Phys. Comm. 6 (1) 2022 : 1 - 6

Physics Communication

http://journal.unnes.ac.id/nju/index.php/pc

Analysis of Four-tier Diagnostic Test on The Topic of Temperature and Heat in High School

Nur Fitri Asih✉, Suharto Linuwih, Fianti

Postgraduate Universitas Negeri Semarang, Semarang, Indonesia

Article Info

Article history:
Submitted 20 November 2021
Revised 17 February 2022
Accepted 18 February 2022

Keywords:
Level of Problem Difficulty, Differential Power of Questions, Distractor Function, Four-tier Diagnostic Test

Abstract

This study aims to analyze the feasibility of the four-tier diagnostic test items given to class XI high school students. This research is a quantitative descriptive study, which collects student answer sheets as data which is then analyzed for item selection, reliability, level of difficulty, discriminating power, distractor functioning, and readability test. This research was conducted on 30 students of class XI at SMA N 1 Slawi. The results concluded that in the selection of items from the 21 questions made, there were 16 items that were suitable for use and 5 items that were not suitable for use. In the reliability of the questions obtained by the reliability value of rcount of 0.760. In the level of difficulty, there are 5 questions in the easy category, 14 questions in the medium category, and 2 questions in the difficult category. In the differentiating power, the questions in the good category are 8 questions, the sufficient category is 8 questions, and the bad category is 5 questions. In the distractor function there are 10 questions whose distractors function well and 11 questions whose distractors do not function properly. Readability test questions according to student assessment questionnaires, four-tier diagnostic test questions are categorized as good to use.

✉Correspondence:
Postgraduate Universitas Negeri Semarang
Jalan Kelud Utara III No.37, Kota Semarang, Indonesia 50237
e-mail: nurfitriasih@students.unnes.ac.id

p-ISSN 2528-5971
e-ISSN 2528-598X

INTRODUCTION

The importance of the role of assessment is needed in the world of education. Assessment is used as an evaluation tool and a measuring tool to measure the success of the student learning process, especially in physics subject matter. Physics is a branch of natural science that is studied at the high school level and examines natural phenomena. In learning physics, there are still many obstacles that occur in the field. One of the obstacles in physics material that students often face is their understanding of the concepts of temperature and heat. Therefore, a measuring tool to diagnose student's understanding ability is required.

According to Zaleha *et al*. (2017), a diagnostic test is a test that is used to determine precisely and ascertain the weaknesses and strengths of students in certain subjects. One of them is by using a four-tier diagnostic test instrument. The four-tier diagnostic test is a development of the three-tier multiple-choice diagnostic test (Fariyani *et al.*, 2015). The difference with previous developments of the three-tier multiple-choice diagnostic test is the addition of students' confidence levels in choosing reasons (Pujayanto *et al*., 2018). The first level is a multiple choice question with five distractors and one answer key. The second level is the level of student confidence in choosing answers. The third level is the reason students answer questions. The fourth level is the level of students' confidence in giving reasons (Gurel *et al*., 2017). Previous research on three-level multiple-choice diagnostic tests was conducted by Lemma & Ethiopia (2012) and Arslan *et al*. (2012) on diagnosing students' misconceptions.

To measure the feasibility of the four-tier diagnostic test instrument questions that have been made, it is necessary to analyze items which include item selection, item reliability, item difficulty level, distinguishing power, distractor function, and readability test items through student assessment questionnaires. Item analysis or item analysis is an assessment of test questions to obtain a set of questions that have adequate quality (Sudjana, 2013). According to Thorndike and Hagen, the analysis of test items has two important objectives. First, the answers to the questions are diagnostic information to examine the lessons of the class and their learning failures and continue to guide them toward good learning. Second, answers to separate questions and improvement of questions based on answers are the basis for preparing better tests for the following year (Purwanto, 2006). With the item analysis, it is possible to identify good questions and bad questions and which questions can be added to the question bank revised, or discarded (Salmina & Adyansyah, 2017).

According to Arikunto (2007), a test can be said to have a high level of confidence if the test can provide fixed results. Reliability refers to the consistency or stability of the assessment results, so it is used to measure an instrument that has high consistency, is accurate, reproducible, and generalizable. According to Fatimah & Alfath (2019), good questions are those whose level of difficulty can be known as not too difficult and not too easy. The level of difficulty of the items correlates with distinguishing power. If the item has a maximum difficulty level, the discriminatory power will be below. Likewise, if the items are too easy, they will not have distinguishing power.

Good questions must also have a distractor or distractor that works well. According to Sabri & Idris (2013), distractors are classified as wrong answers to multiple-choice questions. The distractor or distractor serves as a tool that can describe whether the items made are good or fail. The purpose of using distractor is to deceive those who are less able or do not know to be distinguished from those who are able (Thoha, 1994). According to Arikunto (2013) good distractors will be chosen by students who are less intelligent and not chosen by smart students. If the distractor is chosen by most of the smart students, then the distractor can be said to be not functioning. The distractor function is said to be good if the distractor function has a great appeal to the test takers who do not understand the concept, which at least 5% of the total test-takers.

In addition to the analysis of item selection, item reliability, level of item difficulty, differentiating power of questions, and distractor functionality, it is also necessary to conduct a readability test so that the four-tier diagnostic test questions used can be analyzed for the level of readability and ease of understanding the language in the questions used.

METHODS

The research method used in this study was a descriptive method with an approach that is by collecting student answer sheets as data for further analysis of item selection, reliability, level of difficulty, discriminating power, and distractor functioning. This study also tested the readability of the questions on students through student assessment questionnaires. The data collection technique in this research was the documentation technique. The diagnostic test questions used were in a four-tier format on temperature and heat materials. This four-tier diagnostic test was tested on class XI SMA students who have received the material on temperature and heat. This research was conducted at SMA N 1 Slawi in the academic year 2021/2022 with a sample size of 30 students.

RESULTS AND DISCUSSION

The four-tier diagnostic test used consists of 21 questions containing temperature and heat

material at the XI high school level. The four-tier diagnostic test questions used were questions that have been validated by expert validators. The diagnostic questions were then tested on 30 students as samples. The item analysis of the four-tier diagnostic test instrument for temperature and heat produced are as follows.

1) Item Selection

To select valid items, the crude product-moment correlation formula was used, then the r_{xy} value obtained was consulted with the product-moment r_{table} with a significant level of 5%. If the value of $r_{count} > r_{table}$, then the items used were valid (Arikunto, 2010). The results of the calculation of item selection were as in Table 1.

Table 1. Results of Item Selection

Question Consistency Category	Question Number	Number of Questions
Valid	1,2,3,4,5,6,7,8,9,12,13,15,16,17,19,20	16
Invalid	10,11,14,18,21	5

Based on Table 1. shows that of the 21 items analyzed for item selection, 16 items were obtained that could be used and 5 items that could not be used.

2) Reliability

Reliability relates to the level of trust. A test was said to have a high level of confidence if the test can give fixed results (Arikunto, 2007). A question was said to be reliable if it gives consistent results even though it was used many times. The results of the calculation of the reliability of the questions calculated using the Cronbach Alpha formula obtained the reliability value of r_{count} of 0.760 which was then distributed to r_{table}. The results of the calculation of the reliability of the question were declared reliable. According to Sujarweni (2014), the instrument was said to be reliable if the Cronbach alpha obtained from the analysis using the IBM SPSS Statistics software > 0.6. According to Tavakol & Dennick (2011), standardized and reliable instruments cannot be directly used anywhere, anytime, and on any subject but need to be re-tested every time they were used. Supahar *et al*. (2017), concluded that the instrument was said to be consistent (reliable) if the instrument was carried out from time to time but has the same value. In a previous study by McClary & Bretz (2012), the four-level multiple-choice diagnostic test he developed resulted in a reliability value of 0.41. Bradshaw & Templin (2014) also developed a diagnostic test with an average reliability of 0.988. This shows that the four-tier diagnostic test can be used as an assessment tool.

3) Level of Difficulty

Good questions were questions that were not too easy and not too difficult (Arikunto, 2007). Questions that were too easy do not stimulate students to enhance their efforts, while questions that were too difficult cause students to become discouraged and do not have the enthusiasm to try again. The results of the calculation of the difficulty level of the questions can be seen in Table 2.

Table 2. Results of the Difficulty Level of Questions

Level of Difficulty Category	Question Number	Number of Questions
Easy	3,4,17,18,21	5
Medium	2,5,6,7,8,9,10,12,13,14,15,16,19,20	14
Difficult	1,18	2

Based on Table 2. shows that of the 21 items used, the level of difficulty obtained was 5 items in the easy category, 14 items in the medium category, and 2 items in the difficult category.

4) Distinguishing Power

According to Rusilowati (2014), the discriminatory power of a question was the ability of an item to distinguish between students who have mastered the material and students who have not/less/have not mastered the material. The results of the calculation of the discriminatory power of the questions can be seen in Table 3.

Table 3. Results of Discriminatory Power of Questions

Distinguishing Power Category	Question Number	Number of Questions
Very well	-	0
Good	1,6,7,9,13,16,17,20	8
Sufficient	2,3,4,5,8,12,15,19	8
Bad	10,11,14,18,21	5

Based on Table 3. the results of the differentiating power of the questions show that of the 21 questions used, the distinguishing power of 8 items was in the good category, 8 items in the sufficient category, and 5 items in the bad category.

5) Distractor Function

Multiple-choice questions consist of questions and alternative answers that must contain the correct answer key and distractor or distracting answers. The distractor function analysis was intended to determine whether the available distractors were functioning or not. According to Arikunto (2013) states, the distractor function was said to be good if the distractor function has a great appeal to test takers who do not understand the concept of at least 5% of the total test-takers. The results of the distractor function analysis can be seen in Table 4.

Table 4. Results of the Distractor or Distractor Functional Analysis

Question Number	Distractor Function										Description
	Answer					Reason					
	A	B	C	D	E	A	B	C	D	E	
1	√	√	√	√	√	√	√	√	√	√	function
2	√	√	√	√	√	√	√	-	-	-	lack of function
3	√	√	-	-	√	√	√	√	-	-	lack of function
4	√	-	-	-	-	√	-	-	√	√	not function
5	-	√	√	-	√	√	√	√	√	√	lack of function
6	√	√	√	√	√	√	√	√	√	√	function
7	√	√	√	√	√	√	√	√	√	√	function
8	√	√	√	√	√	√	√	√	√	√	function
9	√	√	√	√	√	√	√	√	√	-	lack of function
10	-	√	√	-	-	√	√	√	√	√	lack of function
11	√	√	-	√	√	√	√	√	-	√	lack of function
12	√	√	√	√	√	√	√	√	√	√	function
13	√	√	√	√	√	√	√	√	√	√	function
14	√	√	√	-	-	√	-	√	√	-	lack of function
15	√	√	√	√	√	√	√	√	√	√	function
16	√	√	√	√	√	√	√	√	√	√	function
17	√	√	-	-	-	√	√	√	√	√	lack of function
18	-	-	√	-	-	-	-	√	√	-	not function
19	√	√	√	√	√	√	√	√	√	√	function
20	√	√	√	√	√	√	√	√	√	√	function
21	√	√	-	-	-	√	-	√	√	-	not function

Based on Table 4. it can be seen that the results of the analysis of the functioning of the distractors or distractors contained 10 items that functioned well from the 21 items used. Items whose distractors or distractors function properly are number 1, 6, 7, 8, 12, 13, 15, 16, 19, and 20.

6) Readability Test
The readability test was carried out on class XI SMA students who had already obtained the material on temperature and heat. The readability test aims to determine the level of readability and the level of ease of understanding language or sentences in the four-tier diagnostic test questions that ware made. The readability test was measured using a student assessment questionnaire given to 30 students of SMA N 1 Slawi as respondents. The student assessment questionnaire made consists of 9 statement points that ware used to measure the readability test. The contents of the student assessment questionnaire statements include readability of sentences in the assessment, ease of question sentences to be understood, accuracy of the composition and length of sentences in the assessment, ease of understanding assessment statements, absence of question sentences in giving rise to multiple interpretations, legibility of tables and figures in assessment, ease of understanding pictures and tables in the assessment, the suitability of the number of questions in the assessment, the suitability of the time allotted to answer and solve the questions. The results of the readability test based on student assessment questionnaires on the four-tier diagnostic test instrument can be seen in Table 5.

Table 5. Readability Test Results Based on Student Assessment Questionnaires

No.	Rated Aspect	Percentage	Category
1	Readability of sentences in the assessment	78.62%	Good
2	Ease of question sentences to understand	70.34%	Good
3	The accuracy of the arrangement and length of sentences in the assessment	69.66%	Good Enough
4	Easy to understand assessment statement	67.59%	Good Enough
5	The absence of a question sentence in causing multiple interpretations	68.97%	Good Enough
6	Readability of tables and figures in assessment	82.76%	Good
7	Ease of understanding figures and tables in the assessment	79.31%	Good
8	The suitability of the number of questions in the assessment	78.62%	Good
9	Appropriate time allotted to answer and solve questions	71.72%	Good

Based on Table 5 it can be concluded that the four-tier diagnostic test instrument on the readability test according to the student assessment questionnaire was categorized as good to use.

CONCLUSION

Based on the research that has been done, it can be concluded that the four-tier diagnostic test instrument for temperature and heat material given to high school students in class XI was feasible to use. The results of the analysis of the four-tier diagnostic test instrument used to have 15 items with good quality and good readability tests.

ACKNOWLEDGEMENT

The researcher expresses his gratitude to SMA N 1 Slawi for providing the opportunity to carry out research, the experts who have provided an assessment of the four-tier diagnostic test instrument, and to all parties who have helped so that this research runs smoothly.

REFERENCES

Arikunto, S. (2007). *Dasar-dasar Evaluasi Pendidikan (Edisi revisi)*. Bumi Aksara.

Arikunto, S. (2010). *Prosedur Penelitian: Suatu Pendekatan Praktek (Edisi revisi)*. Rineka Cipta.

Arikunto, S. (2013). *Dasar-dasar Evaluasi Pendidikan (Edisi 2)*. Bumi Aksara.

Arslan, H. O., Cigdemoglu, C., & Moseley, C. (2012). A Three-Tier Diagnostic Test to Assess Pre-Service Teachers' Misconceptions about Global Warming, Greenhouse Effect, Ozone Layer Depletion, and Acid Rain. *International Journal of Science Education*, *34*(11), 1667–1686. https://doi.org/10.1080/09500693.2012.680618

Bradshaw, L., & Templin, J. (2014). Combining Item Response Theory and Diagnostic Classification Models: A Psychometric Model for Scaling Ability and Diagnosing Misconceptions. *Psychometrika*, *79*(3), 403–425. https://doi.org/10.1007/s11336-013-9350-4

Fariyani, Q., Rusilowati, A., & Sugianto, S. (2015). PENGEMBANGAN FOUR-TIER DIAGNOSTIC TEST UNTUK MENGUNGKAP MISKONSEPSI FISIKA SISWA SMA KELAS X. *Journal of Innovative Science Education*, *4*(2), 41–49. http://journal.unnes.ac.id/sju/index.php/jise

Fatimah, L. U., & Alfath, K. (2019). Analisis Kesukaran Soal, Daya Pembeda dan Fungsi Distraktor. *Jurnal Komunikasi Dan Pendidikan Islam*, *8*, 37–64.

Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *Research in Science and Technological Education*, *35*(2), 238–260. https://doi.org/10.1080/02635143.2017.1310094

Lemma, A., & Ethiopia. (2012). Diagnosing the Diagnostics : Misconceptions of Twelfth Grade Students on Selected Chemistry Concepts in Two Preparatory Schools In Eastern Ethiopia. *ASEAN Journal of Community Engagement*, *2*(2), 16–31.

McClary, L. M., & Bretz, S. L. (2012). Development and assessment of a diagnostic tool to identify organic chemistry students' alternative conceptions related to acid strength. *International Journal of Science Education*, *1*, 1–25. https://doi.org/10.1080/09500693.2012.684433

Pujayanto, Budiharti, R., Radiyono, Y., Nuraini, N. R. A., Putri, H. V., Saputro, D. E., & Adhitama, E. (2018). Developing Four Tier Misconception Diagnostic Test About Kinematics. *Cakrawala Pendidikan*, *37*(2), 237–249.

Purwanto, M. N. (2006). *Prinsip-prinsip dan Teknik Evaluasi Pengajaran*. Remaja Rosdakarya.

Rusilowati, A. (2014). *Pengembangan Instrumen Penilaian*. Unnes Press.

Sabri, S., & Idris, S. (2013). Item Analysis of Student Comprehensive Test for Research in Teaching Beginner String Ensemble Using Model Based Teaching Among Music Students in Public Universities. *International Journal of Education and Research*, *1*(12), 1–14.

Salmina, M., & Adyansyah, F. (2017). Analisis Kualitas Soal Ujian Metematika Semester Genap Kelas XI SMA Inshafuddin Kota Banda Aceh. *Numeracy*, *4*(1), 37–47.

Sudjana, N. (2013). *Penilaian Hasil Proses Belajar Mengajar*. Remaja Rosdakarya.

Sujarweni, V. W. (2014). *Metode Penelitian: Lengkap, Praktis, dan Mudah Dipahami*. Pustaka Baru Press.

Supahar, Rosana, D., Ramadani, M., & Dewi, D. K. (2017). The Instrument for Assessing The Performance of Science Process Skills Based on Nature of Science (NOS). *Cakrawala Pendidikan*, *36*(3), 435–445.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Thoha, M. C. (1994). *Teknik Evaluasi Pendidikan*. PT. Raja Grafindo.

Zaleha, Samsudin, A., & Nugraha, M. G. (2017). Pengembangan Instrumen Tes Diagnostik VCCI Bentuk Four-Tier Test pada Konsep Getaran. *Jurnal Pendidikan Fisika Dan Keilmuan (JPFK)*, *3*(1), 36–42. https://doi.org/10.25273/jpfk.v3i1.980