MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, Jan-June 2026, Pages: 759-771

Bridging Educational Gaps: IRT Model to Assess Mathematics Skills in Islamic Boarding Schools

Rosid Bahar[1]✉, Khusnul Khotimah[2], Clarence Fulca Sabrina[3], Iqbal Maulana [4], Nelsa Charisma Putri [5], Yahya Abdillah[6]

Sekolah Tinggi Agama Islam Idrisiyyh, Tasikmalaya, Indonesia.

email: rosidbahar@stai.idrsiyyah.ac.id, khusnulkhotimah@stai.idrisiyyah.ac.id, clarencefulcasabrina@stai.idrisiyyah.ac.id, iqbalmaulana@stai.idrisiyyah.ac.id, nelsacharismaputri@stai.idrisiyyah.ac.id, yahyaabdillah@stai.idrisiyyah.ac.id

Abstract

Purpose – This study aims to analyze the results of a mathematics competency test among students in pesantren (Islamic boarding schools) using the Logistic Parameter Model. The problem addressed is the lack of attention to general subjects, particularly mathematics, within the pesantren environment

Design/methods/approach – Using an ex post facto approach and descriptive quantitative methods, the study involved 1,242 students from Islamic boarding schools in Garut, Sukabumi, and Tasikmalaya, West Java, Indonesia. Data analysis was conducted using Item Response Theory (IRT) with R Studio software.

Findings – The results indicate that the 2-parameter logistic model (2 PL) had the most fit items, with 32 items, of which 27 were classified as good, or approximately 77%. The analysis also highlights the need for improvement in the construct of each question item, especially in arithmetic and algebra topics.

Research implications/limitations – The results of this study can be used as a reference for teachers in assessing mathematics learning in Islamic boarding schools. Starting from improving learning, to good measuring instruments, and analyzing it with the latest analysis methods to be more comprehensive.

Originality/value – This study illustrates that the instruments that have been used by teachers have not measured what should be measured, especially for students. The difficulty of students is working on questions that should be considered easy by teachers. Therefore, teachers must have a proper instrument that is able to measure the abilities of students, not based on assumptions felt by teachers, especially in general sciences such as mathematics.

Keywords: Mathematics Competency; Item Response Theory; Islamic Boarding School; Logistic Parameter Model;

Introduction

Pondok Pesantren (Islamic Boarding School) is Indonesia's oldest educational institution, and it even fought for the independence of the unitary state of the Republic of Indonesia (Bahar, Munadi, & Rosnawati, 2023; Fitriani, Murdowo, & Liritantri, 2023; Masqon, 2014; Muafiah, Sofiana, & Khasanah, 2022). According to history, the first *pesantren* were established as a place of study for Muslim students (*santri*) to focus on Islamic religious studies to become spreaders of their home region's religion (Effendy, Rohmatika, & Musyarofah, 2023; Gusmian & Abdullah, 2022; Isbah, 2020; Wekke & Hamid, 2013). Islamic boarding schools in Indonesia have evolved into three typological parts, which are as follows: *Salafiyah* (traditional), *khalafiyah* (contemporary), and integrated types. The first type of Islamic boarding school teaches solely traditional Islamic religious knowledge through classical books. The second type is *pesantren*, which not only teach Islam but also other sciences including English and Arabic in everyday life. The third type of Islamic boarding school delivers religious and general knowledge while using classical methods such as Salafiyah or modern ones such as Khalafiyah in the teaching process (Nurtawab & Wahyudi, 2022; Ridhwan, Nurdin, & Samad, 2018; Rouf, 2016; Wulandari, Lathifah, & Setyaningsih, 2022).

The number of *pesantren* that provide formal education, such as elementary schools, junior high schools, high schools, and universities, shows the current growth of *pesantren*. This development also encouraged the establishment of Law No. 18 of 2019 concerning *pesantren*, where *pesantren* can uphold the true teachings of Islam, which are reflected in the character of tolerance, moderation, balance, and community empowerment (Nurtawab & Wahyudi, 2022; The Minister of Religious Affair, 2019). One of the contents of this law is the implementation of at least 5 general subjects, including Mathematics, Pancasila and Citizenship Education (PPKn), Indonesian Language, Natural Sciences, and Social Sciences. It seems that we can start with this rule by evaluating the results of mathematics subject exams at *pesantren*.

The development of *pesantren* in recent years has encouraged many parents from different backgrounds to send their children to these institutions. This motive could be because the parents are also graduates of an Islamic boarding school, or because the parents desire to enroll their children in a boarding school that also offers school education (Supriatna, 2018). These motivations influence the increasingly diversified student input in *pesantren*. *Pesantren*, like a workshop, must have service competence that can help students with varying academic backgrounds and motivations.

One of the most significant issues regarding the impact of student input is that many students still pay little attention to general

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, Jan-June 2026, Pages: 759-771

subjects such as mathematics (Agustyaningrum, Sari, Abadi, & Mahmudi, 2021; Ayuwanti, Marsigit, & Siswoyo, 2021; Ramdhani, Suryadi, & Prabawanto, 2021; Yusnita, 2011). Islamic boarding school managers are well aware of this since students who live in *pesantren* are not under direct parental control, therefore changes in students' enthusiasm for studying and motivation can occur at any time. Some students do not pay attention to mathematical subjects since their initial purpose for enrolling in Islamic boarding school was parental duress, or because they just wanted to study Islamic religion but their parents expected their children to participate in school learning.

The significance of student input pushes mathematics subject teachers to develop test equipment capable of measuring students' mathematical abilities as a whole for various groupings. As a result, one of the efforts made is to give competency exam test equipment for end-of-year assessment activities, which are then examined to generate suitable questions that can be used again and again to measure students' mathematical abilities. Before the passage of the Islamic boarding school law, test kits were administered to students living in *pesantren*, but schools followed the national curriculum of the Ministry of Education and Culture of the Republic of Indonesia. Aside from that, the test equipment was only assessed using conventional test theory, which was deemed to have analytical flaws (Hambleton & Swaminathan, 1985; Hergesell, 2022; Istiyono, Dwandaru, Permatasari, & Aristiawan, 2020). The results of this study are deemed insufficient because they only provide information about the level of difficulty and differences in power based on the sample used (Retnawati, 2016). On the other hand, the analysis is still deemed inadequate because it did not include students from pure *pesantren* that use a mathematics curriculum that adheres to Islamic boarding school laws. As a result, extensive analysis is required, which can yield a wealth of information while also having an impact on test equipment that can evaluate students. This analysis employs an item response theory (IRT) approach with a logistic model with one, two, or three parameters. This approach can offer information function value and decide the applicability of items to the model based on the findings of each logistic parameter analysis.

Item response theory analysis is essentially a development of classical test theory. Classical test theory is thought to include shortcomings such as the connection between measurement findings and samples and the presence of pure values (Hambleton, Swaminathan, & Rogers, 1991; Parmaningsih & Saputro, 2021). To compensate for these shortcomings, modern theory or what is better known as item response theory emerged (Muranaka, Fujino, & Imura, 2023; Zanon, Hutz, Yoo, & Hambleton, 2016). The premise behind this theory is that it can provide information indicating that the potential of the correct answer is not influenced by the sample but by the subject's condition during the measurement. Thus, measurement outcomes or subject characteristics can be anticipated using abilities or traits. The theory is predicated on numerous fundamental assumptions, including unidimensionality, local independence, and parameter invariance (Abal, Sánchez González, & Attorresi, 2023; Hambleton et al., 1991).

The unidimensionality of a test can be explained by the fact that the items in the test instrument only evaluate one ability (Abdelhamid, Bassiouni, & Gómez-Benito, 2021; Albanese, Egger, Bütikofer, Armijo-olivo, & Ha, 2020). However, in practice, this assumption is impossible to meet, as evidenced by the majority of question items' unidimensionality. Meanwhile, local independence measures the ability of a response to an item not to influence or be influenced by responses to other items (Chen, Li, Liu, & Ying, 2018; Noventa, Spoto, Heller, & Kelava, 2019). The third assumption involves parameters referring to the characteristics of the test items that do not depend on the distribution of participants' abilities (Marsigit et al., 2020; Retnawati, 2014). Item response theory, like classical test theory, provides additional information after the assumptions are met. This information includes not just the amount of difficulty and different strengths, but also *Pseudoguessing*, or student-guessed answers. *Pseudoguessing* can alternatively be defined as the situation of accurately answering a question based on chance rather than aptitude (Vierula et al., 2021).

These parameters can provide extra information on the results of mathematical ability tests conducted in *pesantren*. As a result, the results of the question item analysis can be saved in a collection of questions to be used in *pesantren* by the goal of the next instrument, which is to create a new test tool capable of measuring the abilities of Islamic boarding school students.

This project is expected to produce a good set of mathematical instruments that may be used to assess the mathematical ability of Islamic boarding school students. This hope is significant since *Pesantren* have now become formal institutions on par with other formal schools, and they require more attention to compete with other institutions on the same level.

Literature review

Assessment

Assessment in mathematics education is a complex process aimed at evaluating various student competencies and skills. In practice, assessment goes beyond simply measuring learning outcomes; it also supports learning itself through diverse approaches. Mathematics assessments need to encompass aspects of conceptual understanding, procedural abilities, and problem-solving skills demonstrated in various contexts and formats (Mutarutinya et al., 2024). One of the main approaches to mathematics assessment is formative assessment, designed to provide feedback throughout the learning process. This form of assessment includes performance-based tasks that require students to demonstrate understanding through practical application. Furthermore, self- and peer-assessments allow students to evaluate their own and their peers' work, thus encouraging deeper reflection (Mutarutinya et al., 2024).

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, Jan-June 2026, Pages: 759-771

Meanwhile, summative assessment is used to assess learning outcomes at the end of a specific unit or learning period. Standardized exams are the most common form of summative assessment used for accountability purposes. These exams often fail to fully capture students' conceptual understanding (Zheng, Fancsali, Ritter, & Berman, 2019). In terms of the competencies assessed, mathematics assessments encompass not only technical skills but also higher-order thinking competencies. Corrêa & Haslam (2020) suggest that assessments should reflect students' abilities in problem-solving, reasoning, conceptual understanding, and procedural fluency. Furthermore, it is also important to evaluate critical thinking skills, modeling, and the application of mathematical concepts in real-world contexts (De Zeeuw, Craig, & You, 2013). This requires assessment designs that focus not only on correct answers but also on the thought processes behind them.

Item Response Theory

Item Response Theory (IRT) is a statistical framework used to model the relationship between an individual's latent abilities—such as ability or attitude—and their responses to test items. IRT has been widely applied in education, psychology, health measurement, and the social sciences, particularly for developing and analyzing test instruments and questionnaires (De Ayala, 2018; Mazza, Punzo, & Mc-Guire, 2014).This approach allows for more accurate evaluation of instrument quality compared to classical approaches.

IRT models generally use a logistic function to describe the probability of an individual correctly answering an item based on their level of latent ability. The three most commonly used logistic models are: the One-Parameter Logistic Model (1PL), or Rasch model, which only considers item difficulty (Kean, Bisson, Brodke, Biber, & Gross, 2018); the Two-Parameter Logistic Model (2PL), which adds a discrimination parameter to measure how well items differentiate between individuals with different abilities (Azevedo & Migon, 2012; Lo, 2008); and the Two-Parameter Logistic Model (2PL). and the Three-Parameter Logistic Model (3PL), which includes a guessing parameter to accommodate the possibility of students randomly answering correctly (Kalinowski, 2019; Liao, Ho, Yen, & Cheng, 2012).

IRT is widely used in educational testing, particularly in the design, calibration, and scaling of assessments. One of its main advantages is its ability to generate adaptive tests, which automatically adjust the difficulty level of items to the participant's ability (Hori, Fukuhara, & Yamada, 2022; Paek & Cole, 2019). In other fields such as psychology and health, IRT contributes to increasing measurement precision while reducing respondent burden and study costs (Kean et al., 2018).

Evaluating model fit is crucial in IRT applications. For 3PL models, visual techniques such as bin plots are used to assess the extent to which the model reflects the empirical data (Kalinowski, 2019).Furthermore, more flexible models such as the OPLM have been shown to have a better fit than more restrictive models such as Rasch (Khalid, Hussain, Hussain, & Riaz, 2011). This evaluation helps in selecting the most appropriate model for a particular assessment context.

Parameter estimation in IRT also continues to evolve to improve accuracy. Algorithms such as Equal Area Logistic (EARL) are used to produce stable and precise estimates. Furthermore, Bayesian-based approaches and nonparametric methods have been developed to handle non-normal latent ability distributions and improve overall model accuracy (Azevedo & Migon, 2012; Finch & Edwards, 2016). With these various approaches, IRT is a valuable tool in developing valid and reliable assessments.

Research method

Types of research

This is an ex post facto study using a descriptive quantitative approach. Ex Post Facto study is experimental research that does not provide therapy because it is based on facts about a previously occurring event or phenomenon and allows for changes in behavior (Asare & Amo, 2023; Bunari et al., 2023; Logan, 2022).

The event in this case was the results of the mathematics subject competency test, which was administered during the end-of-year assessment activities at the Islamic boarding school. A quantitative technique is utilized to achieve precise findings based on statistical computations, allowing conclusions to be drawn from the entire population (King, Keohane, & Verba, 1994; Sciberras & Dingli, 2023; Watson, 2015).

Subject

This study included 1242 students in grades 1, 2, and 3 at the Ulya Islamic Boarding School Muadallah Muallimin level, which is equivalent to junior high school in Islam. The study takes place in three *pesantren* spread across three cities in West Java, Indonesia: Garut, Sukabumi, and Tasikmalaya. Students were divided into 21 classes, each with its own set of characteristics. The rationalization of subject selection is based on the Cluster Sampling Technique, namely by determining the sample if the object to be studied or data source is very broad. This sample provision is based on the population area determined in two stages. The first stage identifies the sample area, while the second stage, which is similarly based on cluster sampling, determines individuals inside that area (Berndt, 2020; Etikan, 2017; Rahman, Tabash, Salamzadeh, Abduli, & Rahaman, 2022).

Data

Data was obtained from student responses during end-of-year assessment activities for the Mathematics Competency Test. The competency test is in the form of a mathematical ability test instrument, and it consists of 35 multiple-choice questions divided into four categories: A, B, C, and D.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, Jan-June 2026, Pages: 759-771

Data Analysis

Before starting data analysis, three assumptions must be met: unidimensionality, local independence, and parameter invariance (Saepuzaman, Istiyono, & Haryanto, 2022). The unidimensional assumption can be made by looking at the eigenvalues in the inter-item variance-covariance matrix using factor analysis, and unidimensionality can be met by simply measuring one dimension (Retnawati, 2014; Saepuzaman et al., 2022). The second assumption is local independence, namely an assumption test whose verification method depends on the results of the unidimensionality analysis. This means that if the unidimensional test is met, then local independence is also fulfilled (Demars, 2018; Okrey, 2013). The third assumption is that item parameters and ability parameters are invariant. This assumption is supported by estimating item parameters in groups of test participants with varying characteristics or classifications. Gender-determined group characteristics in this study (Marsigit et al., 2020). Parameter invariance was measured using three parameters: difference power (a), the level of difficulty (b), and pseudo-guessing (c). The points in the scatter diagram for each parameter provide evidence of this assumption. When the points approach the gradient, the parameters are considered to be invariant. Ability parameter invariance is accomplished in the same way that item parameter invariance is accomplished. The classification used to estimate ability parameters is using the upper and lower item sections.

The data was then analyzed to determine the suitability of the model using three logistic parameters namely 1 PL, 2 PL, and 3PL. The function in IRT can be employed if it meets one of the logistic parameters, which means that these three parameters will be used if they match the test equipment well (Arslan, Alkan, & Elhan, 2023; Hambleton et al., 1991). SPSS software, R Studio, and Microsoft Excel will be used to assist with the process.

Result

The first step is to put the unidimensional assumption to the test using factor analysis and the SPSS 24.0 software. Before performing the factor analysis test, the data was examined for sample adequacy using the Kaiser Meyer Olkin Measure of Sampling Adequacy (KMO-MSA) test and the Bartlett test to ensure data homogeneity. Table 1 shows these findings.

Table 1. KMO and Bartlett test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.751
Bartlett's Test of Sphericity	Approx. Chi-Square	1675.663
	df	595
	Sig.	.000

Table 1 shows that the sample size of 1242 persons in the study was adequate and homogeneous. This is evidenced in the KMO-MSA value > 0.5 and the Bartlett Test value < 0.5, allowing for factor analysis (Alam & Singh, 2023; Jahanshahi et al., 2023). The data was then analyzed for the unidimensional assumption using exploratory factor analysis (EFA).

The factor analysis results are acquired from the Total Variance Explained output, which presents the factors that were successfully extracted and have an Eigenvalue > 1. The Scree Plot in Figure 1 shows these results.

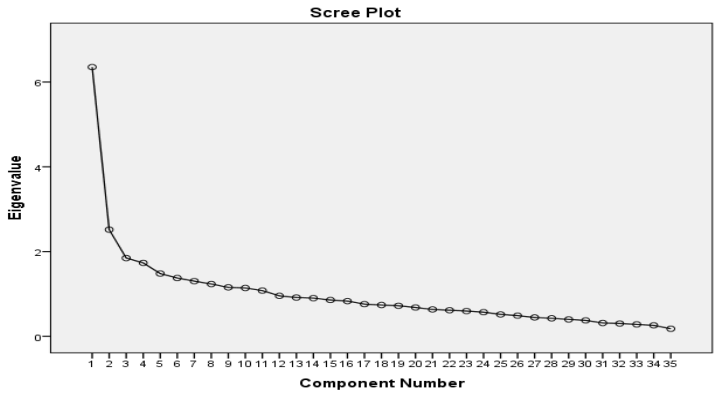


Figure 1. Scree Plot of Factor Analysis Results

According to the Scree Plot (Fig.1), there is one dominant factor with an eigenvalue of 6.351 or 18.14%. This result is twice as large as the second eigenvalue, 2.51 or 7.193%, indicating that the underman assumption is met (Saepuzaman et al., 2022). Furthermore, the total percentage of these 11 factors is 60.62%, indicating that these 11 components can explain 60.62% of the mathematical competency test questions. This proportion also meets the required minimum for taking the number of variables, which is 50% (Fergadiotis et al., 2023; Soland, 2023). This provides more evidence that the mathematics competency test instrument has met the unidimensional assumption test.The assumption of local independence is then tested. It is automatically met in this assumption test. This is because each item in the test instrument is unrelated to the others, implying that the items

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, Jan-June 2026, Pages: 759-771

utilized in the study are unrelated within groups of varying ability θ. Local independence emerges naturally from unidimensionality in this case (Deutscher, Kallen, Werneke, Mioduski, & Hayes, 2023; Retnawati, 2014; Wambua, Mwaura, & Dinga, 2023).

The third step is to put the parameter invariance assumption to the test. Estimating item parameters and test taker ability parameters are used to validate this assumption. The classification of even and odd items is used to measure item parameters, whereas the classification of odd and even responders is used to assess participant abilities. Produce item parameter estimates and ability estimates based on differentiating power (a), level of difficulty (b), and pseudo-guessing (c) using the R Studio software tool. The findings of this analysis are depicted in Figures 2, 3, 4, and 5 as scatter diagrams (Scatter Plots).

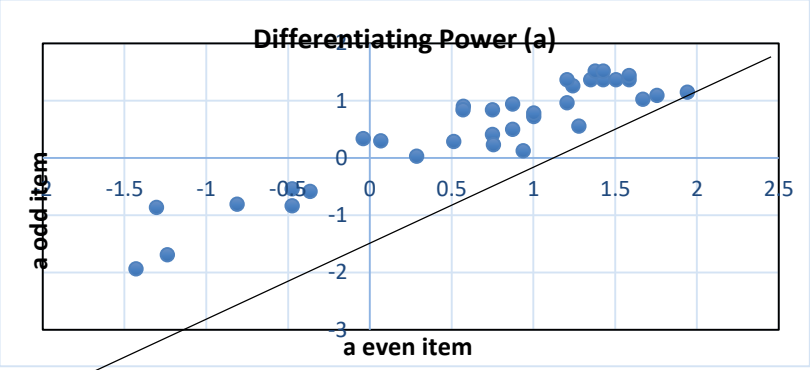


Figure 2. Parameter invariance for item groups and even item groups

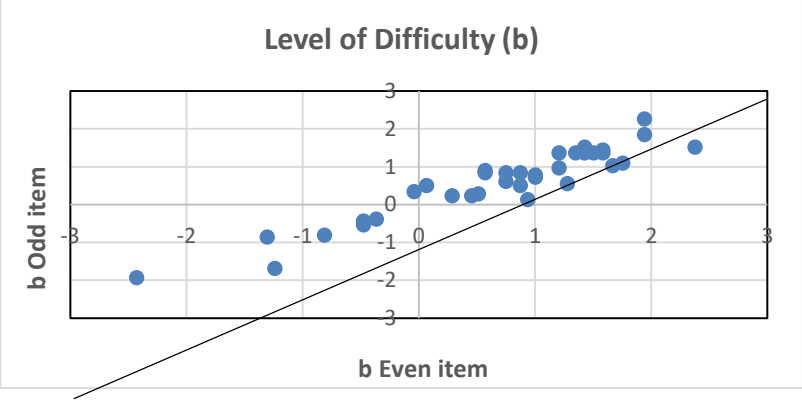
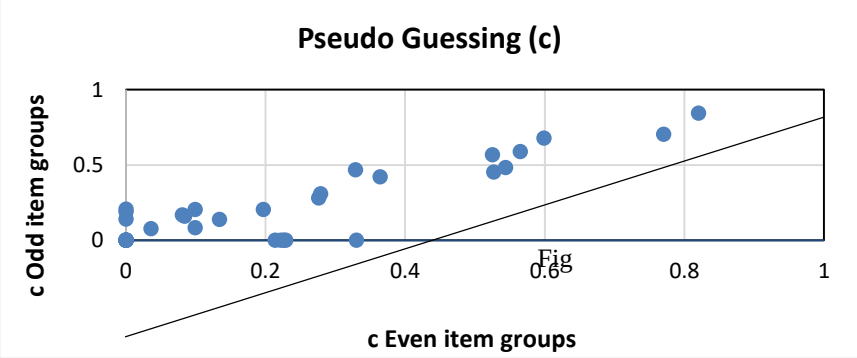


Figure 3. Parameter invariance for *b*-item groups and even item groups



Fig

Figure 4. Parameter invariance for *c* item groups and even item groups

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, Jan-June 2026, Pages: 759-771

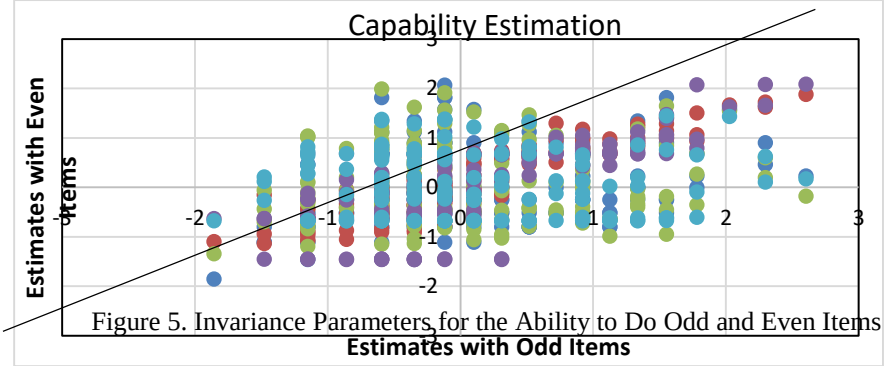


Figure 5. Invariance Parameters for the Ability to Do Odd and Even Items

The points for each item parameter and ability parameter were found to be close to the slope line based on the scatter plot images. This demonstrates that the item parameters and ability parameters depending on parameters a, b, and c do not differ across odd and even item groups. In other words, the assumption of parameter invariance is met.

The next stage is testing the model fit using the R Studio software program by looking at the results of the item fit analysis. The item fit test method used is the P.S_X2 method because the items in the instrument are considered to be few (Arlinwibowo, Retnawati, & Kartowagiran, 2021). Selain Aside from that, empirical evidence to establish the model's suitability is if the p-ValueP.S_X2 value is > 0.05 (Chalmers & Ng, 2017). The results of the analysis are shown in Table 2.

Table 2. Suitability of Mathematics Competency Test Item Items 2021 based on Logistic Model 1, 2, 3 Parameters

Item	(1 PL)		2 PL		3 PL	
	p.S_X2	Information	p.S_X2	Information	p.S_X2	Information
1	0.31	Suitable	0.32	Suitable	0.26	Suitable
2	0.33	Suitable	0.17	Suitable	0.03	Unsuitable
3	0.17	Suitable	0.18	Suitable	0.13	Suitable
4	0.38	Suitable	0.39	Suitable	0.06	Suitable
5	0.08	Suitable	0.07	Suitable	0.02	Unsuitable
6	0.33	Suitable	0.28	Suitable	0.40	Suitable
7	0.78	Suitable	0.65	Suitable	0.21	Suitable
8	0.76	Suitable	0.63	Suitable	0.34	Suitable
9	0.42	Suitable	0.59	Suitable	0.30	Suitable
10	0.40	Suitable	0.21	Suitable	0.37	Suitable
11	0.12	Suitable	0.15	Suitable	0.37	Suitable
12	0.19	Suitable	0.19	Suitable	0.36	Suitable
13	0.08	Suitable	0.06	Suitable	0.01	Unsuitable
14	0.09	Suitable	0.12	Suitable	0.18	Suitable
15	0.03	Unsuitable	0.18	Suitable	0.00	Unsuitable
16	0.37	Suitable	0.48	Suitable	0.09	Suitable
17	0.29	Suitable	0.28	Suitable	0.02	Unsuitable
18	0.64	Suitable	0.36	Suitable	0.70	Suitable
19	0.59	Suitable	0.36	Suitable	0.47	Suitable
20	0.71	Suitable	0.16	Suitable	0.18	Suitable
21	0.45	Suitable	0.35	Suitable	0.45	Suitable
22	0.01	Unsuitable	0.00	Unsuitable	0.00	Unsuitable
23	0.71	Suitable	0.70	Suitable	0.16	Suitable
24	0.02	Unsuitable	0.01	Unsuitable	0.02	Unsuitable
25	0.36	Suitable	0.07	Suitable	0.00	Unsuitable
26	0.26	Suitable	0.19	Suitable	0.40	Suitable
27	0.38	Suitable	0.44	Suitable	0.74	Suitable
28	0.47	Suitable	0.51	Suitable	0.30	Suitable
29	0.20	Suitable	0.32	Suitable	0.01	Unsuitable
30	0.86	Suitable	0.22	Suitable	0.00	Unsuitable
31	0.28	Suitable	0.21	Suitable	0.02	Unsuitable
32	0.04	Unsuitable	0.11	Suitable	0.00	Unsuitable
33	0.40	Suitable	0.31	Suitable	0.00	Unsuitable
34	0.18	Suitable	0.22	Suitable	0.00	Unsuitable
35	0.01	Unsuitable	0.00	Unsuitable	0.00	Unsuitable
The Total of Suitable Models	30		32		20	

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, Jan-June 2026, Pages: 759-771

Based on the results of the analysis presented in Table 3, the 3-parameter model is a model that generates many suitable items; thus, the 3-parameter logistic model is the model that can be used for item analysis.

Discussion

This article discusses everything presented in the findings, from testing assumptions to interpreting model fit results. First, the unidimensional assumption is met. This is, of course, typical, because the government presents multiple basic materials or competencies in one instrument or competency test package for mathematical subjects for grade 1 Muadallah Muallimin students at the senior level, which is equivalent to grade VII of junior high school. These basic abilities measure only one dimension, namely mathematical ability, even though the test presented contains sub-materials such as systems of equations and linear inequalities in one variable, comparisons of equal and inverse values, social arithmetic, gradients of lines, area and perimeter of quadrilaterals, and presentation of data in statistics (Minister of Education and Culture Republik Indonesia, 2018). Thus, in this scenario, it is perfectly reasonable if mathematics competency test data yields data that can quantify only one factor.

Second, the assumption test for local independence was not performed statistically. This is strongly related to the initial assumption test findings, namely unidimensionality. Because the assumption test generates 11 factors but only 1 dominating factor, there is no guarantee that these factors are correlated (Retnawati, 2014). In other words, this assumption test is considered to have been fulfilled

Third, the assumption test is performed by demonstrating the invariance of item parameters and abilities. These two invariants result in a scatter plot with no variance in each item and ability. This is quite possible because the data is collected from *pesantren*, which include students of different levels of ability. As a result, it is not surprising that in *pesantren*, placement tests are frequently administered to incoming students before the start of teaching and learning activities. This is primarily to provide services to students for them to develop their interests, talents, potential, and personal circumstances for them to obtain suitable guidance and counseling (Hariastuti, 2008; Nuraini, Tawil, & Subiyanto, 2019; Rahmawati, Suwarjo, & Utomo, 2019). However, this is where the shortcomings in measuring placement are found, as there is no standard. This is because the instruments utilized are still general in nature, covering all abilities such as English, Arabic, mathematics, and religious knowledge. Furthermore, the instrument was not fully examined and is still being developed without extensive theory. As a result, the instruments utilized cannot completely assess students' mathematical aptitude. As a result, as stated in the introduction, the goal of this research is to obtain instrument items that can be used to test students' mathematical ability evenly across groups and motivations.

After all assumption tests are met, the next analysis is the model suitability test. Based on the study results, it was discovered that for each logistic parameter, 30 items were regarded as suitable utilizing the 1 PL model, 32 items for the 2 PL model, and 20 items for the 3 PL model. This suggests that the 2-parameter model was picked for examining item characteristics because the greatest number is in the 3 PL model (Retnawati, 2014).

Before the item characteristics are analyzed further, first an estimate of the parameters and abilities is presented with an interpretation of "good" or "not good" as a representation that the item will later be used or reused in measuring class 1 mathematics ability. Conditions used for item classification namely: parameter *a* (different power) is limited between 0.00 to +2.00 (Demars, 2018; Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Nima, Cloninger, Persson, Sikström, & Garcia, 2020), parameter b (difficulty level) ranges from -2.00 to 2.00, (Huriaty, 2019; KEZER, 2021; Tsigilis, Krousorati, Gregoriadis, & Grammatikopoulos, 2023).

Table 4. Characteristics of Mathematics Lesson Competency Test Equipment Based on Item Response Theory Model 3 Logistic Parameters

Question Item	*a*	b	Criteria
1	0.233	-4.034	Not good (b < 0)
2	0.267	1.368	Good
3	0.227	1.909	Good
4	0.106	1.573	Good
5	0.082	1.188	Good
6	0.247	-6.95	Not Good (b < 0)
7	0.084	-3.841	Not Good (b < 0)
8	0.25	1.231	Good
9	0.134	6.315	Good
10	0.839	0.748	Good
11	0.768	0.437	Good
12	1.163	0.998	Good
13	1.389	0.333	Good
14	0.605	1.101	Good
15	2.471	0.282	Not Good (a>2)
16	1.351	0.531	Good
17	0.307	0.154	Good

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, Jan-June 2026, Pages: 759-771

18	0.944	1.328	Good
19	0.683	1.873	Good
20	1.753	0.599	Good
21	0.666	2.214	Good
22	1.046	0.325	Good
23	1.491	0.519	Good
24	1.542	0.274	Good
25	2.539	0.418	Not good (a > 2)
26	0.78	1.716	Good
27	1.103	1.271	Good
28	0.149	1.468	Good
29	1.858	0.13	Good
30	1.388	0.412	Good
31	0.062	-2.525	Not good (b < 0)
32	-0.047	-0.193	Not good (b < 0)
33	-0.243	0.113	Good
34	-0.276	-1.124	Not good (b < 0)
35	-0.301	0.839	b Good
The total of good questions			27

According to Table 4, the questions in the Mathematics Learning Competency Test assessment instrument totaled 27 and were considered good. These items will be saved in the question bank and utilized again in the future mathematical competency test, whereas those that are considered not good will be evaluated based on the indicators employed.

Table 4 lists Comparison Material as one of the items that are not good. Students find this material difficult since one of the necessary materials is division or fractional numbers (arithmetic). Students sometimes struggle with division material while attempting to reduce these numbers, especially when the Comparison is in reverse value form as seen in Figure 6.

A car traveling at 60 km/h takes 3 hours and 30 minutes. If the car travels at 90 km/h, the time required to traverse the same distance is...

A. 1 hour 15 minutes **C. 2 hours 20 minutes**

B. 2 hours15 minutes **D. 2 hours 30 minutes**

Figure 6. Comparison Question Items

For some, questions like this may seem like normal things that students in grade 1 or junior high school should be able to solve, but the reality is that questions like this are difficult for students, especially students who do not understand numbers (arithmetic) during their basic education. This problem is consistent with the findings of other studies, which show that students are still weak in reading text but skilled in arithmetic and that there are students who are skilled in understanding text but weak in arithmetic. Even though both skills are essential, especially when dealing with difficult questions (Basri, Purwanto, As'ari, & Sisworo, 2019; Pongsakdi et al., 2020).

In this question, students still cannot understand that the question is an inverse comparison of values. The structure of words that are difficult for students to understand exemplifies this issue. This is consistent with study findings indicating that the problem of "words" in writing tools is a significant issue at the primary and secondary school levels of education (Verschaffel, Schukajlow, Star, & Van Dooren, 2020). Students may struggle to understand the structure of this question. Students, on the other hand, continue to regard this question as difficult since their critical thinking processes in problem-solving remain low. According to a previous study, the critical thinking skills of Indonesian junior high school students are still lacking, particularly in analysis and evaluation (Basri et al., 2019).

The area and perimeter of a square are the next two questions under consideration. This is a level 6 question with a typical question form in which you must first assess what is known to answer it. Figure 7 shows an example of this question.

Mr. Kardi has a rectangular rice field with an area of 50 m^2. The length and width of the rice field are (4x+ 2) m and (2x + 1) m. The length of Mr. Kardi's rice field is ....

Figure 7. Question Items for Calculating the Area of a Square in Algebraic

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, Jan-June 2026, Pages: 759-771

Students must solve this issue by first determining the value of the variable provided in the form of an equation, and then substituting the variable into the length of the rectangle. This may seem obvious, yet it is the reality for students at *pesantren*. There are many basic questions in mathematics, however, many issues are extremely difficult for students to solve. In comparison, this question is regarded as tough not only in *pesantren* but also in public schools. The results of research conducted by (Nurainah, Maryanasari, & Nurfauziah, 2018) stated that in working on questions, students understood the concept, however, due to a lack of accuracy, the answers were less precise. Others write answers immediately without providing information about what is known and asked, resulting in incorrect conceptual application. Algebra material is perceived as difficult not only at the elementary school level but also at the university level, therefore a deeper rational comprehension is required at each school level (Powell, Gilbert, & Fuchs, 2019).

Furthermore, in terms of understanding the material, students are still lacking in understanding the material. Research conducted by (Jailani, Retnawati, Apino, & Santoso, 2020) also states that mathematics is very difficult for students due to a lack of recognition of mathematical patterns, mathematical terms, mathematical symbols, and mathematical concepts, and especially a lack of interest in solving mathematical problems (Jailani et al., 2020). This must be considered while developing tools for mathematics problems, particularly for students in *pesantren*, which continue to disregard general knowledge such as mathematics.

Conclusion and recommendation

This study uses a descriptive quantitative approach to examine the items on the test equipment of the Mathematics Competency Test in *pesantren*. The research findings revealed that the model's suitability in the test set was demonstrated in the 2 logistic parameters (2 PL) model with a total of 32 items that matched, and additional examination yielded 27 items in the good category. More emphasis should be placed on the item instrument's structure, particularly in the "words" portion. Word descriptions are essential to give students comprehension and clarity.

These findings are greatly useful to math teachers at *pesantren*. Teachers can use the findings of this study to develop better, more inventive learning innovations, particularly assessments. This research is, of course, still limited in the middle of the intense discussion about the regulation of *pesantren*. It is intended that future research into Islamic residential schools will be more intensive. Particularly in broad fields such as mathematics or other topics covered by these requirements. It is hoped that additional research will be conducted in analyzing the question items, particularly in mathematics. This is crucial to conduct to objectively assess students' abilities.

Declarations

Thank you to all parties who have helped in this research so that it can be completed as a scientific article. This research was not funded by any institution, but was purely the initiative of the authors. Thank You.

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

Abdelhamid, G. S. M., Bassiouni, M. G. A., & Gómez-Benito, J. (2021). Assessing cognitive abilities using the WAIS-IV: An item response theory approach. *International Journal of Environmental Research and Public Health*, *18*(13), 6835. https://doi.org/10.3390/ijerph18136835

Agustyaningrum, N., Sari, R. N., Abadi, A. M., & Mahmudi, A. (2021). Dominant factors that cause students' difficulties in learning abstract algebra: A case study at a university in Indonesia. *International Journal of Instruction*, *14*(1), 847–866. https://doi.org/10.29333/iji.2021.14151a

Alam, A., & Singh, A. (2023). Groundwater quality assessment using SPSS based on multivariate statistics and water quality index of Gaya, Bihar (India). *Environmental Monitoring and Assessment*, *195*(6), 687. https://doi.org/10.1007/s10661-023-11294-7

Albanese, E., Egger, M., Bütikofer, L., Armijo-olivo, S., & Ha, C. (2020). Construct validity of the Physiotherapy Evidence Database (PEDro) quality scale for randomized trials : Item response theory and factor analyses. *Research Synthesis Methods*, *11*(2), 227–236. https://doi.org/https://doi.org/10.1002/jrsm.1385

Arlinwibowo, J., Retnawati, H., & Kartowagiran, B. (2021). Item response theory utilization for developing the student collaboration ability assessment scale in STEM classes. *Ingénierie Des Systèmes d Information*, *26*(4), 409–415. https://doi.org/10.18280/isi.260409

Arslan, Y. K., Alkan, A., & Elhan, A. H. (2023). Comparison of the response time-based effort-moderated IRT model and three-parameter logistic model according to computerized adaptive test performances: a simulation study.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, Jan-June 2026, Pages: 759-771

Communications in Statistics-Simulation and Computation, 1–14. https://doi.org/https://doi.org/10.1080/03610918.2023.2245175
Asare, P. Y., & Amo, S. K. (2023). Developing preservice teachers' teaching engagement efficacy: A classroom managerial implication. *Cogent Education*, *10*(1), 2170122. https://doi.org/10.1080/2331186X.2023.2170122
Ayuwanti, I., Marsigit, & Siswoyo, D. (2021). Teacher-student interaction in mathematics learning. *International Journal of Evaluation and Research in Education*, *10*(2), 660–667. https://doi.org/10.11591/ijere.v10i2.21184
Azevedo, C. L. N., & Migon, H. S. (2012). Bayesian inference in an item response theory model with a generalized student t link function. *AIP Conference Proceedings*, *1490*(1), 49–58. https://doi.org/10.1063/1.4759588
Bahar, R., Munadi, S., & Rosnawati, R. (2023). The Brainstorming Method on Pesantren Students' Mathematical Connection and Metacognition Skills. *Pegem Journal of Education and Instruction*, *13*(3), 228–238. https://doi.org/10.47750/pegegog.13.03.24
Basri, H., Purwanto, As'ari, A. R., & Sisworo. (2019). Investigating critical thinking skill of junior high school in solving mathematical problem. *International Journal of Instruction*, *12*(3), 745–758. https://doi.org/10.29333/iji.2019.12345a
Berndt, A. E. (2020). Sampling methods. *Journal of Human Lactation*, *36*(2), 224–226. https://doi.org/10.1177/0890334420906850
Bunari, B., Fadli, M. R., Fikri, A., Setiawan, J., Fahri, A., & Izzati, I. M. (2023). Understanding history, historical thinking, and historical consciousness, in learning history : An ex post-facto correlation. *International Journal of Evaluation and Research in Education (IJERE)*, *12*(1), 260–267. https://doi.org/https://doi.org/10.11591/ijere.v12i1.23633
Chalmers, R. P., & Ng, V. (2017). Plausible-value imputation statistics for detecting item misfit. *Applied Psychological Measurement*, *41*(5), 372–387. https://doi.org/10.1177/0146621617692079
Chen, Y., Li, X., Liu, J., & Ying, Z. (2018). Robust measurement via a fused latent and graphical item response theory model. *Psychometrika*, *83*(3), 538–562. https://doi.org/10.1007/s11336-018-9610-4
Corrêa, P. D., & Haslam, D. (2020). Mathematical proficiency as the basis for assessment: A literature review and its potentialities. *Mathematics Teaching-Research Journal*, *12*(4), 3–20. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105458124&partnerID=40&md5=5733490a5199b3792694f5b3d366a26d
De Ayala, R. J. (2018). Item Response Theory and Rasch Modeling. In *The Reviewer's Guide to Quantitative Methods in the Social Sciences: Second Edition* (pp. 145–163). https://doi.org/10.4324/9781315755649-11
De Zeeuw, A., Craig, T., & You, H. S. (2013). Assessing conceptual understanding in mathematics. *Proceedings - Frontiers in Education Conference, FIE*, 1742–1744. https://doi.org/10.1109/FIE.2013.6685135
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
Deutscher, D., Kallen, M. A., Werneke, M. W., Mioduski, J. E., & Hayes, D. (2023). Reliability, validity, and efficiency of an item response theory–based balance confidence patient-reported outcome measure. *Physical Therapy*, *103*(7), pzad058. https://doi.org/10.1093/ptj/pzad058
Effendy, L., Rohmatika, A. H., & Musyarofah, N. (2023). An interest of santri in agriculture in vegetable farming communities in Tarogong Kidul Garut, West Java, Indonesia. *Universal Journal of Agricultural Research*, *11*(4), 723–730. https://doi.org/10.13189/ujar.2023.110406
Etikan, I. (2017). Sampling and sampling methods. *Biometrics & Biostatistics International Journal*, *5*(6). https://doi.org/10.15406/bbij.2017.05.00149
Fergadiotis, G., Casilio, M., Dickey, M. W., Steel, S., Nicholson, H., Fleegle, M., … Hula, W. D. (2023). Item response theory modeling of the verb naming test. *Journal of Speech, Language, and Hearing Research*, *66*(5), 1718–1739. https://doi.org/10.1044/2023_JSLHR-22-00458
Finch, H., & Edwards, J. M. (2016). Rasch Model Parameter Estimation in the Presence of a Nonnormal Latent Trait Using a Nonparametric Bayesian Approach. *Educational and Psychological Measurement*, *76*(4), 662–684. https://doi.org/10.1177/0013164415608418
Fitriani, R. A., Murdowo, D., & Liritantri, W. (2023). Applying the psychological of space in Islamic Boarding School (Case study: Pesantren Al Mahshyar Nurul Iman). *Journal of Islamic Architecture*, *7*(3), 437–444. https://doi.org/10.18860/jia.v7i3.17436
Gusmian, I., & Abdullah, M. (2022). Knowledge transmission and Kyai-Santri network in pesantren in Java Island during the 20th century: A study on Popongan manuscript. *Jurnal Akidah & Pemikiran Islam*, *24*(1), 159–190. https://doi.org/10.22452/afkar.vol24no1.5
Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-1988-9
Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
Hariastuti, R. T. (2008). *Dasar-dasar bimbingan dan konseling*. Surabaya: Unesa University Press.
Hergesell, A. (2022). Using Rasch analysis for scale development and refinement in tourism: Theory and illustration.

Journal of Business Research, *142*, 551–561. https://doi.org/https://doi.org/10.1016/j.jbusres.2021.12.063
Hori, K., Fukuhara, H., & Yamada, T. (2022). Item response theory and its applications in educational measurement Part II: Theory and practices of test equating in item response theory. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(3). https://doi.org/10.1002/wics.1543
Huriaty, D. (2019). Analisis karakteristik parameter butir berdasarkan model logistik 3 parameter. *Lentera: Jurnal Pendidikan*, *14*(2), 33–40. https://doi.org/10.33654/jpl.v14i2.885
Isbah, M. F. (2020). Pesantren in the changing Indonesian context: History and current developments. *QIJIS (Qudus International Journal of Islamic Studies)*, *8*(1), 65. https://doi.org/10.21043/qijis.v8i1.5629
Istiyono, E., Dwandaru, W. S. B., Permatasari, A. K., & Aristiawan, A. (2020). Developing computer based test to assess students' problem-solving in physics learning. *Journal of Physics: Conference Series*, *1440*(1), 012060. https://doi.org/10.1088/1742-6596/1440/1/012060
Jahanshahi, R., Yasaghi, Z., Mirzaei, F., Ghasemi, S., Sanagoo, A., Jouybari, L., & Foji, S. (2023). Burden of adult neurofibromatosis 1 questionnaire: translation and psychometric properties of the Persian version. *Orphanet Journal of Rare Diseases*, *18*(1), 1–6. https://doi.org/10.1186/s13023-023-02681-x
Jailani, Retnawati, H., Apino, E., & Santoso, A. (2020). High school students' difficulties in making mathematical connections when solving problems. *International Journal of Learning, Teaching and Educational Research*, *19*(8), 255–277. https://doi.org/10.26803/ijlter.19.8.14
Kalinowski, S. T. (2019). A Graphical Method for Displaying the Model Fit of Item Response Theory Trace Lines. *Educational and Psychological Measurement*, *79*(6), 1064–1074. https://doi.org/10.1177/0013164419846234
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
KEZER, F. (2021). The effect of item pools of different strengths on the test results of computerized-adaptive testing. *International Journal of Assessment Tools in Education*, *8*(1), 145–155. https://doi.org/10.21449/ijate.735155
Khalid, M. N., Hussain, T., Hussain, M., & Riaz, M. (2011). Item fit under rasch model and one parameter logistic model. *European Journal of Social Sciences*, *20*(4), 604–606. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-79955373216&partnerID=40&md5=7e97c8ac8edbcd61d6c134424cea875b
King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press.
Liao, W.-W., Ho, R.-G., Yen, Y.-C., & Cheng, H.-C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality*, *40*(10), 1679–1694. https://doi.org/10.2224/sbp.2012.40.10.1679
Lo, S.-C. (2008). Equal area estimation of three-parameter logistic model for item response theory. *Proceedings of the 3rd IEEE Asia-Pacific Services Computing Conference, APSCC 2008*, 1447–1452. https://doi.org/10.1109/APSCC.2008.31
Logan, W. A. (2022). *The ex post facto clause: Its history and role in a punitive society*. Oxford University Press.
Marsigit, M., Retnawati, H., Apino, E., Santoso, R. H., Arlinwibowo, J., Santoso, A., & Rasmuin, R. (2020). Constructing mathematical concepts through external representations utilizing technology: An implementation in IRT course. *TEM Journal*, *9*(1), 317–326. https://doi.org/10.18421/TEM91-44
Masqon, D. (2014). Dynamic of pondok pesantren as indegenous Islamic education centre in Indonesia. *EDUKASI: Jurnal Penelitian Pendidikan Agama Dan Keagamaan*, *12*(1). https://doi.org/10.32729/edukasi.v12i1.78
Mazza, A., Punzo, A., & Mc-Guire, B. (2014). KernSmoothIRT: An R package for kernel smoothing in item response theory. *Journal of Statistical Software*, *58*(6). https://doi.org/10.18637/jss.v058.i06
Muafiah, E., Sofiana, N. E., & Khasanah, U. (2022). Pesantren education in Indonesia: Efforts to create child-friendly pesantren. *Ulumuna*, *26*(2), 447–471. https://doi.org/10.20414/ujis.v26i2.558
Muranaka, S., Fujino, H., & Imura, O. (2023). Evaluating the psychometric properties of the fatigue severity scale using item response theory. *BMC Psychology*, *11*(1), 1–11. https://doi.org/10.1186/s40359-023-01198-z
Mutarutinya, V., Balimuttajjo, S., Mjenda, M. A., Wakhata, R., Mukuka, A., & Njiku, J. (2024). Assessment in Mathematics Education: An Inquiry Into Methods and Skills Assessed in the Era of Globalization. In *Impacts of Globalization and Innovation in Mathematics Education* (pp. 115–143). https://doi.org/10.4018/979-8-3693-2873-6.ch006
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
Noventa, S., Spoto, A., Heller, J., & Kelava, A. (2019). On a generalization of local independence in item response theory based on knowledge space theory. *Psychometrika*, *84*(2), 395–421. https://doi.org/10.1007/s11336-018-9645-6
Nurainah, N., Maryanasari, R., & Nurfauziah, P. (2018). Analisis kesulitan kemampuan koneksi matematis siswa SMP Kelas VIII pada materi bangun datar. *JPMI (Jurnal Pembelajaran Matematika Inovatif)*, *1*(1), 61–68.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, Jan-June 2026, Pages: 759-771

Nuraini, P., Tawil, T., & Subiyanto, S. (2019). The impact of Islamic-based career information service to improve career aspirations of students. *Islamic Guidance and Counseling Journal*, *2*(1), 26. https://doi.org/10.25217/igcj.v2i1.242
Nurtawab, E., & Wahyudi, D. (2022). Restructuring Traditional Islamic Education in Indonesia: Challenges for Pesantren Institution. *Studia Islamika*, *29*(1), 55–81. https://doi.org/10.36712/sdi.v29i1.17414
Okrey, G. J. (2013). Item response theory. In *The Routledge Handbook of Language Testing* (pp. 350–363). Routledge. https://doi.org/10.4324/9780203181287-36
Paek, I., & Cole, K. (2019). Using R for Item Response Theory Model Applications. In *Using R for Item Response Theory Model Applications*. https://doi.org/10.4324/9781351008167
Parmaningsih, T. J., & Saputro, D. R. S. (2021). Rasch analysis on item response theory: Review of model suitability. *AIP Conference Proceedings*, *2326*(1), 20017. AIP Publishing LLC. https://doi.org/https://doi.org/10.1063/5.0040305
Permendikbud. (2018). *Perubahan Atas Peraturan Menteri Pendidikan Dan Kebudayaan Nomor 24 Tahun 2016 Tentang Kompetensi Inti Dan Kompetensi Dasar Pelajaran Pada Kurikulum 2013 Pada Pendidikan Dasar Dan Pendidikan Menengah*.
Pongsakdi, N., Kajamies, A., Veermans, K., Lertola, K., Vauras, M., & Lehtinen, E. (2020). What makes mathematical word problem solving challenging? Exploring the roles of word problem characteristics, text comprehension, and arithmetic skills. *ZDM - Mathematics Education*, *52*(1), 33–44. https://doi.org/10.1007/s11858-019-01118-9
Powell, S. R., Gilbert, J. K., & Fuchs, L. S. (2019). Variables influencing algebra performance: Understanding rational numbers is essential. *Learning and Individual Differences*, *74*(June), 101758. https://doi.org/10.1016/j.lindif.2019.101758
Presiden Republik Indonesia. *Undang-Undang Republik Indonesia Nomor 18 Tahun 2019 Tentang Pesantren*. , (2019).
Rahman, M. M., Tabash, M. I., Salamzadeh, A., Abduli, S., & Rahaman, M. S. (2022). Sampling techniques (probability) for quantitative social science researchers: A conceptual guidelines with examples. *SEEU Review*, *17*(1), 42–51. https://doi.org/10.2478/seeur-2022-0023
Rahmawati, A. H., Suwarjo, & Utomo, H. B. (2019). The effect of basic skills counseling as vital skills in peer counseling to Indonesian students. *Universal Journal of Educational Research*, *7*(9), 1874–1881. https://doi.org/10.13189/ujer.2019.070905
Ramdhani, S., Suryadi, D., & Prabawanto, S. (2021). Hambatan belajar matematika di pondok pesantren. *Jurnal Analisa*, *7*(1), 46–55.
Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Yogyakarta: Nuha Medika.
Retnawati, H. (2016). Validitas reliabilitas dan karakteristik butir. *Yogyakarta: Parama Publishing*.
Ridhwan, Nurdin, A., & Samad, S. A. A. (2018). Dynamics of Islamic education in the land of Bugis: Growth, development and typology pesantren in Bone. *IOP Conference Series: Earth and Environmental Science*, *175*(1), 012158. https://doi.org/10.1088/1755-1315/175/1/012158
Rouf, M. (2016). Memahami tipologi pesantren dan madrasah sebagai lembaga pendidikan Islam Indonesia. *TADARUS*, *5*(1), 68–92. https://doi.org/http://dx.doi.org/10.30651/td.v5i1.345
Saepuzaman, D., Istiyono, E., & Haryanto, H. (2022). Characteristics of fundamental physics higher-order thinking skills test using item response theory analysis. *Pegem Journal of Education and Instruction*, *12*(4), 269–279. https://doi.org/10.47750/pegegog.12.04.28
Sciberras, M., & Dingli, A. (2023). Quantitative research. In *Investigating AI readiness in the maltese public administration* (pp. 43–115). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-19900-4_11
Soland, J. (2023). Item response theory models for difference-in-difference estimates (and whether they are worth the trouble). *Journal of Research on Educational Effectiveness*, 1–31. https://doi.org/10.1080/19345747.2023.2195413
Supriatna, D. (2018). Motivasi orang tua memilih pondok pesantren untuk anaknya. *Intizar*, *24*(1), 1–18. https://doi.org/10.19109/intizar.v24i1.1951
Tsigilis, N., Krousorati, K., Gregoriadis, A., & Grammatikopoulos, V. (2023). Psychometric evaluation of the preschool early numeracy skills test–brief version within the item response theory framework. *Educational Measurement: Issues and Practice*, *42*(2), 32–41. https://doi.org/10.1111/emip.12536
Verschaffel, L., Schukajlow, S., Star, J., & Van Dooren, W. (2020). Word problems in mathematics education: a survey. *ZDM - Mathematics Education*, *52*(1), 1–16. https://doi.org/10.1007/s11858-020-01130-4
Vierula, J., Talman, K., Hupli, M., Laakkonen, E., Engblom, J., & Haavisto, E. (2021). Development and psychometric testing of reasoning skills test for nursing student selection : An item response theory approach. *Journal of Advanced Nursing*, *77*(5), 2549–2560. https://doi.org/10.1111/jan.14799
Wambua, R., Mwaura, P., & Dinga, J. (2023). Psychometric properties of a test anxiety scale for use in computer-based testing in Kenya. *The International Journal of Assessment and Evaluation*, *31*(1), 1–18. https://doi.org/10.18848/2327-7920/CGP/v31i01/1-18
Watson, R. (2015). Quantitative research. *Nursing Standard*, *29*(31), 44–48. https://doi.org/10.7748/ns.29.31.44.e8681
Wekke, I. S., & Hamid, S. (2013). Technology on language teaching and learning: A Research on Indonesian Pesantren.

Procedia - Social and Behavioral Sciences, *83*, 585–589. https://doi.org/10.1016/j.sbspro.2013.06.111

Wulandari, D., Lathifah, Z. K., & Setyaningsih, S. (2022). Exploring internal quality assurance system for pesantren in Indonesia. *Journal of Higher Education Theory and Practice*, *22*(16), 126–132. https://doi.org/10.33423/jhetp.v22i16.5606

Yusnita, E. (2011). Pembelajaran kontekstual berlatar pondok pesantren pada materi garis dan sudut di kelas VII MTs. *Prosiding Seminar Nasional Penelitian, Pendidikan Dan Penerapan MIPA, Fakultas MIPA, Universitas Negeri Yogyakarta*, 11–18.

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*(1), 18. https://doi.org/10.1186/s41155-016-0040-x

Zheng, G., Fancsali, S. E., Ritter, S., & Berman, S. R. (2019). Using instruction-embedded formative assessment to predict state summative test scores and achievement levels in mathematics. *Journal of Learning Analytics*, *6*(2), 153–174. https://doi.org/10.18608/jla.2019.62.11