

The quality of test on National Examination of Natural science in the level of elementary school

Nurul Hamidah, Edi Istiyono

Department of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

Article Info

Article history:

Received Jul 1, 2021

Revised Feb 28, 2022

Accepted Apr 4, 2022

Keywords:

Item difference power

Item difficulty level

Item quality

National Examination

ABSTRACT

This research analyzed the quality of National Examination questions on Natural science subjects in elementary schools. The sampling technique used was simple random sampling. Response data were obtained from 250 grade VI elementary school students who responded on the tryout questions. This study was a descriptive study with a quantitative approach to describe the quality of the elementary school science test try-out items. The quality of this question is described quantitatively which includes an index of difficulty level and discriminatory power of questions. The data was collected using a test technique using a test instrument consisting of 30 multiple-choice questions. The analysis was carried out using the BILOG-MG version 4 application to obtain the parameters of distinguishing power and grain difficulty level. Based on the model fit analysis, it was found that the most appropriate model to use for the analysis of the science tryout questions for elementary schools was item response theory (IRT) with the 2PL model. The study concluded that the National Examination questions on science subjects in elementary schools had met the criteria for the level of difficulty and distinguishing power so that they could be said to be in a good category.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Nurul Hamidah

Department of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta
Depok, Sleman, Yogyakarta 55281, Indonesia

Email: nurulhamidah.pep.2017@student.uny.ac.id

1. INTRODUCTION

Natural science is one of the subjects taught in elementary schools. Science learning in schools aims to develop curiosity and a positive attitude towards science, technology, and society, develop process skills to investigate the environment, solve problems and make decisions. As well as developing knowledge and understanding of scientific concepts that will be useful and can be applied in everyday life. Science is included in the subjects tested in the elementary school National Examination. One way to monitor the quality and standards of science teaching and learning in schools is through the assessment of student learning outcomes. Assessment is the main and effective thing to find out students' understanding after they have participated in a series of learning activities because it is impossible for one student and another student to have the same understanding of what they have learned [1]. The purpose of the assessment is to allow students to show what they have learned, find out how learning improves over time, motivate students, and classify students in class rankings [2]. Assessment is the process of identifying information and interpretations about student learning to provide information about student achievement and progress and set the direction for continuous teaching and learning [3]. Assessments must be carried out by teachers to monitor the process, progress, and improvement of student learning outcomes on an ongoing basis [4]. Teaching materials and strategies as well as the stages of assessment and evaluation must be continued

according to the existing situation determined before teaching [5]. Therefore, it is important to design an appropriate assessment, one of which is by preparing an assessment tool that will be used.

The tool that can be used in the assessment activities is a set of tests. The test aims to determine the learning achievements and abilities of students in certain fields [6]. Tests are assessment tools that are often used to obtain information about individuals in various behaviors and tests in schools are carried out in the context of assessing three learning domains, namely cognitive, affective, and psychomotor learning domains [7]. The cognitive domain is still a top priority in the assessment of learning in schools. Cognitive assessment is usually carried out at the end of the lesson, in the middle of the school year, and at the end of the school year. This is by the form of the test based on its function, namely formative and summative tests. Formative tests are carried out at the end of each lesson to determine the achievement of students in certain competencies. While the summative test is a test that is carried out at the end of the semester or the end of the education unit which functions to report student learning outcomes within a certain period. One of the summative tests used in Indonesia is the National Examination.

The National Examination in primary schools uses a multiple-choice form of test. Multiple-choice questions are very objective instruments for measurement [8]. The use of multiple-choice questions is usually driven by the need for teacher efficiency and the provision of rapid feedback aimed at encouraging effective learning because multiple-choice questions require selecting the correct answer from a set of alternatives [9]. The advantages of multiple-choice tests for teachers are that multiple-choice questions are easier in statistical analysis and can measure a wide range of abilities [10]. Multiple-choice test formats are used in a variety of contexts, from low-level assessments for daily assessments to high-level assessments that determine professional progress [11]. With various advantages, these multiple-choice questions are used in National Examination in Indonesia, especially for elementary schools. On that basis, the regional policymakers compiled a National Examination tryout in the form of a multiple-choice test. This National Examination tryout is one type of summative test that is useful to determine the readiness of students in facing the National Examination. With the hope that the tryout can provide practice questions whose characteristics are almost the same as the real National Examination. However, the situation at school is that most teachers do not understand how the quality of the items in the tryout is prepared. They only assume that when many students do not reach the minimum completeness criteria (*Kriteria Kelulusan Minimum/KKM*) then the exam questions are considered difficult questions. This is what lies behind the need for an analysis of the characteristics of the items on the National Examination tryout questions.

Analysis of the characteristics of the items is a way to measure the quality of the items on the test by looking at how appropriate the item is for the test taker and how well the item measures the test taker's ability [12]. Item analysis provides two types of information, namely difficulty items or facility items, which help teachers find out whether the test items are appropriately used according to the level of test-takers and item discrimination which allows teachers to see whether each item in the test can provide consistent information about students' abilities [13]. Item analysis has several objectives according to [14], namely: i) To help assess the value or quality of the test; ii) Can help in the next test revision; iii) Can be used to create test files for future testing; iv) Leads to an increase in skills in making tests; v) Provide diagnostic value and assist in planning future learning activities; vi) Provides a basis for discussing test results; and vii) Can be a learning experience for students, if students help or are notified of the results of item analysis. The importance of analyzing the characteristics of test items is that achievement tests are very important as a measuring tool to be used in the evaluation process [15]. So, to develop a compatible test it must be considered with the characteristics of the test.

Schools must have valid and reliable tests to assess students covering areas of curriculum content [16]. A measurement has high validity if it produces accurate data about the measured variable by the measurement objectives and the reliability shows the firmness or consistency of a measuring instrument [17]. The assessment instrument must be valid and reliable [18]. In addition to being valid and reliable, in preparing for the test, it is necessary to pay attention to the characteristics of the items. The characteristics of the items include item difficulty level, discriminating power, and guessing. The characteristics of these items can be analyzed through classical test theory (CTT) and item response theory (IRT) methods.

CTT is a simple model that describes how measurement error affects the observed score [19]. In CTT item parameters are seen from the index of difficulty level and distinguishing power. The difficulty level is the average score of the item according to the proportion of examinees who answered the item correctly and discriminatory power is a parameter in item selection of how effectively the items 'distinguish between examinees who are relatively high on the criterion of interest and those who are relatively low [20]. The item difficulty index in the CTT is between 0.0 and 1.0 starting from easy items with higher p-values and difficult items with lower p-values [21]. The difference is the point-biserial correlation [22]. CTT has the disadvantage that the characteristics of the items depending on the condition of the test taker. If the test taker is in the high ability group, the item has an easy level of difficulty and vice versa. To overcome this weakness, IRT was developed.

There are three models used to estimate item characteristics. Model 1 logistic parameter (1PL) provides an estimate of the relationship between the level of difficulty (b) and ability. In the 2 logistic parameter (2PL) model there is an additional item parameter, namely item discriminating power (a). In the 3 logistic parameter model (3 PL) grain characteristics are seen from the level of difficulty (b), discriminating power (a), and guessing (c) [23]. The intense pulsed light (IPL) model is a simple model where the ability of students is determined from the level of difficulty of the test items because the discriminating power has a constant value, and the guessing parameter is set to zero so that the characteristics of the items in the IPL model are determined by the value of the level of difficulty. The 2PL uses parameter a (distinguishing power) and parameter b (level of difficulty) to estimate the ability of test-takers (θ) and parameter guessing (c) is set to zero, so that the quality of the questions is determined by the parameters of discriminating power and level of difficulty.

The use of CTT and IRT has not been widely understood by the teachers at elementary school science tryout questions at the regional level. The most item analyzes, which included determining item difficulty and discrimination, and distractor analysis were not carried out because they would take a long time to be done manually [24]. This lack of understanding prepares for the questions not to go through the stages of analyzing the quality of the questions quantitatively first so that the quality of the questions produced is not known, especially on the characteristics of each item.

The purpose of this study was to determine the quality of the elementary school science tryout items. There are three research questions: i) How is the quality of the elementary school Natural science tryout questions if an analysis of the characteristics of the items is carried out with CTT?; ii) How is the quality of the elementary school Natural science tryout questions if an analysis of the characteristics of the items is carried out with IRT?; iii) Which model gives the most suitable results for the analysis of Natural science tryout items? Therefore, this article aims to determine the quality of the Natural science tryout questions according to the characteristics of the items based on their parameters with CTT and IRT, to compare the parameters obtained by CTT and IRT analysis, and to determine the number of items fit with each model to provide a model that is most suitable to be used in the analysis of the tryout question.

2. RESEARCH METHOD

This study was a descriptive study with a quantitative approach to describe the quality of the elementary school science test try-out items. Descriptive research provides an accurate description or description of the status or characteristics of a situation or phenomenon [25]. This study explored facts about the quality of the National Examination tryout questions on Natural science subjects in elementary schools. The quality of this question is described quantitatively which includes an index of difficulty level and discriminatory power of questions. The sampling technique used was simple random sampling. This technique is used based on the similarity of characteristics of grade VI students in elementary schools so that all students can be used as sampling for the study. With simple random sampling technique, it was obtained 250 grade 6 elementary school students. Data were collected based on the result of tests from 250 respondents using research instruments (35 items of multiple-choice test) prepared by the researchers.

The multiple-choice test was chosen as a matter of the national science exam in elementary schools because the multiple-choice instrument was able to measure students' abilities objectively. Student response data were analyzed using CTT and IRT with the help of the BILOG-MG version 4 application. Analysis of CTT to see the level of difficulty and differentiation of items. Good test items are not too easy and not too difficult, so an analysis of the difficulty level of the items [26] provides the equation used in finding the value of the difficulty index (p) as in (1).

$$\text{Difficulty index } (p) = \frac{\text{The number of students who answered the questions correctly}}{\text{the total number of students}} \quad (1)$$

The second analysis is the distinguishing power of the questions. Distinguishing power (D) is the ability of questions to distinguish students' abilities from high abilities and low abilities. The equation used in the analysis of discriminating power [27] as in (2).

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0 \sqrt{p(1-p)}}{S_x} \quad (2)$$

Where \bar{X}_1 indicates the average crude score on the test for all students who answered the item correctly, while \bar{X}_0 indicates the average crude score on the test for all students who answered the item incorrectly, and S_x is the standard deviation and p shows the proportion of students who answered the test

correctly. The percentage value of the correct answer is used to determine the item difficulty level criteria. There are three categories of item difficulty levels in CTT, namely difficult, moderate, and easy. The criteria for the level of difficulty are described in detail in Table 1.

The distinguishing power of the grain can be seen from the Pt-Biserial Correlation value. There are three categories of distinguishing points based on the value of ρ -bis, namely good, good enough, and not good. The categories for distinguishing items are detailed in Table 2.

Table 1. Category of item difficulty level in CTT

Level of difficulty	Category
< 0.30	Difficult
0.30-0.70	Moderate
>0.70	Easy

Table 2. Grain distinguishing power category in CTT

ρ -bis Value	Category
<0.20	Not good
0.20-0.29	Moderate
0.30-0.70	Good

The IRT used 1PL model to see the level of difficulty (b). The IPL model is a simple model where the ability of students is determined from the level of difficulty of the test items because the discriminating power has a constant value and the guessing parameter is set to zero. The mathematical formula used in the 1PL model is as in (3).

$$P_i = \frac{e^{Di(\theta-b)}}{1+e^{Di(\theta-b)}} \quad (3)$$

The formula description shows the starting item 1,2,3, ..., n. The P_i is the probability that the test taker has the ability (θ) to answer item 1, b_i is the parameter of the difficulty level of item 1, e is a constant number approaching 2.718, and D represents the scale parameter [14]. The next IRT used is the 2-parameter logistic model (2PL) to see the difficulty level (b) and the distinguishing power of item (a). The 2PL model uses parameter a (distinguishing power) and parameter b (difficulty level) to estimate the ability of the test taker (θ). In this model, the parameter c (guess) is set to zero. The mathematical formula used in the 2PL model is as in (4).

$$P_i = \frac{e^{Dai(\theta-b)}}{1+e^{Dai(\theta-b)}} \quad (4)$$

In the formula, i show 1,2,3, ..., n. The P_i is the probability that the test taker has the ability (θ) to answer item 1, a_i shows the parameter of difference in item 1, b_i is the parameter of the difficulty level item 1, e is a constant number approaching 2.718, and D is a scale parameter [28]. The discriminatory power in the IRT was obtained showing the ability of the questions to distinguish high and low students' abilities. There are four categories of distinguishing points in IRT (poor, good enough, good, and very good). Grain quality is good if it has a good minimum differentiating power category. The categories used to describe the distinguishing power of items can be related to the range of parameter values are shown in Table 3 [29].

Table 3. Item distinguishing power category in the IRT

Category	Value Range
Not Good	< 0.20
Moderate	0.20-0.29
Good	0.30-0.39
Very Good	>0.40

The level of difficulty in the IRT is seen from the threshold value (b). There are five categories of problem difficulty levels based on the b value obtained starting from the very difficult, difficult, medium, easy, and very easy categories. The quality of the questions is said to be good if the difficulty level of the questions meets various levels by the proportions. Categories to see the level of difficulty with the criteria are shown in Table 4 [30].

Table 4. Item difficulty level category in IRT

Threshold value	Category
$b > 2$	Very difficult
$1 < b \leq 2$	Difficult
$-1 \leq b \leq 1$	Moderate
$-1 > b \geq -2$	Easy
$b > -2$	Very easy

3. RESULTS

The tryout questions for the 6th-grade science National Examination for elementary schools were made based on the grid issued by the National Education Unit (*Badan Nasional Sertifikasi Profesi/BNSP*). The science material tested consists of 17 main materials, namely the characteristics of living things and animal classification, the relationship between living things in the ecosystem, the reproduction and adjustment of living things, environmental preservation, parts of the plant or animal body and their functions, the life cycle of several animals, the human skeleton, the system. organs in humans, objects, force and motion, forms of energy and their changes, temperature and heat, sound and light, electricity and magnetism, natural resources, the water cycle, and the solar system. Each subject matter must be distributed in at least one item. There are 35 items in the sixth-grade Natural science National Examination tryout questions for elementary schools.

The distribution of science material is prepared based on material from class IV, class V, and class VI with the proportion of 20% class IV, 30% class V material, and 50% class VI material. Each material must be distributed in the question grid. The grid is used as a guide for making question items and answer keys. The results of the distribution of the material in items are shown in Table 5.

The distribution of material to items shows that each material is well distributed in each item so that this tryout question can be said to measure what should be measured by the material being taught. Next is the item analysis to determine the characteristics of the items seen from the item parameters including the difficulty level of the items and the distinguishing power analyzed by CTT and IRT.

Table 5. Description of the material in the item

No.	Subject matter	Item number	Total
1.	The relationship between living things in the ecosystem	1, 2, 3, 4, 5	5
2.	Plant/animal body parts and their functions	6	1
3.	The life cycle of some animals	7, 8, 9	3
4.	Organ systems in humans	10, 11, 12, 13, 14, 15, 17	7
5.	The reproduction and adaptation of living things	16	1
6.	Object	18, 19, 20, 21	4
7.	Force and motion	22, 23	2
8.	Energy forms and their changes	24, 25	2
9.	Temperature and heat	26, 27	2
10.	Sound and light	28	1
11.	Electricity and magnetism	29, 30, 31	3
12.	Natural resources	32, 33	2
13.	Water cycle	34	1
14.	Solar system	35	1

3.1. Classical test theory

Item parameters using CTT can determine the level of difficulty and distinguishing power. The item difficulty level is known by the percentage of test-takers who answered the questions correctly. Therefore, the BILOG-MG output looks at the pattern completion task (PCT) value. While the discriminating power of items is theoretically calculated based on the biserial correlation value so that the analysis results can be seen from the biserial correlation column. The characteristics of the difficulty level of the Natural science National Examination practice questions consist of moderate, easy, and difficult categories. The proportions of easy, medium, and difficult categories are also well distributed. The distinguishing characteristics of the questions generated are in the bad, good enough, and good categories. The characteristics of each item based on the category of difficulty level and distinguishing power are presented in Table 6.

Difficult levels of items are classified into easy, medium, and difficult items. The easiest item is item number 2 with a PCT value of 95.6. This means that 95.6% of students can answer correctly for this item. The most difficult item is item number 32 with a PCT value of 19.3. This means that only 19.3% of students were able to answer these items correctly. Two items have negative distinguishing power, namely items 21 and 29. This means that these items were answered correctly by students with low abilities and were

answered incorrectly by students with high abilities so that these items could not differentiate between students' abilities. The results of the difficulty level analysis of each item need to be grouped according to easy, medium, and difficult categories so that the proportions for each category can be seen. Of the 35 items, it can be classified based on the degree of difference in the grain as shown in Table 7.

Table 6. Characteristics of items based on CTT

No.	Name	PCT (%)	Difficulty level category	Correlation Biserial	Distinguishing power category
1.	ITEM0001	0.48	Moderate	0.27	Moderate
2.	ITEM0002	0.96	Easy	0.32	Good
3.	ITEM0003	0.84	Easy	0.41	Good
4.	ITEM0004	0.24	Difficult	0.07	Not good
5.	ITEM0005	0.84	Easy	0.54	Good
6.	ITEM0006	0.42	Moderate	0.27	Moderate
7.	ITEM0007	0.84	Easy	0.24	Moderate
8.	ITEM0008	0.72	Easy	0.22	Moderate
9.	ITEM0009	0.48	Moderate	0.24	Moderate
10.	ITEM0010	0.32	Moderate	0.23	Moderate
11.	ITEM0011	0.26	Difficult	0.31	Good
12.	ITEM0012	0.29	Difficult	0.01	Not good
13.	ITEM0013	0.57	Moderate	0.29	Moderate
14.	ITEM0014	0.80	Easy	0.18	Not good
15.	ITEM0015	0.86	Easy	0.26	Moderate
16.	ITEM0016	0.58	Moderate	0.12	Not good
17.	ITEM0017	0.53	Moderate	0.29	Moderate
18.	ITEM0018	0.33	Moderate	0.15	Not good
19.	ITEM0019	0.76	Easy	0.34	Good
20.	ITEM0020	0.35	Moderate	0.30	Moderate
21.	ITEM0021	0.34	Moderate	-0.03	Not good
22.	ITEM0022	0.84	Easy	0.31	Good
23.	ITEM0023	0.30	Difficult	0.19	Not good
24.	ITEM0024	0.46	Moderate	0.34	Good
25.	ITEM0025	0.62	Moderate	0.32	Good
26.	ITEM0026	0.88	Easy	0.13	Not good
27.	ITEM0027	0.59	Moderate	0.39	Good
28.	ITEM0028	0.74	Easy	0.26	Moderate
29.	ITEM0029	0.21	Difficult	-0.11	Not good
30.	ITEM0030	0.78	Easy	0.42	Good
31.	ITEM0031	0.22	Difficult	0.44	Good
32.	ITEM0032	0.19	Difficult	0.08	Not good
33.	ITEM0033	0.56	Moderate	0.20	Moderate
34.	ITEM0034	0.34	Moderate	0.16	Not good
35.	ITEM0035	0.27	Difficult	0.18	Not good

Table 7. Classification of items based on item difficulty level

No.	Level of difficulty	Item number	Item total
1.	Easy	2, 3, 5, 7, 8, 14, 15, 19, 19, 22, 26, 28, 30	13
2.	Moderate	1, 6, 9, 10, 13, 16, 17, 18, 20, 21, 24, 25, 27, 33, 34	15
3.	Difficult	4, 11, 12, 23, 29, 31, 32, 35	7

Based on the results of the difficulty level classification, it was found that 13 items had a difficulty Easy level. The number of items that have a moderate difficulty level is 15 items. While the remaining seven items are in the difficult category. Some of the questions are in the medium and easy categories. Only a small number of questions are in the difficult category. After being grouped by category, the percentage distribution of the difficult level is then made. This is necessary to see whether the division of the categories of the difficulty level is by the appropriate percentage. The percentage result for the easy category is 37%, for the medium category is 43% and for the difficult category is 20%. The results of the analysis of the distinguishing power of each item showed that there were items that had good, moderate, and bad distinguishing power. Based on its distinguishing power, an item is accepted if it has a minimum distinguishing power of moderate. Therefore, it is necessary to group items that have good distinguishing power and bad distinguishing power. The classification of distinguishing power is shown in Table 8.

Table 8 shows that most (77.14%) of the questions had a good distinguishing power category, so the questions could be accepted. Meanwhile, the remaining 22.86% of questions have no good distinguishing power. Items numbered 4, 18, 21, 26, 29, 32, 34, and 35 have poor discriminatory power, meaning they have a low level of discrimination. These points are less able to distinguish students with low abilities and students with high abilities.

Table 8. Category of item acceptance based on distinguishing power

No.	Receipt of items	Item number	Total	Percentage
1.	Discernment good	1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 22, 23, 24, 25, 27, 28, 30, 31, 33	27	77.14%
2.	Discrimination not good	4, 18, 21, 26, 29, 32, 34, 35,	8	22.86%

3.2. Item response theory

The IRT 1PL model is used to see the quality of the science National Examination practice questions based on the level of the Difficult category. The results of the analysis using BILOG-MG for analysis with IRT 1PL were obtained at the PH2 output. To find out the value of the difficulty level of the item seen from the threshold value. This parameter indicates the difficulty level category from very easy, easy, medium, difficult, and very difficult. The results of the calculation of the item difficulty level parameters based on the IRT IPL model can be seen in Table 9.

Table 9. Item difficult rate based on IRT model 1 IPL

Item	Threshold	Level of difficulty	Item	Threshold	Level of difficulty
001	0.13	Moderate	019	-2.44	Easy
002	-6.27	Very easy	020	1.27	Difficult
003	-3.36	Very easy	021	1.38	Difficult
004	2.35	Very difficult	022	-3.36	Very easy
005	-3.48	Very easy	023	1.80	Difficult
006	0.71	Moderate	024	0.33	Moderate
007	-3.42	Very easy	025	-1.06	Easy
008	-2.00	Easy	026	-4.10	Very easy
009	0.17	Moderate	027	-0.78	Moderate
010	1.53	Difficult	028	-2.17	Easy
011	2.13	Very difficult	029	2.77	Very difficult
012	1.88	Difficult	030	-2.57	Very Easy
013	-0.57	Moderate	031	2.62	Very difficult
014	-2.82	Very easy	032	2.96	Very difficult
015	-3.74	Very easy	033	-0.47	Moderate
016	-0.64	Moderate	034	1.34	Difficult
017	-0.23	Moderate	035	2.09	Very difficult
018	1.45	Difficult			

Based on Table 9, the threshold value varies from -6.27 to 2.96. A value of -6.27 indicates a very easy item and a value of 2.96 indicates a very difficult item. This is by the category of the difficulty level required in the IRT. Items in the very easy category, namely 2, 3, 5, 7, 14, 15, 22, 26, 30 with a total of nine items. Items in the easy category are 8, 19, 25, 28 with a total of four items. Items in the moderate category are 1, 6, 8, 13, 16, 17, 24, 27, 33 with a total of 9 items. The items in the difficult category are 10, 12, 18, 20, 21, 23, 34 with a total of 7 items, and the remaining items in the very difficult category are 4, 11, 29, 32, 32, 35 with a total of six items. The results of the difficult level analysis showed that the practice questions for the national science exam content consisted of very easy, easy, moderate, difficult, and very difficult questions. It is necessary to classify to determine the number of questions in each category. The number of questions from each category is then percentage to determine the proportion of the very easy category, the easy category, the moderate category, the difficult category, and the very difficult category. The distribution with difficult levels is very easy, moderate, difficult, and very difficult with the following percentages.

Figure 1 shows that most (26%) questions were in the moderate category. Only a few (17%) were in the very difficult category and 11% were in the easy category. Other questions are 20% in the difficult category and 26% in the very easy category. In theory, these results have shown a good proportion of the difficulty level of a problem. Next is the analysis using the 2PL measurement model (logistics parameter) to determine the level of difficulty and distinguishing power of grains. The level of difference is seen from the threshold value, the moderation of the grain distinction is seen from the slope value. The results of the analysis using BILOG-MG for analysis with IRT 2PL were obtained at the output of PH2 with the results as shown in Table 10.

⊘ Very Easy — Easy ▨ Medium ■ Difficult ⊠ Very Difficult

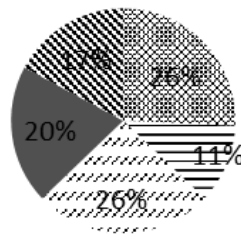


Figure 1. Distribution of item difficulty level with IRT model 1 IPL

Table 10. Item difference level and distinguishing power based on IRT model 2PL

Item	Slope	Discernment	Threshold	Item difference level	Item	Slope	Discernment	Threshold	Item difference level
001	0.41	Very good	0.09	Moderate	019	0.50	Very good	-1.61	Very good
002	0.64	Very good	-3.31	Easy	020	0.42	Very good	0.95	Very good
003	0.66	Very good	-1.80	Easy	021	0.20	Not good	2.04	Not good
004	0.26	Moderate	2.721	Very difficult	022	0.53	Very good	-2.08	Very good
005	0.89	Very good	-1.55	Easy	023	0.32	Good	1.71	Good
006	0.39	Good	0.56	Moderate	024	0.50	Very good	0.21	Very good
007	0.43	Very good	-2.50	Easy	025	0.48	Very good	-0.72	Very good
008	0.41	Very good	-1.53	Easy	026	0.39	Good	-3.24	Good
009	0.36	Good	0.13	Moderate	027	0.55	Very good	-0.49	Very good
010	0.36	Good	1.29	Difficult	028	0.46	Very good	-1.51	Very good
011	0.42	Very good	1.60	Difficult	029	0.20	Not good	4.08	Not good
012	0.21	Moderate	2.60	Very difficult	030	0.66	Very good	-1.38	Very good
013	0.43	Very good	-0.43	Moderate	031	0.63	Very good	1.44	Very good
014	0.35	Good	-2.44	Easy	032	0.26	Moderate	3.37	Moderate
015	0.45	Very good	-2.63	Easy	033	0.32	Good	-0.45	Good
016	0.27	Moderate	-0.71	Moderate	034	0.27	Moderate	1.47	Moderate
017	0.42	Very good	-0.19	Moderate	035	0.34	Good	1.89	Good
018	0.28	Moderate	1.53	Difficult					

Based on Table 10, it is found that the level of difficulty of the questions based on the threshold value is in the range of -3.31 and 4.08. The value of -3.31 indicates the threshold value for very easy questions and the value of 4.08 indicates the threshold value for the very difficult category. Based on this range, the difficulty level of the questions is in the very difficult, difficult, moderate, easy, and very easy categories. The items are very difficult, namely 4, 12, 21, 29, 32, as many as five items. Difficult items are seven items, namely 10, 11, 18, 23, 31, 34, 35. Moderate items are 10 items, namely 1, 6, 9, 13, 19, 20, 24, 25, 27, 33. Items are easy as many as nine items, namely 3, 5, 8, 14, 15, 22, 26, 28, 30. And for very easy items there are four items, namely 2, 7, 16, 17. The number of items based on the level of difficulty is then presented to see the proportion in each category. This aims to find out whether the questions made meet the proportion requirements for each category of difficulty level. The distribution of the percentage level of difficulty can be shown in Figure 2.

Figure 2 shows that most of the questions are in the moderate category, which is 29%. Questions that are in the easy category 26%. Problems in the difficult 20% category and only a small portion, namely 11% in the very easy category and 14% in the very difficult category. Based on Table 10, the slope value is in the range of 0.20 to 0.89. The value of 0.20 shows the lowest distinguishing power with the unfavorable category and the value of 0.89 shows the highest differentiating power with the very good category. Of the 35 items, it can be classified based on the distinguishing power of the items as shown in Table 11.

Very Easy Easy Medium Difficult Very Difficult

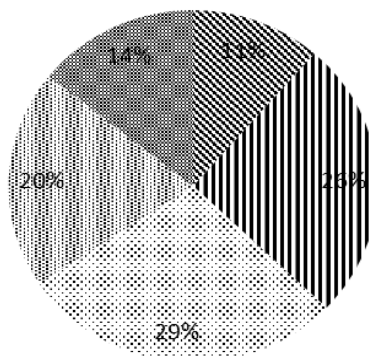


Figure 2. Distribution of difficulty levels of items with IRT model 2 IPL

Table 11. Item classification based on item distinguishing power in IRT model 2PL

No.	Discernment	Item number	Total
1.	Very good	1, 2, 3, 5, 7, 8, 11, 13, 15, 17, 19, 20, 22, 24, 25, 27, 28, 30, 31	19
2.	Good	6, 9, 10, 14, 23, 26, 33, 35	8
3.	Moderate	4, 12, 16, 18, 32, 34	6
4.	Not good	21, 29	2

Based on Table 11, 19 items have a very good distinguishing power category. The distinguishing power of very good is in the range of 0.41 to 0.89. The number of items in a good category, there are eight items located on the slope value with a range of 0.29 to 0.32. and for the moderate category, there are six items in the slope value range from 0.21 to 2.71. Only a small proportion have a poor category, namely two items with a slope value range of 0.14 and 0.20. The results of the quality analysis of the science exam practice questions with the CTT and IRT 1PL and 2PL models gave slightly different results. The item difficulty level on the IRT was more varied than the CTT. For discriminating power results also provide better results. The comparison of item parameter values for CTT and IRT is presented in Table 12.

Table 12. Comparison of grain parameter values for CTT and IRT

Item parameters	Value	Estimation method		
		CTT	IRT	
Level of difficulty	Mean	0.54	-0.47	-0.05
	Standard deviation	0.24	2.40	1.93
Discernment	Mean	0.24	-	0.42
	Standard deviation	0.14	-	0.15

Based on Table 12, the IRT difficulty level parameter for the 2PL model has the lowest average value. This suggests that the 1PL model provides the lowest possible index of item difficulty. On the other hand, the discriminatory power measured in the IRT revealed that the 2-parameter logistic model provided higher parameter values compared to the classical theory test (CTT). However, all parameters of the average difficulty level are in the moderate category and the distinguishing power values are in the moderate category. After knowing the characteristics of each item, the next step is to determine which model is most suitable for analyzing these items. Model fit is seen from the number of items that fit the model. For the CTT, it can be seen from the value of the distinguishing power of the items, if the distinguishing power of the items is in the moderate to a very good category, then these items can be accepted. Moderating in IRT, the fit of the model is tested using the Chi-Square value in the CHISQ column and p which indicates the probability. Items are said to fit the model if the probability value is $p > 0.05$ (significance level). Based on the results of the output on BILOG-MG, it can be concluded that the suitability of the items with the model is as shown in Table 13.

Table 13 explains that the number of items fit for each model is different. The use of CTT, IPL model, 2PL gives different results. For some items, for example, items numbered 4, 18, 26, 32, 34, and 35 do not fit in the CTT but the items fit the IPL and 2PL models. Items number 5 and 27 fit in CTT and 2PL models but do not fit in IPL models. Moderate item 29 does not fit in all models; therefore item 29 can be said to be of poor quality. Based on the results of the suitability of the model, the model that gave the highest value for the practice of the elementary school science exam was the 2PL model with the fit of the items to the 94.28 model. Therefore, this model is most suitable for use.

Table 13. Item fit with model

Model	Model fit	Item number	Total	Percentage
CTT	Item accepted (fit)	1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 22, 23, 24, 25, 27, 28, 30, 31, 33	27	77.14%
	Item not accepted (not fit)	4, 18, 21, 26, 29, 32, 34, 35,	8	22.86%
IPL	Fit	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 30, 31, 32, 33, 34, 35	32	91.42%
	Not fit	5, 27, 29	3	8.58%
2PL	Fit	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35	33	94.28%
	Not fit	21, 29	2	5.71%

4. DISCUSSION

The results of calculating the quality of science exam practice questions with CTT gave PCT results between 19.3 and 95.6. These results indicate the level of difficulty is in the level of difficult to easy. PCT 19.3 showed that only 19.3% of students were able to answer the questions correctly or the difficulty index was 0.19. For the easiest question, the PCT score is 95.6%, indicating that 95.6% of students answered the question correctly or the index of difficulty was 0.96. Based on the difficulty level of each item, 37% of the items are in the easy category, 43% of the items are in the medium category and 20% of the items are in a low category. When the difficulty index is equal to 0 it indicates that all students answered the questions incorrectly on the item, and if the difficulty index is equal to 1, it indicates that all students answered correctly for the item [31]. In this study, no items were having a difficulty level of 0 and no items having a difficulty level of 1. The ideal level of difficulty in multiple-choice items was between 0.33 and 0.77 [32]. Based on this theory, the number of items that have an ideal level of difficulty is 15 items. And the remaining items are still in the very easy and very difficult categories so that ideally it still needs to be improved both on the subject matter or the answer choices. Based on the distribution of easy, medium, and difficult items, the questions are not appropriate because ideally the questions are arranged in order from the easiest to the most difficult questions [33].

The use of the IRT model in analyzing student responses shows that there is an effect of the diversity of participant characteristics on test items (multiple choice) in terms of the accuracy of the item parameter estimates and individual ability parameters [34]. With IRT for the 1PL model, it gives a difficulty level of -6.27 to 2.77 in the very easy to a very difficult category. A value closer to -2 logit indicates an easier item, and a value close to +2 logit indicates a more difficult item [35]. To classify items - test items into good, moderately good, or bad items, the following criteria are used for good test items, the value of the discrimination parameter must be greater than or equal to 1 while the difficulty parameter value must be from 0.5 to +1, the item any test with a score above +1 will be considered difficult and if it is less than 0.5 it is considered an easy item [36]. In the results of the analysis, seven items are smaller than -2 and there are five items whose value exceeds +2 So it can be said that the 12 questions do not meet the requirements seen from the level of item difficulty.

The use of IRT model 2PL gives a value of difficulty level between -3.31 and 4.08 with very easy to very difficult categories. The difference in value is only due to a different scale. However, the easiest items and the most difficult items are the same for estimation with IRT model IPL and 2PL model. The easiest item is item number 2 and the most difficult item is item number 29. There is a slight difference in the value of the difficulty level with the IPL model. This is by the results of the study [37] that the analysis of the difficulty level of the values changes with the increase in the model used. There is some variation in item difficulty levels between IRT and CTT when compared [38]. The estimation with CTT for the most difficult item is item number 29 and the easiest item is item number 32. The results of this study is in line with previous research [39] that there is a statistically significant mean difference between the difficulty of CTT-based items and IRT.

For the parameter of discriminating power of items, estimation with CTT gives the result that the items are in the category of discriminatory power of poor to very good. Items with good discriminating power are items that tend to be answered correctly by students who do well on the test, in this case, high-ability students [40]. Items with a low discrimination index indicate the presence of ambiguous words in the item and items with a negative index should be examined further to determine why a negative value is obtained [41]. There are eight items with poor discriminatory power. The item discriminatory power parameters based on estimates with the IRT model 2PL yielded only two items with poor discriminating power categories. The discriminatory power of items in the CTT is highly dependent on the ability of the test taker. The estimation of discriminating power parameters is more accurate when using a sample of low-ability individuals [42].

Model fit analysis was used to determine the most appropriate model used to analyze the test. The results of item analysis provided by CTT and IRT were almost the same, but IRT provided additional item statistics and a more sophisticated calculation method to minimize measurement errors [43]. The results of another study [44] showed that to analyze the parameters of the items in terms of item difficulty, item discrimination, and responses given by students for each item, the CTT, and IRT (2PL) models can be used. Supporting these findings, the results of this study show that the 2PL model IRT provides more fit items than the 1PL and CTT model IRT so that this model is the most appropriate for estimating the item parameters in elementary school science practice questions. These findings similar with previous research [45] that the estimation of item characteristics with CTT removes more items than the estimation of item characteristics with IRT. IRT is mostly used to model student responses to test items and educational test assessments, teachers should be encouraged to use IRT in developing test items [46]. By conducting item suitability analysis, in the preparation of the questions, one of the most suitable models can be selected so that it will produce items with the best quality.

5. CONCLUSION

The quality of the items on the National Examination practice for science subjects is in a good category. The items are in the easy to difficult difficulty level category and only two questions have negative discriminating power. The quality of the items seen from the IRT IPL model is also in the good category. Difficulty levels vary from very easy to very difficult. The distinguishing power of the elementary school science exam practice questions is good. The item is categorized by the difficulty level of the item difficult in the CTT, it will also correspond to the IRT classification which is almost the same for the item difficulty level. The same can be said for the discriminatory categorization of items between the CTT and IRT approaches. Of course, this research has been done by comparing the CTT and IRT with various logistic parameter models. This study provides information about the characteristics of the elementary school science exam practice items that have not been previously analyzed. This study contributes to the compilers of the following year's tryout to consider the characteristics of the items used both in terms of CTT and IRT. Recommendations for compilers of exam questions or practice exams in elementary schools, before the questions are used, it is better to conduct trials with several respondents. This research is limited to the analysis of the quality of the multiple-choice National Examination practice questions. This study has not investigated further the function of the distractor, the student's abilities, and the test information obtained. The analysis carried out is still limited to the CTT and IRT of the IPL and 2PL models so that it can be further analyzed with the 3PL model and for more detailed item quality parameters.

ACKNOWLEDGEMENTS

The authors would like to thank you for funding this research. This work was supported by the teachers in primary school in Banyumas Regency.

REFERENCES




- [1] Winarno, M. Zuhri, Mansur, I. Sutomo, and K. Widhyahrini, "Development of assessment for the learning of the humanistic model to improve evaluation of elementary school mathematics," *International Journal of Instruction*, vol. 12, no. 4, pp. 49–64, 2019, doi: 10.29333/iji.2019.1244a.
- [2] C. Alonso-Fernández, I. Martínez-Ortiz, R. Caballero, M. Freire, and B. Fernández-Manjón, "Predicting students' knowledge after playing a serious game based on learning analytics data: A case study," *Journal of Computer Assisted Learning*, vol. 36, no. 3, pp. 350–358, 2020, doi: 10.1111/jcal.12405.
- [3] B. Jia, D. He, and Z. Zhu, "Quality and Feature of Multiple-Choice Questions in Education," *Problems of Education in the 21st Century*, vol. 78, no. 4, pp. 576–594, 2020, doi: 10.33225/pec/20.78.576.
- [4] W. Maba, I. B. K. Perdata, and I. N. Astawa, "Constructing assessment instrument models for teacher's performance, welfare and education quality," *International Journal of Social Sciences and Humanities*, vol. 1, no. 3, pp. 88–96, 2017, doi: 10.29332/ijssh.v1n3.59.

- [5] C. Türk and H. Kalkan, "Developing an Achievement Test for Astronomy Education," *Journal of Studies in Education*, vol. 5, no. 3, p. 89, 2015, doi: 10.5296/jse.v5i3.8099.
- [6] Z. Zainuddin, "Students' learning performance and perceived motivation in gamified flipped-class instruction," *Computers and Education*, vol. 126, no. 1, pp. 75–88, 2018, doi: 10.1016/j.compedu.2018.07.003.
- [7] H. Putranta and Supahar, "Development of physics-tier tests (PysTT) to measure students' conceptual understanding and creative thinking skills: A qualitative synthesis," *Journal for the Education of Gifted Young Scientists*, vol. 7, no. 3, pp. 747–775, 2019, doi: 10.17478/jegys.587203.
- [8] A. Dehnad, H. Nasser, and A. F. Hosseini, "International conference on current trends in ELT a comparison between three-and four-option multiple-choice questions," *Procedia - Social and Behavioral Sciences*, vol. 98, no. 1, pp. 398–403, 2014, doi: 10.1016/j.sbspro.2014.03.432.
- [9] E. Ulitzsch, M. von Davier, and S. Pohl, "A Multiprocess Item Response Model for Not-Reached Items due to Time Limits and Quitting," *Educational and Psychological Measurement*, vol. 80, no. 3, pp. 522–547, 2020, doi: 10.1177/0013164419878241.
- [10] K. Yamamoto, H. J. Shin, and L. Khorramdel, "Multistage Adaptive Testing Design in International Large-Scale Assessments," *Educational Measurement: Issues and Practice*, vol. 37, no. 4, pp. 16–27, 2018, doi: 10.1111/emip.12226.
- [11] H. Putranta and Jumadi, "Physics teacher efforts of Islamic high school in Yogyakarta to minimize students' anxiety when facing the assessment of physics learning outcomes," *Journal for the Education of Gifted Young Scientists*, vol. 7, no. 2, pp. 119–136, 2019, doi: 10.17478/JEGYS.552091.
- [12] D. S. Christian, A. C. Prajapati, B. M. Rana, and V. R. Dave, "Evaluation of multiple choice questions using item analysis tool: a study from a medical institute of Ahmedabad, Gujarat," *International Journal Of Community Medicine And Public Health*, vol. 4, no. 6, p. 1876, 2017, doi: 10.18203/2394-6040.ijcmph20172004.
- [13] S. Lin, "Item Analysis of English Grammar Achievement Test," *Mandalay University of Foreign Languages Research Journal*, vol. 9, no. 1, pp. 13–20, 2018.
- [14] P. U. Osadebe and M. O. Jessa, "Development of social studies achievement test for assessment of secondary school students," *European Journal of Open Education and E-learning Studies*, vol. 3, no. 1, pp. 104–124, 2018.
- [15] I. Himelfarb, "A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating," *Journal of Chiropractic Education*, vol. 33, no. 2, pp. 151–163, 2019, doi: 10.7899/JCE-18-22.
- [16] P. U. Osadebe, "Construction of Economics Achievement Test for Assessment of Students," *World Journal of Education*, vol. 4, no. 2, pp. 58–64, 2014, doi: 10.5430/wje.v4n2p58.
- [17] H. Okonkwo *et al.*, "A blinded clinical study using a subepidermal moisture biocapacitance measurement device for early detection of pressure injuries," *Wound Repair and Regeneration*, vol. 28, no. 3, pp. 364–374, 2020, doi: 10.1111/wrr.12790.
- [18] H. D. Gibbs, E. F. Ellerbeck, B. Gajewski, C. Zhang, and D. K. Sullivan, "The Nutrition Literacy Assessment Instrument is a Valid and Reliable Measure of Nutrition Literacy in Adults with Chronic Disease," *Journal of Nutrition Education and Behavior*, vol. 50, no. 3, pp. 247–257.e1, 2018, doi: 10.1016/j.jneb.2017.10.008.
- [19] R. Gortler, J. P. Fox, I. Eekhout, M. W. Heymans, and J. W. R. Twisk, "Missing item responses in latent growth analysis: Item response theory versus classical test theory," *Statistical Methods in Medical Research*, vol. 29, no. 4, pp. 996–1014, 2020, doi: 10.1177/0962280219897706.
- [20] R. A. Feinberg and M. von Davier, "Conditional Subscore Reporting Using Iterated Discrete Convolutions," *Journal of Educational and Behavioral Statistics*, vol. 45, no. 5, pp. 515–533, 2020, doi: 10.3102/1076998620911933.
- [21] C. R. Reynolds, R. A. Altmann, and D. N. Allen, "Item Analysis: Methods for Fitting the Right Items to the Right Test," *Mastering Modern Psychological Testing*, vol. 1, no. 1, pp. 263–289, 2021, doi: 10.1007/978-3-030-59455-8_7.
- [22] D. G. Bonett, "Point-biserial correlation: Interval estimation, hypothesis testing, meta-analysis, and sample size determination," *British Journal of Mathematical and Statistical Psychology*, vol. 73, no. S1, pp. 113–144, 2020, doi: 10.1111/bmsp.12189.
- [23] L. A. Shepard, W. R. Penuel, and J. W. Pellegrino, "Using Learning and Motivation Theories to Coherently Link Formative Assessment, Grading Practices, and Large-Scale Assessment," *Educational Measurement: Issues and Practice*, vol. 37, no. 1, pp. 21–34, 2018, doi: 10.1111/emip.12189.
- [24] D. A. Tan, C. C. Cordova, I. P. B. Saligumba, and L. L. B. Segumpan, "Development of Valid and Reliable Teacher-Made Tests for Grade 10 Mathematics Denis," *International Journal of English and Education*, vol. 8, no. 1, pp. 62–83, 2014.
- [25] H. K. Mohajan, "Qualitative Research Methodology in Social Sciences and Related Subjects," *Journal of Economic Development, Environment and People*, vol. 7, no. 1, p. 23, 2018, doi: 10.26458/jedep.v7i1.571.
- [26] Z. Kaya and S. Tan, "New trends of measurement and assessment in distance education," *Turkish Online Journal of Distance Education*, vol. 15, no. 1, pp. 206–217, 2014, doi: 10.17718/tojde.30398.
- [27] C. Fu, D. L. Xu, and M. Xue, "Determining attribute weights for multiple attribute decision analysis with discriminating power in belief distributions," *Knowledge-Based Systems*, vol. 143, no. 1, pp. 127–141, 2018, doi: 10.1016/j.knsys.2017.12.009.
- [28] E. Latipah, H. C. A. Kistoro, and H. Putranta, "The Effects of Positive Parenting toward Intolerance in Pre-School Children," *International Journal of Early Childhood Special Education*, vol. 12, no. 2, pp. 137–146, 2020, doi: 10.9756/INT-JECSE/V12I2.201065.
- [29] C. K. Jaggi, L. E. Cárdenas-Barrón, S. Tiwari, and A. A. Shafi, "Two-warehouse inventory model for deteriorating items with imperfect quality under the conditions of permissible delay in payments," *Scientia Iranica*, vol. 24, no. 1, pp. 390–412, 2017, doi: 10.24200/sci.2017.4042.
- [30] H. Retnowati, *Item response theory and its application*. Nuha Medika (in Indonesian), 2014.
- [31] J. M. Azevedo, E. P. Oliveira, and P. D. Beites, "Using Learning Analytics to evaluate the quality of multiple-choice questions: A perspective with Classical Test Theory and Item Response Theory," *International Journal of Information and Learning Technology*, vol. 36, no. 4, pp. 322–341, 2019, doi: 10.1108/IJILT-02-2019-0023.
- [32] H. C. A. K. Kistoro, B. K. Kartowagiran, E. L. Latipah, N. N. Naim, and H. P. Putranta, "Implementation of Islamic religious learning strategies in children with autism in Indonesia," *Specijalna Edukacija i Rehabilitacija*, vol. 19, no. 4, pp. 227–246, 2020, doi: 10.5937/SPECEDREH19-28813.
- [33] M. Kusumawati and S. Hadi, "An analysis of multiple choice questions (MCQs): Item and test statistics from mathematics assessments in senior high school," *Research and Evaluation in Education*, vol. 4, no. 1, pp. 70–78, 2018, doi: 10.21831/reid.v4i1.20202.
- [34] D. Almaleki, "Examinee Characteristics and their Impact on the Psychometric Properties of a Multiple Choice Test According to the Item Response Theory (IRT)," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 6889–6901, 2021, doi: 10.48084/etasr.4056.
- [35] J. Mailool, B. Kartowagiran, T. H. Retnowati, S. Wening, and H. Putranta, "The effects of principal's decision-making, organizational commitment and school climate on teacher performance in vocational high school based on teacher perceptions," *European Journal of Educational Research*, vol. 9, no. 4, pp. 1675–1687, 2020, doi: 10.12973/EU-JER.9.4.1675.




- [36] O. O. Adedoyin and T. Mokobi, "Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items," *International Journal of Asian Social Science*, vol. 3, no. 4, pp. 992–1011, 2013, [Online]. Available: <https://archive.aessweb.com/index.php/5007/article/view/2471>.
- [37] H. B. Yılmaz, "A comparison of IRT model combinations for assessing fit in a mixed format elementary school science test," *International Electronic Journal of Elementary Education*, vol. 11, no. 5, pp. 539–545, 2019, doi: 10.26822/iejee.2019553350.
- [38] J. L. Pimentel and M. L. A. Villaruz, "Comparison of item difficulty estimates in a basic statistics test using ltm and CTT software packages in R," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 3, pp. 367–372, 2020, doi: 10.14569/ijacsa.2020.0110346.
- [39] S. Soysal and E. Yılmaz Koğar, "An Investigation of Item Position Effects by Means of IRT-Based Differential Item Functioning Methods," *International Journal of Assessment Tools in Education*, vol. 8, no. 2, pp. 239–256, 2021, doi: 10.21449/ijate.779963.
- [40] A. Gero and Y. Stav, "Summative assessment based on two-tier multiple-choice questions: Item discrimination and engineering students' and teachers' attitudes," *International Journal of Engineering Education*, vol. 37, no. 3, pp. 830–840, 2021.
- [41] Z. A. Ashraf and K. Jaseem, "Classical and Modern Methods in Item Analysis of Test Tools," *International Journal of Research and Review*, vol. 7, no. 5, pp. 397–403, 2020, doi: 10.5281/zenodo.3938796.
- [42] H. S. Al-zboon, "The Effect of the Multiple-Choice Test Length on Estimating the Item Parameters and the Test Information Function According to the Three-Parameter Logistic Model in the Item Response Theory," *Journal of Economics and Sustainable Development*, vol. 11, no. 24, pp. 42–50, 2020, doi: 10.7176/jep/11-24-05.
- [43] A. A. Bichi, R. Embong, R. Talib, S. Salleh, and A. Bin Ibrahim, "Comparative Analysis of Classical Test Theory and Item Response Theory using Chemistry Test Data," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 5 C, pp. 1260–1266, 2019, doi: 10.35940/ijeat.E1179.0585C19.
- [44] B. Subali, Kumaidi, N. S. Aminah, and B. Sumintono, "Student achievement based on the use of scientific method in the natural science subject in elementary school," *Jurnal Pendidikan IPA Indonesia*, vol. 8, no. 1, pp. 39–51, 2019, doi: 10.15294/jpii.v8i1.16010.
- [45] M. A. Ayanwale, J. O. Adeleke, and T. I. Mamadelo, "An Assessment of Item Statistics Estimates of Basic Education Certificate Examination through Classical Test Theory and Item Response Theory approach," *International Journal of Educational Research Review*, vol. 3, no. 4, pp. 55–67, 2018, doi: 10.24331/ijere.452555.
- [46] N. Esomonu and O. J. Okeke, "French Language Diagnostic Writing Skill Test for Junior Secondary School Students: Construction and Validation Using Item Response Theory," *International Journal of Education and Social Science Research*, vol. 04, no. 02, pp. 334–350, 2021, doi: 10.37500/ijessr.2021.4227.

BIOGRAPHIES OF AUTHORS



Nurul Hamidah    is a graduate of the Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta, Jalan Colombo No. 1 Yogyakarta, Yogyakarta 55281, Indonesia. She is now an Elementary School Teacher in Banyumas Regency, Central Java, Indonesia. Her research focuses on education and learning in primary schools. She can be contacted at email: nurulhamidah.pep.2017@student.uny.ac.id.



Edi Istiyono    is a professor in the field of educational research and evaluation in the field of physics studies. He is a senior lecturer at Universitas Negeri Yogyakarta, Jalan Colombo No. 1 Yogyakarta, Yogyakarta 55281, Indonesia. He studied physics majoring at Universitas Negeri Yogyakarta in 1992 and continued his master's education at Universitas Gadjah Mada in 1999 and graduated from research and evaluation doctoral education at Universitas Negeri Yogyakarta in 2014. His research is in accordance with his expertise in the field of physics education. He can be contacted at email: edi_istiyono@uny.ac.id.