An Application of Classical Test Theory for Item Characteristic Analysis of Chemical Literacy Instruments

Sri Nurfadillah Ningsih[1], Edi Istiyono[2], Widihastuti[3]

[1,2,3] Program Studi Penelitian dan Evaluasi Pendidikan, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

ABSTRAK

Kemampuan literasi sains siswa di Indonesia sangat memprihatinkan, penyebabnya adalah siswa tidak terbiasa mengerjakan soal-soal yang berbasis literasi sains maka perlu adanya instrument penilaian terstandar berbasis literasi kimia yang mampu mengukur taraf ketercapaian identifikasi ilmiah dan menginterpretasi pembelajaran sains/kimia. Tujuan penelitian ini adalah mendeskripsikan karakteristik butir instrument penilaian berbasis literasi kimia materi laju reaksi dengan pendekatan teori tes klasik. Jenis penelitian adalah deskriptif kuantitatif. Subjek penelitian sebanyak 80 siswa kelas XI MIPA SMA Negeri 4 Takalar. Data dianalisis menggunakan program ITEMAN 4.3. Berdasarkan hasil analisis diproleh koefisien reliabilitas sebesar 0,834, tingkat kesukaan 87% butir soal berada pada kategori sedang, dan 13% berada pada kategori sukar. Uji daya beda, sebanyak 73% butir soal berada pada kategori sangat baik, dan untuk uji distaktor, 93% dari keseluruhan butir soal telah menunjukkan distraktor yang berfungsi dengan baik, sehingga dapat disimpulkan bahwa secara keseluruhan butir soal memiliki reliabilitas yang tinggi dan karakteristik butir yang baik.

ABSTRACT

The science literacy skills of students in Indonesia are very concerning, the cause is that students are not accustomed to working on questions based on science literacy, so it is necessary to have a standardized assessment instrument based on chemical literacy that can measure the level of achievement of scientific identification and interpret science/chemistry learning. This study aimed to describe the characteristics of chemical literacy-based assessment instrument items of reaction rate material with a classical test theory approach. The type of research is descriptive quantitative. The research subjects were 80 class XI MIPA SMA Negeri 4 Takalar students. Data were analyzed using the ITEMAN 4.3 program. Based on the results of the analysis, a reliability coefficient of 0.834 was obtained, the level of favorability of 87% of the items was in the moderate category, and 13% was in the difficult category. In the difference test, as many as 73% of the items are in the very good category, and for the distractor test, 93% of all items have shown distractors that function properly, so it can be concluded that overall the items have high reliability and good item characteristics.

1. INTRODUCTION

People who have knowledge and can apply their knowledge to solve problems in real life are called science-literate people (Bond, 1989). Therefore, the achievement of a science-literate generation has become a demand of the times (Atika et al., 2019; DeBoer, 2000; Dewi et al., 2022; Zahro, 2019). Literacy competence is one of the most important pillars of science and technology development, especially in education (Chasanah et al., 2022; Rusilawati, 2018). This is also in line with Kemendikbud (2017) that one of the fundamental abilities that can succeed in Indonesia's development in the 21st century is to have science literacy competencies. Chemistry is also known as the center of science to be the basis for science technology and industry (Chang, 2010) so chemical literacy is the basis or part of science literacy (Rahayu, 2017; Mozeika & Bilbokaite, 2010). There are many benefits felt when students are trained with chemical literacy-based learning, one of which is that students will get used to solving real-life problems with scientific thinking and methods, not easily consumed by hoax news, and will contribute positively to the surrounding environment (Imnasari et al., 2018). Science literacy views the importance of thinking and

*Corresponding author
E-mail addresses: srinurfadillah.2022@student.uny.ac.id (Sri Nurfadillah Ningsih)

action skills that involve mastery of thinking in recognizing and addressing social issues (Pratiwi et al., 2019; Mayasari & Paidi, 2022). This is reinforced by Agustiawan & Puspitasari (2019) that science literacy is directly correlated with the formation of a generation that has strong scientific thinking and attitudes in acting and making decisions based on thoughts that can be justified through scientific evidence. In chemical literacy, there are four domains including chemical content knowledge, chemistry in context, higher-order learning skills, and affective aspects.(Cigdemoglu & Geban, 2015; OECD, 2015).

However, the current problem is the acquisition of Indonesia's science literacy level during the 12 years of participation is always ranked in the bottom fifth. Nonetheless, an understanding of science literacy plays a very important role in assessing the quality of education (Fuadah et al., 2017). Based on the Trends in International Mathematics and Science Study (TIMSS) data, the science literacy scores of students in Indonesia for 5 measurements were 492, 510, 471, 426, and 397 respectively. As for the PISA report published on 3 December 2019, Indonesia's ranking in terms of reading ability is 72 out of 77 countries, and science ability, ranked 70 out of 78 countries (OECD, 2018). Windyariani et al (2017) explained that the understanding of science education that aims to develop students' science literacy seems to be less well understood by science teachers. The learning process and evaluation methods used are still traditional and focus on understanding concepts, so students have not been trained in science literacy skills (National Research Council, 1996; Ridwan et al., 2013). This is reinforced by Utami et al (2022) who explained that the low ability of science literacy is due to students not being accustomed to solving tests or questions related to science process skills. Furthermore, Wahyuni & Yusmaita (2020) also mentioned that the cause of the low science literacy scores of Indonesian students in PISA is that students are not accustomed to solving discourse-based problems, besides that the assessment and the learning process at school also tend not to support students in floating science literacy skills. So far, test instruments have only focused on content, not on science literacy such as the application of science in everyday life, critical thinking to solve problems, and some science process skills (Putri, 2020). Based on the results of research 78% of chemistry teachers do not understand the true meaning of chemical literacy (Fahmina et al., 2019). Teachers still think that chemical literacy competence is the ability to read chemical literature, both from printed books and from the internet. This misunderstanding of chemical literacy results in the use of chemical literacy-based instruments that are still underused.

The fact that the level of chemistry ability of students varies requires a standardized instrument based on chemical literacy that can measure the level of achievement of Scientific Identification, and interpret science learning (Chasanah et al., 2022). One of the elements that must be considered in the assessment is to try and ensure that the assessment results accurately describe the ability of students (Tim Pusat Penilaian Pendidikan, 2019). An assessment is called accurate if the assessment results contain as little error as possible (Guntur, 2014). To get results that accurately describe the ability of students, the quality of the test instruments used must be valid, reliable, and have good item parameters. So it is necessary to analyze the items empirically (Purwati et al., 2021). Empirical item analysis aims to obtain information about the characteristics of each item (Elviana, 2020; Gusmizain, 2022). Analysis of items empirically can be done with two approaches, namely with the classical test theory approach and modern theory or item response theory (Retnawati, 2016) Item analysis in this study uses a classical test theory (CTT) approach.

Classical test theory, also known as classical pure score theory (Allen & Yen, 1979), stems from the application of a simple mathematical model relating the observed score (X), the true score (T), and the error score (E) (Mardapi, 2012). If written with a mathematical statement, the sentence becomes

$$X = T + E \qquad\qquad (1)$$

Measurement error in classical test theory is considered as unsystematic or random error. This error is the difference between the observed score obtained and the theoretically expected observed score. Systematic measurement error does not include measurement error.

According to Allen & Yen (1979: 56), "*classical true-score theory is a simple, quite useful model that describes how errors of measurement can influence observed scores*". Allen & Yen state that classical test theory is a simple, fast, and useful theory that describes how standard errors of measurement can affect observed scores. Classical test theory has been widely used in the process of item analysis. This is because CTT analysis is based on data that is not too large (Nurcahyo, 2017). Besides having a simple formula, classical test theory also has assumptions that are easy to understand (Mardapi, 2012: 53) namely 1) the instrument only measures one dimension; 2) there is no relationship between the pure score and the error score; 3) There is no relationship between the error in the first measurement and the error in the second measurement; 4) There is no relationship between the pure score on the first measurement and the error on the second measurement; 5) There is no relationship between the pure score on the second measurement and the error on the first measurement; and 6) The average measurement error in the population is zero. The basic assumptions of classical test theory can be expanded in various formulas that have benefits in the implementation of psychological measurements (Lababa, 2018). Validity, reliability,

difficulty level, differentiation, and effectiveness of distractors, are important formulas in item selection that can be obtained through classical test theory analysis (Setyawarno, 2018)

Various facts and theories that exist become a strong reason for the importance of studies related to analyzing the quality of test instrument items that not only measure aspects of chemical content but are also able to measure students' chemical literacy competencies. This article will focus on describing the characteristics of chemistry test items consisting of reliability, difficulty level of questions, differentiation, and the functioning of distractors on reaction rate material using the classical test theory approach.

2. METHOD

This study is part of development research that focuses on evaluating the quality of instruments that have been developed before they are used in a wider scope. This type of research is quantitative descriptive research. The research subjects were 80 students of class XI MIPA SMA Negeri 4 Takalar considering that the participants had learned the reaction rate material. The instrument used was a chemical literacy-based assessment instrument for reaction rate material. Chemical literacy competency-based test instruments are constructed based on aspects of content, HOLs (High-order learning skills), and attitudes. It consists of 15 items which are divided into 3 stimulus contexts, namely the marine field, the environmental context, and the technology field. This instrument has been proven valid. Response data with dichotomous scoring was analyzed using the classical test theory approach with the help of the ITEMAN 4.3 program to obtain information about reliability, difficulty level, differentiation, and distractors.

3. RESULT AND DISCUSSION

Result

Based on the results of the analysis of 15 items with the help of the ITEMAN 4.3 program, reliability information, difficulty index, discrimination index, and distractors. The reliability testing aims to assess the extent to which the instrument remains consistent. A quality instrument will provide the same measurement results and consistent answers every time the instrument is used. The question items analyzed for reliability using the ITEMAN 4.3 program are categorized based on the following criteria Arikunto (2016) which are presented in Table 1.

Table 1. Reliability Categories (Arikunto, 2016)

Reliability Coefficient	Correlation	Interpretation/Reliability
R ≤ 0.20	Very Low	Very irregular/Very Bad
0.20 < r ≤ 0.40	Low	Not Fixed/Bad
0.40 < r ≤ 0.60	Medium	Fairly Fixed/Good Enough
0.60 < r ≤ 0.80	High	Fixed/Good
0.80 < r ≤ 1.00	Very High	Very Stable/Very Good

The results of the item reliability analysis on the science literacy-based chemistry test instrument for the XI MIPA class reaction rate material analyzed using the classical test theory approach with the ITEMAN 4.3 program can be seen in Table 2.

Table 2. Reliability Results using the ITEMAN 4.3 Program

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	0.834	1.604	0.698	0.554	0.712	0.822	0.713	0.832

Based on Table 2, it can be seen that overall the test items have very high reliability/consistency, which is 0.834 with a standard error measurement of 1.604. It can be concluded that the instrument will provide the same measurement results and have consistent answers if tested many times. A good test is a test that can be trusted (reliable) if it provides fixed or consistent results when tested many times (Widoyoko, 2018).

In classical test theory, the level of item difficulty is determined based on criteria Allen & Yen (1979) can be seen in Table 3.

Table 3. Difficulty Index Categories (Allen & Yen, 1979)

No	Difficulty Index	Category
1	>0.7	Easy
2	0.3-0.7	Medium
3	<0.3	difficult

The results of the analysis of the level of difficulty of questions on the chemical literacy-based chemistry test instrument for class XI MIPA analyzed using the classical test theory approach with the ITEMAN 4.3 application can be seen in Table 4.

Table 4. Level of Difficulty of Item 1

N	P	Total Rpbis	Total Rbis	Alpha w/o
80	0.388	0.369	0.469	0.830

The level of difficulty can be seen from the P (difficulty) value of the ITEMAN output. Based on Table 4 and referring to the categorization of the level of difficulty Allen & Yen (1979) it can be seen that the difficulty index of question number 1 of 0.388 is in the medium difficulty category. The same analysis was carried out on all question items whose results can be seen in Table 5.

Table 5. Difficulty Index

No	Index (P)	Category
1	0.388	Medium
2	0.537	Medium
3	0.550	Medium
4	0.487	Medium
5	0.550	Medium
6	0.313	Medium
7	0.463	Medium
8	0.313	Medium
9	0.487	Medium
10	0.313	Medium
11	0.250	Difficult
12	0.300	Difficult
13	0.338	Medium
14	0.325	Medium
15	0.325	Medium

Based on Table 5, it can be concluded that of the 15 items measuring chemical literacy skills, there are 13 items that show results with moderate difficulty categories and 2 items with difficult categories.

The categorization of discrimination in this study is based on the criteria for the discrimination index Garvin & Ebel (1980) which is presented in Table 6.

Table 6. Discrimination Index Category (Garvin & Ebel, 1980)

Discrimination index	Category
>0.4	Very good
0.3-0.39	Good
0.2-0.29	Item requires revision
< 0.19	Item must be eliminated

The results of the analysis of the discrimination of items using the classical test theory approach in the ITEMAN application are presented in Table 7.

Table 7. The Output of Item 1 of Discrimination

N	P	Total Rpbis	Total Rbis	Alpha w/o
80	0.388	0.369	0.469	0.830

The index of item discrimination can be seen from the Rpbis (Discrimination) value from the ITEMAN output. Based on Table 7 and referring to the categorization of discrimination index Garvin & Ebel (1980). Then it can be seen that the discrimination index of question number 1 of 0.369 is the good discrimination category. The same analysis was carried out on all question items whose results can be seen in Table 8.

Table 8. Item Discrimination Index

Item Number	Index	Category
1	0.369	Good
2	0.400	Good
3	0.404	Very Good
4	0.413	Very Good
5	0.470	Very Good
6	0.644	Very Good
7	0.472	Very Good
8	0.515	Very Good
9	0.421	Very Good
10	0.414	Very Good
11	0.474	Very Good
12	0.500	Very Good
13	0.394	Good
14	0.571	Very Good
15	0.369	Good

Based on Table 8, it can be seen that of the 15 items on the measurement of chemical literacy in reaction rate material, there are 11 items (73%) that show results with a very good discrimination category and there are 4 items (27%) showing results with good categories, namely on items 1, 2, 13, 15 and there are no items with negative discrimination.

In classical test theory, the classification of distractors is done based on the standard set by Fernandes (1984), which states that a distractor is considered effective if it is selected by at least 2% of test-takers. Distractors that do not meet this threshold need to be replaced with alternative distractors that may be more attractive to test-takers

Table 9. Distractor Categories (Fernandes, 1984)

Distractor Index	Category
≥2%	Good
<2%	Revised

The results of the item distractor analysis using the classical test theory approach in the ITEMAN application are presented in Table 10.

Table 10. Distractor analysis result

Item Number	Alternative Answer									
	A		B		C		D		E	
	Index	%	Index	%	Index	%	Index	%	Index	%
1	0.263	26.3	0.138	13.8	0.388*	38.8	0.125	12.5	0.087	8.7
2	0.537*	53.7	0.175	17.5	0.063	6.3	0.100	10.0	0.125	12.5
3	0.150	15	0.113	11.3	0.163	12.3	0.550*	55	0.025	2.5
4	0.087	8.7	0.138	13.8	0.175	17.5	0.113	11.3	0.487*	48.7
5	0.550*	55	0.138	13.8	0.075	7.5	0.125	12.5	0.113	11.3
6	0.313*	32.3	0.313	31.3	0.175	17.5	0.188	18.8	**0.013**	**1.3**
7	0.200	20.0	0.462*	46.2	0.225	22.5	0.087	8.7	0.025	2.5
8	0.212	21.2	0.075	7.5	0.313*	31.3	0.338	33.8	0.063	6.3
9	0.163	16.3	0.487*	48.7	0.138	13.8	0.175	15.5	0.037	3.7
10	0.237	23.7	0.125	12.5	0.212	21.2	0.313*	31.3	0.113	11.3
11	0.275	27.5	0.225	22.5	0.113	11.3	0.138	13.8	0.250*	25
12	0.287	28.7	0.175	17.5	0.212	21.2	0.300*	20.0	0.025	2.5
13	0.188	28.8	0.338*	22.8	0.225	22.5	0.125	12.5	0.125	12.5
14	0.075	7.5	0.138	23.8	0.275	27.5	0.188	18.8	0.325*	32.5
15	0.163	16.3	0.325*	32.5	0.150	15	0.263	26.3	0.100	10.0

By referring to Fernandes (1984) categorization of distractors, it can be concluded that of the 15 items on the Measurement of Science Literacy in the XI Science class reaction rate material, there is 1 item, namely the 6th item that has an option or answer choice that needs revision, namely on option E because the percentage of students who choose the option is low, which is below 2%. The following will present option statistics for item number 6.

Table 11. Option Statistics for Question Number 6

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	25	0.313	0.644	0.843	10.040	4.010	Maroon	**KEY**
B	25	0.313	-0.231	-0.302	4.400	2.614	Green	
C	14	0.175	-0.264	-0.390	3.571	1.505	Blue	
D	15	0.188	-0.209	-0.303	4.067	1.907	Olive	
E	1	0.013	-0.083	-0.283	3.000	0.000	Gray	
Omit	0							
Not Admin	0							

When viewed from Table 11, it can be seen that the answer option E is the option with the lowest N value, which is only chosen by 1 learner so the option needs to be revised again. As for the 14 items or 93% of the items have shown the answer choice options with distractors or checkers that work well, because the percentage of students who choose the option is above 2%.

Discussion

Assessment of literacy skills is important and needs to be improved by teachers in exploring students' weaknesses and strengths, accurate assessment results can be used as a basis for developing strategies to improve and develop student learning (Abell & Siegel, 2011; Xu & Brown, 2016; Looney et., 2017) In this study, the chemical literacy assessment instrument consisting of 15 multiple choice questions, was prepared based on indicators of chemical literacy according to (Schwartz et al., 2006) and indicators of reaction rate material, taking into account the answers given by students of class XI MIPA SMAN 4 Takalar, and by using classical analysis (CTT), the reliability value, difficulty level, differentiability, and distractor functionality were obtained. The purpose of this study is to determine how much consistency or constancy of the items that have been prepared, and whether this assessment instrument has good item characteristics so that it can be used to measure students' chemical literacy competencies.

Reliability is the consistency of a series of measuring instruments in measuring some groups with different characteristics (Arikunto, 2017; Prastika, 2021). The quality of the items from the aspect of item reliability in the classical test theory approach uses Cronbach's alpha formula (KR-20). Based on the results of the analysis, a very high reliability coefficient of 0, 834 was obtained so that it can be concluded that the state of the instrument in various forms, namely the test results will remain the same if carried out by different people (inter-rater), The test results remain the same if done by the same person at different times (retesting), the test results are the same if done by different people at the same time with different tests (parallel form) and the test results remain the same using various constructive questions (internal consistency). This is supported by a theory that states that high reliability will show consistency between two measurement scores on the same object, even though it uses different measuring devices and different scales. (Mehrens & Lehmann, , 1973; Reynold, Livingstne, & Wilson, 2010).

Reliability is also related to measurement error or in analysis known as standard error measurement (SEM). Based on the analysis results, the low SEM value is only 1.604. High reliability indicates that the error in obtaining measurement results is small. The higher the reliability of the instrument, the less likely there is an error in measurement, while low reliability indicates greater measurement error (Retnawati, 2016). According to Rohmad & Sarah (2021), Measurement errors can arise due to various factors, including the characteristics of the instrument used, such as errors in designing and carrying out measurements that do not follow standardized rules, inadequate quality of questions in the instrument, cooperation during carrying out tests or filling out instruments, multi-interpretive items, psychological conditions of participants who respond to the instrument, lack of motivation of participants, the measurement environment is less supportive or a combination of all these factors.

The level of difficulty of the item shows the possibility of how many respondents can answer a question item correctly (Arikunto, 2017). The level of item difficulty states whether the item is in the difficult, moderate, or easy category. Based on the results of the analysis of 15 items of Measurement of

Chemical Literacy Ability class XI, there are 13 items that show results with moderate difficulty level categories and 2 items with difficult categories. This information proves that about 87% of the test items have a good level of difficulty. This is supported by Chasanah et al (2022); Allen & Yen (1979); Mardapi (2012) that a good question is a question that is not too easy and also not too difficult for students to answer correctly. As for 13% of the overall items show results with difficult categories, namely in the 11th item with a difficulty index of 0.250 and the 12th with a difficulty index of 0.3000. This means that only a few students can answer correctly on items 11 and 12.

If analyzed further, the item that has the smallest difficulty index value is item number 11, which is only 0.250. This shows that out of 80 test takers/students only 20 or 25% of students can answer correctly. Item number 11 is presented in Figure 1.

11. The reactions that occur in the combustion of motor vehicle fuel are ....

	Reaction between CO and NO	Reaction between CH_4 and NO_2
A.	$CO_{(g)} + NO_{(g)} \rightarrow CO_{2(g)} + N_{2(g)}$	$CH_{4(g)} + NO_{2(g)} \rightarrow CO_{2(g)} + N_{2(g)} + H_2O_{(g)}$
B.	$2CO_{(g)} + 2NO_{(g)} \rightarrow 2CO_{2(g)} + 2N_{2(g)}$	$CH_{4(g)} + NO_{2(g)} \rightarrow CO_{2(g)} + N_{2(g)} + 2H_2O_{(g)}$
C.	$2CO_{(g)} + 2NO_{(g)} \rightarrow 2CO_{2(g)} + N_{2(g)}$	$CH_{4(g)} + 2NO_{2(g)} \rightarrow CO_{2(g)} + N_{2(g)}$
D.	$CO_{(g)} + NO_{(g)} \rightarrow CO_{2(g)} + N_{2(g)}$	$CH_{4(g)} + 2NO_{2(g)} \rightarrow CO_{2(g)} + N_{2(g)} + 2H_2O_{(g)}$
E.	$2CO_{(g)} + 2NO_{(g)} \rightarrow 2CO_{2(g)} + N_{2(g)}$	$CH_{4(g)} + 2NO_{2(g)} \rightarrow CO_{2(g)} + N_{2(g)} + 2H_2O_{(g)}$

Figure 1. Item Number 11

The factors that cause students' difficulties in solving question number 11 are; first, most students are still lacking in the chemical literacy content domain, students have not been able to analyze and interpret the context or events presented scientifically, students have not been able to equalize chemical reactions, third students do not understand the concept of reduction and oxidation reactions that occur in combustion reactions by motor vehicle fuel. Based on the results of the study, it is explained that the distribution of the order of questions in the test affects the scores of test takers (Yasar, 2017). Tests that are sorted from those with easy to difficult difficulty levels or vice versa cause lower test scores than when the distribution of difficulty levels is randomly arranged (MacNichol, 1960). In this study, the instrument developed, the distribution of item difficulty was randomized.

Discrimination refers to a question's capacity to differentiate between students with high abilities and those with low abilities (Arikunto, 2016). Table 3, shows that of the 15 items measuring the ability of science literacy in reaction rate material, all items can perfectly distinguish students with low abilities and students with high abilities so this instrument is suitable for use. According to Reynolds et al, (2010) explain that a good item is an item that can accurately distinguish between test takers of different abilities. The question's differentiation power is unidirectionally correlated with validity, the greater the index of differentiation power, the greater the item validity coefficient. This is evidenced in the research of Zein et al (2013); Nurhalimah et al (2022); Setiawan (2013) that there is a significant relationship between the power of difference and the validity of the items that have a high index of the power of difference also has a high validity value.

The final analysis of classical test theory is distractor analysis. In multiple choice questions, alternative answers that are not correct can be used as distractors for students who do not know the correct answer alternative. Distractors will provide information to educators about students who master the subject matter and students who do not master the subject matter. Based on the results of the analysis, it is found that there is still 1 item that has an option or answer choice that needs to be revised because the percentage of students who choose the option is below 2%, namely in the 6th item presented in Figure 2. However, 93% of all question items have distractors or checkers that function properly because the percentage of students who choose each option is above 2%. The effectiveness of an exception/distractor can be said to function properly if the test distractor has a great attraction for test followers who do not understand the concept or do not master the material. Execution effectiveness is a measurement of alternative answers to multiple-choice questions (Arikunto, 2016).

Figure 2. Item 6 of the chemical literacy-based assessment instrument

4. CONCLUSION

Science literacy-based chemistry questions on reaction rate material were analyzed for item characteristics using classical test theory with the help of the ITEMAN 4.3 program, a reliability coefficient of 0.834 was obtained, meaning that all items were very reliable, 87% of all items had a difficulty level in the medium category and 13% in the difficult category. This information proves that 87% of all test items have been able to describe the function of student abilities and are included in the category of good questions because the questions are not too difficult and not too easy for students to answer correctly. The differential power obtained about 73% of the items with very good differential power and 27% with good differential power. This information shows that all items can accurately distinguish students with low abilities and students with high abilities. As for the distractor test, 93% of the items have shown answer choice options with distractors or distractors that function properly, and 1.3% of the items show one of the answer choice options needs to be revised because the distractor is not functioning properly.

5. ACKNOWLEDGE

6. REFERENCES

Abell, S. K., & Siegel, M. A. (2011). *Assessment literacy: What science teachers need to know and be able to do?* In D. Corrigan, J. Dillon, R. Gunstone (Eds.). The professional knowledge based of science teaching. Netherlands: Springer. https://doi.org/10.1007/978-90-481-3927-9

Agustiawan, H., & Puspitasari, E. D. (2019). Pembentukan karakter siswa SMP melalui literasi sains. *Symposium of Biology Education (Symbion)*, https://doi.org/10.26555/symbion.3548

Arikunto, S. (2016*). Prosedur penelitian suatu pendekatan praktik.* Jakarta: Rineka Cipta.

Arikunto, S. (2017). *Pengembangan instrument penelitian dan penilaian program*. Yogyakarta: Pustaka Pelajar.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*, Monterey, CA: Brooks/Cole.

Atika, A. R., Westhisi, S. M., & Zahro, I. F. (2019). Pelatihan literasi sains untuk meningkatkan kemampuan berpikir ilmiha pada guru Pendidikan anak usia dini. *Jurnal Pendidikan Anak Usia Dini Undiksha*. 7(3), 266-271. https://doi.org/10.23887/paud.v7i3.22297

Bond, D. (1989). In pursuit of chemical literacy: A place for chemical reactions. *Journal of Chemical Education*, *66*(2). https://doi.org/10.1021/ed066p157

Chang, R. (2010). *Chemistry (Tenth ed).* New York: McGraw-Hill.

Chasanah, N., Widodo, W., & Suprapto, N. (2022). Pengembangan Instrumen Asesmen Literasi Sains Untuk Mendeskripsikan Profil Peserta Didik. *PENDIPA Journal of Science Education*, *6*(2). https://doi.org/10.33369/pendipa.6.2.474-483

Cigdemoglu, C., & Geban, O. (2015). Improving students' chemical literacy levels on thermochemical and thermodynamics concepts through a context-based approach. *Chemistry Education Research and Practice*, *16*(2). https://doi.org/10.1039/c5rp00007f

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Wiston, Inc.

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching*, 37(6), 582-601. 10.1002/1098-2736(200008)37:63.0.CO;2-L

Dewi, C. A., Rahayu, S., Muntolib., Parlan. (2022). Pentingnya mengoptimalkan literasi kimia melalui pembelajaran barbasis isu-isu sosiosaintifik di abad ke-21. *Proceeding Seminar Nasional.* 348-359, https://proceeding.unnes.ac.id/index.php/snipa/article/view/1371

Elviana. (2020). Analisis butir soal evaluasi pembelajaran Pendidikan agama islam menggunakan program anates. *Jurnal Mudarrisuna.* 10(2). 59-74. http://dx.doi.org/10.22373/jm.v10i2.7839

Fahmina, S. S., Indriyanti, N. Y., Setyowati, W. A. E., Masykuri, M., & Yamtinah, S. (2019). Dimension of Chemical Literacy and its Influence in Chemistry Learning. *Journal of Physics: Conference Series*, *1233*(1). https://doi.org/10.1088/1742-6596/1233/1/012026

Fernandes, H.J.X. (1984). *Evaluation of educational programs.* Jakarta: National Education Planning Evaluation and Curriculum Development.

Fuadah, H., Rusilowati, A., & Hartono. (2017). Pengembangan alat evaluasi literasi sains untuk mengukur kemampuan literasi sains siswa sertema perpindahan kalor dalam kehidupan. *Lembaran Ilmu Kependidikan*, *46*(2). https://doi.org/10.15294/lik.v46i2.11350

Garvin, A. D., & Ebel, R. L. (1980). Essentials of educational measurement. *Educational Researcher*, *9*(9). https://doi.org/10.2307/1175572

Guntur. (2014). Penilaian berbasis kinerja (*performance-based assessment)* pada pendidikan jasmani. Jurnal Pendidikan Jasmani Indonesia. 10(1). 15-22.

Gusmizain,A. (2022). Karakteristik butir soal tes mata kuliah matriks & ruang vector mahasiswa matematika. *Jurnal Evaluasi Pendidikan*, 13(2). 127-130. https://doi.org/10.21009/jep.v13i2.28328

Imnasari, M., Sudarmin., Sumarni, W. (2018). Analisis literasi kimia peserta didik melalui pembelajaran inkuiri terbimbing bermuatan etnosains. *Jurnal Inovasi Pendidikan Kimia*. 12(2), 2201-221, https://journal.unnes.ac.id/nju/index.php/JIPK/article/view/15480

Kemendikbud. (2017). *Panduan Gerakan literasi nasional*. Jakarta Timur

Lababa, D. (2018). Analisis butir soal dengan teori tes klasik: sebuah pengantar. *Jurnal Ilmiah Iqra*. 2(2). 29-37. 10.30984/jii.v2i2.538

Looney, A., Cumming, J., Kleij, F. V. D., & Harris, K. (2017). Reconceptualizing the role of teachers as assessors: teacher assessment identity*. Assessment in education principals, Policy & Practice*. 25(2), 442-467. https://10.1080/0969594X.2016.1268090

MacNichol, K., (1960). Effects of varying order of item difficulty in an unspeeded verbal test. Unpublished manuscript, *Educational Testing Service*, Princeton, NJ

Maradpi, D. (2012). *Pengukuran penilaian & evaluasi pendidikan*. Yogyakarta: Nuha Medika.

Mayasari, T., & Paidi. (2022). Analisis kemampuan literasi sains siswa kelas XI SMA Negeri di kota Yogyakarta mata Pelajaran biologi ditinjau dari kefavoritan sekolah. *Jurnal Edukasi Biologi*. 8(1). https://doi.org/10.21831/edubio.v8i2.18212

Mehrens, W.A., & Lehmann, I.J. (1973). *Measurement and evaluation in education and psychology*. New York: Holt, Rinehart, and Winston. Inc.

Mozeika, D., & Bilbokaite, R. (2010). Teaching and learning method for enhancing 15-16-year-old students' knowledge as one of scientific literacy aspect in chemistry: Results based on research and approbation. *The International Journal of Education Researchers,* 3(1), 1-16, http://ijer.eab.org.tr/1/3/1_mozeika.pdf

National Research Council, (1996). *Discipline-based education research, Understanding and improving learning in undergraduate science and engineering*, Washington DC.,

Nurcahyo, F. A. (2017). Aplikasi IRT dalam analisis aitem tes kognitif. *Buletin Psikologi*, *24*(2). https://doi.org/10.22146/buletinpsikologi.25218

Nurhalimah, S., Hidayati, Y., Rosidi, I., & Hadi, W. P. Hubungan antara validitas item dengan daya pembeda dan tingkat kesukarann soal pilihan ganda pas. *Jurnal Natural Science Educational Research*, 4(3). https://doi.org/10.21107/nser.v4i3.8682

OECD. (2015). PISA 2015 Results in Focus *: Snapshot of performance in science, reading and mathematics*. OECD Publishing Online.

OECD. (2018). *PISA 2018 Results combined executive summaries volume I, II & III.* PISA, OECD Publishing

Prastika, Y. D. (2021). Pengaruh validitas, reliabilitas dan tingkat kesukaran terhadap kualitas butir soal ekonomi menggunakan *software* anates di SMKN 2 bangkalan. STKIP PGRI Bangkalan

Pratiwi et al. (2019). Pembelajaran IPA abad 21 dengan literasi sains siswa. *Jurnal Materi dan Pembalajaran Fisika*. 9(1). https://doi.org/10.20961/jmpf.v9i1.31612

Purwati, H., Retnawati, H., Jailani, J., & Retnowati, T. H. (2021). Analisis karakteristik butir soal ujian nasional matematika SMP/MTs berdasarkan pendekatan teori tes klasik. *Jurnal Sains Dan Edukasi Sains*, *4*(2). https://doi.org/10.24246/juses.v4i2p46-51

Putri, R. K. (2020). Pengembangan Instrumen Tes Literasi Sains Siswa Pada Topik Keanekaragaman Makhluk Hidup. *Jurnal Pendidikan dan Pembelajaran Biologi*. 4(1), 71-78. https://doi.org/10.33369/diklabio.4.1.71-78

Rahayu, S. (2017). Mengembangkan Keterampilan Tingkat Tinggi Siswa melalui Pembelajaran Kimia Berkonteks Socioscientific Issues (SSI) dan Nature of Science (NOS*). Prosiding Seminar Nasional Kimia Dan Pembelajarannya (SNKP).*

Retnawati, H. (2016). *Analisis kuantitatif instrument penelitian*. Yogyakarta: Parama Publishing.

Reynolds, C.R., Livingstone & Wilson. (2010). *Measurement and assessment in education*. New Jersey: Pearson Education, Inc.

Rusilawati, A. (2018). Asesmen Literasi Sains: Analisis Karakteristik Instrumen dan Kemampuan Siswa Menggunakan Teori Tes Modern Rasch model. *Prosiding Seminar Naional Fisika Universitas Riau*. 3. 1-14. https://snf.fmipa.unri.ac.id/wp-content/uploads/2019/03/0.-300B-2-15NI.pdf

Rohmad., & Sarah, S. (2021). *Pengembangan instrument angket*. Yogyakarta: Penerbit K-Media.

Satiawan, A. (2014). Simulation study of item validity testing and item discrimination index. *Proceeding of International Conference on Electrical Engineering*, 99-104. :10.11591/eecsi.1.355

Setyawarno, D. (2018). Penggunaan aplikasi software iteman (*item and test analysis*) untuk analisis butir soal pilihan ganda berdasarkan teori tes klasik. *Jurnal Ilmu Fisika dan Pembelajarannya*. 2(2). 11-21. https://doi.org/10.19109/jifp.v1i1.866

Shwartz, Y., Ben-zvi, R., & Hofstein, A. (2006b). The Use of Scientific Literacy Taxonomy for Assessing the Development of Chemical Literacy Among High School Students. *Chemistry Education Research and Practice*, 7(4), 203–225. 10.1039/B6RP90011A

Tim Pusat Penilaian Pendidikan. (2019). *Panduan penilaian kinerja (performance assessment).* Pusat Penilaian Pendidikan: Jakarta.

Utami, S. H. A., Marwoto, P., & Sumarni, W. (2022). Analisis kemampuan literasi sains pada siswa sekolah dasar ditinjau dari aspek konten, proses, dan konteks sains. *Jurnal Pendidikan Sains Indonesia*, *10*(2). https://doi.org/10.24815/jpsi.v10i2.23802

Wahyuni, A., & Yusmaita, E. (2020). Perancangan instrumen tes literasi kimia pada materi asam dan basa kelas XI SMA/MA. *Edukimia*, *2*(3). https://doi.org/10.24036/ekj.v2.i3.a186

Widoyoko, E. P. (2018). *Teknik penyusunan instrumen penelitian pendidikan kimia.* Yogyakarta: Pustaka Pelajar.

Windyariani, S., Setiono, S., & Sutisnawati, A. (2017). Pengembangan Model Asesmen Literasi Sains Berbasis Konteks Bagi Siswa Sekolah Dasar. *Seminar Nasional Pendidikan Berkemajuan Dan Menggembirakan*, *2*.

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching dan Teacher Education*. 58, 149-162. 10.1016/j.tate.2016.05.010

Yasar, M. (2017). Comparison of item difficulty levels obtained as per two different methods. European *Journal of Education Sciences*. 4(3). http://dx.doi.org/10.19044/ejes.v4no3a4

Zahro, I. F., Atika, a. R., Westhisi, S. M. (2019). Strategi pembelajaran literasi sains untuk anak usia dini. *Jurnal Ilmiah Potensia*. 4(2). https://doi.org/10.33369/jip.4.2.121-130

Zein, A., Fadillah, M., & Novianti, R. (2013). Hubungan antara validitas butir, reliabilitas, tingkat kesukaran dan daya pembeda soal ujian semester genap bidang studi biologi kelas XI SMA/MA negeri di kota padang tahun Pelajaran 2010/2011. *Prosiding Semirata FMIPA Universitas Lampung*