estimation of latent traits and reducing response burden (Kean et al., 2018)(Thomas, 2011). Thus, this study aims to comprehensively measure the numeracy skills of santri in West Java pesantren by applying modern IRT-based measurement principles aligned with transformative learning and 21st-century educational challenges. Based on these results, it is necessary to develop a mapping of the numeracy abilities of Islamic boarding school students based on IRT in the context of West Java Islamic boarding school education, which has not received attention in the literature or national assessment practices.

METHODS

This study uses a quantitative descriptive approach to describe and analyze the numeracy skills of students in Islamic boarding schools (pesantren) in West Java. The study locations included MTsS Cipulus (Purwakarta), SMP Islam Terpadu Yaspida (Sukabumi), and SMP Terpadu Al-Amin (Tasikmalaya), with a total of 383 students participating. These three institutions were selected purposively because they represent the varied characteristics of modern and salafiyah Islamic boarding schools in West Java, offer junior high school/Islamic junior high school (SMP/MTs) education levels aligned with the focus of numeracy assessment, and demonstrate readiness to implement standardized test-based studies. This selection also took into account affordability, research permits, and the institutions' commitment to supporting the assessment of students' numeracy skills.

Quantitative research was chosen because of its systematic handling of numerical data, allowing for objective measurement and analysis of variables using statistical methods through R Studio. This software is widely used in academics due to its robust statistical capabilities, extensive package ecosystem, open-source nature, and flexibility in handling complex analyses (Buchholz, 2021; England, 2021; Sosa et al., 2010; Tang & Ji, 2014; Velec & Huang, 2014). The selected Islamic boarding schools reflect a variety of curricula and learning approaches, thus providing a more comprehensive picture of the students' numeracy skills. Purposive sampling was used to ensure that participants met criteria relevant to the research focus, ensuring that the data obtained aligns with the desired analysis objectives.

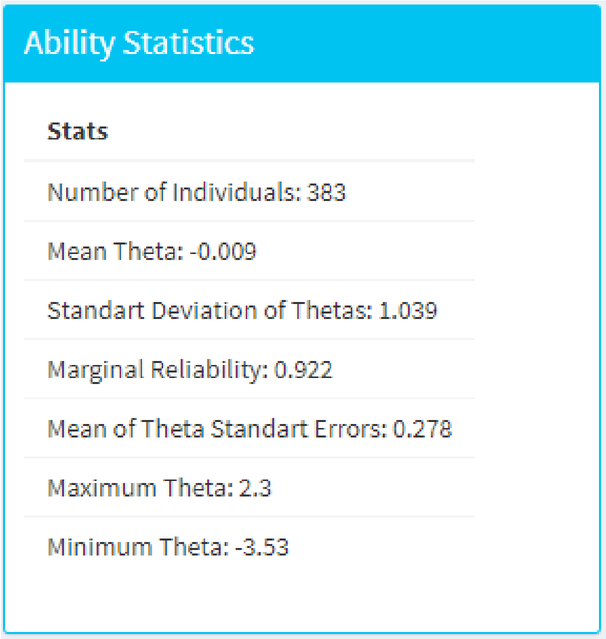
Data collection was conducted through a numeracy test comprising including 2 matching questions, 9 multiple choice questions, 6 complex multiple-choice questions, 6 short essay questions, and 2 essay questions. Content validity through expert judgment and construct validity using dimensionality analysis and model suitability, while reliability is assessed using marginal reliability with the IRT approach. The instrument was designed to measure various aspects of numeracy understanding, based on constructs encompassing content, context, and cognitive processes (Rodríguez et al., 2024). Data analysis was conducted using Item Response Theory (IRT) with R Studio using the IRTGUI package (Yildiz, 2021), which allows for more accurate and objective measurement of numeracy ability than classical approaches. IRT was chosen for its ability to reliably evaluate item characteristics and individual abilities reliably (Mayer & Bryan, 2024). In this study, Model 1 PL was used based on recommendations obtained from R Studio output, thus only measuring the level of difficulty. This study indicates that the level of difficulty provides in-depth insights for developing numeracy assessments that align with the learning needs of students in Islamic boarding schools (Zhu et al., 2025).

RESULTS AND DISCUSSION

Results

The first step in the analysis was testing instrument validity and reliability. Validity was assessed through content and construct validity (Roebianto et al., 2023; Sugiarta et al., 2024). Content validity, evaluated via expert judgment using the Content Validity Index (CVI), yielded scores between 0.8 and 0.9 per item, indicating strong validity (F. Wang & Sahid, 2024). Construct validity, analyzed using Exploratory Factor Analysis (EFA) in SPSS, showed all items had factor loadings above 0.3, confirming their alignment with the numeracy construct (Hair et al., 2010).

The second step was reliability testing. Reliability was estimated using R Studio with the IRTGUI package (Yildiz, 2021). The results are presented in Figure 1.



Ability Statistics

Stats

Number of Individuals: 383

Mean Theta: -0.009

Standart Deviation of Thetas: 1.039

Marginal Reliability: 0.922

Mean of Theta Standart Errors: 0.278

Maximum Theta: 2.3

Minimum Theta: -3.53

Figure 1. Reliability Estimation

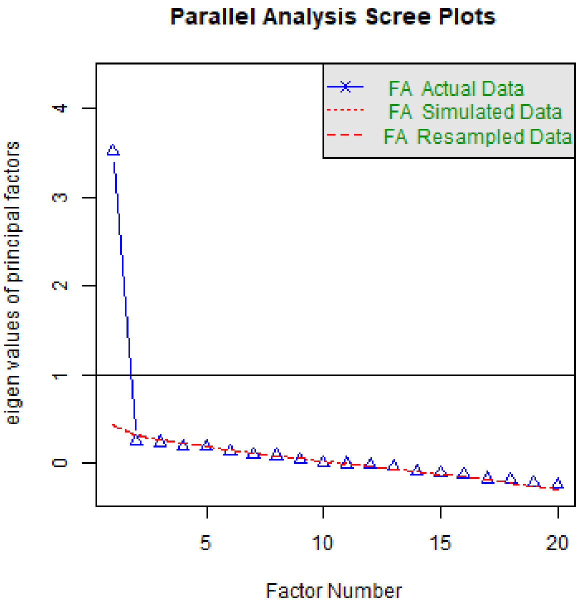


Figure 2. Scree Plots

Table 1. Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	6.467	25.867	25.867
2	2.552	10.209	36.076
3	1.604	6.417	42.493
4	1.420	5.679	48.172
5	1.394	5.576	53.747
6	1.136	4.545	58.293
7	1.081	4.323	62.616

The instrument's reliability coefficient shows a very good level of consistency, with a coefficient of 0.922 indicating high reliability, so the instrument is considered suitable for measuring the constructs being studied consistently and accurately(Ellis, 2013; Peterson & Kim, 2013).

The third step involved numeracy skill testing, preceded by assumptions testing for unidimensionality and local independence (Abal et al., 2023) The verification of unidimensionality assumption can actually be conducted simultaneously with construct validity using factor analysis (EFA) in SPSS, with the following outputs: Total Variance Explained (Table 1) and Scree Plot (Figure 2).

Unidimensionality was confirmed via EFA in SPSS, with the first factor's eigenvalue of 6.467 explaining 25.867% of the variance, dominant over other factors (Nima et al., 2020). The scree plot (Figure 2) further supported unidimensionality.

The results of the unidimensionality test facilitate the subsequent assumption testing for local independence, as the local independence assumption is automatically satisfied when the unidimensionality assumption is met (Santoso et al., 2024).

Based on the assumption test results, the numeracy ability analysis can appropriately be conducted using Item Response Theory analysis. Subsequently, to assess the students' numeracy skills, a model fit analysis was first performed for each instrument item. This step is crucial to determine the most suitable model for measuring the students' abilities. The model fit test results from the R Studio program are presented in Table 2.

Table 2. Fit Indexes

Model	AIC	BIC	loglikelihood
Rasch	11874.51	11963.02	-5916.26
2PLM	11903.31	12071.90	-5911.66
3PLM	11936.29	12189.17	-5908.15

Based on Table 2, the lowest values are observed for the Rasch model or 1PL (One-Parameter Logistic) model (Desjardins & Bulut, 2018). This indicates that the instrument can only measure students' numeracy proficiency through item difficulty parameters alone. Furthermore, among the 25 items analyzed using the 1PL/Rasch model, 18 items demonstrated good characteristics, as evidenced by their logit values falling within the acceptable difficulty range of -2.00 to 2.00 (Demars, 2018). This demonstrates that the majority of test items exhibit proportional difficulty levels and are suitable for effectively assessing students' numeracy competencies.

All prerequisite aspects for conducting the numeracy ability analysis have been satisfied. Subsequently, the analysis of students' numeracy competencies will employ the Maximum Likelihood Estimation (MLE) method. This method is widely used for estimating θ (theta) and has been proven to be asymptotically normally distributed under specific conditions (Sinharay, 2015). The theta (θ) values can be obtained from the R Studio program output, yielding a logit scale range of -3.53 to 2.30 for the 383 students. Using Microsoft Excel, these logit scales were converted to a 1-100 scale through direct proportion transformation, with the conversion results presented in **Table 3.**

Table 3. Conversion Result

Student	Competency (Theta)	Convertion of Conpetency to a 0-100 scale
1	-2,45	38
2	-3,18	34
3	-2,45	38
...	....	....
383	4,8645	74

The results will be categorized into four categories: Proficient, Competent, Basic, and Requires Special Intervention (Ministry of Education and Culture, 2021).

Discussion

The content validity of this study was established through expert evaluations involving four specialists in the fields of Islamic Studies, Indonesian Language, and Mathematics. The Content Validity Index (CVI) for each item ranged from 0.8 to 0.9, indicating strong validity (F. Wang & Sahid, 2024; Yao et al., 2008). The analysis proceeded with construct validity testing using Exploratory Factor Analysis (EFA). The results revealed that all items exhibited factor loadings above 0.30, indicating sufficiently strong relationships between the items and the numeracy construct (Trendafilov & Hirose, 2022; Zizler & Ittyipe, 2023). The determination of factor quantity through eigenvalue analysis and scree plot verification ensured the resulting numeracy dimension structure demonstrated both stability and theoretical coherence (Lloret-Segura et al., 2014).

The reliability test was conducted using an Item Response Theory (IRT) approach with *R Studio* and IRTGUI, yielding a reliability coefficient of 0.922, which is classified as very high (Mohd Salleh et al., 2023; Na et al., 2024). These results indicate that the instrument possesses strong and stable internal consistency for use across various *Pondok Pesantren* (Islamic boarding schools), enabling equitable measurement of students' numeracy skills regardless of regional background (Iqbal & Ul Islam, 2024). High validity and reliability serve as a critical foundation to ensure data quality in quantitative research (Horner et al., 2004; Nunnally, 1975).

Furthermore, to strengthen the analysis, the instrument was further examined using IRT. The results indicate that the Rasch model (1PL) was the most suitable compared to the 2PL and 3PL models, as evidenced by the lowest AIC, BIC, and log-likelihood values (Sen & Cohen, 2024). This model accommodates linear measurement based on item difficulty levels and does not rely on item discrimination parameters, making it ideal for the heterogeneous context of *pesantren* (Islamic boarding schools) (Andrich, 1988). Of the 25 items, 18 fell within the difficulty range of -2.00 to 2.00 logits, which is considered optimal for proportionally measuring numeracy skills. The remaining items serve as a foundation for further development, confirming that this instrument is suitable for designing numeracy learning interventions for santri (Islamic boarding school students).

A thorough examination of students' numeracy skills requires a comprehensive understanding of the assessment framework used as the measurement instrument. In this context, the Asesmen Kompetensi Minimum (AKM) or Minimum Competency Assessment, developed by Indonesia's Ministry of Education, Culture, Research, and Technology, serves as the primary reference for capturing students' numeracy proficiency. The AKM emphasizes that numeracy extends beyond mere computational ability; rather, it represents applied mathematical thinking skills relevant to daily life. To achieve this, the AKM incorporates three assessment dimensions: content, context, and cognitive processes. These dimensions not only structure the test items but also reflect the complexity of thinking required to cultivate mathematically literate citizens (Purnomo et al., 2022). Furthermore, AKM's emphasis on contextual questions and high-level problem-solving aligns with the principles of Transformative Learning, which encourages students to critically reinterpret real-world situations. In particular, numeracy questions that require reflective reasoning, such as interpreting data in new contexts or solving non-routine problems, are often the most challenging for students. This highlights the need to strengthen Transformative-based learning that fosters critical reflection, strategy adjustment, and the ability to envision alternative solutions.

The content dimension of numeracy assessment encompasses four core domains: number, measurement and geometry, data and uncertainty, and algebra. The context dimension in AKM's numeracy assessment covers three primary categories: personal, sociocultural, and scientific, designed to reflect real-life situations where students apply mathematics (Organization for Economic Co-operation and Development, 2019). Meanwhile, the cognitive dimension in AKM evaluates how students understand, apply, and reason when solving mathematical problems (Siswaningsih et al., 2023). In this study, the content and context dimensions will be adapted to incorporate religious and pesantren (Islamic boarding school) settings, reflecting the distinctive daily life of santri. Consequently, the content and context are specifically framed within a religiously grounded environment—the *pesantren*. The integration of religious content and *pesantren* contexts provides a unique landscape for numeracy application. Examples include calculating inheritance shares (*ilmu warits*) through fractional number concepts, determining the Ka'bah's area using geometric principles, and similar activities. This approach enhances the meaningfulness of numeracy assessments, as they are rooted in students' routines—steeped in religious values and the distinctive communal life of *pesantren*.

These three components were then analyzed, forming the basis of the students' competencies. The initial analysis began with the content component presented in Figure 3.

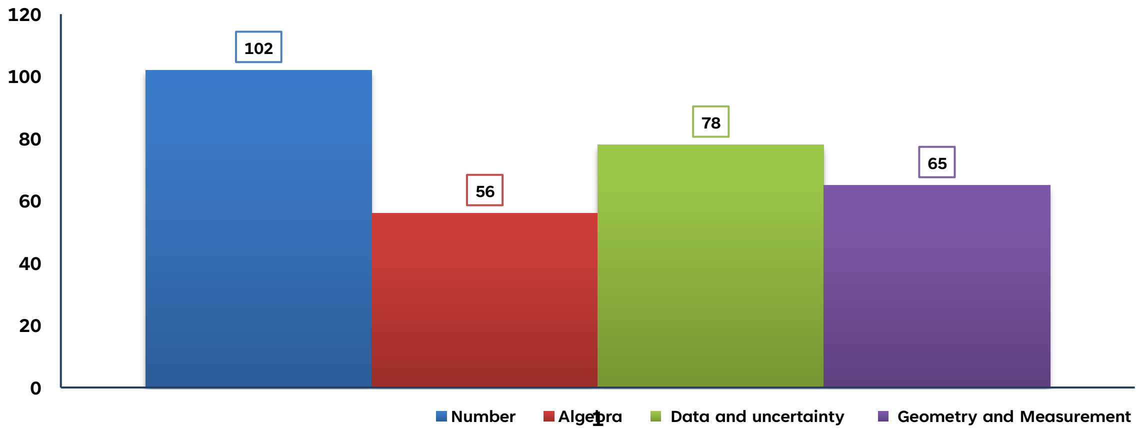


Figure 3. Numeracy Skill Distribution by Content Component

Analysis of 383 students revealed dominant mastery in number concepts (102 students), while algebra showed the lowest performance (56 students). Geometry (65 students) and data and uncertainty (78 students) demonstrated intermediate proficiency. This indicates an uneven skill distribution. These findings contrast with prior studies identifying geometry as the weakest area (Andriatna et al., 2024). Subsequent Analysis: Context Component with Core Indicators (Personal, Socio-Cultural, and Scientific). The analytical results are presented in Figure 4.

Analysis of 383 students revealed that the majority demonstrated stronger comprehension of numeracy problems with religious contexts. Specifically: 116 students successfully solved problems rooted in personal religious experiences, 149 students understood problems with pesantren-based socio-cultural contexts, Only 36 students showed proficiency in scientific-context numeracy problems.

These findings suggest that religious context integration in assessments aligns better with pesantren students' backgrounds compared to scientific contexts, which remain challenging for most students. The cognitive dimension encompassing comprehension, application, and reasoning is presented in Figure 5.

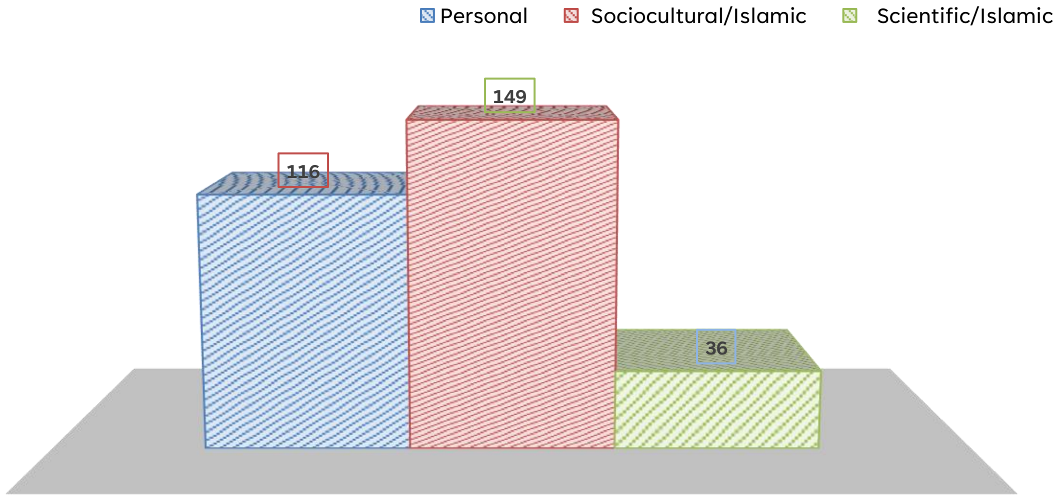


Figure 4. Numeracy Skill Distribution by Context Component

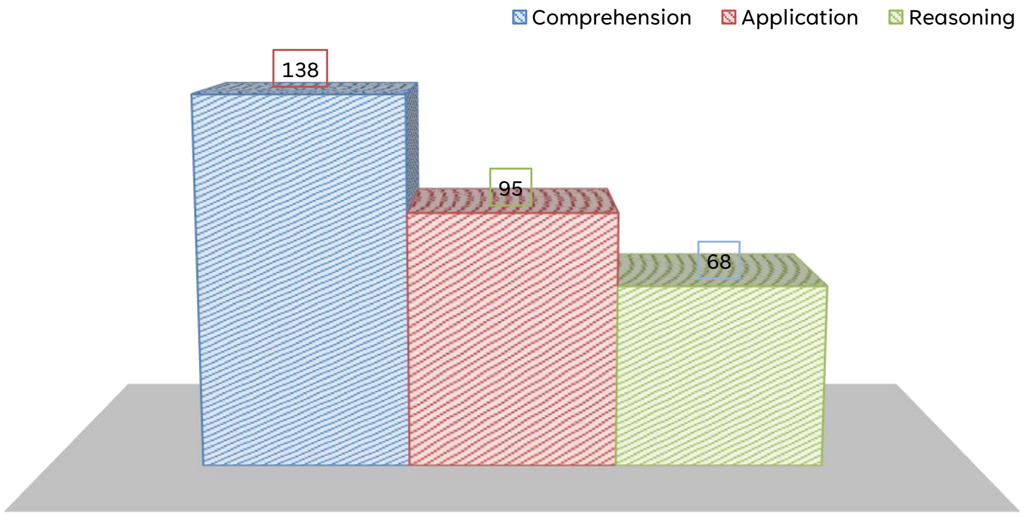


Figure 5. Numeracy Skill Distribution by Cognitive Component

Analysis of student’s cognitive abilities revealed distribution across three thinking levels. Among 383 students: 138 students demonstrated mastery at the comprehension level (highest category), 95 students could apply numeracy concepts in problem-solving, only 68 students reached the reasoning level. These findings indicate that most students remain at basic-to-intermediate mathematical thinking levels, highlighting the need for strengthened instructional strategies to advance higher-order thinking skills, particularly in numeracy reasoning.

This analysis evaluates santri numeracy skills using individual ability parameters derived from the Rasch measurement model within Item Response Theory (IRT). Based on R Studio output, the santri ability estimates—represented as theta (θ) values—

range from -3.53 to 2.30 logits across the 383 participants. Theta (θ) values serve as the core component in IRT modeling, designed to examine the probabilistic relationship between an individual's latent ability and their item responses. This model provides more precise ability estimates by simultaneously accounting for item difficulty parameters and guessing tendencies (Haberman, 2007; Sideridis & Alahmadi, 2022). *Santri* with positive theta values demonstrate above-average proficiency, while negative values indicate below-average performance. This logit range reflects considerable diversity in numeracy skills among *santri*, therefore, the representative approaches is needed to streamline interpretation.

To facilitate data presentation and interpretation for diverse educational stakeholders—including teachers, madrasah administrators, and policy researchers—the theta logit scale was converted to a 1–100 scale using a proportional comparison method implemented in Microsoft Excel (Ekstrand et al., 2022). This conversion was implemented to make santri ability measurements more communicable and directly comparable to conventional test scores. The transformation process preserves the ordinal nature of theta data, where inter-score differences continue to reflect proportional ability intervals from the original scale. Consequently, interpretations of santri numeracy outcomes become more intuitive while retaining the statistical integrity of the IRT model employed (Tennant & Conaghan, 2007). The 0 to 100 scale represents the standard grading system used in Indonesian schools and madrasahs, making it familiar to educators, students, and administrators in Indonesia.

The required santri competency results are systematically categorized into four proficiency levels: proficient, competent, basic, and requires special intervention (Ministry of Education and Culture, 2021). The competency results are presented in Figure 6.

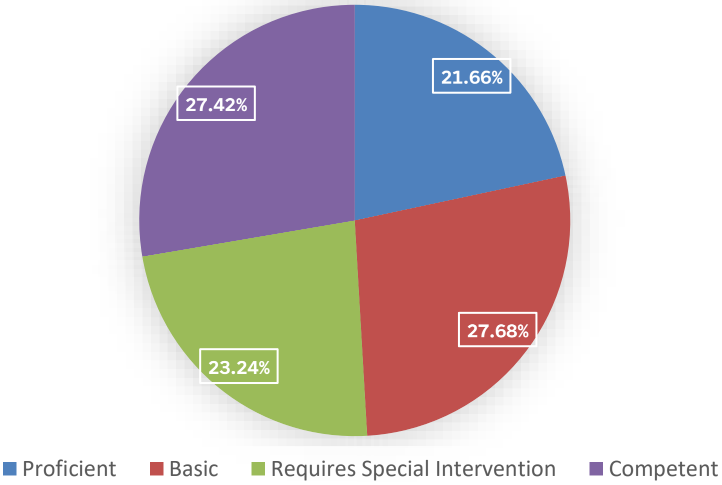


Figure 6. Numeracy Proficiency Categories

The results demonstrate that the santri exhibit sufficient numeracy competency, as evidenced by: Only 23.24% classified as requiring special intervention, and 76.76% achieving proficiency levels ranging from Basic to Advanced. These results indicate that a substantial majority of santri have achieved adequate numeracy proficiency, while a distinct subgroup requires targeted instructional attention.

The proportion of santri requiring special intervention (23.24%) is notably lower than percentages reported in comparable numeracy studies across Indonesia. For example, A regional study of 150 junior high school students in East Java revealed a markedly different proficiency distribution: 62% of them are at Surface-Level Numeracy Proficiency, 30% are at intermediate level, and only 8% reached advanced level (Sa'dijah et al., 2023).

This success can be attributed to the transformative learning approach based on numeracy in a contextual form implemented in *pesantren*, which emphasizes conceptual understanding through practice or direct comprehension, integrated with religious values. (Selvianiresa & Prabawanto, 2017) stated that integrating everyday situations into mathematics learning can help students connect new concepts with their prior experiences. When students understand that their knowledge is relevant and applicable in real-life situations, learning becomes more engaging and fosters intrinsic motivation through a contextual approach (Puglisi & Domènech-Gil, 2023).

The results of this study also serve as a new impetus and represent an interesting trend at present, considering that the PISA research results indicate that Indonesia has improved its ranking in mathematical literacy skills. Indonesia's position in 2022 rose by 5 to 6 ranks compared to 2018, reflecting a positive improvement among the participating PISA countries (Ariyanti et al., 2024). The results of this study also indicate that the numeracy learning approach implemented not only emphasizes cognitive aspects but also builds connections between mathematical knowledge and the life experiences of students. This aligns with the principles of transformative learning based on experience and reflection, as well as being oriented towards assessment (Halupa, 2015).

The transformation of learning through numeracy assessment offers a promising breakthrough in enhancing the quality of education in *pesantren*. This approach shifts the function of assessment from merely a measurement tool to an essential part of the learning process that fosters reflection, active participation, and deep understanding. This study demonstrates that transformative assessment can enhance students' understanding of numeracy concepts in a contextual and meaningful way. These findings align with the perspective of Diale et al., (2023) and Hunde et al., (2025), which states that such assessments can revolutionize teaching practices by fostering creativity and flexibility in learning.

From the perspective of instrument quality, the analysis results indicate that the numeracy assessment instruments used in this study have met the criteria for validity and reliability. By employing the one-parameter logistic Rasch model, the instrument has proven capable of measuring numeracy skills objectively and consistently. The Rasch model was chosen for its ability to estimate individual abilities fairly, without being influenced by the characteristics of items that are too easy or too difficult. In other words, each student is assessed based on their actual abilities, not merely on their score. This approach is also in line with the perspective of Bond (2015), which states that the Rasch model allows for the creation of linear instruments and supports meaningful interpretation of results within the context of learning. Furthermore, the use of item response theory such as Rasch is considered highly relevant in the context of transformative assessment because it facilitates fair and adaptive data-driven evaluations (Boone et al., 2014).

In a global context, various transformative learning approaches have been developed to enhance the effectiveness of assessment, one of which is Lesson Study for Transformative Assessment (LSforTA). Although this approach was not directly utilized in our research, LSforTA serves as a relevant example of how assessment practices can be improved through teacher collaboration and continuous reflection. Hunde et al., (2025) demonstrates that the implementation of LSforTA has a positive impact on teachers' teaching practices, enhances student engagement, and improves the quality of assessments used. These findings reinforce the argument that the success of transformative assessment is significantly determined by teachers' openness to innovation and their reflective abilities in responding to students' learning needs. In our research, similar principles—namely reflection, engagement, and meaning-making—also serve as the foundation for designing contextual and empowering numeracy assessment processes for students.

Overall, this research provides strong evidence that transformative learning based on numeracy assessment is an effective approach to developing students' numeracy skills. By positioning assessment as an integral part of the learning process rather than merely an endpoint, students are encouraged to engage in reflective and meaningful learning. This success demonstrates that the transformation of learning is not just a discourse, but can be realized through strategic and transformative assessment design. These findings serve as an important foundation for designing curricula and teaching strategies in Islamic boarding schools that are relevant to the demands of 21st-century numeracy.

CONCLUSION

This research concludes that the assessment process of students' numeracy skills has been conducted using valid and reliable instruments, in accordance with the standards for developing quantitative measurement tools. Through analysis based on Item Response Theory (IRT), it was found that the Rasch model (1 Parameter Logistic/1PL) is the most suitable model for measuring students' numeracy skills, as it fairly and linearly considers the difficulty level of the test items. The distribution of students' numeracy skills indicates that 21.67% are categorized as proficient, 27.68% as competent, 27.42% as basic, and 23.24% require special intervention. These findings suggest that although a portion of the students possess good numeracy competencies, there remains a significant proportion that requires special attention in strengthening their numeracy skills.

Overall, the results of this study affirm that transformative learning based on numeracy assessment is highly applicable in the context of Islamic boarding schools. Structured assessments grounded in religious contexts are not only relevant but also capable of capturing students' competencies more authentically, while simultaneously promoting data-driven pedagogical improvements in pesantren.

ACKNOWLEDGMENT

The researcher would like to express his gratitude to the Directorate of Islamic Religious Higher Education (Diktis) of the Indonesian Ministry of Religion for fully funding this research through the Litapdimas 2025 program.

REFERENCES

Abal, F. J. P., Sánchez González, J. F., & Attorresi, H. F. (2023). Adaptation of the Bergen instagram addiction scale in Argentina: Calibration with item response theory. *Current Psychology*. https://doi.org/10.1007/s12144-023-04257-1

Amusa, J. O., Ayanwale, M. A., Oladejo, A. I., & Ayedun, F. (2022). Undergraduate Physics Test Dimensionality and Conditional Independence: Perspective from Latent Traits Model Package of R Language. *International Journal of Assessment* *and Evaluation*, *29*(2), 47–62. https://doi.org/10.18848/2327-7920/CGP/V29I02/47-61

Andriatna, R., Sujadi, I., Budiyono, Kurniawati, I., Wulandari, A. N., & Puteri, H. A. (2024). Junior high school students' numeracy in geometry and measurement content: Evidence from the minimum competency assessment result. *AIP Conference Proceedings*, *3046*(1), 020036. https://doi.org/10.1063/5.0194570

Andrich, D. (1988). *Rasch models for measurement: Sage publications*. Sage Publications.

Ariyanti, L., Hanurawan, F., & Ramli, M. (2024). Development of problem mind mapping-based learning model (PMM-BL) integration patterns to improve critical thinking skills in primary school students. *Edelweiss Applied Science and Technology*, *8*(6), 2905–2919. https://doi.org/10.55214/25768484.v8i6.2631

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences* (Vol. 10). Springer.

Buchholz, S. W. (2021). Quantitative designs for practice scholarship. In *Research for Advanced Practice Nurses, Fourth Edition: From Evidence to Practice* (pp. 143–172). https://doi.org/10.1891/9780826151339.0009

Demars, C. E. (2018). Classical test theory and item response theory. *The Wiley Handbook of Psychometric Testing*, *1–2*, 49–73. https://doi.org/10.1002/9781118489772.ch2

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.

Dewi, M. P., & Wajdi, M. B. N. (2023). Pesantren Laws as an Accelerator for Community Economic Development. *EDUTEC : Journal of Education And Technology*, *6*(4), 290–298. https://doi.org/10.29062/edu.v6i4.772

Diale, B. M., Victor-Aigbodion, V., & Eseadi, C. (2023). Transformative assessment in digitalised post-COVID-19 education: Implications for higher education teachers. In *Fostering Diversity and Inclusion Through Curriculum Transformation* (pp. 101–112). https://doi.org/10.4018/978-1-6684-6995-8.ch006

Díez-Palomar, J., Ramis-Salas, M., Močnik, I., Simonič, M., & Hoogland, K. (2023). Challenges for numeracy awareness in the 21st century: making visible the invisible. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1295781

Ekstrand, J., Westergren, A., Årestedt, K., Hellström, A., & Hagell, P. (2022). Transformation of Rasch model logits for enhanced interpretability. *BMC Medical Research Methodology*, *22*(1), 332. https://doi.org/10.1186/s12874-022-01816-1

Ellis, J. L. (2013). A standard for test reliability in group research. *Behavior Research Methods*, *45*(1), 16–24. https://doi.org/10.3758/s13428-012-0223-z

England, A. (2021). Quantitative and Qualitative Research Methods. In *Research for Medical Imaging and Radiation Sciences* (pp. 71–96). Springer International Publishing. https://doi.org/10.1007/978-3-030-79956-4_5

Estrada-Mejia, C., de Vries, M., & Zeelenberg, M. (2016). Numeracy and wealth. *Journal of Economic Psychology*, *54*, 53–63. https://doi.org/10.1016/j.joep.2016.02.011

Gashaj, V., Thaqi, Q., Mast, F. W., & Roebers, C. M. (2023). Foundations for future math achievement: Early numeracy, home learning environment, and the absence of math anxiety. *Trends in Neuroscience and Education*, *33*, 1–18. https://doi.org/10.1016/j.tine.2023.100217

Goos, M., Bennison, A., Forgasz, H., & Yasukawa, K. (2024). Research in Numeracy Education. In *Research in Mathematics Education in Australasia 2020–2023* (pp. 59–84). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-1964-8_4

Haberman, S. J. (2007). The information a test provides on an ability parameter. *ETS Research Report Series*, *2007*(1), i–16. https://doi.org/10.1002/j.2333-8504.2007.tb02060.x

Hair, J. F., Black, W. C., & Babin, B. J. (2010). *Multivariate data analysis: A global perspective* (7th ed.). Prentice Hall.

Halupa, C. M. (2015). Transformative learning: Theory and practice for faculty and students. In *Transformative Curriculum Design in Health Sciences Education* (pp. 1–39). https://doi.org/10.4018/978-1-4666-8571-0.ch001

Horner, Robert H, Todd, Anne W, Lewis-Palmer, Teri, Irvin, Larry K, Sugai, George, & Boland, Joseph B. (2004). The School-Wide Evaluation Tool (SET): A Research Instrument for Assessing School-Wide Positive Behavior Support. *Journal of Positive Behavior Interventions*, *6*(1), 3–12. https://doi.org/10.1177/10983007040060010201

Hunde, A. B., Abate, M. T., & Wedajo, A. L. (2025). Lesson Study as a Tool for Improving Teachers' Transformative Assessment Practices. *SAGE Open*, *15*(2). https://doi.org/10.1177/21582440251333483

Iqbal, J., & Ul Islam, T. (2024). Comparison of statistical models for individual's ability index and ranking. *Educational Research and Evaluation*, *29*(3–4), 171–185. https://doi.org/10.1080/13803611.2024.2325454

Kean, J., Bisson, E. F., Brodke, D. S., Biber, J., & Gross, P. H. (2018). An Introduction to Item Response Theory and Rasch Analysis: Application Using the Eating Assessment Tool (EAT-10). *Brain Impairment*, *19*(1), 91–102. https://doi.org/10.1017/BrImp.2017.31

Khoirurrijal, M. F., Karim, A. R., Zaini, A., & Salik, M. (2023). Pesantren and the Human Development Index in Indonesia Post Law Number 18 of 2019. *Santri: Journal of Pesantren and Fiqh Sosial*, *4*(1), 45–66. https://doi.org/10.35878/santri.v4i1.697

Lim, H., & Wells, C. S. (2022). [RETRACTED ARTICLE] irtplay : An R Package for Unidimensional Item Response Theory Modeling. *Journal of Statistical Software*, *103*(12). https://doi.org/10.18637/jss.v103.i12

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: una guía práctica, revisada y actualizada. *Anales de Psicología*, *30*(3), 1151–1169. https://doi.org/10.6018/analesps.30.3.199361

Mayer, J. D., & Bryan, V. M. (2024). On personality measures and their data: A classification of measurement approaches and their recommended uses. *Personality and Social Psychology Review*, *28*(3), 325–345.

Meguellati, S., Samia, A., Ferhat, A., Djelloul, A., & Khalifa, Z. A. (2024). A Critical Analysis of the Use of Classical Test Theory (CTT) in Psychological Testing: A Comparison with Item Response Theory (IRT). *Pakistan Journal of Life and Social Sciences*, *22*(2), 9442–9449. https://doi.org/10.57239/PJLSS-2024-22.2.00715

Ministry of Education and Culture. (2021). *Framework asesmen kompetensi minimum*.

Mohd Salleh, K., Sulaiman, N. L., & Gloeckner, G. (2023). Exploring test concept and measurement through validity and reliability process inTVET research: Guideline for the novice researcher. *Journal of Technical Education and Training*, *15*(1), 257–264. https://doi.org/10.30880/jtet.2023.15.01.022

Na, C., Clarke-Midura, J., Shumway, J., van Dijk, W., & Lee, V. R. (2024). Validating a performance assessment of computational thinking for early childhood using item response theory. *International Journal of Child-Computer Interaction*, *40*. https://doi.org/10.1016/j.ijcci.2024.100650

Nima, A. Al, Cloninger, K. M., Persson, B. N., Sikström, S., & Garcia, D. (2020). Validation of subjective well-being measures using item response theory. *Frontiers in Psychology*, *10*(January), 1–33. https://doi.org/10.3389/fpsyg.2019.03036

Nuha, M. F. A. U., Muklason, A., & Agustiawan, Y. (2024). Enhancing Administrative Efficiency in Pondok Pesantren: Exploring the Acceptance of E-Santren App System for Administrative Tasks. *Procedia Computer Science*, *234*, 795–804. https://doi.org/10.1016/j.procs.2024.03.096

Nunnally, J. C. (1975). Psychometric Theory— 25 Years Ago and Now. *Educational Researcher*, *4*(10), 7–21. https://doi.org/10.3102/0013189X004010007

Organization for Economic Co-operation and Development. (2019). *PISA 2018 assessment and analytical framework*. OECD. https://doi.org/10.1787/b25efab8-en

Peterson, R. A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology*, *98*(1), 194–198. https://doi.org/10.1037/a0030767

Puglisi, D., & Domènech-Gil, G. (2023). Enabling lifelong learning by using multiple engagement tools. *Proceedings of the International CDIO Conference*, 633–643. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85177062697&partnerID=40&md5=ec71d45908d5fc0a985c92a3a1f805fa

Purnomo, H., Sa'dijah, C., Hidayanto, E., Sisworo, Permadi, H., & Anwar, L. (2022). Development of Instrument Numeracy Skills Test of Minimum Competency Assessment (MCA) in Indonesia. *International Journal of Instruction*, *15*(3), 635–648. https://doi.org/10.29333/iji.2022.15335a

Rodríguez, S. P., van der Velden, R., Huijts, T., & Jacobs, B. (2024). Identifying literacy and numeracy skill mismatch in OECD countries using the job analysis method. *Oxford Economic Papers*, *76*(3), 859–876. https://doi.org/10.1093/oep/gpad045

Roebianto, A., Savitri, S. I., Aulia, I., Suciyana, A., & Mubarokah, L. (2023). Content validity: Definition and procedure of content validation in psychological research. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, *30*(1), 5–18. https://doi.org/10.4473/TPM30.1.1

Sa'dijah, C., Purnomo, H., Abdullah, A. H., Permadi, H., Anwar, L., Cahyowati, E. T. D.,

& Sa’diyah, M. (2023). Students’ numeracy skills in solving numeracy tasks: Analysis of students of junior high schools. *AIP Conference Proceedings*, *2569*, 040011. https://doi.org/10.1063/5.0113664

Santos, G. A., & Pechliye, M. M. (2023). Aprendizagem numa perspectiva reflexiva e transformadora em um curso de ciências biológicas: Evidências do processo. *CICIC 2023 - Decima Tercera Conferencia Iberoamericana de Complejidad, Informatica y Cibernetica En El Contexto de the 14th International Multi-Conference on Complexity, Informatics, and Cybernetics, IMCIC 2023 - Memorias*, 91–94. https://doi.org/10.54808/CICIC2023.01.91

Santoso, A., Retnawati, H., Pardede, T., Apino, E., Rafi, I., Rosyada, M. N., Kassymova, G. K., & Wenxin, X. (2024). From Investigating the Alignment of A Priori Item Characteristics Based on the CTT and Four-Parameter Logistic (4-PL) IRT Models to Further Exploring the Comparability of the Two Models. *Practical Assessment, Research and Evaluation*, *29*(14), 1–28. https://doi.org/10.7275/pare.2043

Selvianiresa, D., & Prabawanto, S. (2017). Contextual Teaching and Learning Approach of Mathematics in Primary Schools. *Journal of Physics: Conference Series*, *895*(1), 012171. https://doi.org/10.1088/1742-6596/895/1/012171

Sen, S., & Cohen, A. S. (2024). An Evaluation of Fit Indices Used in Model Selection of Dichotomous Mixture IRT Models. *Educational and Psychological Measurement*, *84*(3), 481–509. https://doi.org/10.1177/00131644231180529

Sideridis, G., & Alahmadi, M. (2022). Estimation of Person Ability under Rapid and Effortful Responding. *Journal of Intelligence*, *10*(3), 67. https://doi.org/10.3390/jintelligence10030067

Sinharay, S. (2015). The Asymptotic Distribution of Ability Estimates. *Journal of Educational and Behavioral Statistics*, *40*(5), 511–528. https://doi.org/10.3102/1076998615606115

Siswaningsih, W., Susetyo, B., Ariesta, A. S., & Rahmawati, T. (2023). Implementation of Minimum Competency Assessment (MCA) Containing Ethnoscience on the Topic of Electrolyte and Non-Electrolyte Solutions. *AIP Conference Proceedings*, *2642*. https://doi.org/10.1063/5.0113856

Sosa, M. M., García, M. R., & Piña, R. U. (2010). R: A not much spread and very useful tool for clinical research. *Revista Cubana de Investigaciones Biomedicas*, *29*(2). https://www.scopus.com/inward/record.uri?eid=2-s2.0-79851482129&partnerID=40&md5=e9345416c62ecc8515c5e3aaa4b849b0

Sugiarta, I. M., Ariawan, I. P. W., Ardana, I. M., Divayana, D. G. H., Sukawijana, I. K. G., & Sugiharni, G. A. D. (2024). Validity and reliability of the discrepancy evaluation instrument for measuring inequality in the online learning. *International Journal of Evaluation and Research in Education (IJERE)*, *13*(6), 3952. https://doi.org/10.11591/ijere.v13i6.28106

Tang, H., & Ji, P. (2014). Using the Statistical Program R Instead of SPSS To Analyze Data. In *ACS Symposium Series* (Vol. 1166, pp. 135–151). https://doi.org/10.1021/bk-2014-1166.ch008

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, *57*(8), 1358–

1362. https://doi.org/10.1002/art.23108

Thomas, M. L. (2011). The Value of Item Response Theory in Clinical Assessment: A Review. *Assessment*, *18*(3), 291–307. https://doi.org/10.1177/1073191110374797

Tout, D. (2020). Evolution of adult numeracy from quantitative literacy to numeracy: Lessons learned from international assessments. *International Review of Education*, *66*(2–3), 183–209. https://doi.org/10.1007/s11159-020-09831-4

Trendafilov, N., & Hirose, K. (2022). Exploratory factor analysis. In *International Encyclopedia of Education: Fourth Edition* (pp. 600–606). https://doi.org/10.1016/B978-0-12-818630-5.10015-6

Us, K. A., Musyaffa, A. A., & Ilyas. (2023). The quality of education in Islamic boarding schools in the revolutionary era 4.0 in middle of Covid-19 (Study of Islamic boarding schools in Jambi City). *AIP Conference Proceedings*, *2805*(1), 040001. https://doi.org/10.1063/5.0168158

Velec, M., & Huang, S. H. (2014). Quantitative methodologies and analysis. In *Research for the Radiation Therapist: From Question to Culture* (pp. 87–126). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054262642&partnerID=40&md5=088dbbc5f2953da65a6b9df9e88e711e

Wang, F., & Sahid, S. (2024). Content validation and content validity index calculation for entrepreneurial behavior instruments among vocational college students in China. *Multidisciplinary Reviews*, *7*(9), 2024187. https://doi.org/10.31893/multirev.2024187

Wang, V. X. (2018). Critical theory and transformative learning. In *Critical Theory and Transformative Learning*. https://doi.org/10.4018/978-1-5225-6086-9

Yao, G., Wu, C., & Yang, C. (2008). Examining the content validity of the WHOQOL-BREF from respondents' perspective by quantitative methods. *Social Indicators Research*, *85*(3), 483–498. https://doi.org/10.1007/s11205-007-9112-8

Yildiz, H. (2021). IRTGUI: An R Package for Unidimensional Item Response Theory Analysis With a Graphical User Interface. *Applied Psychological Measurement*, *45*(7–8), 551–552. https://doi.org/10.1177/01466216211040532

Zaqiah, Q. Y., Hasanah, A., & Heryati, Y. (2024). The role of steam education in improving student collaboration and creativity: A case study in Madrasah. *Jurnal Pendidikan Islam*, *10*(1), 101–112. https://doi.org/10.15575/jpi.v10i1.35207

Zhu, P., Chen, C.-C., Wang, Q., Luke, M. M., & Liu, Y. (2025). Item Response Theory Analysis and Measurement Invariance Testing of the Cultural Humility and Enactment Scale. *Measurement and Evaluation in Counseling and Development*, *58*(1), 27–46. https://doi.org/10.1080/07481756.2024.2344000

Zizler, P., & Ittyipe, S. (2023). On Nonnegative Loading Matrices: Two-Factor Case. *Applied Mathematics E - Notes*, *23*, 477–483. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174570074&partnerID=40&md5=d435640a5a0ac8144ebc1051b2f653dd