An Analysis of Science Test Items

by Suwarto Suwarto

Submission date: 20-Aug-2023 04:47AM (UTC-0700)
Submission ID: 2148262944
File name: 12._SUWARTO_2.docx (191.36K)
Word count: 5056
Character count: 27735

An Analysis of Science Test Items

Prof. Dr. Suwarto, M.Pd

Universitas Veteran Bangun Nusantara, Sukoharjo, Indonesia,
e-mail: suwartowarto@yahoo.com

Abstract*: The research objective is to describe: the characteristics of science tests based on classical test theory and modern test theory. The research design is quantitative and descriptive. The objects of research are science achievement tests during Covid-19, science teachers, school principals, and deputy principals. The data was obtained from the responses of 280 students to all answer sheets of class VIII SMP MTA Gemolong Sragen as the population of this study. The answer keys for science questions and one package of science questions (50 multiple choice items) were obtained from the science teacher. Research techniques with interviews and documentaries. Data analysis was carried out using the Quest program. Research results: (1). Characteristics of science test based on classical test theory: Content validity is not met, test reliability is 0.960, item difficulty category in percentage is easy: moderate: difficult = 10%: 84%: 6%, item discriminatory category in percentage is poor: enough: good: very good = 2%: 4%: 14%: 80%, so the dominant distractor is very good, while the distractor function in percentage is not effective: effective = 0.70%: 99.30%, so the distractor dominant is effective. (2). The characteristics of the science test are based on modern test theory: the Threshold category for the science test in percentage is very difficult: difficult: medium: easy: very easy = 0%:12%:78%:10%:0%, so the dominant science test Threshold is moderate. The percentage of match between the IPA test items and the Racsh Model is 88%.*

Keywords*: Item difficulty, item discrimination, distractor function, reliability*

Introduction

Every education at certain times during an educational period always evaluates (Sultana, 2018). This means that the teacher always evaluates the results that have been achieved by students at certain times during the education period. Assessment of student learning outcomes must be carried out continuously in other words the teacher must continuously follow the learning outcomes that have been achieved by students from time to time. The teacher as an educator is to provide feedback to students about their progress and help improve their learning development.

Tests are one of the most effective measurement tools used by teachers to measure the quantity and quality of their learning. Crocker and Algina (1986) describe the test as a standard procedure for obtaining a sample of behavior from a particular domain. Tests are well-crafted instruments that, in total, measure realistic learning outcomes that represent expected behavioral traits. Etsey (2004) suggests that comprehensive learning objectives include observable behavior, conditions under which the intended behavior must be manifested, and a level of performance deemed sufficient to demonstrate mastery of learning outcomes helps in assessing knowledge and concepts that lead to cognitive, affective, and psychomotor development of students.

Tests are more widely used to evaluate student learning outcomes in terms of the cognitive domain. In the assessment of teaching Sciences, the cognitive aspect is often used as a benchmark for achieving science language learning outcomes. This can be seen in the final assessment of science language learning which only assesses cognitive aspects because the test items used only measure mastery of knowledge of the material being taught. To evaluate students' science achievement, the teacher usually gives students several questions in a test. The teacher can carry it out after each material chapter is finished or at the end of the semester. The test is called an achievement test. Achievement test is a test that is the focus of measurement is the learning objective. Achievement test is an assignment instrument in education that is very important as a source of information for decision-making. It is one of the most widely used measurement tools to determine student learning outcomes in teaching-learning processes or educational programs. It is important for teachers, schools, and educational institutions to do this to find out how far students have achieved the expected learning goals. Therefore, the researcher concluded that the achievement test was in the form of a planned test to reveal the maximum performance that had been taught. Teachers, schools, or other educational institutions can use achievement test results to make decisions or provide feedback to improve the teaching and learning process. In formal education activities in class, achievement tests can be in the form of daily tests, formative tests, summative tests, and college entrance exams (Suwarto, 2013).

A summative test is an assessment activity that generates scores, which are then used to determine student achievement. This test is carried out if the unit of learning experience or all of the subject matter has been completed. A summative test is used to determine the classification of awards at the end of a course or program (Putri, 2017 & Sugianto, 2017). On the other hand, formative tests are used to track how students are progressing in their studies and provide them with feedback that they can use to improve their performance as teachers and students. Formative tests help students better understand their strengths and limitations and how they can improve in those areas while also allowing teachers to see where their students are having difficulty and take quick action to help them.

A teacher as a test developer must know the basics of preparing a good achievement test to obtain valid and reliable measurement results. Learning, teaching, and content knowledge must all be in sync for a test score to be valid. When this occurs, the test value is actual. This is supported by Mulianah & Hidayat, 2013; Suwarto, 2016, 2021, 2023 and Cheng, Yang & Du, 2019, to obtain an actual score, practical tests are needed to identify accurately. A good test should consist of good items that meet the criteria of the test and offer actual information with minimum error. High-quality test results are the key that can explain actual learning outcomes. According to Suwarto (2021, 2023), a test is said to be a good test, and must meet the characteristics of a good test. This is; validity test, reliability test, item difficulty, item discrimination and especially for multiple choice tests, have an effective distractor for each item. Analyzing test items is needed to determine the level of validity and reliability of the assessment. As a result, the quality of the test will affect the test results. The quality of each item affects the quality of the test. The teacher should focus on the quality of the item items, so the teacher needs to do item analysis because by analyzing the item items, the teacher can identify the quality of each item, know which item fits the criteria, which item should be deleted, and which item should be deleted. revised.

During the Corona Virus Disease 2019 (Covid-19) pandemic from December 2019 until now, all teaching and learning processes, including exams, are temporarily carried out at home. This needs to be done to minimize mass physical contact to break the chain of spreading the virus. Therefore, through distance learning using cellphones, PC (personal computers), and laptops, evaluations and tests are carried out. A media is considered very effective in preventing the spread of Covid-19 in the educational environment. The teacher gives tests that are sent via cell phones or laptops to students or parents. Then students work on assignments or tests from home. During the pandemic, science achievement tests were carried out by subject teachers independently due to distance limitations, so that good tests made by subject teachers need to be investigated. Based on interviews conducted by researchers, the science achievement test during Covid-19 was conducted by a science teacher. The test is not piloted; even the science teacher made a test without making a grid in advance according to the syllabus. Therefore, the purpose of this study is to determine the validity, reliability, item difficulty , item discrimination, and the effectiveness of distractors based on classical test theory, as well as to see how the characteristics of tests based on modern test theory. This research is expected to provide input and examples for science teachers, educators, test developers, and other parties who make achievement tests. In addition, this research was conducted to provide a reference for similar research in the future.

According to Suryabrata (2005) and Fernandez (1984), the development of a specificity achievement test has an area, subject test, test objectives, material to be made for the test, type of test, and number of items in the test. Then, they designed a test grid that included specific objectives, specific values, and indicators. In building the test, they print the test items according to the grid that has been made, after that, the test must be validated qualitatively, professional judgment, quantitative validation, theoretical validation, material, construction, language validation, and content validity. Then, they revised the test according to the reviewers' input, and after that, all the good test items were assembled into a test. After completing a test, it must be tested on a group of students. To analyze, there is a classical test theory: item difficulty, item discrimination, distractor function, reliability, and content validity. After that, the test items are selected based on the results of the test analysis (classical test theory: accepted, revised, and rejected or modern test theory: threshold value, accepted or rejected, and suitability of the Rasch model or one-parameter logistic model). Finally, the test items that passed were compiled into a standardized test. Then the tests are printed and distributed to students or in schools.

The characteristics of a good test will be focused on quantitative item analysis. Richard & Sheila. (1999) explained that quantitative item analysis is an item study based on empirical data from the test being tested. There are two kinds of quantitative item analysis, namely analysis based on classical test theory and modern test theory. Item analysis based on classical test theory is a study of questions through information obtained from student answers to improve the quality of questions using classical test theory. This technique has several advantages, namely cheap, easy, can be implemented quickly, and simple. The characteristics of the test are the item difficulty, the item discrimination, distractors, validity, and reliability (Suwarto, 2021, 2023). Existing research, among others. Huda and Wahyuni's research (2019). Knowing the characteristics of science try-out questions based on Classical Test Theory (CTT) using the Iteman program.

Research by Hamimi, Zamharirah & Rusydy (2020), research to determine the quality level of questions consisting of validity, level of difficulty, reliability, deception, and discriminating power of math questions at SMP Negeri 1 Susoh. The test questions are made directly by the math teacher, who first creates a question grid based on competency standards and basic competencies. The form of test questions given to students is in the form of multiple choice and essay. However, the researcher only analyzed multiple choice. The test was made by the Mathematics MGMP (Musyawarah Guru Mata Pelajaran). The results of the study show that the questions used are relatively invalid because there are still many questions that have low and very low validity. The solution, the problem is not used. In addition, the questions studied also have a low level of reliability. However, these questions have a relatively good level of difficulty, with test results showing that most of the questions have a moderate level of difficulty. The questions have good discriminating power.

The similarity between Richard & Sheila and Hamimi, Zamharirah & Rusydy's research is that both quantitative analyzes were carried out with the Iteman program. The Iteman program is only able to analyze qualitatively based on classical test theory. Iteman's program is not capable of quantitative analysis based on modern test theory. Has been done by author is a quantitative analysis based on classical test theory and based on modern test theory, using the Quest program.

Research methods

The research design is quantitative and descriptive. Quantitative research because researchers calculate the characteristics of the test (item difficulty, item discrimination, the functioning of the distractor, test reliability, threshold value, Infit Meansquare, Outfit t, and item fit) using the Quest program. This research is descriptive because the researcher describes the characteristics of the test. The research location is at SMP MTA Gemolong Sragen. The objects of research were science achievement tests during Covid-19, science teachers, school principals, and vice principals. There is a set of science tests consisting of 50 multiple-choice questions. Data were obtained from students' responses to all answer sheets of class VIII students as the population of this study. There are 280 student answer sheets. Answer keys to science questions and a package of science questions were obtained from the science teacher. Researchers have conducted unstructured interviews in the form of open questions as a data collection technique. This is based on the research methods used by researchers, which depend heavily on the understanding of researchers and information data obtained from observations and interviews. The researcher asked permission from the administration and the school principal to conduct research at SMP MTA Gemolong Sragen. Second, the researcher asked the science teacher for class VIII to get information about the school program curriculum, and data for class VIII students, and asked how the science achievement test was made during the Covid-19 pandemic. Data analysis was performed using the Quest program.

Research Results and Discussion

Designing a science achievement test during Covid-19, no stages. It was made by a science teacher who teaches directly to his students. Based on the researcher's interview with him, he immediately made the test without making a grid. So, he makes it straight about adapting what he teaches in class over some time. He was simply copying and pasting from previous tests that the MGMP, himself, and other science teachers had made. In addition, he did not attempt to analyze test characteristics such as item difficulty, item discrimination, and the functioning of the distractor. And for the whole test is also not analyzed such as the validity test and reliability test.

Analysis Results Based on Classical Test Theory

The lowest item difficulty was 0.225 on item 49 and the highest item difficulty index was 0.739, namely items 1 and 27. Based on the item difficulty, it can be concluded that the most difficult item on the science achievement test made by the science teacher was item

49 while the items the easiest are items 1 and 27. A summary of the difficulty level of items by category in the science achievement test is presented in Table 1.

Table 1. Summary of Item Difficulty (p) from the Science Achievement Test

Category	Item Number	Total	Percentage
Easy $(0.70 < p \leq 1.00)$	1,12,27,31,35	5	10
Moderate $(0.30 \leq p \leq 0.70)$	2,3,4,5,6,7,8,9,10,11,13,14,15,16,17,18,19,20,21,22,23,24,25,26,28,29,30,32,33,34,36,37,38,39,40,41,43,44,46,47,48,50	42	84
Difficult $(0.00 \leq p < 0.30)$	42,45,49	3	6
Total		50	100

The Science test has 5 easy items with a percentage of 10%, 42 medium items with a percentage of 84%, and 3 difficult items with a percentage of 6%. Based on these results, the item difficulty is more dominant on medium items, so the researcher concludes that the item does not have proportional item difficulty, even though the ideal test should consist of 25% easy questions, 50% medium questions, and 25% questions difficult (Kunandar, 2013 & Suwarto, 2021, 2023). Roid & Haladyna (1982) stated that a test that does not have a proportional item difficulty level cannot reveal the actual competence of students. The test is also more dominant on moderate items, Brown (2004) confirms that items that are well made should not be too easy or difficult, the test must be balanced so that a science teacher can obtain information about students' natural science competencies. In contrast, Haider et al. (2012) argue that the category of moderate items can indicate that students have a good understanding of answering the test because more than half of the students answered the items correctly. The difficulty level of these test items can be compared to other studies that examine the difficulty level of summative test items (Mulianah & Hidayat, 2013; Maharani & Putro, 2020; Saputra, Retnawati & Yusron, 2021), even though the test conditions are not the same. Previous studies have found that the difficulty level of questions has more moderate items than the others.

Table 2. Summary of Item Discrimination ($r_{Pt.Biser}$) Science Achievement Test

Category	Item Number	Total	Percentage
Bad $(r_{Pt.Biser} \leq 0,19)$	6	1	2
Sufficient $(0.20 < r_{Pt.Biser} < 0.29)$	42,49	2	4
Good $(0.30 < r_{Pt.Biser} < 0.39)$	1,30,33,34,46,47,50	7	14
Very Good $(0.40 \leq r_{Pt.Biser})$	2,3,4,5,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,31,32,35,36,37,38,40,41,43,44,45,48	40	80
Total		5 0	100

This shows that the test contains more well-constructed items than poorly constructed items, but there is an imbalance between easy, medium, and difficult items. Items that are difficult and unbalanced are thought to be due to the Covid-19 pandemic which requires students to work at home, so students can ask friends for help, or can browse the internet.

All of these can affect the item difficulty index. The lowest item discrimination was 0.13 in item 6 and the highest item discrimination was 0.74 in item 28. The summary of item discrimination by category in the science learning achievement test conducted by science teachers is presented in Table 2. The item discrimination of this test is Good. Based on the Quest program, it shows that 1 item is bad with a percentage of 2%, 2 items are accepted with a percentage of 4%, 7 items are good with a percentage of 14%, and 40 items are very good with a percentage of 80%. These results indicate that 1 bad item must be dropped and 2 acceptable items must be revised (Dichoso & Joy, 2020). This result is good because 80% of the items are very good and 14% of the items are good (Dichoso & Joy, 2020). This means that most of the questions can be used to measure students' actual science competence. These items can also distinguish high achievers, moderate achievers, and low achievers. This is following Suwarto (2021, 2023) where the greater item discrimination implies that the item is increasingly able to distinguish between low-achieving and high-achieving students. This is to detect individual differences among students. The results of this test can be compared with other studies (Boopathiraj & Chellamani, 2013; Singh et al., 2014; Saputra et al., 2021), although the test conditions are not similar. The researchers found good item discrimination. Meanwhile, different results were found from previous studies such as (Sa'adah, 2017; Toksöz & Ertunç, 2017; Rehman, Aslam & Hassan, 2018; Manalu, 2019; Karim, Sudiro & Sakinah, 2021) which reported that items with poor discriminating power, then the item cannot differentiate between high achieving students and low achieving students.

The distractor is a multiple-choice answer that is wrong. Its function is to make students confused or miscalculate when choosing the correct answer among the alternatives provided. The distractor is said to be effective, if it is selected by more than 5% of the respondents, in a decimal number of 0.050. The distractor is said to be ineffective if it is chosen by less than 5% of the respondents or in a decimal number of 0.050. (Suwarto, 2021, 2023). Based on effective distractors and ineffective distractors the science achievement tests made by the science teacher are as follows. The percentage of ineffective distractors from the science achievement tests was 0.70%. The percentage of effective distracters on the science achievement tests is 99.30%. This test distractor has 1 ineffective distractor (0.7%) of 150 distractors that must be revised and 149 effective distractors (99.30%) of 150 distractors. The results of the percentage of effective distractors in this study were almost the same as in previous studies, namely Maharani & Putro, 2020. They found 80% of distractors were effective. This result can be explained that item discrimination can affect the deceptive index. Most of the science achievement items can distinguish between high and low achievers which can be assumed that a high item discrimination can lead to an effective deceptive index (Kheyami, Jaradat, Al-Shibani, & Ali, 2018). They also said that the ideal number of distractors was at least 3 items. The results of this study are more effective distractors so that the quality of the items is getting better. The results of the science achievement test research, almost all items have an effective distractor. It is assumed that the science teacher designed the test himself so that the teacher already knows the characteristics of the students.

The reliability of the science achievement test made by the science teacher was 0.960. This shows that the test items are very reliable. A test with a high level of reliability is classified as a good test (Sa'adah, 2017). In addition, good tests can be used for subsequent testing. The results of this study also show the extent to which science achievement test measurements remain consistent after being repeated on subjects and under the same conditions (Rudyatmi & Rusllowati, 2017). This reliable test is almost the same as previous studies (Anggreyani, 2009; Mulianah & Hidayat, 2013; Pascual, 2016; Sugianto, 2017; Manalu, 2019; Saputra et. al., 2021) although the test conditions are not the same. They found a reliable test. The estimated reliability of the test can be trusted because it is far above the reliability coefficient limit of 0.700. Several factors affect the estimation of reliability, including group homogeneity, time allocation, and test duration. In addition, another factor affecting the estimated reliability is the number of items that are classified as difficult (Crocker and Algina, 1986).

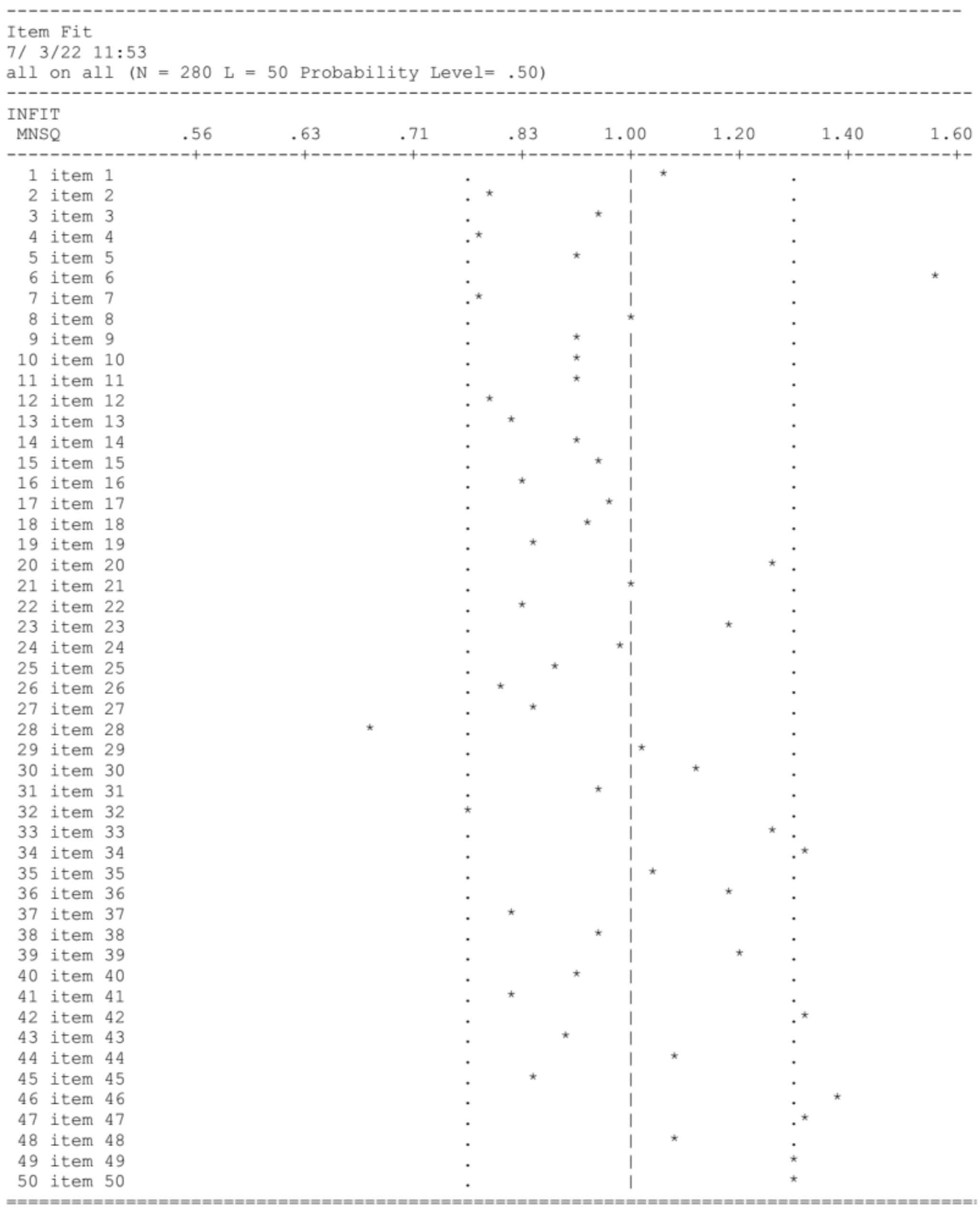
The analysis based on the classical test theory above has a weakness, namely the characteristics of the items depend on the group of test takers who are subjected to the items. In classical statistical test theory, questions such as the difficulty index of questions depend on the group of test takers, if the test is done by clever students, the questions are easy (the level of difficulty of the item becomes large) and vice versa, if the test is done by students who are not good at it, the questions become difficult (level of difficulty). the difficulty becomes small). Therefore, the characteristics of the questions are inconsistent or change depending on the ability of the students taking the test. Analysis based on classical test theory has a weakness because the characteristics of the test depend on the high group and the low group. Thus, this shows that when analyzing tests based on classical test theory, the characteristics of the tests are inconsistent or change depending on student achievement (Hambleton, Swaminathan & Rogers, 1991). Therefore, the researcher continues to analyze the characteristics based on modern test theory to analyze the characteristics of the test.

Analysis Results Based on Modern Test Theory

Table 3. Category Summary *Threshold* (*b*) the science achievement tests

Category	Item Number	Total (%)
Very Difficult ($b > 2$)	-	0 (0%)
Difficult ($1 < b \leq 2$)	6,21,42,45,48,49	6 (12%)
Moderate ($-1 \leq b \leq 1$)	2,3,4,5,7,8,9,10,11,13,14,15,16,17,18,19,20,22,23,24,25, 26,28,29,30, 32,33,34,36,37,38,39,40,41,43,44,46,47,50	39 (78%)
Easy ($-1 > b > -2$)	1,12,27,31,35	5 (10%)
Very Easy ($b < -2$)	-	0 (0%)
	Total	50 (100%)

Test characteristic analysis based on modern test theory uses one-parameter logistics (1PL) because the Quest program can only analyze the one-parameter logistic model (Adams & Khoo, 1996). Based on Table 3, the percentage of Threshold science achievement test = very difficult: difficult: moderate: easy: very easy = 0%:12%:78%:10%:0%.



```
----------------------------------------------------------------------------------------
Item Fit
7/ 3/22 11:53
all on all (N = 280 L = 50 Probability Level= .50)
-----------------------------------------------------------------------------------------
INFIT
 MNSQ           .56       .63       .71       .83      1.00      1.20      1.40      1.60
-----------------+---------+---------+---------+---------+---------+---------+---------+-
  1 item 1                                .              |  *           .
  2 item 2                                . *            |              .
  3 item 3                                .          *  |              .
  4 item 4                                .*             |              .
  5 item 5                                .         *    |              .
  6 item 6                                .              |              .            *
  7 item 7                                .*             |              .
  8 item 8                                .              *              .
  9 item 9                                .         *    |              .
 10 item 10                               .         *    |              .
 11 item 11                               .         *    |              .
 12 item 12                               . *            |              .
 13 item 13                               .   *          |              .
 14 item 14                               .         *    |              .
 15 item 15                               .           *  |              .
 16 item 16                               .    *         |              .
 17 item 17                               .            * |              .
 18 item 18                               .          *   |              .
 19 item 19                               .     *        |              .
 20 item 20                               .              |            * .
 21 item 21                               .              *              .
 22 item 22                               .    *         |              .
 23 item 23                               .              |        *     .
 24 item 24                               .             *|              .
 25 item 25                               .       *      |              .
 26 item 26                               .  *           |              .
 27 item 27                               .     *        |              .
 28 item 28                      *         .              |              .
 29 item 29                               .              |*             .
 30 item 30                               .              |     *        .
 31 item 31                               .           *  |              .
 32 item 32                               *              |              .
 33 item 33                               .              |            * .
 34 item 34                               .              |              .*
 35 item 35                               .              | *            .
 36 item 36                               .              |        *     .
 37 item 37                               .   *          |              .
 38 item 38                               .           *  |              .
 39 item 39                               .              |         *    .
 40 item 40                               .         *    |              .
 41 item 41                               .   *          |              .
 42 item 42                               .              |              .*
 43 item 43                               .        *     |              .
 44 item 44                               .              |   *          .
 45 item 45                               .     *        |              .
 46 item 46                               .              |              .   *
 47 item 47                               .              |              .*
 48 item 48                               .              |   *          .
 49 item 49                               .              |              *
 50 item 50                               .              |              *
=========================================================================================
```

Figure 1. Fit map items for the Science Achievement Test

Based on Figure 1, it can be seen that the 6 items on the Science achievement test are not fit because the asterisk statistics are out of fit which is between the two dotted vertical lines, namely: items 6, 28, 34, 42, 46, and 47, while 44 other items fit (Adams & Khoo, 1996). The percentage of compatibility of the Science test items with the Racsh Model = 44/50x100% = 88%.

Table 4. Analysis of Accepted and Rejected Science Achievement Test Items

Category (Criteria)	Item Number	Sum (%)
Accepted (*Outfit t* ≤ 2.00)	1,2,3,4,5,7,8,9,10,11,12,13,14,16,17,18,19,21,22,23 ,25,26,27,28,29, 30,31,32,33,35,36,37,38,39,40,41,43,44,45,48	40 (80%)
Rejected (*Outfit t* > 2.00)	6,15,20,24,34,42,46,47,49,50	10 (20%)
	Total	50 (100%)

Based on Table 4, the percentage of science test questions that passed = accepted: rejected = 80%:20%. Meanwhile, if you look at the Item Fit Map for the Science achievement test, you can see Figure 1.

Conclusions and suggestions

Characteristics of the science test based on classical test theory: Content validity was not met, test reliability was 0.960, item difficulty category in percentage was easy: moderate: difficult = 10%:84%:6%, item discrimination category in percentage was bad: enough: good: very good = 2%: 4%: 14%: 80%, so the dominant is very good, while the function of the distractor in percentage is ineffective: effective = 0.70%: 99.30%, so the distractor is effective dominant. The characteristics of the Science test are based on modern test theory: the Threshold category of the Science test in percentage terms is very difficult: difficult: moderate: easy: very easy = 0%:12%:78%:10%:0%, so the Science test Threshold is moderate. The percentage of compatibility of the Science test items with the Racsh Model is 88%.

Suggestions that can be given item difficulty levels should be made 25 percent easy, 50 percent moderate, and 25 difficult. Thus, the ability of students who are low, medium, and high can all be measured. Content validity should be fulfilled, so that the Science test has items that can measure what should be measured (all aspects that must be measured are represented in the Science test items). The items that make up the Science test should conform to the Rasch Model.

References

Adams, R. J. & Khoo, S. T. (1996). *Quest the Interactive Test Analysis System*. Australia: The Australian Council for Educational Research Ltd.

Anggreyani, A. (2009). *Penerapan Teori Uji Klasik dan Teori Respon Butir dalam Mengevaluasi Butir Soal*. Departemen Statistika Fakultas Matematika Dan Ilmu Pengetahuan Alam Institut Pertanian Bogor.

Boopathiraj, C. & Chellamani, K. (2013). Analysis of Test Items on Difficulty Level and Discrimination Index in The Test pfor Research in Education. *International Journal of Social Science & Interdisciplinary Research 2*(2), 189-193. Retrieved from: indianresearchjournals.com.

Brown, H. D. (2004). *Language Assessment: Principle and Classroom Practice*. United States of America: Pearson Education.

Cheng, Y., Yang, Y. & Du, D. (2019). A class of asymptotically optimal group testing strategies to identify good items. *Discrete Applied Mathematics Journal*. 260, 109–116.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: CBS College.

Dichoso, A. A. & Joy M. R. J. (2020). Test Item Analyzer using Point-Biserial Correlation and P-Values. *International Journal of Scientific & Technology Research, 9*(4). 2122-2126

Etsey, Y. K. (2004). "Assessing performance in schools: Issues and practice," Ife Psychologia, vol. 13, no. 1, 123-135

Fernandez. H. J. X. (1984). *Testing and Measurement*. Jakarta: National Education Planning Evaluation and Curriculum Development.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. London: Sage Publication.

Hamimi, L., Zamharirah, R. & Rusydy. (2020). Analisis Butir Soal Ujian Matematika Kelas VII Semester Ganjil Tahun Pelajaran 2017/2018. *Mathema Journal*. *2*(1). 57-66.

Huda, N. & Wahyuni, T. S. (2019). Analisis Butir Soal IPA Try Out USBN Tahun Ajaran 2018/2019 dalam Kaitannya dengan Level Kognitif. *Jurnal Pendidikan dan Pembelajaran Dasar*. *12*(1), 29-39.

Kheyami, D., Jaradat, A., Al-Shibani, T., & Ali, F. A. (2018). Item Analysis of Multiple-Choice Questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos Univesity Medical Journal, 18*(1), 68-74.

Kunandar, K. (2013). *Penilaian Autentik: Penilaian Hasil Belajar Peserta Didik Kurikulum 2013*. RajaGrafindo Persada.

Maharani, A. V. & Putro, N. H. P. S. (2020). Item Analysis of English Final Semester Test. *Indonesian Journal of EFL and Linguistics*. *5*(2), 2020. 492-504.

Manalu, D. (2019). An Analysis of Students Reading Final Examination by Using Item Analysis Program on Eleventh Grade of SMA Negeri 8 Medan. *Journal of English Teaching & Applied Linguistics, 1*(1), 13-19.

Mulianah, S. & Hidayat, W. (2013). Pengembangan Tes Berbasis Komputer. *Kuriositas, 2*(6), 27- 43.

Pascual, G. R. (2016). Analysis of The English Achievement Test for ESL Learners in Northern Philippines. *International Journal of Advanced Research in Management and Social Sciences, 5*(12),1-5. retrieved from www.garph.co.uk.

Putri, B. D. T. (2017). the Validity Analysis of English Summative Test of Junior High School. *Journal of Languages and Language Teaching*. *5*(1), 6-11.

Rehman, A., Aslam, A. & Hassan, S. H. (2018). Item Analysis of Multiple-Choice Questions. *Pakistan Oral & Dental Journal, 38*(2), 291-293.

Richard & Sheila. (1999). *Item Analysis for Criterion-Referenced Tests*. New York: Research Foundation of SUNY/Center for Development of Human Services.

Roid, G. H. & Haladyna, T. M. (1982). *A Technology for Test-Item Writing*. London: Academic Press, Inc.

Rudyatmi, Ely & Rusllowati, A. (2017). *Evaluasi Pembelajaran*. Semarang: Faculty of Mathematics and Science Unnes.

Sa'adah, N. (2017). The Analysis of English Mid-Term Test Items based on the Criteria of a Good Test at the First Semester of the Eighth Grade Students of Mts. Mathalibul

Huda Mlonggo in The Academic Year Of 2016/2017. *Journal Edulingua, 4*(1),45-58.

Saputra, A. N. S., Retnawati, H. & Yusron, E. (2021). Analysis Difficulties and Characteristics of Item Test of on Biology National Standard School Examination. *Advances in Social Science, Education and Humanities Research*, *542*, 8-14.

Singh, J. P., Kariwal P., Gupta S.B., & Shrotriya V.P. (2014). Improving Multiple Choice Questions (MCQs) through item analysis: An assessment of the assessment tool. *International Journal of Sciences & Applied Research, 1*(2), 53-57. Retrived from: www.ijsar.in.

Sugianto, A. (2017). Validity and Reliability of English Summative Test for Senior High School. *Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature, 3*(2), 22-38. P-ISSN: 2460-0938. E-ISSN: 2460-2604.

Sultana, N. (2018). Test Review of the English Public Examination at the Secondary Level in Bangladesh. *Language Testing in Asia, 8*(16), 1-9.

Suryabrata, S. (2005). *Pengembangan Alat Ukur Psikologis*. Yogyakarta: C.V Andi Offset.

Suwarto. (2013). Difficulty, Difference, and Reliability Level of New Student Selection Test for Veteran Bangun Nusantara Sukoharjo University. *National Seminar on Science Education*. 652- 658.

Suwarto. (2013). *Pengembangan Tes Diagnostik Dalam Pembelajaran*. Yogyakarta: Pustaka Pelajar.

Suwarto. (2016). The Biology Test Characteristic of 7th Grade by The Period of The Odd Term. *Jurnal Penelitian Humaniora*. *17*(1), 1-8.

Suwarto. (2021). The Characteristics of Indonesia a Second- Semester Final Test for Eight-grade Students. *Turkish Online Journal of Qualitative Inquiry*. *12*(9), 356-370.

Suwarto, S., Suyahman, S., Meidawati, S., Zakiyah, Z., & Arini, H. (2023). The COVID-19 pandemic and the characteristic comparison of English achievement tests. *Перспективы науки и образования*, (2 (62)), 307-329.

Toksöz, S. & Ertunç., A. (2017). Item Analysis of a Multiple-Choice Exam. *Advances in Language and Literary Studies, 8(*6), 141-146.

An Analysis of Science Test Items

ORIGINALITY REPORT

5%	4%	4%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Thresia Trivict Semiun, Maria Wihelmina Wisrance, Merlin Helentina Napitupulu. "English Summative Test: The Quality of Its Items", English Education:Journal of English Teaching and Research, 2022 Publication	1%
2	www.schuhfried.com Internet Source	1%
3	repository.upi.edu Internet Source	1%
4	www.aeaafrica.org Internet Source	1%
5	Taza Nur Utami, Hartono. "Mobile game as a media learning mathematics in the Covid-19 pandemic", AIP Publishing, 2022 Publication	1%
6	eprints.nottingham.ac.uk Internet Source	1%
7	es.scribd.com Internet Source	1%

8	jurnal.unimed.ac.id Internet Source	1%
9	Ibnu Rafi, Heri Retnawati, Ezi Apino, Deni Hadiana, Ida Lydiati, Munaya Nikma Rosyada. "What might be frequently overlooked is actually still beneficial: Learning from post national-standardized school examination", Pedagogical Research, 2023 Publication	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		