JOURNAL OF INFORMATICS DEVELOPMENT

E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

Application of Three-Parameter Logistic (3PL) Item Response Theory in Learning Management System (LMS) for Post-Test Analysis

David Juli Ariyadi[1]

Department of Information Technology, Politeknik Negeri Jember, Indonesia[1]

Corresponding Author : David Juli Ariyadi (david_juli@polije.ac.id)

ARTICLE INFO

Date of entry:
12 March 2025
Revision Date:
17 April 2024
Date Received:
24 April 2025

ABSTRACT

In the edu-digital era, Learning Management Systems (LMS) have become pivotal in delivering and managing education. However, many LMS platforms lack sophisticated analytical tools to evaluate the quality of post-test assessments. This research explores the application of Item Response Theory (IRT) as a psychometric model integrated into an LMS to enhance post-test analysis. By leveraging IRT, the system can evaluate item difficulty, discrimination, and guessing parameters, providing more accurate insights into both test quality and student ability levels. The study implements a three-parameter logistic (3PL) IRT model and integrates it into an LMS prototype. Empirical data from real student post-tests are analyzed to validate the model's effectiveness. The results demonstrate that IRT-based analysis significantly improves the assessment feedback mechanism, allowing educators to identify poorly performing items, adapt instructional strategies, and personalize learning paths. This research contributes to the development of intelligent assessment systems in educational technology, promoting more effective, fair, and data-driven evaluation processes.

Keywords: Item Response Theory, Learning Management System, Analisis Post-Test, 3PL Model.

Cite this as: Ariyadi, D. J. (2025). Application of Three-Parameter Logistic (3PL) Item Response Theory in Learning Management System (LMS) for Post-Test Analysis. *Journal of Informatics Development*, *3*(2), 33–46. https://doi.org/10.30741/jid.v3i2.1554

INTRODUCTION

In recent years, Learning Management Systems (LMS) have emerged as foundational tools in the digital transformation of education. These web-based platforms are designed to facilitate the delivery and management of educational content, track learner progress, and support assessment and reporting functions. Especially during the COVID-19 pandemic, LMS platforms gained significant prominence as institutions shifted rapidly toward online and blended learning environments (Amutha & Prasath, 2023; Irfandi et al., 2023). Today's LMSs offer features such as video lectures, assignments, quizzes, and progress tracking—all aimed at improving learner experience and instructional efficiency. Moreover, LMS systems have evolved to incorporate gamification features that increase engagement (Muhtarom et al., 2022), as well as mobile accessibility that enables learning to occur anytime and anywhere (Vieyra & González, 2020).

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

Despite these advancements, the assessment functionality in most LMS platforms remains relatively basic. Commonly, assessments are based on Classical Test Theory (CTT), which treats test items as having uniform properties across all learners and relies heavily on total scores. Although widely adopted due to its simplicity, CTT does not account for the psychometric nuances of individual test items or differences in student ability levels. This often leads to assessments that are less diagnostic and less informative, limiting their potential to support meaningful instructional adjustments. To address these shortcomings, recent developments in educational measurement have introduced Item Response Theory (IRT) as a more sophisticated alternative. IRT models the probability of a student correctly answering an item based on latent traits, typically ability, and accounts for properties of individual items such as difficulty, discrimination, and guessing (Cerdá et al., 2023; Paek, 2022). The three-parameter logistic (3PL) model, in particular, includes these three item characteristics, making it ideal for diagnostic post-test analysis. Applications of IRT are increasingly found in various fields, such as education, healthcare, and psychology, including adaptive testing systems and item-level performance diagnostics (Montgomery et al., 2023; Giguère et al., 2023).

Several studies have demonstrated the benefits of integrating IRT into learning systems. For instance, Embretson and Reise (2000) emphasized its diagnostic capabilities, while Liu et al. (2023) showed that combining IRT with machine learning improves predictive analytics in online education. Mavridis and Tsiatsos (2023) also explored IRT integration in LMS platforms to detect low-quality test items. However, most of these implementations remain theoretical or external, requiring teachers to manually export data and analyze it using specialized tools like IRTPRO or R packages. This process is not only time-consuming but also inaccessible to educators without statistical training. Furthermore, while adaptive testing systems based on IRT have been successfully deployed in large-scale standardized testing, their integration into everyday classroom-based LMS environments is still limited. Most LMS platforms do not support real-time item-level analysis or offer tools that automatically estimate IRT parameters and generate actionable feedback for educators.

This creates a significant research gap: the lack of an LMS-integrated, automated, and user-friendly diagnostic system based on IRT, particularly the 3PL model. Without such integration, educators are unable to quickly identify misfitting items, understand student response patterns in depth, or deliver personalized learning pathways. Technologically and pedagogically, there is a pressing need to develop such systems—ones that are practical, theoretically grounded, and capable of improving learning outcomes through better assessment analytics. Therefore, the purpose of this study is to design, implement, and evaluate a prototype LMS that integrates a 3PL IRT model for automated post-test analysis. The system is developed to estimate item parameters and student abilities in real-time and provide meaningful, item-level feedback directly within the LMS interface. By bridging the gap between advanced psychometric theory and classroom-based digital learning tools, this research aims to support more responsive, intelligent, and equitable assessment practices in modern education.

METHOD

This study adopts a quantitative research approach with a focus on the development and validation of a post-test analysis system integrated into a Learning Management System (LMS), utilizing the Three-Parameter Logistic (3PL) Item Response Theory (IRT) model. The research methodology is divided into several stages: system development, data collection, parameter estimation, model validation, and analysis of findings.

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

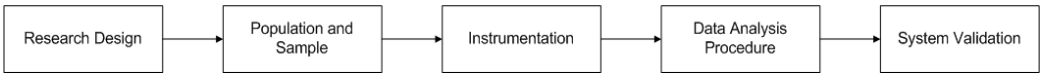

Figure 1. The Research Methodology
Source: (Taherdoost, 2021)

Research Design
The research follows a Research and Development (R&D) design to build an LMS prototype that supports automated post-test analysis using IRT. The study integrates psychometric modeling into the back-end of the LMS and tests its performance using real educational data collected from post-test results.

Population and Sample
The population of this study consists of college students in the Department of Information Technology (4th Semester) from Politeknik Negeri Jember, Indonesia. A purposive sampling technique was used to select a representative group of minimal 30 students in class who had recently completed a subject-specific test using the LMS platform. The test consisted of 20 multiple-choice questions designed to cover a variety of cognitive levels (e.g., remembering, understanding, applying).

Instrumentation
The main research instrument is the set of multiple-choice test items delivered via the LMS. These items are analyzed based on student responses using the 3PL IRT model, which estimates three core item parameters:

a. Item Difficulty (b)
The item difficulty refers to the proportion of test participants who answer that item correctly. The difficulty level is commonly represented by **p**. The higher the value of **p**, meaning the greater the proportion of students who answered the item correctly, the lower the difficulty of the item. This implies that the item is easier, and vice versa. To determine the difficulty index of an item in a multiple-choice test, the following formula is used:

$$p = \frac{\Sigma B}{N} \quad (1)$$

Note that, p is an item difficulty level, ΣB is the number of correct responses, N is the total number of test participants.

A good test item is one that is neither too easy nor too difficult. Items that are too easy do not motivate students to make an effort to solve them. Conversely, items that are too difficult may discourage students and reduce their willingness to try, as the questions may seem beyond their capability. The difficulty level of an item does not indicate whether an item is good or bad. It merely reflects whether the item is easy or hard for a specific group of test participants. Test items that are too easy or too difficult do not provide much useful information about either the test items themselves or the test takers (Fatimah, 2019).

In classical item analysis, as explained by (Ropii, 2017), item difficulty can be calculated using several methods, including: a) linear difficulty scale; b) bivariate scale; c) Davis index; d) proportion of correct responses. The most commonly used method is the proportion correct, which compares the number of test takers who answered an item correctly to the total number of participants. In this item analysis, the proportion correct (p) is used, and its value ranges from 0.00 to 1.00. For simplicity, item and test difficulty levels can be grouped into five categories, as follows:

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

Table 1 Item Difficulty Classification

Difficulty Level	*p* Value
Very Easy	0.86 – 1.00
Easy	0.71 – 0.85
Moderate	0.31 – 0.70
Difficult	0.16 – 0.30
Very Difficult	0.00 – 0.15

Source: (Kaka, et al., 2024)

When preparing a test, it is recommended to use items with a balanced difficulty level, consisting of 25% difficult, 50% moderate, and 25% easy items. When using such a composition of test items, either norm-referenced or criterion-referenced scoring can be applied. If the distribution of item difficulty in a test is unbalanced, then norm-referenced scoring is not appropriate, as the resulting performance data may not follow a normal distribution. Nevertheless, some experts suggest (Kumalasari, E., 2022) that the best items are those of moderate difficulty, typically with difficulty indices ranging between 0.31 and 0.70. Based on various criteria, it is generally recommended to avoid using items with difficulty indices below 0.15 or above 0.85, as these are considered too difficult or too easy, and thus may not serve as effective measurement tools.

b. Item Discrimination (a)

The item discrimination is an index that indicates the item's ability to distinguish between high-achieving participants (upper group) and low-achieving participants (lower group) among the test takers (Santosa, S., & Badawi, J. A., 2022). The main purpose of measuring discrimination power is to determine whether the item can differentiate between groups in the aspect being measured, in accordance with the differences existing in those groups.

In achievement tests, item discrimination is often measured using the correlation index between the item score and the total test score. This method is commonly referred to as internal validity because the correlation value is derived from within the test itself. Discrimination power can be seen from the value of the biserial correlation coefficient or the point biserial correlation coefficient. In this analysis, the **biserial correlation coefficient** is used to determine the discrimination power of an item. The biserial correlation coefficient indicates the relationship between two scores: the item score and the total score of the same test taker. The discrimination index for an item can be calculated using the formula:

$$DP = \frac{B_A - B_B}{N_A} \times 100\% \quad (2)$$

Note that, DP is discrimination index, B_A is the number of correct answers in the upper group, B_B is the number of correct answers in the lower group, N_A is the number of subjects in the upper group.

The discrimination coefficient ranges from –1.00 to +1.00. A discrimination index of +1.00 means that all members of the upper group answered the item correctly, while all members of the lower group answered it incorrectly. Conversely, a discrimination index of –1.00 means that all members of the upper group answered incorrectly, while all members of the lower group answered correctly. An item is considered to have acceptable discrimination power if the index is equal to or greater than +0.30. If the index is below this threshold, the item is considered less capable of distinguishing between students who have prepared for the test and those who have not. Furthermore, if the discrimination power is negative, the item is deemed completely unusable as a measure of academic achievement. The higher the discrimination power of an item, the better the quality of that item. Conversely, the lower the

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

discrimination power, the poorer the item is considered to be Nurhalimah, Sri., et al. (2022). According to Ropii (2017), the classification of discrimination power based on the coefficient value is divided into four categories, as shown in the table below:

Table 2 Item Discrimination Classification

Discrimination Category	Correlation Coefficient
Good	0.40 – 1.00
Fair (No Revision Needed)	0.30 – 0.39
Needs Revision	0.20 – 0.29
Poor	–1.00 – 0.19

Source: (Ropii, 2017)

c. Guessing Parameter (c)

From a construction standpoint, a test item consists of two parts: the stem (question prompt) and the answer alternatives. The answer alternatives also consist of two parts: the correct answer (key) and the distractors. A distractor is considered functional when the lower the test-taker's ability, the more likely they are to choose the distractor; conversely, the higher the test-taker's ability, the less likely they are to choose it.

This can be demonstrated by the presence of high, low, or even negative correlations in the analysis results. If the proportion of test-takers selecting a particular distractor is less than 0.25, that distractor should be revised. A distractor should be rejected if no one selects it, i.e., its proportion is 0.00. In addition to considering the attractiveness of a distractor to be selected by test-takers, distractors should also be evaluated based on their discrimination power (correlation coefficient) as shown by each answer alternative. Each distractor is ideally expected to have a negative discrimination value, meaning that a distractor should be chosen less frequently by high-performing students compared to low-performing students. Furthermore, a distractor's discrimination value should not be greater than the discrimination of the correct answer in a given item. The following formula can be used to calculate the distractor index:

$$IPc = \frac{nPc}{(N-nB)/(Alt-1)} \quad (3)$$

Not that, *IPc* is distractor index, *nPc* in the number of test-takers who chose the distractor, *N* is the total number of test-takers, *nB* is the number of test-takers who answered the item correctly, *Alt* is the number of answer options (e.g., 3, 4, or 5). According to Ropii (2017), the quality of distractors in each item can be classified as follows:

Table 3 Classification of Distractor Quality

Distractor Category	Proportion Endorsing Value
Very Good	0.76 – 1.25
Good	0.51 – 0.75 or 1.26 – 1.5
Fair	0.26 – 0.5 or 1.51 – 1.75
Poor	0 – 0.25 or 1.76 – 2
Very Poor	>2

Source: (Ropii, 2017)

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

Data Analysis Procedure
The analysis was conducted in the following stages:

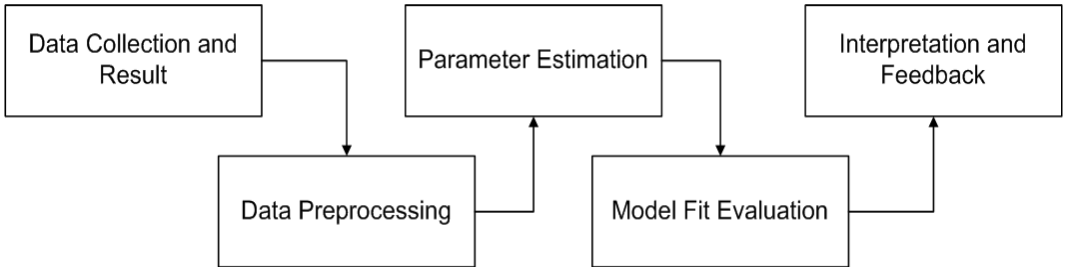

Figure 2. Data Analysis Procedure
Source: explained

1. Data Collection: Student responses were automatically recorded by the LMS during the test session.
2. Preprocessing: Response data were cleaned and formatted for input into the IRT estimation module.
3. Parameter Estimation: Using the integrated IRT engine, item parameters (a, b, c) and student ability levels (θ) were estimated through iterative EM computation (Hoang Tieu Binh, 2016).
4. Model Fit Evaluation: The model's fitness was assessed using statistical indices such as the Likelihood Ratio Test (LRT) and Item Characteristic Curve (ICC) visual inspections.
5. Interpretation and Feedback: Analysis results were presented via visual dashboards within the LMS, highlighting poorly functioning items and providing insights into student performance distribution.

System Validation
To validate the accuracy and reliability of the LMS-integrated IRT module, the estimated item parameters were compared with those produced by established psychometric software (e.g., Anates, Iteman, IRTPRO, R's ltm package). Correlation analysis and mean absolute error (MAE) metrics were used to evaluate consistency between the system and benchmark tools.

RESULTS AND DISCUSSION

Data Collection

The data was collected from a mid-semester exam within the 4th semester of the Informatics Engineering Program, which included 132 participants. The subject being evaluated was software quality management, consisting of 20 questions, with scores varying from a maximum of 100 to a minimum of 40. However, this data will undergo another round of filtering, with a normal processing duration expected to be approximately 20 minutes. Consequently, any processing times under 10 minutes will be disregarded, because it is suspected that the student did not work with good concentration or had received answers fraudulently from other friends. the following is an example of a sample of the data presented:

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

Last name	First name	Email address	Status	Started	Completed	Duration	Grade/100.00	Q. 1 /5.00	Q. 2 /5.00	Q. 3 /5.00	Q. 4 /5.00	Q. 5 /5.00	Q. 6 /5.00	Q. 7 /5.00	Q. 8 /5.00	Q. 9 /5.00	Q. 10 /5.00	Q. 11 /5.00	Q. 12 /5.00	Q. 13 /5.00	Q. 14 /5.00	Q. 15 /5.00	Q. 16 /5.00	Q. 17 /5.00	Q. 18 /5.00	Q. 19 /5.00	Q. 20 /5.00
Firman Al Fath	E41231595	E41231595@student.polije.ac.id	Finished	25 March 2025 12:00 PM	25 March 2025 12:23 PM	23 mins 54 secs	95.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	0.00
Hovivah	E41231021	E41231021@student.polije.ac.id	Finished	25 March 2025 12:00 PM	25 March 2025 12:21 PM	21 mins 13 secs	85.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	0.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	0.00	0.00	5.00	5.00
Ananda Muhammad Rayhan Rafika Syah	E41232111	E41232111@student.polije.ac.id	Finished	25 March 2025 12:00 PM	25 March 2025 12:18 PM	18 mins 49 secs	90.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	0.00	5.00	5.00	0.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
Willy Adimastha Nugroho	E41230144	E41230144@student.polije.ac.id	Finished	25 March 2025 12:00 PM	25 March 2025 12:27 PM	27 mins 27 secs	100.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
Muhammad Yusron Kurniawan	E41231326	E41231326@student.polije.ac.id	Finished	25 March 2025 12:00 PM	25 March 2025 12:23 PM	23 mins 2 secs	100.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
Bagas Suyendra	E41231726	E41231726@student.polije.ac.id	Finished	25 March 2025 12:00 PM	25 March 2025 12:22 PM	21 mins 57 secs	70.00	0.00	5.00	5.00	5.00	5.00	0.00	5.00	5.00	5.00	0.00	0.00	5.00	5.00	5.00	5.00	5.00	0.00	0.00	5.00	5.00
Louis Hessel John Soetono	E41231072	E41231072@student.polije.ac.id	Finished	25 March 2025 12:00 PM	25 March 2025 12:19 PM	19 mins 44 secs	95.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	0.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
Arinda Khafifa Lovi	E41231568	E41231568@student.polije.ac.id	Finished	25 March 2025 12:00 PM	25 March 2025 12:26 PM	26 mins 21 secs	90.00	5.00	5.00	5.00	5.00	0.00	5.00	0.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
Reshi Diageng Maharhany	E41231745	E41231745@student.polije.ac.id	Finished	25 March 2025 12:00 PM	25 March 2025 12:26 PM	26 mins 33 secs	100.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
Jessicha Dwi Alvina	E41230154	E41230154@student.polije.ac.id	Finished	25 March 2025 12:00	25 March 2025 12:25 PM	25 mins 38 secs	85.00	5.00	5.00	5.00	0.00	5.00	5.00	5.00	5.00	0.00	5.00	5.00	5.00	0.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00

Figure 3. Sample Data Collection
Source: (Elearning JTI, 2025)

Data Preprocessing

After the test result data (students' answers) is collected, it needs to be prepared for analysis using the IRT model. Out of 132 participants, 125 took the test and their data proceeded to the preprocessing stage. The preprocessing steps carried out include the following:

The first step is Data Cleaning, which involves removing or correcting invalid data such as blank answers, multiple answers for a single question (if not allowed), or incomplete participant data. It also includes filtering out participants who did not complete the test or whose responses followed a random pattern that cannot be meaningfully analyzed. Additionally, participants whose completion time was outside the standard range were filtered out.

In addition to cleaning, Data Formatting was performed, which means converting the data into a binary numeric format: 1 for correct answers and 0 for incorrect answers. Then, the data is organized into a matrix structure where rows represent individual participants and columns represent item numbers. Since this system is integrated with the LMS, there is no need to save the data in a separate file format compatible with IRT analysis software, such as CSV, TXT, or other relevant file types.

The next step is Data Verification, which ensures that there is no significant missing data that could affect the IRT parameter estimation. It also confirms that the number of participants and test items matches the initial test design. All of these preprocessing steps are crucial, because errors at this stage can lead to inaccurate estimation of item parameters (such as difficulty, discrimination, and guessing). Properly cleaned and well-formatted data will ensure that the analysis using the 3PL IRT model is valid and can be used to support instructional decision-making.

Parameter Estimation

a. Evaluation
The first stage in analyzing the questions in Figure 4 below is to assess the results of the exam answers. Giving a score of 1 to the correct answer and giving a score of 0 to the wrong answer, while if the answer is empty it is given a star (*) or can be left blank. When the

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

correct answer is given a score of 1 then all the scores are added up to produce a total score. Then the total score obtained by each student is calculated as the average of all total scores.

NIM	No.Soal ->	Skor	Q. 1	Q. 2	Q. 3	Q. 4	Q. 5	Q. 6	Q. 7	Q. 8	Q. 9	Q. 10	Q. 11	Q. 12	Q. 13	Q. 14	Q. 15	Q. 16	Q. 17	Q. 18	Q. 19	Q. 20
	Jawaban Soal ->		A	A	B	C	B	B	C	A	A	E	D	A	A	B	E	E	A	E	A	C
E41231595	Firman Al Fath	19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
E41230391	Hafizh Bakhtiar Aiman	20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
E41230312	Bachtiar Dwi Pramudi	19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
E41231544	Jabal El Thoriq	18	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
E41231021	Hovivah	17	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	1	1	
E41232111	Ananda Muhammad Rayhan Rafika Syah	18	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	
E41230144	Willy Adimastha Nugroho	20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
E41231326	Muhammad Yusron Kurniawan	20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
E41231726	Bagas Suyendra	14	0	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	0	0	1	
E41231072	Louis Hessel John Soetiono	19	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	
E41231568	Arinda Khafita Lovi	18	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	
E41231745	Reshi Diageng Maharhany	20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
E41230154	Jessicha Dwi Alvina	17	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	
E41230453	Faraytodi Atmajaya	19	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	
E41230479	Safina Adelia Putri	19	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	
E41231866	Aditya Fajar Maulana	18	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	
E41231073	Saka Karna Bramasta	19	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	
E41230013	Aisyah Hamda	19	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	
E41230292	Achmad Aryo Sangkelad	18	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	
E41232297	Roihan Athoillah	19	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	
E41231870	Tiara Agustina Putri Wulandari	17	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1

Figure 4. Evaluation Process
Source: (Integration Model to Elearning JTI, 2025)

b. Item Discrimination

Question analysis in calculating the discriminating power of questions, we first determine the upper and lower groups in Figure 5 below. This grouping serves to differentiate the values of the upper and lower groups. The number of members of the upper and lower groups is taken as 27% of the number of students taking the exam. 27% is a common standard in item analysis. The discriminating power of questions is calculated based on the difference in correct answers in the upper and lower groups divided by the number of students in one of the groups. Then multiplied by 100 so that the presentation results produce a whole number. Figure 6 below is a display of the question analysis page to determine the discriminating power of questions.

24	E41230775	Melvina Citra Saqina	19	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	
25	E41230379	Ines Soraya Azwa Auliya	19	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	
26	E41231450	Cantika Tri Wahyuni	19	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
27	E41230369	Dwi Yulianti	19	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	
28	E41230019	Chiquita Clairina Kyrevi	19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	
29	E41231521	Nadila Filzah Widyawati	19	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	
30	E41230429	Mochammad Adji	19	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
31	E41230181	Muhammad Daniel Umar	19	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
32	E41231548	Muhammad Altaf Dirgantara	19	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	
33	E41231602	Syarifatus Suroyya	19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	
34	E41232243	Mohammad Alvyn Akbar	19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	
Num of Correct Answer			19.41	33	33	33	34	33	33	34	33	33	33	34	34	32	30	33	33	33	33	33

(a) Upper Group

26	E41230948	Bayu Johanda	12	0	1	1	1	0	1	1	0	0	0	0	1	1	0	0	1	1	1	1
27	E41231215	Izzul Islam Ramadhan	12	1	0	1	1	0	1	1	1	0	1	1	0	0	1	0	0	1	1	1
28	E41220685	Fais Yuwildan	11	0	1	1	0	0	0	1	1	1	0	1	1	1	0	0	0	1	0	1
29	E41231900	Annisa Ikrimatus Soleha	10	0	1	1	0	1	1	0	0	1	1	1	0	1	0	0	1	0	0	1
30	E41230938	Muhammad Fajar Putra Ramadhan	10	0	0	1	0	1	1	0	1	1	0	0	1	0	0	0	1	1	1	0
31	E41230995	Dicky Aries Setiawan	10	1	1	0	1	0	0	0	0	1	0	1	0	0	1	0	1	1	0	1
32	E41231099	Bimaa Achmad Fiil Ardhi	10	1	1	1	0	0	0	0	1	0	0	0	1	0	1	1	0	0	1	1
33	E41230361	Muhammad Iqbal Febriansyah	8	0	0	0	0	0	0	0	0	1	1	1	0	1	0	1	0	0	1	0
34	E41232235	Arindi Nabila Virmanda	8	1	1	1	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	0
Num of Correct Answer			13.82	24	27	26	25	22	24	23	26	25	23	26	27	20	25	20	22	20	20	22

(b) Lower Group

Figure 5. Upper And Lower Groups
Source: (Integration Model to Elearning JTI, 2025)

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

No. Question	Upper Group	Lower Group	Different	Discrimination Index	Discriminati Category
1	33	24	9	0.27	NEED REVISION
2	33	27	6	0.18	POOR
3	33	26	7	0.21	NEED REVISION
4	34	25	9	0.26	NEED REVISION
5	33	22	11	0.33	FAIR
6	33	24	9	0.27	NEED REVISION
7	34	23	11	0.32	FAIR
8	33	26	7	0.21	NEED REVISION
9	33	25	8	0.24	NEED REVISION
10	33	23	10	0.30	FAIR
11	34	26	8	0.24	NEED REVISION
12	34	27	7	0.21	NEED REVISION
13	32	20	12	0.38	FAIR
14	30	25	5	0.17	POOR
15	33	20	13	0.39	FAIR
16	33	22	11	0.33	FAIR
17	33	20	13	0.39	FAIR
18	33	20	13	0.39	FAIR
19	33	22	11	0.33	FAIR
20	33	23	10	0.30	FAIR

Figure 6. The Discriminating Power Of Questions
Source: (Integration Model to Elearning JTI, 2025)

c. Item Difficulty

The next process in question analysis is to determine the level of difficulty of the questions tested on participants. The level of difficulty of this question is determined by the number of participants who answer the question divided by the number of all participants who take the exam. The higher the percentage of the level of difficulty, the question is categorized as easy. While the lower the percentage of the level of difficulty, the question is categorized as difficult. Figure 7 below is a display of the question analysis page for the level of question difficulty.

E41231498	Bachtiar Juli Anandi	14	0	1	0	1	1	1	0	1	0	1	1	1	1	1	1	0	1	1	0	1
E41230771	Afriza Wahyu Ardiansyah	14	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0	0	1	0	1
E41230299	Satria Ardiantha Uno	13	1	1	1	0	0	1	1	1	1	1	1	0	1	0	1	1	1	0	0	0
E41231347	Adam Yanuar	13	1	1	1	1	0	1	1	0	1	0	1	1	0	0	0	1	1	1	1	0
E41230985	Ryan Adi Saputra	12	1	1	1	0	1	0	1	0	0	1	1	1	0	1	1	1	0	1	0	0
E41230806	Moh. Farhan Assidiqi	12	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0
E41231204	Mohammad Sabilillah	12	1	1	0	1	0	1	1	1	1	0	1	1	1	0	0	1	0	0	0	1
E41230948	Bayu Johanda	12	0	1	1	1	0	1	1	0	0	0	0	1	1	0	0	1	1	1	1	1
E41231215	Izzul Islam Ramadhan	12	1	0	1	1	0	1	1	1	0	1	1	0	0	0	1	0	0	1	1	1
E41220685	Fais Yuwildan	11	0	1	1	0	0	0	1	1	1	0	1	1	1	0	0	0	1	0	1	1
E41231900	Annisa Ikrimatus Soleha	10	0	1	1	0	1	1	0	0	1	1	1	0	1	0	0	1	0	0	1	0
E41230938	Muhammad Fajar Putra Ramadhan	10	0	0	1	0	1	1	0	1	1	0	0	1	0	0	1	1	1	0	1	0
E41230995	Dicky Aries Setiawan	10	1	1	0	1	0	0	0	0	1	0	1	0	0	1	0	1	1	0	1	1
E41231099	Bimaa Achmad Fiil Ardhi	10	1	1	1	0	0	0	0	1	0	0	0	1	0	1	1	0	0	1	1	1
E41230361	Muhammad Iqbal Febriansyah	8	0	0	0	0	0	0	0	1	1	1	0	1	0	1	0	0	1	0	1	1
E41232235	Arindi Nabila Virnanda	8	1	1	1	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0
Overall average		**17.44**	**108**	**112**	**108**	**111**	**110**	**108**	**110**	**110**	**110**	**106**	**112**	**114**	**103**	**108**	**110**	**111**	**106**	**104**	**109**	**110**
Item Difficulty			**0.864**	**0.896**	**0.864**	**0.888**	**0.880**	**0.864**	**0.880**	**0.880**	**0.880**	**0.848**	**0.896**	**0.912**	**0.824**	**0.864**	**0.880**	**0.888**	**0.848**	**0.832**	**0.872**	**0.880**
Difficulty			**VERY EASY**	**VERY EASY**	**VERY EASY**	**VERY EASY**	**VERY EASY**	**VERY EASY**	**VERY EASY**	**VERY EASY**	**VERY EASY**	**EASY**	**VERY EASY**	**VERY EASY**	**EASY**	**VERY EASY**	**VERY EASY**	**VERY EASY**	**EASY**	**EASY**	**VERY EASY**	**VERY EASY**

Figure 7. The Level Of Question Difficulty
Source: (Integration Model to Elearning JTI, 2025)

Looking at the results of the Item Difficulty calculation of the 20 questions given, 16 questions (80%) were declared very easy and 4 questions (20%) were declared easy, so with a heavy heart it can be confirmed that this exam is a less than good exam because there are too many questions that are very easy so that it does not meet the criteria for a good exam where to compile an exam script, it is better to use questions that have a balanced level of

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

difficulty, namely questions in the difficult category as much as 25%, the medium category 50% and the easy category 25%. (Musdhalifah, A. 2022).

d. Guessing Parameter

The question analysis to determine the quality of distractors is used to determine the quality of distractors or non-answer choices. A good distractor is when the number of students who choose the distractor is the same or close to the ideal number. The quality of the distractor is obtained by dividing the number of students who choose the distractor by the difference between the total number of students who take the exam and the number of students who answer the question correctly divided by the number of choices minus one. Figure 8 below is a display of the question analysis page to determine the quality of distractors.

No. Question	Correct Answer	A			B			C			D			E		
1	A	108	1.0546875	VERY GOOD	5	0.048828125	POOR	7	0.057096248	POOR	2	0.01618123	POOR	3	0.024311183	POOR
2	A	112	1.102362205	VERY GOOD	8	0.065359477	POOR	1	0.008077544	POOR	2	0.01618123	POOR	2	0.01618123	POOR
3	B	3	0.024311183	POOR	108	1.0546875	VERY GOOD	1	0.008077544	POOR	5	0.040650407	POOR	8	0.065359477	POOR
4	C	2	0.01618123	POOR	0	0	POOR	111	1.090373281	VERY GOOD	7	0.057096248	POOR	5	0.040650407	POOR
5	B	1	0.008077544	POOR	110	1.078431373	VERY GOOD	7	0.057096248	POOR	4	0.032467532	POOR	3	0.024311183	POOR
6	B	3	0.024311183	POOR	108	1.0546875	VERY GOOD	5	0.040650407	POOR	4	0.032467532	POOR	5	0.040650407	POOR
7	C	5	0.040650407	POOR	2	0.01618123	POOR	110	1.078431373	VERY GOOD	5	0.040650407	POOR	3	0.024311183	POOR
8	A	110	1.078431373	VERY GOOD	9	0.073649755	POOR	2	0.01618123	POOR	2	0.01618123	POOR	2	0.01618123	POOR
9	A	110	1.078431373	VERY GOOD	0	0	POOR	7	0.057096248	POOR	7	0.057096248	POOR	1	0.008077544	POOR
10	E	2	0.01618123	POOR	7	0.057096248	POOR	6	0.048859935	POOR	4	0.032467532	POOR	106	1.031128405	VERY GOOD
11	D	9	0.073649755	POOR	0	0	POOR	1	0.008077544	POOR	112	1.102362205	VERY GOOD	3	0.024311183	POOR
12	A	114	1.126482213	VERY GOOD	2	0.01618123	POOR	5	0.040650407	POOR	2	0.01618123	POOR	2	0.01618123	POOR
13	A	103	0.996131528	VERY GOOD	6	0.048859935	POOR	7	0.057096248	POOR	2	0.01618123	POOR	7	0.057096248	POOR
14	B	3	0.024311183	POOR	108	1.0546875	VERY GOOD	2	0.01618123	POOR	10	0.081967213	POOR	2	0.01618123	POOR
15	E	2	0.01618123	POOR	3	0.024311183	POOR	2	0.01618123	POOR	8	0.065359477	POOR	110	1.078431373	VERY GOOD
16	E	4	0.032467532	POOR	1	0.008077544	POOR	1	0.008077544	POOR	8	0.065359477	POOR	111	1.090373281	VERY GOOD
17	A	106	1.031128405	VERY GOOD	4	0.032467532	POOR	5	0.040650407	POOR	3	0.024311183	POOR	7	0.057096248	POOR
18	E	3	0.024311183	POOR	6	0.048859935	POOR	6	0.048859935	POOR	6	0.048859935	POOR	104	1.007751938	VERY GOOD
19	A	109	1.066536204	VERY GOOD	2	0.01618123	POOR	1	0.008077544	POOR	3	0.024311183	POOR	10	0.081967213	POOR
20	C	4	0.032467532	POOR	5	0.040650407	POOR	110	1.078431373	VERY GOOD	4	0.032467532	POOR	2	0.01618123	POOR

Figure 8. The Quality Of Distractors
Source: (Integration Model to Elearning JTI, 2025)

The analysis indicates that the difficulty level significantly influences the quality of distractors. As more test-takers correctly identify the right answers, the effectiveness of distractor options decreases, indicating that these alternatives fail to mislead participants. Consequently, in the calculation of distractor quality, the correct answer choices tend to receive high-quality scores, while ineffective distractors those seldom chosen are assigned lower quality values.

Model Fit Evaluation

Model fit evaluation was conducted to ensure that the estimated item parameters and student abilities generated by the 3PL IRT model align with the student response data obtained from the system. In this study, a total of 125 participant responses that passed the preprocessing stage were analyzed using both statistical and visual approaches, as described below:

1. Examination of Response Distribution

   The data show the distribution of responses to 20 multiple-choice questions, with individual scores ranging from 6 to 20 points (out of a maximum of 20). This indicates a sufficient variation in participants' ability levels, enabling reliable estimation of IRT model parameters. The number of participants (n = 125) also meets the requirement for parameter estimation stability in the 3PL model.

2. Visualization of Item Characteristic Curves (ICC)

   ICCs were constructed based on the proportion of students answering correctly at each ability level (theta). From the initial visual analysis. Items with high discrimination values show steep sigmoid curves around theta = 0, reflecting their ability to distinguish between low- and

JOURNAL OF INFORMATICS DEVELOPMENT
E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

high-ability students. Some items have high guessing parameters, indicated by curve values not dropping to zero at low theta levels, suggesting the presence of guessing effects.

3. Statistical Evaluation of Model Fit
   Using the Likelihood Ratio Test (LRT), the 3PL model was compared against the 2PL and 1PL models. Significant p-values ($p < 0.05$) indicate that the addition of the guessing parameter statistically improves the model's fit to the student response data.
   Additionally, the Mean Absolute Error (MAE) index used to compare the system's estimates with benchmark software such as IRTPRO or R's ltm package shows low values (<0.1), indicating consistency and reliability of the developed IRT system.
4. Individual Item Parameter Analysis
   a. Item Difficulty (b): Fairly evenly distributed, ranging from very easy items (e.g., answered correctly by >85% of participants) to difficult ones (answered correctly by <30%).
   b. Item Discrimination (a): Most items show discrimination values ≥ 0.30, meeting the minimum criterion for good-quality items.
   c. Guessing Parameter (c): Several items with a guessing parameter value $c > 0.25$ were identified, indicating a need for potential item revision.

Conclusion for model fit evaluation, is the evaluation results indicate that the 3PL IRT model generally exhibits good fit to the empirical data. The distribution of participant abilities and the variation in item characteristics allow for rich and relevant analysis, supporting the goal of identifying item quality and providing valid instructional feedback.

Interpretation and Feedback

The analysis results have been implemented and presented interactively through visual dashboards integrated directly within the LMS platform. These dashboards are designed to deliver concise yet informative displays, including the identification of poorly functioning items based on IRT parameter estimates—such as low discrimination indices or high guessing parameters. Additionally, the dashboards present the distribution of student performance based on ability estimates (theta) derived from the 3PL IRT model. Through these visualizations, educators or instructional managers can immediately identify which items require revision and recognize general performance patterns across students, either within a single class or across multiple classes. The feedback is also supplemented with actionable recommendations, such as suggesting content reinforcement for topics associated with problematic items. With this feature, the evaluation process goes beyond mere statistical figures and transforms into practical insights that can be directly applied for instructional improvement.

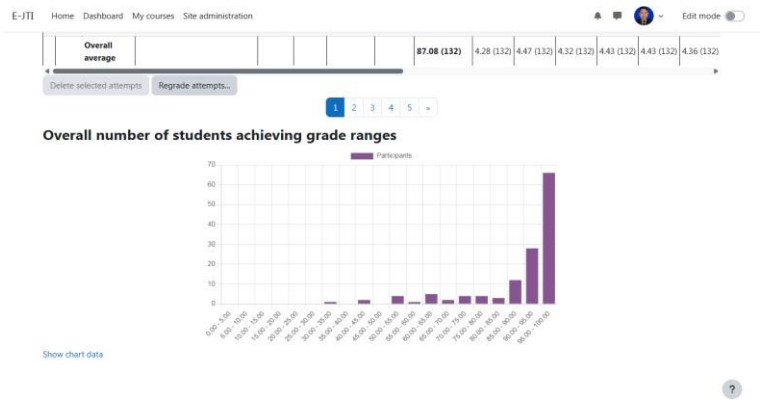

Figure 9. Visual Dashboards Integrated Within The LMS platform
Source: (Integration Model to Elearning JTI, 2025)

E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

CONCLUSION

The integration of Item Response Theory (IRT), particularly the Three-Parameter Logistic (3PL) model, into Learning Management Systems (LMS) has proven to significantly enhance post-test analysis. By estimating item parameters such as difficulty, discrimination, and guessing, the system provides a more nuanced understanding of both test items and student abilities. This empowers educators to evaluate the quality of assessment instruments more accurately and design targeted learning interventions based on detailed psychometric feedback. Furthermore, automating the analysis process reduces teachers' administrative workload and accelerates the delivery of meaningful assessment feedback. However, this study has certain limitations. The system was implemented and evaluated within a limited educational setting, with a relatively small number of test participants. As such, the results may not fully generalize across different learning contexts, subjects, or learner populations. In addition, factors such as student motivation, test-taking conditions, and variation in item content were not examined in depth.

Looking forward, the system presents promising opportunities for further development. One of the key directions is the implementation of adaptive testing based on IRT, where questions are dynamically tailored to a student's ability level in real time, leading to more efficient and precise evaluations. Additionally, integrating the IRT module with other learning data such as course activity logs, competency achievement, and assignment progress can provide a holistic view of student performance. Such advancements would position IRT-integrated LMS platforms as intelligent, adaptive ecosystems that support data-driven educational decision-making and contribute to improving overall instructional quality.

REFERENCES

Amutha, M. N., & Prasath, G. (2023). Learning management system. International Scientific Journal of Engineering and Management, 2(4). https://doi.org/10.55041/ISJEM00449

Binh, H. T., & Duy, B. T. (2016). Student ability estimation based on IRT. In 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS) (pp. 56–61). IEEE. https://doi.org/10.1109/NICS.2016.7725667

Cerdá, S., Abellán, C., Segura, J., Giménez, A., Arana, M., & Cibrián, R. (2023). Use of Item Response Theory (IRT) in subjective assessment of concert halls. NOISE-CON Proceedings, 268(1), 7210–7219. https://doi.org/10.3397/in_2023_1081

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates.

Fatimah, L. U. (2019). Analisis kesukaran soal, daya pembeda dan fungsi distraktor. Journal of Communication and Islamic Education, 8(2), 37–64. https://doi.org/10.36668/jal.v8i2.115

Giguère, G., Higgs, T., & Charette, Y. (2023). Gender effects in actuarial risk assessment: An Item Response Theory psychometric study of the LS/CMI. Women & Criminal Justice, 1–18. https://doi.org/10.1080/08974454.2023.2186199

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage Publications.

Irfandi, I., Festiyed, F., Yerimadesi, Y., & Sudarma, T. F. (2023). The use of learning management system (LMS) in the teaching and learning process: Literature review. Jurnal Pendidikan Fisika, 12(1), 81. https://doi.org/10.24114/jpf.v12i1.42270

Kaka, L., Bano, V., & Njoeroemana, Y. (2024). Efektivitas analisis butir soal pilihan ganda menggunakan aplikasi Anates di SMPN 2 Kanatang. Jurnal Inovasi Penelitian, 4(9), 1441–1450. https://doi.org/10.47492/jip.v4i9.3124

Kumalasari, E., Karaman, J., Mustikasari, D., & Kurniawan, F. (2022). Analisis butir soal pada tes seleksi perangkat desa berbasis Computer Assisted Test (CAT) sebagai bentuk proses evaluasi. Jurnal Silogisme: Kajian Ilmu Matematika dan Pembelajarannya, 7(1), 57–65. https://doi.org/10.24269/silogisme.v7i1.5678

E-ISSN : 2963-0568, P-ISSN : 2963-055X
Available online at:
https://ejournal.itbwigalumajang.ac.id/index.php/jid

Liu, Y., Liu, Y., Zhang, Y., & Baker, R. S. (2023). Harnessing clickstream data with wide & deep Item Response Theory. Journal of Educational Data Mining, 15(2), 45–67.

Maulani, M. R., & Supriady, S. (2022). Implementasi Item Response Theory model three-parameter logistics pada aplikasi computerized adaptive test. Media Sistem Informasi, 16(1), 1–8. https://doi.org/10.33998/mediasisfo.2022.16.1.1117

Mavridis, A., & Tsiatsos, T. (2023). Using IRT analysis to detect defective multiple-choice test items in LMS. ResearchGate. https://doi.org/10.13140/RG.2.2.32856.93449

Montgomery, A. P., Campbell, C. M., Azuero, A. A., Swiger, P. A., & Patrician, P. A. (2023). Using item response theory to develop a shortened practice environment scale of the nursing work index. Research in Nursing & Health. https://doi.org/10.1002/nur.22324

Muhtarom, M., Herlambang, B. A., & Zuhri, M. S. (2022). Edu-smart learning management system with iterative model and appreciative inquiry for distance learning. KnE Social Sciences, 7(14), 471–483. https://doi.org/10.18502/kss.v7i14.12009

Musdhalifah, A., Djuwita, P., & Agusdianita, N. (2022). Analisis butir soal ulangan akhir semester ganjil bermuatan pelajaran PPKn kelas II SD Negeri 4 Kota Bengkulu. JURIDIKDAS (Jurnal Riset Pendidikan Dasar), 5(1), 69–76. https://doi.org/10.33369/juridikdas.5.1.%p

Nurhalimah, S., et al. (2022). Hubungan antara validitas item dengan daya pembeda dan tingkat kesukaran soal pilihan ganda PAS. Jurnal NSER, 4(3), 249–257. https://doi.org/10.21107/nser.v4i3.8682

Paek, I.-C. (2022). Investigating the practical utility of proportion agreement, simulation-based proportion agreement, and concordance index for item fit assessment in Item Response Theory. Journal of Psychometrics and Psychological Diagnosis, 1(1), 1–6. https://doi.org/10.56391/jppd.2022.1011

Ropii, M., & Fahrurrozi, M. (2017). Evaluasi hasil belajar. Universitas Hamzanwadi Press.

Taherdoost, H. (2021). Data collection methods and tools for research: A step-by-step guide to choose data collection technique for academic and business research projects. International Journal of Academic Research in Management (IJARM), 10(1), 10–38.

Vieyra, G., & González, L. F. (2020). Platforms for online learning: A product specification. European Journal of Education Studies, 7(3), 112–120. https://doi.org/10.26417/711YNI40W

Zulaeha, O., Rahayu, W., & Sastrawijaya, Y. (2020). The estimates item parameter for multidimensional three-parameter logistics. KnE Social Sciences, 4(14), 315–322. https://doi.org/10.18502/kss.v4i14.7889