JPPIPA 9(10) (2023)

Jurnal Penelitian Pendidikan IPA

Journal of Research in Science Education

http://jppipa.unram.ac.id/index.php/jppipa/index

Analysis of Item Characteristics of Natural Sciences National Examinations for Junior High School Based on the Classical Test Theory Approach

Mariana[1*], Djaffar Lessy[1], Dinar Riaddin[1], Muhammad Rizal Hardiansyah[2], Corneli Pary[3]

[1]Mathematics Education, FITK, IAIN Ambon, Maluku, Indonesia
[2]Tadris IPA, FITK, IAIN Ambon, Maluku, Indonesia
[3]Biology Education, FITK, IAIN Ambon, Maluku, Indonesia

Received: April 18, 2023
Revised: September 15, 2023
Accepted: October 25, 2023
Published: October 31, 2023

Corresponding Author:
Mariana
marianaiain17@gmail.com

DOI: 10.29303/jppipa.v9i10.3698

Abstract: This study aims to analyze the characteristics of the items based on the classical test theory approach. This research is exploratory research with a quantitative approach. The data used in this study were data from the 2015 Science Subject National Examination for Junior High School in Daerah Istimewa Yogyakarta (DIY) province. The samples taken in this study were 9054 students. Test takers are grouped into high groups and low groups based on the total score obtained. Estimation of item parameters using Excel and QUEST. The results of the analysis show that based on the classical test theory approach, the average item difficulty level $\bar{P}$) in the high group is in the easy category, while in the low group, it is in the medium category, the average item differential index $(\bar{r}_{pbis})$ shows that the test ability to distinguish students with high abilities and low abilities in both the high and low groups is in the unfavorable category; and the reliability coefficient of the test (ρ) in the high group has high reliability, and the low group has perfect reliability.

Keywords: Classical test theory; Item characteristics; National examinations; Natural science

Introduction

Evaluation is part of a series of activities to improve the quality and productivity or performance of an institution in implementing its program (Mardapi, 2012). Measurement and assessment are important steps or actions that must be taken to carry out evaluation activities because all evaluation actions must be preceded by measurement and assessment activities. Measurement is the process of giving numbers that are expected to show students' abilities about a subject; While assessment is the activity of interpreting or describing measurement results.

In the measurement process, a measuring instrument is required. This measuring instrument provides information about a person's position in the measured attribute, so that to obtain measurement results that can describe the actual measurement results, a measuring instrument with a high level of validity and reliability is needed. Correspondingly, (Retnawati, 2016; Widiatmo et al., 2019) the condition of an instrument is said to be suitable for use as a measuring instrument if it is based on empirical facts and theoretical reasons to produce conclusions. A good instrument is an instrument that can produce data and provide accurate information so that the information obtained from the measurement results describes the actual ability of students.

In particular, educators can measure the extent of students' abilities and understanding of learning using tests. Giving tests is an essential factor in achieving learning goals (Ali & Istiyono, 2022; Anggoro et al., 2019; Cohen & Swerdlik, 2009; Kaplan & Saccuzzo, 2009). A test is said to be good if it has good items on the questions given. There are many types of questions in the test: complete, multiple-choice, true-false, open-

How to Cite:
Mariana, M., Lessy, D., Riaddin, D., Hardiansyah, M.R., & Pary, C. (2023). Analysis of Item Characteristics of Natural Sciences National Examinations for Junior High School Based on the Classical Test Theory Approach. *Jurnal Penelitian Pendidikan IPA*, *9*(10), 7837–7844. https://doi.org/10.29303/jppipa.v9i10.3698

ended, short answer, and descriptive or persuasive. To measure the quality of the items listed in the test, item analysis is essential to ensure that the quality of the items is relevant to standards and the quality of constructive alignment in designing the test.

An important goal of analyzing questions is to produce good quality questions by reviewing the questions given, to help identify deficiencies in the test. This can confirm students' abilities and understanding of the material studied (Anastasi & Urbina, 1997; Aiken, 1997; Hussin et al., 2018). Analysis of test items can help improve understanding and can predict several criteria, such as the validity and reliabilty of a test (Murphy & Davidshofer, 2005). One test that is worthy of analyzing the quality of the questions is the national examination test. The quality of national exam questions plays a very important role in efforts to identify students' mastery of competencies and their difficulties (Retnawati et al., 2017).

The results of the national examination in 2015 until the last in 2019 are still used as standards in determining the graduation of students, therefore the question preparation team must be able to compile instrument question items so that they have a high level of validity and have good differentiation. A good condition is not to have a tendency to influence respondents to choose certain responses. In addition, the data must also be reliable, so that the instrument is able to produce data that is also reliable (Sugiyono, 2011); Ardi et al., 2023). The results of the 2015 SMP/MTs National Examination in DIY province placed science subjects with the second lowest score after Mathematics, with an average science score of 48.08 and an average score in mathematics of 45.06. Therefore, it is necessary to analyze what is best used so as to be able to measure the ability of students.

The activity of analyzing question items is the process of collecting, summarizing, and using information from students' answers to make decisions about each assessment. The implementation of the analysis can be done in two ways, namely based on the approach of classical test theory (CTT) and item response theory (IRT). The CTT calculates scores based on the items that subjects answer correctly on the test (French, 2001; Hu et al., 2021). Classical test theory (CTT) has been widely developed for 20 decades (Embretson & Reise, 2000). Although CTT has the disadvantage of being examinee sample dependent and item sample dependent (Fan, 1998; Hambleton, Ronald K Swaminathan, 1985; Hambleton, R. K Swaminathan & Rogers, 1991; Hambleton et al., 2000; Lord, 1980) to this day CTT is still mainstream among psychologists and educationalists, as well as other fields of behavioral studies. Analysis of question items based on empirical data with a classical theory approach, namely item difficulty level, difference index, and reliability test.

The difficulty level of the item symbolized by P is one of the parameters of the question item which is very important and useful in terms of analyzing a test. This is because by looking at the grain parameters, it will be known how good the quality of a question item is. If P_I is close to 0, then the problem is too difficult, while if P_I is close to 1, it indicates the problem is very easy, so it must be discarded. This happens because these items are not able to distinguish the abilities of a student and another student (Retnawati, 2016).

The difficulty level of a question item is the opportunity to correctly answer a question at a certain ability level which is generally or usually expressed in the form of an index. This difficulty index is generally expressed in the form of proportions whose magnitude ranges from 0.00–1.00. Allen & Yen (1979) state that generally the item difficulty index of a question should be located in the interval 0.3 – 0.7. At this interval, information about the student's abilities will be maximally obtained. When designing the index of difficulty of a test device, it is very necessary to consider the purpose of preparing the test device.

The calculation of this difficulty index is performed for each item question. To determine the difficulty index of an item on a multiple-choice test device (dichotomous scoring), the following equation is used (Nitko, 1996).

$$\mathrm{P}_i = \frac{\sum B}{N} \tag{1}$$

with:
P_i = proportion of correct answers to a particular question item (difficulty level)
$\sum B$ = the number of test takers who answered correctly
N = The number of test takers who answered the item

The greater the P value, that is, the greater the proportion of the test takers in answering correctly, the question is considered easy. The smaller the P-value, the more difficult the problem. A good question should not be too easy or too difficult (Aldyza & Andani, 2018).

In classical test theory, the difficulty level of the questions depends on the ability of the examinees. For highly skilled examinees, question points are easy. For low-ability examinees, the question points become difficult. In the easy question items, it appears that the examinee's ability is high. Meanwhile, on difficult question items, the ability of the examinee becomes low. Therefore, the difficulty level of the question items does not fully describe the size of the characteristics of the actual question items, but rather the average ability of the group of examinees.

Question differentiating power is an index value that indicates the ability of question items to distinguish groups of high-ability and low-ability examinees. The distinguishing power of an item of this question is based on the test results of one group so the discriminating power does not necessarily apply to other groups. The differentiating power index ranges from -1.00 to 1.00. The higher the value of the discriminating power of the question, the better the problem.

To determine the distinguishing power of the problem, a discrimination index, a biserial correlation index, a biserial point correlation index, and an alignment index can be used. In the item analysis in this study, only the biserial point correlation index was used. Its correlation coefficient for a test item is determined by the formula:

$$r_{pbis} = \left[\frac{\bar{X}_1 - \bar{X}}{s_x}\right]\sqrt{\frac{p_1}{1-p_1}} \tag{2}$$

Where:

r_{pbis} is biserial point correlation coefficient, X_i is a continuous variable,$\bar{X}_1$ is the average X score for the test taker who answered correctly the item, $\bar{X}$ is the average X score, s_x is the standard deviation from the X score, and p_1 is the proportion of the test takers answering the item correctly.

In an item in the test, the differential power index is called good if the value is greater than or equal to 0.3. The distinguishing power index of an item that is small in value will cause the item to be unable to distinguish students with high ability and students with low ability (Etobro & Fabinu, 2017; Jailani & Almukarramah, 2020; (Retnawati et al., 2017).

In an instrument used for data collection, the reliability of the test result scores is very useful information in test development. Mehrens & Lehmann, (1973) states that reliability is a degree of consistency between two measurement results on the same object, even though using different measuring devices and different scales.

In education, measurements cannot be directly made on the traits or characters being measured. This characteristic or character is abstract and can be measured through an indicator. This makes it difficult to obtain a stable measuring instrument to measure a person's characteristics. This stability is said to be reliable. Prove the reliability of measuring instruments in the form of a value that can be done using statistical calculations. This value is commonly called the reliability coefficient.

The reliability coefficient can be interpreted as an index of the reliability or stability of measurement results. A reliable measuring instrument will provide a stable measurement result (Lawrence, 2019) and consistency (Mehrens & Lehmann, 1973). This means that a measuring instrument is said to have a high-reliability coefficient when it measures the same thing, so the results are still the same or close to the same even though it is measured at different times.

Allen & Yen (1979) stated that a test is reliable if the observation score has a high correlation with the actual score. Furthermore, it is stated that the reliability coefficient is the correlation coefficient between two intensity scores obtained from the measurement results using parallel tests. Thus, a test is said to be reliable if the measurement results obtained are close to the actual conditions of the test takers.

The reliability (ρ) of a test is generally expressed numerically in terms of coefficients of -1.00 ≤ ρ ≤ ±1.00. A high coefficient indicates high reliability. Conversely, if the coefficient of a test score is low then the reliability of the test is low. If the reliability is perfect, then the reliability coefficient is ±1.00. The expectation is that the reliability coefficient is positive.

Reliability is also related to measurement error. High reliability means that it shows a small error in obtaining measurement results. The greater the reliability of an instrument, the smaller the measurement error, and vice versa, the smaller the reliability of the score, the greater the measurement error. Reliability can be calculated by the alpha coefficient, in (Crocker & Algina, 1986) defined as follows:

$$\rho = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma_i{}^2}{\sigma_x{}^2}\right] \tag{3}$$

Information:
n = number of question items, ,,
$\sigma_i{}^2$= score variance per question item,
$\sigma_x{}^2$ = total score variance

Based on the problems that have been described, one solution to determine the quality of national examination tests is to analyze the characteristics of the test items using a classical test theory approach, so that items can be produced that have good quality and are reliable.

Method

This research is an exploratory study that uses a quantitative approach in order to estimate item parameters on the test device. For data adequacy, this study used data from the results of the National Examination for Natural Science Subjects of Junior High School DIY Province in 2015. The instrument used was the Science National Examination questions, which consisted of 40 multiple-choice questions. The number of test takers was 16765 people, then for needs analysis, a sample of 9054 students was taken. Analysis of item parameters difficulty (P) was performed with Excell, and analysis of different power indexes (r_{pbis}) and score reliability (ρ) using the QUEST program.

Result and Discussion

The results of this study are intended to describe the characteristics of UN question items for junior high school science subjects for the 2015 academic year. The scores of students science scores in the National Science Examination can be seen in Table 1. The maximum score obtained by the examinee is 40.00, the minimum score is 8.00, and the average score of participants can answer the questions correctly as many as 27 questions, with a median score (middle score) of 28.00 and a standard deviation score of 7.83.

Table 1. Student score statistics of national examinations natural science subjects SMP/MTs DIY

Statistics	Student Grade Score
Mean score	27.06
Standard deviation	7.83
Minimum score	8.00
Maximum score	40.00
Median	28.00

To analyze the characteristics of the items, the data is divided into high and low groups based on the total score obtained. Furthermore, the 40 items on the national science exam were then analyzed for their difficulty level, item discriminating power index, and reliability index, the results of which can be seen in Table 2. Meanwhile, to look at the characteristics of the questions based on the number of difficulty levels and the discriminating power of the questions in the high group and low test takers can be seen in Figures 1 and 2.

Table 2. High Group and Low Group Item Parameter Summary

Group	Item parameters	Question criteria
High group	$\bar{P}$ = 0.904	Easy
	$\bar{r}_{pbis}$ = 0.133	Not good
	ρ = 0.990	High reliability
Low group	$\bar{P}$ = 0.416	Medium
	$\bar{r}_{pbis}$ = 0.143	Not good
	ρ = 1.000	Perfect reliability

From table 2, it can be seen that the characteristics of the items in the high group have an average level of item difficulty that is quite easy but not good at differentiating the abilities of the test takers, but on the other hand in the low group, the average item difficulty level is in the medium category and also has poor grain discrepancy. however, both groups have high-reliability coefficients.

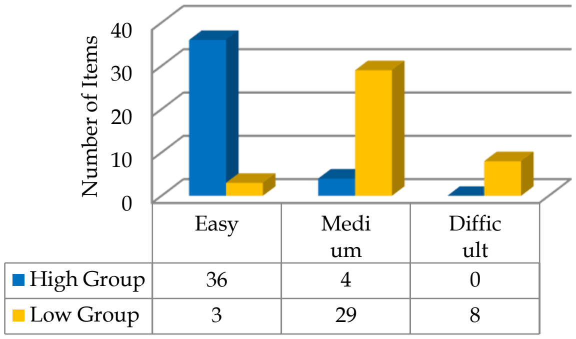


Figure 1. High group and low group item difficulty

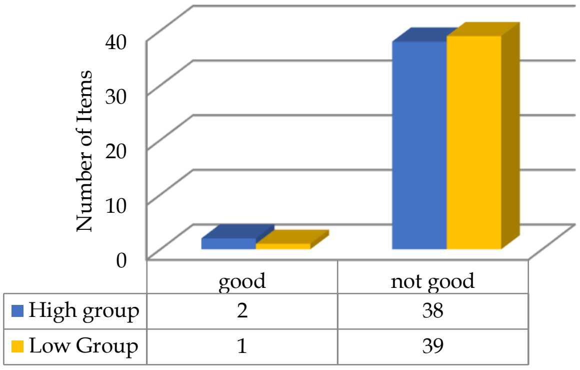


Figure 2. High group and low group difference power

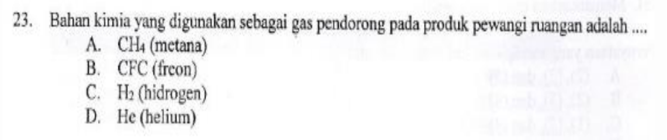
Based on Figures 1 and 2 above, it is known that the test takers in the high group answered the test questions, while the test takers in the low difficulty group were in the medium category, and there were even 8 questions in the difficult category. This follows research that has shown that questions should not be too easy or too difficult, because questions that are too easy can reduce students' interest (stimulation) in learning, whereas questions that are too difficult can make students less interested in learning enthusiastic (Aldyza & Andani, 2018).

Characteristics of Test Items in High Groups

Based on Table 3, the results of the analysis of item characteristics in the high group obtained the difficulty level of the question items that were in the difficult category did not exist, questions that were classified as medium as many as 4 points (10%), namely questions number 3, 23, 36, 38 ($0.3 \leq P \leq 0.7$), and questions that were classified as easy as many as 36 points (90%) ($P > 0.7$) which means that almost all test takers in the high group could answer the questions correctly. In general, the average difficulty level of the grain ($\bar{P}$) in the high group of 0.904 is in the easy. If further explored questions in the high group that has a moderate level of difficulty (0.583) one of them is chemistry problem number 23.

Table 3. Item Parameters Based on Classical Test Theory in High and Low Groups

item	High Group				Low Group			
	p	Category	r_{pbis}	Category	P	Category	r_{pbis}	Category
1	0.979	easy	0.120	not good	0.360	medium	0.080	not good
2	0.998	easy	0.010	not good	0.680	medium	0.280	not good
3	0.553	medium	-0.040	not good	0.345	medium	0.170	not good
4	0.998	easy	0.040	not good	0.337	medium	0.180	not good
5	0.951	easy	0.110	not good	0.307	medium	0.150	not good
6	0.993	easy	0.040	not good	0.434	medium	0.190	not good
7	0.982	easy	0.070	not good	0.360	medium	0.160	not good
8	0.982	easy	0.090	not good	0.618	medium	0.190	not good
9	0.918	easy	0.200	not good	0.273	difficult	0.080	not good
10	0.892	easy	0.170	not good	0.371	medium	0.170	not good
11	0.984	easy	0.110	not good	0.404	medium	0.130	not good
12	0.960	easy	0.130	not good	0.260	difficult	0.100	not good
13	0.993	easy	0.060	not good	0.403	medium	0.210	not good
14	0.952	easy	0.190	not good	0.220	difficult	0.060	not good
15	0.938	easy	0.170	not good	0.270	difficult	0.090	not good
16	0.965	easy	0.160	not good	0.306	medium	0.080	not good
17	0.945	easy	0.110	not good	0.392	medium	0.160	not good
18	0.865	easy	0.160	not good	0.329	medium	0.100	not good
19	0.949	easy	0.140	not good	0.379	medium	0.130	not good
20	0.838	easy	0.230	not good	0.197	difficult	0.120	not good
21	0.962	easy	0.100	not good	0.392	medium	0.110	not good
22	0.883	easy	0.190	not good	0.431	medium	0.110	not good
23	0.583	medium	0.320	good	0.273	difficult	0.070	not good
24	0.949	easy	0.120	not good	0.783	easy	0.160	not good
25	0.925	easy	0.150	not good	0.454	medium	0.160	not good
26	0.977	easy	0.090	not good	0.636	medium	0.200	not good
27	0.957	easy	0.070	not good	0.729	easy	0.200	not good
28	0.999	easy	0.000	not good	0.778	easy	0.330	good
29	0.859	easy	0.240	not good	0.351	medium	0.120	not good
30	0.994	easy	0.060	not good	0.521	medium	0.230	not good
31	0.975	easy	0.080	not good	0.567	medium	0.210	not good
32	0.861	easy	0.240	not good	0.274	difficult	0.080	not good
33	0.856	easy	0.230	not good	0.341	medium	0.080	not good
34	0.836	easy	0.070	not good	0.416	medium	0.150	not good
35	0.936	easy	0.180	not good	0.333	medium	0.120	not good
36	0.581	medium	0.320	good	0.191	difficult	0.050	not good
37	0.963	easy	0.070	not good	0.585	medium	0.230	not good
38	0.689	medium	0.200	not good	0.555	medium	0.050	not good
39	0.948	easy	0.170	not good	0.319	medium	0.100	not good
40	0.793	easy	0.160	not good	0.455	medium	0.140	not good
Average	0.904	easy	0.133	not good	0.416	medium	0.143	not good
Reliability	0.990			Very high	1.000			Perfect



23. Bahan kimia yang digunakan sebagai gas pendorong pada produk pewangi ruangan adalah ....

A. CH_4 (metana)
B. CFC (freon)
C. H_2 (hidrogen)
D. He (helium)

Figure 3. Question Number 23 about the Name of Chemical Compounds. (Source, Junior high school natural science national examination questions, 2015)

From Figure 3, it can be analyzed that the factors that cause students to have difficulty answering questions are methane, hydrogen, and helium which are gaseous and are part of CFCs. Students do not master the nomenclature of compounds and their parts according to their actual functions.

Based on the results of the analysis for the different power indexes of the question items, it was found that the question items that had good difference power were

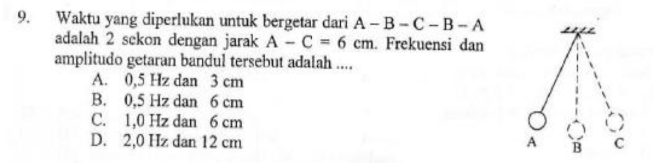
2 points (5%), namely questions number 23 and 36 (r_{pbis} > 0.3), and the question items that were classified as bad as many as 38 points (r_{pbis}< 0.3), which means that the high group of questions could not distinguish the abilities of the upper and lower group participants. In general, the average grain difference power ($\bar{r}_{pbis}$) in the high group of 0.133 is in a bad (not good) category.

The estimated reliability of the test score is seen from the reliability of the QUEST output item of 0.990 which shows the level of accuracy and consistency of participants in the high group in answering the questions is very good. Qualified for reliability 0.990 > 0.700 (high reliability).

Characteristics of Test Items in the Low Group

Analysis of the characteristics of items in the low group can be interpreted as the difficulty level of the question items that are in the difficult category as many as 8 points (20%), namely questions number 9, 12, 14, 15, 20, 23, 32, and 36 (p < 0.3), as many as 29 questions (72.5%) which have a moderate level of difficulty (0.3 ≤ p ≤ 0.7), and as many as 3 points (7.5%) which are relatively easy, namely questions number 24, 27, and 28 (p > 0.7). In general, the average difficulty level of grain ($\bar{P}$) in the low group of 0.416 is in the medium category.

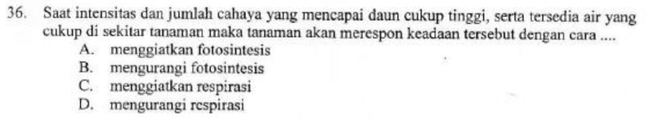
Furthermore, the analysis in the low group obtained the results of 8 questions that are classified as difficult consisting of 4 physics questions, namely numbers 9, 12, 14, and 15; Chemistry questions consist of 2 questions, namely numbers 20 and 23, and 2 biology questions, namely questions number 32 and 36. The following is displayed one item that is classified as difficult, namely point 9:



9. Waktu yang diperlukan untuk bergetar dari A – B – C – B – A adalah 2 sekon dengan jarak A – C = 6 cm. Frekuensi dan amplitudo getaran bandul tersebut adalah ....
   A. 0,5 Hz dan 3 cm
   B. 0,5 Hz dan 6 cm
   C. 1,0 Hz dan 6 cm
   D. 2,0 Hz dan 12 cm

Figure 4. Question Number 9 about Simple Pendulum Swings (Source, Junior High School Science National Examination Questions, 2015).

Based on the results of the analysis, the factors causing students' difficulty in solving the above questions *are lack of understanding of concepts, namely (1) students do not understand the concept of frequency, (2) students do not understand the concept of amplitude*, so students will be fooled into answering other answer choices besides the Answer Key (A). The next question that has the most difficulty (0.191) is biology problem number 36:



36. Saat intensitas dan jumlah cahaya yang mencapai daun cukup tinggi, serta tersedia air yang cukup di sekitar tanaman maka tanaman akan merespon keadaan tersebut dengan cara ....
   A. menggiatkan fotosintesis
   B. mengurangi fotosintesis
   C. menggiatkan respirasi
   D. mengurangi respirasi

Figure 5. Question number 36 about the process of respiration and photosynthesis in plants. (Source, Junior High School Science National examination Questions, 2015).

From Figure 5, the results of the analysis, the factors causing students' difficulties are students' lack of understanding of respiration and photosynthesis in plants. Based on the results of the analysis of the different power indexes of the question items, it was found that the question items that had good difference power were 1 item (2.5%), namely question number 28 (r_{pbis} > 0.3), and the question items that were classified as bad as many as 39 points (97.5%) (r_{pbis} < 0.3), which means that in the low group, the question could not distinguish the ability of the upper group participants and the lower group. In general, the average point difference power ($\bar{r}_{pbis}$) of 0.143 is in the easy category. The estimated reliability of the test score is seen from the reliability of the QUEST output item of 0.990 which shows the level of accuracy and consistency of participants in the low group in answering the questions is very good and meets the reliable requirements of 1.00 > 0.70 (perfect reliability).

Conclusion

Based on the results of the analysis of the characteristics of the UN Science question items in 2015 using a classical test theory approach, the level of difficulty and differentiating power of the questions is influenced by the ability of the group. For the high-ability group, the question points become easy and for the low-ability groups, the question items become difficult. From the 40 questions analyzed, the average difficulty level (P) of the high group in the easy category (0.904) and the low group in the medium category (0.416), the average difference power index ($\bar{r}_{pbis}$) indicating that the ability of the test to distinguish high-ability and low-ability students in both the high group (0.133) and the low group (0.143) was in the not good category, and the reliability coefficient (ρ) of the high group test (0.990) has high reliability and the low group (1,000) has perfect reliability.

Author Contribution
The authors jointly contributed to this work.

Funding
This research uses independent costs.

Conflicts of Interest
No conflict of interest.

References

Aiken, L. R. (1997). *Psychological testing and assessment* (9th ed.). Allyn and Bacon.

Aldyza, N., & Andani, D. (2018). Analisis Tingkat Kesukaran Butir Soal Ujian Nasional (UN) Ipa Smp Tahun Ajaran 2014/2015 Di Kabupaten Aceh Tenggara. *Jurnal Pendidikan Almuslim, 6*(1). Retrieved from http://jfkip.umuslim.ac.id/index.php/jupa/article/view/330

Ali, A., & Istiyono, E. (2022). An analysis of item response theory using program R. *Al-Jabar : Jurnal Pendidikan Matematika*, *13*(1), 109–123. https://doi.org/10.24042/ajpm.v13i1.11252

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory* (L. S. Wrightsman (ed.)). Brooks/Cole Publishing Company.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed). Prentice Hall.

Anggoro, B. S., Agustina, S., Komala, R., Komarudin, K., Jermsittiparsert, K., & Widyastuti, W. (2019). An Analysis of Students' Learning Style, Mathematical Disposition, and Mathematical Anxiety toward Metacognitive Reconstruction in Mathematics Learning Process Abstract. *Al-Jabar : Jurnal Pendidikan Matematika*, *10*(2), 187–200. https://doi.org/10.24042/ajpm.v10i2.3541

Ardi, A., Hervi, F., & Mudjiran, M. (2023). Validity and Reliability Questionnaire of Students' Critical Thinking Skills in General Biology Course. *Jurnal Penelitian Pendidikan IPA*, *9*(3), 1436–1444. https://doi.org/10.29303/jppipa.v9i3.2761

Cohen, C., & Swerdlik. (2009). *Psychology testing and assessment: An introduction to test and measurement* (7th ed.). McGraw-Hill.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rine Hart, and Winston.

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

Etobro, A. B., & Fabinu, O. E. (2017). Students' perceptions of difficult concepts in biology in senior secondary schools in Lagos state. *Global Journal of Educational Research*, *16*(2), 139. https://doi.org/10.4314/gjedr.v16i2.8

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

French, C. L. A. (2001). *Review of Classical Methods of Item Analysis*. Paper Presented in Annual Meeting of the Southweast Educational Research Association.

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

Hambleton, Ronald K Swaminathan, H. (1985). *Item response theory*. Kluwer Inc.

Hambleton, R. K., Robin, F., & Xing, D. (2000). Item response models for the analysis of educational and psychological test data. In *Handbook of applied multivariate statistics and mathematical modeling*, 553-581. Retrieved from https://doi.org/10.1016/B978-012691360-6/50020-3

Hu, Z., Lin, L., Wang, Y., & Li, J. (2021). The Integration of Classical Testing Theory and Item Response Theory. *Psychology*, *12*, 1397–1409. https://doi.org/10.4236/psych.2021.129088

Hussin, W. N. T. W., Harun, J., & Shukor, N. A. (2018). Problem Based Learning to Enhance Students Critical Thinking Skill via Online Tools. *Asian Social Science*, *15*(1), 14. https://doi.org/10.5539/ass.v15n1p14

Jailani, J., & Almukarramah, A. (2020). Upaya Peningkatan Kualitas Pembelajaran Biologi Melalui Pembelajaran Bermakna Dengan Menggunakan Peta Konsep. *Jurnal Biology Education*, *8*(2), 122–130. https://doi.org/10.32672/jbe.v8i2.2371

Kaplan, R. M., & Saccuzzo, D. P. (2009). *Psychological testing: Principles, applications and issues* (7th ed.). Nelson Education.

Lawrence, M. . (2019). Question to ask when evaluating test. *Practical Assessment, Research, and Education*, *4*. https://doi.org/10.7275/5c28-0n19

Lord, F. . (1980). *Applications of item response theory to practical testing problems* (1st ed). Erlbaum.

Mardapi, D. (2012). *Pengukuran, Penilaian, dan Evaluasi Pendidikan*. Nuha Litera.

Mehrens, W., & Lehmann, I. (1973). *Measurement and evaluation in education and psychology*. Hold, Rinehart and Winston Inc.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing principles and applications* (Pearson (ed.); 6th ed). Education.

Nitko, A. J. (1996). *Educational assesment of students* (2nd ed). Merill an imprint of Prentince Hall Englewood Cliffs.

Retnawati, H. (2016). *Validitas reliabilitas dan karakteristik butir*. Parama Publishing.

Retnawati, H., Kartowagiran, B., Arlinwibowo, J., & Sulistyaningsih, E. (2017). Why are the mathematics national examination items difficult and what is teachers' strategy to overcome it? *International Journal of Instruction*, *10*(3), 257–276. https://doi.org/10.12973/iji.2017.10317a

Sugiyono. (2011). *Metode Penelitian Kuantitatif Kualitatif dan R&D* (1st ed.). Alfabeta.

Widiatmo, T., Jufri, A. W., & Jamaluddin, J. (2019). Analysis of Validation of Instruments to Measure Student's Critical Thinking Ability and Science

Literation. *Jurnal Penelitian Pendidikan IPA*, *5*(2), 212–218. https://doi.org/10.29303/jppipa.v5i2.272