ISEJ Indonesian Science Education Journal

ISEJ: Indonesian Science Education Journal
ISSN: 2716-3350
Vol. 2, No. 1, Januari 2021, Hal 27-37

Analysis Of The Quality Of Items for VIII Class IPA Teachers MTs.N I Kota Bengkulu Study Year 2019/2020

Raden Gamal Tamrin Kusumah[1], Hadiwinarto[2]

[1,] *Mahasiswa Program Studi Doktor Pendidikan FKIP Universitas Bengkulu*
[2] *Fakultas Keguruan dan Ilmu Pendidikan Universitas Bengkulu*

Correspondent e-mail:
[1] raden@iainbengkulu.ac.id
[2] hadiwin@unib.ac.id

Abstract
This study aims to analyze the validity, reliability, level of difficulty, differentiation power, and item distractors made by science subject teachers of class VIII MTs.N 1 Bengkulu, Academic Year 2019/2020. Samples were taken by saturated sampling technique. The instrument in this study was a daily test sheet made by a science subject teacher for grade VIII, which amounted to 20 multiple choice questions. Analysis of the construct validity of multiple choice questions in terms of material, construction and language is appropriate, there are several questions that still need to be improved in terms of construction. Analysis of the validity of the UH 1 items showed that 40% of the questions were valid; reliability of 0.76 means that it has high consistency; the difficulty level of the questions was 40% difficult, 55% moderate and 5% easy questions; difference power 45% bad, 35% sufficient and 20% good; 80% of cheaters are functional. At UH 2 60% of the questions are valid; reliability of 0.78 means it has high consistency; 35% for difficult questions and 65% for medium questions; difference power 30% bad, 45% sufficient and 25% good; 75% of cheaters are functioning.

Keywords: Question Quality, Question Items, Natural Science Lessons

INTRODUCTION

The learning process is a system consisting of several components that are interrelated and interact in achieving learning objectives. One of the most important components in the teaching and learning process is evaluation. Evaluation or evaluation is a systematic process to determine or make decisions to the extent to which program objectives have been achieved (inGronlund, Djaali and Muljono, 2008). In relation to education, Nurkancana and Sunartana (1986) argue that educational evaluation is an action or a process to determine the value of everything related to education.

In evaluation activities, assessment tools or techniques are needed, so that their implementation will be more focused. Evaluation tools in education used to collect data can be in the form of tests or non-tests (Purwanto, 2011). Nurkancana and Sunartana (1986) state that the test is a way to conduct an assessment in the form of a task or a series of tasks that must be done

by a student or a group of students so as to produce a value about student behavior or achievement as learners.

Tests can be arranged in the form of tests (questions) in the form of objective or subjective. According to Purwanto (2011) an objective test is a test in which all the information needed to answer the test is available. The subjective test according to Nurkancana and Sunartana (1986) is a form of test consisting of questions or orders which require answers in the form of relatively long descriptions. The two forms of tests used in this evaluation must be justified, meaning that the test can qualify as a good evaluation tool when viewed from the quality of the items.

To determine the quality of the items, it can be revealed through item analysis. Item analysis can provide detailed information about the state of each item such as the strengths and weaknesses of the items, the complete specification of the questions and the problems contained in the questions such as errors in making the answer keys, too difficult or easy questions and so on. The questions that have been prepared by the teacher, whether it is the end of semester test questions or the daily test questions, need to be analyzed. This refers to the opinion of Suryabrata (1987) as quoted below: Questions that have been written carefully based on various considerations cannot simply be considered good questions. These questions still need to be tested through question analysis.

The good and bad of a test or evaluation tool can be viewed from its validity, reliability, level of difficulty and differentiation (Nurkancana, 1986). A test is called valid or has validity if the test can precisely measure what is being measured. The validity of the items needs to be sought to find out which test items cause the overall item to be bad because it has low validity. Furthermore, the items are said to be valid if they have great support for the total score. The score on the item causes the total score to be high or low. In other words, it can be said that the items have high validity if the scores on the items are parallel to the total scores.

This alignment is defined as correlation, so to determine the validity of the items used the correlation formula (Arikunto, 2008). In this study, an analysis of the validity of the items was carried out using the correlation formula product moment coarse number.

Reliability is the determination of a test when tested on the same subject (Arikunto, 2008). To find out this permanence, basically the parallel results are seen. To analyze the reliability of the testin this study the items Kuder-Richarson 20 formula or abbreviated as KR-20 was used.

The quality or not of the learning result test items can also be seen from the degree of difficulty or the level of difficulty that each item has. Learning result test items can be said to be good items if they are not too difficult and not too easy in other words, the degree of difficulty of the items is moderate or sufficient (Sudijono, 2007). According to Arikunto (2008), a good question is one that is not

too easy or not too difficult. Problems that are too easy do not stimulate students to increase their solving efforts. Conversely, questions that are too difficult will cause students to become discouraged and have no enthusiasm to try again because they are out of reach. The number that shows the difficulty and ease of a problem is called the difficulty index. The magnitude of the difficulty index is between 0.0 and 1.0. This difficulty index shows the difficulty level of the question. A problem with a difficulty index of 0.0 indicates that the item is too difficult, on the other hand an index of 1.0 indicates that it is too easy. In evaluation terms, this difficulty index is given the symbol P (Stands for Proportion).

Good test items must also be able to show their distinguishing power. According to Arikunto (2008), "difference power is the ability of a question to distinguish between smart (high-skilled) students and less (low-skilled) students". Difference power relates to the degree of ability of the items to differentiate the behavior of the test taker in the developed test. A question can be

said to have distinguishing power if the question can be answered by high-ability students and cannot be answered by low-ability students. If a question can be answered by smart or poor students, it means that the question has no distinctive power, likewise if the question cannot be answered by smart students and students lacking, it means that the question is not good because it does not have distinguishing power. According to Sudijono (2007), the distinguishing power of question items can be seen by looking at the size of the discriminatory index number of the items. The item discriminatory index number is a number or numbers that indicate the size of the discriminatory power possessed by an item. Discriminatory power is basically calculated on the basis of dividing students into two groups, namely: the higher group- namely the group of students who are classified as smart, and(the lower group lower group) - namely the group of students who are classified as less intelligent. The item discrimination index is generally given a symbol with the letter D.

Each item on the multiple choice objective test consists of several possible answers, or better known as options or alternatives. The options are usually three or five in number, and of the possible answers attached to each item, one of them is the correct answer (answer key) while the rest is the wrong answer. That wrong answer is what is commonly known as a distractor. The purpose of installing a distractor on each item is so that there are many students taking the test who are interested in choosing it, because they think that the distractor they chose is the correct answer. The more students are fooled, the more the distractor can perform their function as well as possible. Conversely, if no one has chosen a distractor, then the distractor cannot perform its function properly. According to Sudijono (2007) "a distractor is declared to have functioned well if at least 5% of all test participants have been selected."

Analysis of the daily test questions made by teachers in MTs.N 1 Bengkulu has been done, but it is not optimal, because it is only for administrative needs. It has not yet arrived at choosing items that are valid, reliable and have different power and level of difficulty both to be used as standardized test questions. Therefore, it is hoped that the results of this study can determine the quality of the questions made by MTs teachers. Negeri Slawi so that it can be used as feedback for improving the evaluation system in the next class as well as being an input for improving the evaluation program in schools.

METHOD

The form of research used in this research is descriptive. The sample in the study was a daily test question made by a science subject teacher of class VIII MTs.N 1 Bengkulu Academic Year 2019/2020 in the form of multiple choice questions, consisting of 20 items of daily test material 1 and material 2 daily test questions of 20 items. The procedure in this study consists of 2 stages, namely: 1) the preparation stage, 2) the implementation stage and 3) the final stage.

The preparation stage: includes the following steps: (1) conducting initial observations to obtain information on the implementation of daily tests; (2) formulate a multiple choice question review format for the preparation of the construct validity analysis. The implementation stage includes the following steps: (1) collecting research data, namely: daily test sheet questions made by science subject teachers for class VIII 2019/2020 academic year (daily test questions for material 1) and answer keys and collecting answer sheets for eighth grade students of participants daily tests; (2) Analyzing the daily test questions made by the teacher (daily test questions material 1) by calculating the validity, reliability, difficulty level, distinguishing power and item distractors; (3) Interpretation of data from the calculation of validity, reliability, level of difficulty, distinguishing power and item distractor; (4) Based on the results of the analysis and review, a discussion was held with the teacher to make daily test questions on the next material; (5) Collecting daily test sheet questions made by the science subject teacher for class VIII of the 2019/2020 academic year (material 2 daily test questions) and answer keys and collecting answer

sheets for students participating in daily tests; (6) Analyzing the teacher's daily test questions (daily test item 2) by calculating the validity of the items, the reliability of the questions, the difficulty level of the items, the distinguishing power of the teacher's daily test items and the teacher's daily test questions as well as examining multiple choice questions; (7) Interpreting the data from the calculation of validity, reliability, difficulty level, distinguishing power and item distractors; (8) Study of the daily test questions made by science subject teachers of class VIII MTs.N 1 Bengkulu (material 1 and material 2 daily test questions) was carried out with the help of 3 science teachers. The review is carried out using the multiple choice question format. And the final stage, among others: (1) Making conclusions in response to the problem formulation; (2) Prepare a research report.

RESULTS AND DISCUSSION

Based on the results of the analysis that has been carried out, it is known that the construct validity, the item validity, the reliability, the difficulty level, the distinguishing power, and the distractor.

Construct Validity

Te results of the analysis of the construct validity of the UH 1 question showed that there were three questions, namely questions 1, 15 and 20, which did not match the aspects of the review in aspect 5, namely the subject matter was formulated briefly, clearly and firmly. Whereas in the analysis of the construct validity of the UH 2 questions, it is known that one question, namely question number 6 does not match aspect 13, namely the choice of answers in the form of numbers / time arranged according to the order of the size of the numbers or chronologically. So it can be said that the question is quite good when viewedin terms of the construct validity of the questions because most of the questions were in accordance with the aspects of the review, although there were some questions that still needed to be corrected in the question construction aspect.

Validity of Question Items

The results of the analysis of the validity of the UH 1 and UH 2 items can be seen that the validity of the items in UH 1 is 20% very low category, 40% low category, 35% moderate category and 5% high category of 20 questions. Whereas in the UH 2 questions, the very low category was 15%, the low category was 25%, the moderate category was 55% and the high category was 5% from the 20 questions. At UH 1 the questions that were declared valid were item numbers: 2, 3, 5, 6, 13, 14, 15 and 18, the rest were item numbers: 1, 4, 7, 8, 9, 10, 11, 12, 16, 17 , 19, and 20 were declared invalid. Whereas at UH 2 items that were declared valid were: 1, 2, 5, 7, 8, 11, 13, 14, 17, 18, 19, 20 and those that were valid were: 3, 4, 6, 9, 10, 12 , 15, 16.

Reliability Based on the reliability analysis of multiple choice questions made by science teacher class VIII MTs.N 1 Bengkuluat UH 1 and UH 2 can be summarized as follows: the reliability of UH 1 questions is 0.76 in the high category and the reliability of UH 2 questions is 0.78 in the high category so that it can be said that the question is seen in terms of good reliability.

Difficulty Level

Analysis of the difficulty level of the items found that at UH 1 there were 40% of the questions in the difficult category, 55% of the questions in the medium category and 5% of the questions in the easy category. Meanwhile, at UH 2 it is known that 35% of the questions are in the difficult category and 65% of the questions are in the medium category. This can be summarized as follows. At UH 1 the hard items are numbered: 2, 4, 7, 10, 15, 17, 19, 20; seed item numbers: 3, 5, 6, 8, 9, 11, 12, 13, 14, 16, 18 and easy items are only number 1. Whereas at

UH 2 difficult items are numbered: 3, 7, 10, 12, 15 , 16, 20; The seed item numbers: 1, 2, 4, 5, 6, 8, 9, 11, 13, 14, 17, 18.19 and the easy items do not exist.

Distinguishing

Poweranalysis of the distinguishing power of UH 1 and UH 2 items, it is known that UH 1 questions with good category difference are 20%, enough categories are 40% and bad categories are 40%. At UH 2, 25% ofquestionsgood categories, 45% of the items in the fair category and 30% of the questions in the bad category. At UH 1, the different power categories for bad items are numbers: 1, 4, 7, 9, 11, 17, 19, 20; item category is enough item number: 2, 5, 8, 10, 12, 15, 16, 18 and good item category number: 3, 6, 13, 14. Whereas in UH 2, different power category for bad item number: 3, 10 , 12, 15, 16, 20; Enough item categories number: 4, 5, 6, 7, 9, 11, 14, 18, 19 and good item categories number: 1, 2, 8, 13, 17.

Distractors

Results of distractor analysis of UH 1 and UH multiple choice questions 2 it is known that out of the 20 questions presented 80% of the cheaters are functioning and 20% of the distractors are not functioning. Whereas at UH 2, 75% of the distractor was functioning and 25% of the distractor did not function. So it can be said that the problem is quite good when viewed from the question distractor because most of the trickster are functioning.

At UH 1, the functional distractors are the numbers: 2, 3, 4, 5, 6, 7, 9, 10, 11, 13, 14, 15, 16, 18, 19, 20; on the other hand, the non-functioning numbers: 1, 8, 12, 17. Whereas at UH 2, the functional distractors are item numbers: 2, 3, 5, 6, 8, 9, 10, 11, 12, 14, 16, 17, 18, 19, 20 and non-functioning item numbers: 1, 4, 7, 13, and 15.

Discussion of Problem Constructs Validity

Analysis of construct validity relates to the arrangement or appearance of the questions. The 3 aspects examined in the construct validity analysis of multiple choice questions include aspects of material, construction and language (Wahidmuri, 2010: 124). The material aspects studied included (1) the suitability of the questions with basic competencies (requiring a written test for multiple choice forms); (2) the material in question is in accordance with the competence (urgency, relevance, continuity, high daily usability); (3) homogeneous and logical answer choices; (4) there is only one answer key.

In the construction aspect, what is examined includes: (1) the subject matter must be formulated briefly, clearly and firmly; (2) the formulation of the main questions and the choice of answers are only statements that are needed; (3) the subject matter does not provide guidance answer keys; (4) the subject matter is free of multiple negative statements; (5) the choice of answer is homogeneous and logical in terms of material; (6) clear pictures, graphs, tables, diagrams or the like; (7) the length of the answer choices must be relatively the same; (8) the answer choices do not use the statement "all the answers above are wrong / correct" or the like; (9) the choice of answers in the form of numbers or time is arranged according to the order of the number size or chronology; and (10) the item does not depend on the answer to the previous question. Meanwhile, the language / culture aspect consists of: (1) using language in accordance with the rules of the Indonesian language; (2) using communicative language; (3) do not use the locally applicable language and (4) the answer choices do not repeat the same word or group of words, unless they constitute one unified meaning.

Based on the results of the analysis, it is known that in terms of material, 100% of the questions on UH 1 are in accordance with the aspects of the study. When viewed from aperspective construction, there are still a number of questions that need to be revised because they are not in accordance with several aspects of the analysis. These aspects include aspects number 5, namely the subject matter that is compiled is not formulated briefly, clearly and firmly

according to one or two of the teachers who study it, the questions in question are questions number 1, 15 and 20. In terms of language, optional questions double UH 1 as a whole is in accordance with the aspects of the review. At UH 2 on the material aspects, all the questions were in accordance. In the construction aspect, question number 6 is not in accordance with the 13th aspect, namely the choice of answers in the form of numbers / time which are arranged according to the order of the size of the numbers or chronology according to the three teachers who studied them. In terms of language, the UH 2 multiple choice questions are generally in accordance with the aspects of the study. From the results of the analysis, the questions UH 1 and UH 2 can be said to have construct validity, although there are some questions that still need to be fixed in the construction aspect, but in general the questions are quite good. Question Item Validity

Based on the results of the analysis of the validity of the items at UH 1, it was known that the correlation between the item scores and the total score of daily tests was low overall. Of the 20 questions presented, as many as 60% or 12 items were invalid while only 40% or 8 items were valid. The questions that were not valid were numbered questions: 1, 4, 7, 8, 9, 10, 11, 12, 16, 17, 19, 20. At UH 2 it was found that the validity of the questions as a whole was quite high. Of the 20 questions presented, as many as 60% or 12 items were valid, while 40% or 8 items were invalid. Invalid questions are questions number: 3, 4, 6, 9, 10, 12, 15, 16.

Questions are said to be valid or have high validity are questions that can measure the expected competence. Meanwhile, questions that are not valid or have low validity mean that the questions cannot measure the expected competence. According to Arikunto (2008: 65) an evaluation technique is said to have high validity (called valid) if the evaluation technique can fully measure certain expected abilities. So it can be said that the test tool is able to measure students' abilities in achieving the expected competencies in the material. The score on the item / item causes the total score to be high or low. In other words, an item has high validity if the score on the item is parallel to the total score.

In item analysis, the characteristics of the items will be seen and the good ones will be selected. Good items are items whose characteristics meet the requirements as the criteria for good test items. Item analysis is carried out on a large number of test items. Item analysis will abort some of the items analyzed because they do not meet the ability to measure learning outcomes well (Purwanto, 2011). A question that has low or invalid validity means that the question cannot measure the expected competence. In the UH 1 question,question which have a very low category of item validity consisting of questions number 4, number 7, number 11 and number 20, and there are 8 questions that have low category item validity, namely questions number 1, 8, 9, 10, 12, 16, 17, and 19. In UH 2 questions, there were three questions with very low category validity, namely questions 3, 10 and 15, and in the low category there were 5 questions, namely questions number 4, 6, 9, 12 and 16 .

Reliability

From the calculation formula known KR-20 reliability index value. Reliability index ranges from 0 - 1 with 5 criteria. The higher the reliability coefficient of a test, the higher its consistency or accuracy. The reliability index value at UH 1 was 0.76 and at UH 2 was 0.78, which means that the question has high reliability. High reliability as meant in this case includes the accuracy or accuracy of the measurement results and the consistency or stability of the measurement results.

The Difficulty

Level of the test items shows how difficult or easy the test items and the overall test have been. The level of difficulty index is the ratio between correct item answers and the number of item answers (Gronlund in Santos, 2012: 5). The difficulty level analysis shows whether the test

items are classified as too difficult, difficult, moderate, easy or too easy. Through the study and analysis of the level of difficulty of the tests being tested, the feasibility of the test questions can be revealed, both for each test item and the entire test item.

According to Sudjana (2004: 136), a package of questions given to students should have a balance between difficult: enough: easy with a ratio of 3: 4: 3 or 2: 5: 3. Of the 20 questions based on table 3, it is known that at UH 1 there were 8 questions in the difficult category, 11 questions in enough categories and 1 questions in the easy category with a difficult ratio of 8: enough 11: easy 1 or 4: 5.5: 0.5. Whereas at UH 2, there were 7 questions in thecategory difficult and 13 questions in sufficient category with a ratio of 7: enough difficult 13: easy 0 or. 3,5: 6,5: 0. The results of this comparison show that the multiple choice questions UH 1 and UH 2 have an unbalanced proportion. In other words, it can be said that the problem is dominated by the level of difficulty enough at UH 1 and UH 2. In preparing the questions, it is necessary to pay attention to the percentage of the difficulty level of the questions. According to Joesmani (1988), the difficulty level between 25-75% (sufficient category) is an adequate level of difficulty. The lower the percentage of the difficulty level of the questions, the more difficult the questions are, because few test takers answer the questions correctly. A good question according to Arikunto (2008: 207) is one that is not too easy or too difficult. Easy questions do not stimulate students to enhance their solving efforts. Conversely, questions that are too difficult will cause students to become discouraged and not have the enthusiasm to try because they are out of reach.

Distinguishing Power

According to Arikunto (2008), the distinguishing power of a question is the ability of a question to distinguish between smart students and students who are less intelligent. A question that can be answered correctly by a smart student or a student who is less intelligent, then the question is said to be a bad question because it does not have any distinguishing power.

Based on the results of the analysis of the different power of the multiple choice questions UH 1 questions with a good category difference amounting to 20%, enough categories as much as 40% and bad categories as much as 40%. At UH 2, as many as 25% of the items in the good category, 45% of the questions in the enough category and 30% of the questions in the bad category. When viewed from the difference in power, UH 1 and UH 2 questions can be said to be good because most of the questions have a fairly good difference in power where the questions that have good differences mean that these questions can distinguish high-ability and low-ability students. So it can be said that at UH 1and UH 2 have quite good distinction power because most of them have good differentiation power, meaning that the questions are sufficient to differentiate between high-ability and low-ability students. This means that students who are smart answer more questions in question, while students who are less intelligent answer more wrongly.

Even so, there are several questions that still have low difference power at UH 1, namely questions number 1, 4, 7, 9, 11, 17, 19 and 20. In question number 7 the difference is negative, which is -0.09 and in question number 11 it is -0.06. According to Sudijono (2007: 388), if the discrimination index number is negative, it means that the items in question are answered correctly by lower group students than upper class students, or students who are in the smart group answer more wrongly while the group students are less intelligent. even more correct answers.

Distractors

Based on table 5. which shows the results of the distractor analysis of multiple choice questions at UH 1, it is known that 16 questions have a choice of answers to functional questions. This means that the answer choices (not the answer key) have functioned properly, namely as a distraction. The rest, namely 4 questions, namely question number 1 choice A, C, D;

question number 8 choice C; question number 12 of choice E; question number 17 choice D of the distractor doesn't work. At UH 2, it is known that 15 questions have answer choices on functional questions. This means that the choice of answers has served as a distraction. The rest, namely 5 questions, namely question number 1 choice B; question number 2 choice D; question no.7 choice C, question no.13 choice E and question 15 choice E have not functioned properly as a distraction.

According to Aprianto (2008), there are several factors that influence the function of the deceiver, namely if the questions are too easy, the subject matter provides clues to the answer key and students already know the material to be asked too easily. Multiple choice tests that are arranged without paying attention to the homogeneity of the choice of answers will likely not function. Because the test taker will easily guess without thinking long and will immediately answer the answer key, meaning that ignoring other answer choices is a non-homogeneous deceiver. Likewise, if the subject matter provides clues to the correct answer. The instructions for choosing the correct answer make the test taker answer according to the instructions. This will make other alternative answers not work.

Distractors are said to function if they are selected by a large proportion of low-ability students and at least 5% of all high-ability test takers are selected. If more cheaters are chosen by highly skilled students, it can be said that the cheaters are misleading. If the cheater is chosen evenly by the test takers, then the cheater functions. According to Widodo (2010), the cause of

deceit is not chosen by test takers because it looks too misleading. In addition, it is necessary to pay attention to whether the answer choices are not homogeneous or if the students really understand the concept of the material being taught.

As a follow-up to the results of the analysis on the function of the distractor, then for the distractor who has functioned on the question, it can be used in the next semester's daily test, while the distractor that is not functioning needs to be replaced or revised.

CONCLUSION

Based on the results of the research and discussion, it was concluded that the validity of the items at UH 1 was 40% valid and at UH 2 60% was valid; The reliability of UH 1 questions is 0.76 and UH 2 is 0.78 in the high category; The difficulty level of the UH 1 questions is 40% for the difficult category questions, 55% for the moderate category questions, 5% for the easy category questions. At UH 2 35% of the questions in the difficult category and 65% of the questions in the medium category; Difference power of UH 1 questions, 20% good category questions, 40% category enough and 40% is bad. At UH 2, 25% of the questions were in good category, 45% of questions were in good category and 30% of questions were in bad category; The distractor in the UH 1 question was 80% functional and 75% of the distractor was functioning

REFERENSI

Arikunto, Suharsimi. 2008. Basic Basic Educational Evaluation. (Revised Edition). Jakarta: Earth Literacy.

Arjuniwati, A. 2019. Increasing Motivation and Learning Outcomes Through the Application of the Numbered Heads Together (NHT) Learning Model to the Opportunity Material for Class XII Mathematics Subjects. Tambusai Education Journal. Volume 3, No. 1, p. 1–13.

Astalini, & Kurniawan, DA 2019. Instrument Development for Middle School Students' Attitudes Toward Natural Science Subjects. Journal of Science Education (JPS). Volume 7, No. 1, p. 1–7.

Borg, W.R dan Gall, M.D. 1983. Educational Research An Introduction (4th Ed). White Plains : Logman Inc.

Dewantari, A. 2015. Laporan Pengolahan Hasil Angket SNP SMAN 1 Surakarta dan SMAN 2 Surakarta Tahun Pelajaran 2014/2015. Laporan Tidak Dipublikasikan. Pascasarjana FKIP Universitas Sebelas Maret Surakarta.

Dewi, NDL, & Prasetyo, ZK 2016. Development of Science Assessment Instruments to Map Critical Thinking and Practical Skills of Middle School Students. Journal of Science Education Innovation. Volume 2, No. 2, p. 213–222.

Dewi, NR, & Akhlis, I. 2016. The Development of Multicultural Education-Based Natural Science Learning Tools Using Games to Develop Student Character. Unnes Science Education Journal. Volume 5, No. 1, p. 1098–1108.

Djaali & Muljono, Pudji. 2008. Measurement in the field of education. Jakarta: Grasindo.

Halaydina, T.M dan Downing, S.M. 1989. A Taxonomy of Multiple Choice Item Writing Rules. Applied Measurements In Education, 2(1), 37-50.

Hamid, MA 2016. The Development of ICT-Based Student Learning Outcomes Assessment Instruments in Basic Electrical Electronics Learning. Scientific Journal of Electrical Engineering Education. Volume 1, No. 1, p. 37–46.

Hardiani, IN 2017. Development of Social Attitudes Assessment of Social Studies Learning Instruments for Class IV SD. Journal of Educational Partners. Volume 1, No. 6, p. 2550– 0481.

Harlen, W. 1992. The Teaching of Science: Studies in Primary Education. London: David Fulton Publishers.

Hayati, S., & Lailatussaadah. 2016. The Validity and Reliability of the Active, Creative and Fun Learning Knowledge Instrument (PAKEM) Using the Rasch Model. Scientific Journal. Volume 16, No. 2, p. 169–179.

Joesmani. 1988. Measurement and Evaluation in Teaching. Jakarta: Ministry of Education and Culture, Directorate General of Higher Education Project for the Development of Educational Personnel Education Institutions.

Keterampilan Proses Sains. Makalah Tidak Dipublikasikan. Program Studi Ilmu Pengetahuan Alam Konsentrasi Pendidikan Biologi-SL. Sekolah Pasca Sarjana Universitas Pendidikan Indonesia.

Nurkancana, Wayan & Sunartana, PPN 1986. Educational Evaluation. (4th printing). Surabaya: National Business.

Permendiknas No. 20 Tahun 2007. Standar Penilaian Pendidikan. Jakarta : Depdiknas. Permendiknas No. 22 Tahun 2006. Standar Isi Pendidikan. Jakarta : Depdiknas.

Permendiknas No. 23 Tahun 2006. Standar Kompetensi Lulusan. Jakarta : Depdiknas.

Purnomo, P., & Palupi, MS 2016. The Development of Mathematics Learning Outcomes Tests for Solving Problems Related to Time, Distance and Speed for Class V Students. Research Journal (PGSD Special Edition). Volume 20, No. 2, p. 151–157.

Purwanto, N. 2010. Prinsip-Prinsip dan Teknik Evaluasi Pengajaran. Bandung : PT Remaja Rosdakarya.

Purwanto, Ngalim. 2011. Evaluation of Learning Outcomes. Yogyakarta: Student Library.

Ramirez, R.P.B. dan Ganaden, M.S. 2008. Creative Activities and Students' Higher Order Thinking Skills. Journal of Education Quarterly, Vol 66 (1), 22-23.

Sudijono, Anas. 2007. Introduction to Educational Evaluation. Jakarta: PT. Raja Grafindo Persada.

Sudjana, N. 2009. Penilaian Hasil Proses Belajar Mengajar. Bandung : PT Remaja Rosdakarya.

Sudjana, Nana. 2004. Assessment of Results and Teaching and Learning Process. Bandung: Youth Rosdakarya,

Sugiyono. 2017. Quantitative, Qualitative, and R & D Research Methods. Bandung: Alfabeta.

Suharman. 2018. Tests as a Measure of Academic Achievement. Scientific Journal of Islamic Religious Education. Volume 10, No. 1, p. 93–115.

Sujana, IWC 2019. Functions and Objectives of Indonesian Education. ADI WIDYA: Journal of Basic Education. Volume 4, No. 1, p. 29–39.

Suryabrata, Sumadi, 1987. Development of Learning Outcomes Tests. Jakarta: Rajawali Press.

Treagust, D.F. 2006. Development and Use of Diagnostic Test to Evaluate Students Misconception In Science. International Journal of Science Education, 10, 2 pp 159-169.

Trimawati, K., Kirana, T., & Raharjo, R. 2020. Development of Integrated Science Assessment Instruments in Project Based Learning (PJBL) Model to Improve Middle School Students' Critical and Creative Thinking Ability. QUANTUM: Journal of Science Education Innovation. Volume 11, No. 1, p. 36–52.

Tuysuz, C. 2009. Development of Two-Tier Diagnostic Instrument and Assess Students Understanding In Chemistry. Scientific Research and Essay Academic Journals Vol 4 (6) pp. 626-631, June 2009. ISSN 1992-2248.

Uno, HB 2008. Motivation & Measurement Theory. Jakarta: PT Bumi Aksara.

Wahidmuri, et al. 2010. Evaluasi Pembelajaran kompetensi dan praktik. Yogyakarta: Nuha Litara

Yuliati, L. 2008. Model-Model Pembelajaran Fisika. Malang : Universitas Negeri Malang.

Yusuf, M. 2015. Assessment and Evaluation of Education: Pillars of Information Providers and Activities to Control Quality of Education. Jakarta: Prenadamedia Group.