International Journal of Community Service Learning
Volume 8, Issue 3, 2024, pp. 293-303
P-ISSN: 2579-7166 E-ISSN: 2549-6417
Open Access: https://doi.org/10.23887/ijcsl.v8i3.81953

Training to Analyze Learning Result Test Items for High School Teachers

Hari Sugiharto Setyaedhi[1*], Lamijan[2], Bakhrudin All Habsy [3]

[1,2,3] Teknologi Pendidikan, Universitas Negeri Surabaya, Surabaya, Indonesia

ARTICLE INFO

Article history:
Received July 03, 2024
Accepted August 13, 2024
Available online August 25, 2024

Kata Kunci :
Pelatihan, Menganalisis Butir Soal, Hasil Belajar

Keywords:
Training, Analysis Of Question Items, Learning Outcomes

ABSTRAK

Butir soal tes hasil belajar yang telah dikembangkan oleh guru SMA tidak semuanya dianalisis butir soalnya dengan demikian belum dapat dikatakan sebagai tes hasil belajar yang bermutu. Permasalahan para guru di SMA sebagian besar belum memahami bagaimana menganalisis butir soal tes hasil belajar. Tujuan dari pelatihan ini untuk meningkatkan pengetahuan bagi para guru SMA dalam menganalisis butir soal yang berkualitas. Subjek penelitian 25 guru SMA yang terdiri dari perwakilan guru mata pelajaran. Metode penelitiannya adalah kuantitatif dengan Desain one Group Pretest-Posttest. Teknik pengumpulan data menggunakan eksperimen, observasi dan dari skor hasil pretes dan postes. Analisis data yang digunakan pada penelitian ini adalah uji-t berpasangan (Paired Sample t-Test). Olah data penelitian ini menggunakan SPSS. Hasil penelitian sesuai dengan target capaian yaitu meningkatkan kemampuan guru dalam menganalisis butir soal, hal ini dibuktikan dengan peningkatan nilai posttest dibandingkan dengan pretest secara signifikan dengan hasil pada sig. (2 tailed) = 0,001 < 0,05. Sebagai bukti dan telah mengikuti pelatihan para guru diberikan diberikan sertifikat. Implikasi dari kegiatan pengabdian ini adalah meningkatnya kompetensi guru SMA dalam menganalisis butir soal tes hasil belajar.

ABSTRACT

Not all of the learning outcomes test items that have been developed by high school teachers have been analyzed, so they cannot be said to be quality learning outcomes tests. The problem is that most high school teachers do not understand how to analyze learning outcomes test items. The aim of this training is to increase knowledge for high school teachers about analyzing quality questions. The research subjects were 25 high school teachers, consisting of subject teacher representatives. The research method is quantitative with a one-group pretest-posttest design. Data collection techniques use experiments, observation, and pre- and post-test scores. The data analysis used in this research is a paired t-test (paired sample t-test). Process this research data using SPSS. The results of the research are in accordance with the achievement target, namely increasing the teacher's ability to analyze question items. This is proven by a significant increase in posttest scores compared to the pretest, with results in sig. (2 tailed) = 0.001 < 0.05. As proof that teachers have participated in training, they are given certificates. The implication of this service activity is to increase the competence of high school teachers in analyzing learning outcomes test items.

1. INTRODUCTION

In evaluating learning, teachers carry out assessments using tests. The tool used to evaluate learning is a test (Jumrah et al., 2023; Widiyawati et al., 2019). Tests are an important part of the learning assessment process (Adom et al., 2020; Faiz et al., 2022; Masitoh & Aedi, 2020). Tests are tools that can be used to measure the success of learning outcomes (Fahrurrozi & Laili Rahmawati, 2021; Sholihah et al., 2017). Tests play an important role in knowing learning outcomes (Manfaat & Nurhairiyah, 2021; Rizky Ananda Setiyawan & Palupi Sri Wijayanti, 2020). Tests are used to determine the learning outcomes that students have achieved during the learning process (Aisyah et al., 2021; Sa'diyyah et al., 2021). The test is prepared based on students' responses to the questions contained in the test items (Adom et al., 2020; Sa'diyyah et al., 2021). Tests are used for various purposes, such as accepting new students, graduation, selection, and others (Litna et al., 2021; Ma'rifah et al., 2021).

The requirements for a good test are: a) reliable; b) valid; c) objective; d) discriminatory; e) comprehensive; and f) easy to use (Purnamasari, 2019; Quansah et al., 2019). Before being used, the test

*Corresponding author
E-mail addresses: harisetyaedhi@unesa.ac.id (Hari Sugiharto Setyaedhi)

must have several requirements, such as: 1) validity, the degree of accuracy of the measuring instrument for what is to be measured (Setyaedhi et al., 2023; Sugiono et al., 2020); 2) reliability, meaning if the test is tested several times, then the results are relatively the same as long as the object being measured has not changed (Farida & Musyarofah, 2021; Idrus, 2019; Sugiono et al., 2020). 3) Objectiveness ensures that the researcher's subjective factors do not influence the test results. 4) balanced, meaning the difficulty index of the test items must be balanced with the objectives of the test. 5) discriminative, meaning the test must be able to distinguish between students who are clever and not clever, 6) The norm means test results must be simple according to certain standards (Elviana, 2020; Sulistianingsih, 2020; Supriyati & Dudung, 2019; Warju et al., 2020). What is worth paying attention to in a test is that it meets the demands of validity and reliability, namely the accuracy of the measurement results and the consistency of the measurement results (Nengsi & Efrina, 2019; Quansah et al., 2019). A test is said to be of quality if it has been tested for validity and reliability (Alfiatunnisa et al., 2022; Bashooir & Supahar, 2018; Budiantoro et al., 2021; Dewi & Sudaryanto, 2020; Puspasari & Puspita, 2022).

To develop a reliable test tool, the following procedures must be followed: 1) prepare test specifications; 2) write test questions; 3) review test items; 4) pilot test; 5) quantitative item analysis; 6) revise test; 7) prepare test; 8) administer the test; and 9) interpret test results (Magdalena et al., 2020; Muliyani & Huriaty, 2016; Muttaqin & Kusaeri, 2017; Ndiung & Jediut, 2020; Purnomo & Maria Sekar Palupi, 2016). An instrument, when related to the field of education, means a tool that can be used to measure student learning achievements, teachers' teaching and learning processes, and the achievement of a particular program (Adom et al., 2020; Widiyanto, 2018). Tests are most commonly used in measurement activities such as learning outcomes tests, formative tests, summative tests, etc. Tests as measuring tools must be prepared as well as possible according to the guidelines provided. A test is an instrument or tool used to test a particular competency to differentiate it from other competencies. Creating objective tests is not easy (Hodyanto & Saputro, 2018; Wartoni & Benyamin, 2020). As a result, there are many questions that are not appropriate to those studied, which makes the assessment of learning outcomes less good.

There are several things related to question analysis to determine whether a test is good or not. First, the validity of a test is determined by how well the test measures what it is intended to measure. Next is test reliability, which describes the consistency of the test. Third, the level of difficulty of a question is the ratio of the number of correct answers to the number of examinees. Fourth, differentiation power, namely the ability of questions to differentiate between students who have mastered the material and those who have not mastered it, Fifth, the effectiveness of the distractor. With the help of computer applications, question analysis can be carried out effectively and efficiently. Question item analysis is a type of teacher activity that is very useful for interpreting student learning outcomes tests and determining the quality of questions. The aim is, among other things, to analyze systematic information about the test items, namely validity, reliability, difficulty, discrimination, and effectiveness of distractors, to find out whether the test is good or not.

The difficulty index for a question item is the percentage of examinees who answered the question correctly. The difficulty index of an item is usually denoted by p (proportion). The larger the p value, the easier the problem. On the other hand, the smaller p, the more difficult the questions are. The p index is the ratio of the number of questions answered correctly to the number of subjects who answered the questions (Elviana, 2020; Sijabat et al., 2024; Warju et al., 2020). Good questions are questions that are not too easy or too difficult. Discriminative power, or differentiating power, is that a question item is an indicator of the extent to which a question item is able to differentiate between high-achieving groups (upper groups) and low-achieving groups (lower groups) (Elviana, 2020; Supriyati & Dudung, 2019; Warju et al., 2020). Its function is to determine whether or not the question items are accepted. A group of students with high skills and a group of students with low skills. Distractors can work well if they are chosen by less intelligent test-takers. Conversely, the more intelligent test takers are, the less likely they are to choose distractors. The distribution of answer choices is the basis for considering question items. For questions in the form of multiple-choice tests, distractors function well if at least 5% of all test takers choose them. Distractors are used to identify test takers with high skills (Elviana, 2020; Supriyati & Dudung, 2019). Validity means accuracy (Elviana, 2020; Syaifudin, 2020). The validity of an item is the accuracy of measuring that item has to measure what should be measured (Setyaedhi et al., 2023; Sijabat et al., 2024). An item can be said to have high validity if the scores on the item in question are in line with the total score, or in statistical language, if there is a significant positive correlation between the score of the item and the total score. Reliability relates to the accuracy of the instrument in measuring what is being measured, the accuracy of the measuring results, and how accurate it would be if repeated measurements were made (Elviana, 2020; Sijabat et al., 2024). Reliability is the consistency of information obtained from one or more assessments. It can be concluded from this expert opinion from this expert opinion that the concept of reliability refers to the consistency or constancy of measurement results.

Tests as a measuring tool must be specifically adapted to the learning objectives and prepared as well as possible according to the rules given. Educational tests are instruments or tools used to test certain skills to differentiate a competency from other competencies. Therefore, the test instruments are arranged as best as possible. A test considered a good measuring tool must meet testing requirements such as validity and reliability (Lia et al., 2020; Ramadhan et al., 2024). Test quality is an important factor in assessing learning outcomes so that they really measure what is being measured. Assessment is an important part of the learning process. However, teachers do not only develop tests to achieve learning outcome tests but also to test whether the tests used can do their job as a quality measure of learning outcome tests. Quality tests are very helpful in making decisions about student progress. Teachers always hope that the results obtained now will be better than before. A good evaluation will increase students' learning motivation and can be an input for teachers to improve the quality of learning in the future; therefore, it is considered important to make efforts to improve understanding of question item analysis.

However, in reality, this is not in line with the results of the PKM team's interview with the high school, where information was obtained that some of the high school teachers did not understand how to analyze question items so that they could produce quality learning outcomes tests. Thus, this training material was implemented based on the problems identified. in high school in carrying out daily tasks at school. In the world of education, evaluating, analyzing, and compiling quality questions needs to be done before the questions are tested on students as a measuring tool. This aims to ensure that the questions function well and are able to measure student competency accurately. Question item analysis is an evaluation process that aims to evaluate how effective an exam or test question is in measuring students' understanding of the material being taught. This community service provides training to high school teachers, which aims to increase their knowledge of evaluating learning outcomes, analyzing test items, and compiling quality test items. In this way, teachers can develop valid and reliable test items so that they can reflect test takers' learning outcomes after participating in teaching and learning activities.

2. METHOD

This is quantitative research with a one-group pretest-posttest design. Data collection techniques use experimentation and observation, and the pretest and posttest scores are then compared. The data analysis used in this research was a paired t-test (paired sample t-test) using SPSS. The research subjects were 25 high school teachers, representing each subject. Based on the main problem in high school, namely the importance of increasing knowledge in analyzing valid and reliable learning outcome test items in supporting high school learning evaluation, Figure 1 shows the stages of community service.

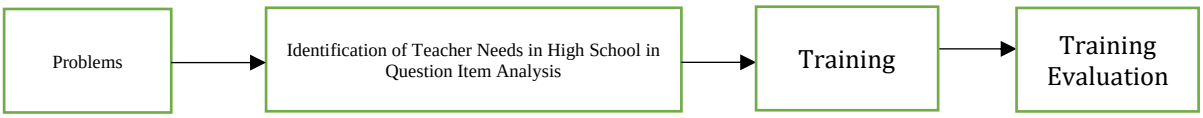

Figure 1. Stages of Community Service

Before the service in the form of training was carried out, the PKM team from the Department of Curriculum and Educational Technology, Surabaya State University, chaired by Hari Sugiharto, had coordinated with the school. Coordination is carried out through interviews and discussions via telephone with related parties such as the education office, school principals, and high school teachers. The purpose of coordination is to find out the problems that exist among high school teachers. The PKM team immediately identified and prepared a PKM plan to overcome conditions in the field. High school teachers complain about their lack of knowledge about developing quality learning outcomes tests. So far, teachers in developing learning outcomes tests have only created questions and then tested them on students.

Training is carried out offline and online, with a schedule agreed upon by both parties. Before the training was carried out, the PKM team carried out a pretest. This was done to determine the teachers' initial abilities in analyzing question items. Test development training is carried out offline at the time of service and online every week. Evaluation is carried out by conducting a posttest at the end of the meeting. The post-test was carried out to find out whether the test development material that had been taught by the Unesa PKM team had been mastered by them.

In this service activity, face-to-face (offline) learning will be held on Tuesday, June 11, 2024, while online learning will be held according to agreement. The methods used are: 1) interactive discussions about problems and solving daily problems faced at school; 2) assignments, namely pretests and posttests; and 3) practice on how to analyze test items. In Table 1, the pretest and posttest grids are displayed.

Table 1. Pretest and Posttest Grid

Basic Competency	Material	Indicator	No Items
Analysis of Question Items	Difficulty of Question Items	Teachers understand the use of the item difficulty index	1, 2
		Teachers know how to find the difficulty index for test items	3
		The teacher is able to interpret the difficulty index of the questions	4, 5
	Different Power of Question Items	Teachers understand the use of the test item discrimination index	6, 7
		Teachers know how to find the index of differentiability of test items	8
		The teacher is able to interpret the different power index of the questions	9, 10
	Distractor Effectiveness	Teachers understand the usefulness of the effectiveness of distracting question items	11, 12
		Teachers know how to find the effectiveness of distracting questions	13
		Teachers are able to interpret the effectiveness of distracting question items	14,15
	Validity of Question Items	Teachers understand the use of the validity of test items	16, 17
		Teachers know how to find the validity of question items	18
		The teacher is able to interpret the validity of the question items	19
	Test Reliability	Teachers understand the use of item reliability	20
		The teacher is able to interpret the validity of the question items	21
		Teachers are able to analyze the causes of low test reliability	22, 23, 24, 25

It was agreed that the material would be carried out online four times via Zoom meeting, resulting in 32 hours of learning. In Table 2, the materials, methods, and presenters are displayed.

.

Table 2. Materials, Methods and Presenters

Materials	Source Person	Hours
Policy Regarding Improving the Quality of Learning Based on Education Reports	Siti Arofah, S.Pd, M.Pd	1 jp
Learning Evaluation	Dr. Lamijan, M.Pd	2 jp
Doing assignments independently	Dr. Lamijan, M.Pd	4 jp
group discussion	Dr. Lamijan, M.Pd	3 jp
Learning through a cognitive approach	Dr. Bakhrudin All Habsy, M.Pd	4 jp
Doing assignments independently	Dr. Bakhrudin All Habsy, M.Pd	3 jp
group discussion	Dr. Bakhrudin All Habsy, M.Pd	3 jp
Practice Analyzing Question Items	Dr. Hari Sugiharto Setyaedhi, M.Si	3 jp
Interpretation of Question Item Analysis	Dr. Hari Sugiharto Setyaedhi, M.Si	3 jp
Group discussion	Dr. Hari Sugiharto Setyaedhi, M.Si	3 jp
Reflection on Implementation of Activities	Committee	1 jp
Amount		**32 jp**

3. RESULT AND DISCUSSION

Result

The school is very supportive of the implementation of the Community Service Program. The school facilitates facilities and infrastructure to support community service. This event was attended by high

school teachers. There were 25 high school teachers who took part in the training, representing all subjects. Figure 2 shows photos of the resource persons. In Figure 3, a photo of one of the resource persons is shown providing material.

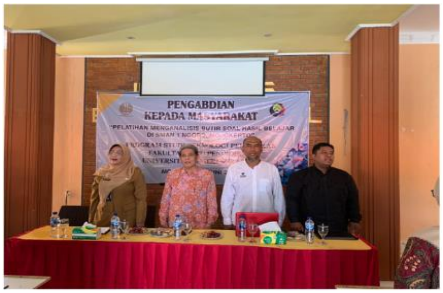

Figure 2. Siti Arofah, Lamijan, Hari Sugiharto, Bakhrudin

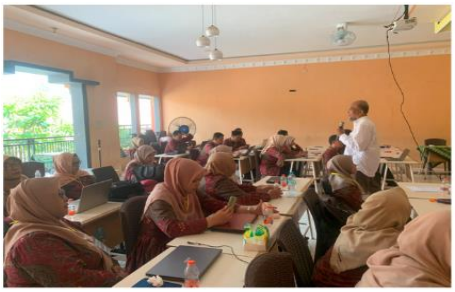

Figure 3. The Resource Person Explains and the Teacher Listens

The training activities began with the opening of the event taking place offline on June 11, 2024, then continued with a pre-test to determine initial skills in analyzing learning outcomes test items, and after that, the delivery of material on the concept of learning evaluation presented by Dr. Lamijan, M.P.D. The second material is about learning through a cognitive approach by Dr. Bakhrudin All Habsy, M.P.D.; after that, the third material was by Dr. Hari Sugiharto Setyaedhi, M.Si., about analyzing learning outcomes test items. The event continued with responses from the teachers in the form of discussions, questions, and answers. The participants' responses were very good, as evidenced by the many questions. Due to the very limited face-to-face time and the material presented being incomplete, as well as the many questions still remaining, the material was continued online. If calculated, almost the majority of participants responded by asking various questions.

The results of this PKM activity produced several outputs. The results of the activities achieved from this PKM activity are: increasing teacher understanding of the concept of a) analysis of learning outcomes test items, b) level of difficulty of item items, c) distinguishing power of item items, d) effectiveness of distractors, e) validity test and f) reliability, b) increasing teachers' understanding of test development procedures, such as guidelines for preparing tests in objective form and descriptions, characteristics of online learning evaluation instruments, c) teachers can apply fun learning evaluation techniques, d) high school teachers get a certificate that is validated by the Dean of the Faculty of Education, Surabaya State University. PKM activities receive an Instagram message that preaches the importance of training so that teachers can analyze test items on learning outcomes so that teachers can develop quality items. In Figure 4, the news capture is shown.
https://www.instagram.com/p/C8Ja92Ivvlx/?igsh=ZnVsbzYyMzF2MHlu&img_index=1

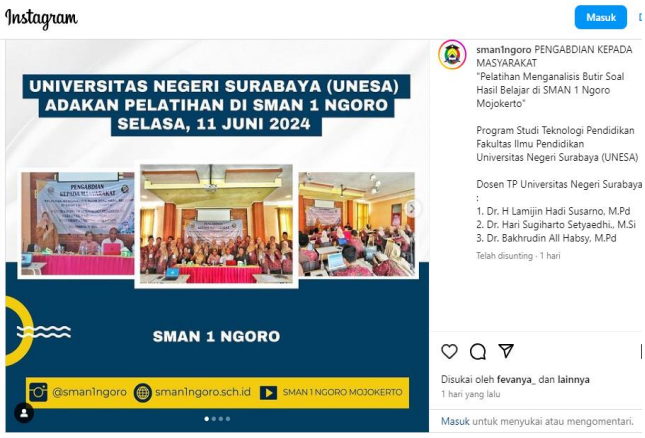

Figure 4. Solo Pos Coverage

This PKM activity was posted on the YouTube channel on June 11, 2024, with the title Community Service. This is the result of "Training to Analyze Question Items." In Figure 5, the news capture is shown.
https://youtu.be/HnFV6rkZYW0?si=RiCg1hkcp_TAKigM

Figure 5. View on YouTube Channel

The results of this PKM activity produced several outputs. The results of the activities achieved from this PKM activity are: a) increasing teacher competency on how to compose quality test questions; b) increasing teacher competency on how to analyze test items. To determine the level of understanding of teachers and to evaluate the implementation of service, it is necessary to hold a post-test. Pretest and posttest scores will be compared using the t test. The main requirement before carrying out a t test is that the data must be homogeneous and normally distributed.

Based on data analysis, The "Test of Homogeneity of Variances" output above shows that the significance value (Sig.) of the pretest and posttest variables is 0.806. Because the Sig value of 0.806 > 0.05, as is the basis for decision-making in the homogeneity test above, it can be concluded that the pretest and posttest are the same or homogeneous. In table 4, the results of data normality calculations are displayed. Based on data analysis, the significance value of Asiymp.Sig (2-tailed) is 0.200 > 0.05. So in accordance with the basis for decision making in the Kolmogorov-Smirnov normality test above, it can be concluded that the data is normally distributed. In table 5, the average pre-test and post-test scores are shown. Mean Pre test and Post test showed in Table 5.

Table 5. Mean Pre Test and Post Test

Descriptive Statistics			
	N	**Mean**	**Std. Deviation**
Pretest	25	39.92	9.755
Posttest	25	70.96	9.076

Summary of descriptive statistical results from the two samples examined, namely the pretest and the posttest. The results before the test reached an average of 39.92, while the results after the test reached an average of 70.96. The 25 teachers who participated in the exam. To show whether there is a significant difference, the results of the paired sample t test must be interpreted. In Table 6, the t-test results are displayed. Based on data analysis, In the "Paired Samples Test" output above, the Sig value is known. (2-tailed) of 0.000 < 0.05. This means that there is an average difference between the pretest and posttest (Hasyim et al., 2021; Tarumasely, 2020). Thus, training in analyzing learning outcomes test items for high school teachers is said to be successful.

Discussion

The community service activities participated in by the high school teachers went very smoothly and met all our expectations. The high school teachers are enthusiastic and motivated, and they hope that this evaluation activity can be implemented in their respective schools. In evaluating learning, a teacher carries out measurements using tests. Research from Arif et al., Mappalesye et al. (Arif et al., 2022; Mappalesye et al., 2021) states that the test instruments developed must be of high quality. For this reason, the tests must be valid and reliable. This statement is supported by several other researchers, such as Rizky Ananda (Rizky Ananda Setiyawan & Palupi Sri Wijayanti, 2020). To create a quality test, test item analysis is needed. This is in line with statements from previous research, such as those of Iskandar and Rizky Ananda, who stated that item analysis is very necessary to develop a good test (Iskandar & Rizal, 2018; Rizky Ananda Setiyawan & Palupi Sri Wijayanti, 2020). This is also in accordance with the statement, which states that a test can be said to be good if an analysis of the question items has been carried out (Putri & Sari, 2020; Setyaedhi et al., 2023). Making tests must be carried out using correct and good procedures to produce good measuring instruments (Aisyah et al., 2021; Ramadhan et al., 2024). Elviana stated that

teachers, in evaluating student learning outcomes, rarely analyze quantitatively (empirically) (Elviana, 2020). This is in line with Setyaedhi's statement (Setyaedhi et al., 2023) and is also in line with Erawati's statement, which states that quantitative test instruments are rarely used (Erawati, 2018). A good test instrument must meet requirements such as: 1) medium difficulty of items; 2) item discrimination; and 3) distractors or distractions. 4) good validity, and 5) high reliability (Arifin, 2017; Setyaedhi et al., 2023; Sijabat et al., 2024).

Based on the data above, the questions with easy and difficult levels of difficulty need to be researched and revised in detail (Widayanti et al., 2021). Learning outcomes test items can be said to be good items if they are neither too difficult nor too easy; in other words, the degree of difficulty of the items is sufficient (Muluki et al., 2020). This is supported by statements from Arikunto and Setyaedhi, who state that good items are questions that are neither too easy nor too difficult (Arikunto, 2018; Setyaedhi et al., 2023). Medium-category questions are good questions, meaning that students with high ability can answer the questions correctly and students with low ability will have difficulty answering the questions. Follow-up actions that can be taken after the questions have been analyzed for their level of difficulty are as follows: 1) Questions in the medium category or good questions are used as a question bank; 2). Items in the category that are too difficult can be discarded or dropped.

Discriminating power is the ability of the items to differentiate students with high ability (understanding the material) from students with low ability (lacking understanding of the material) (Muluki et al., 2020). This is in line with the statement from Qurrota et al., which states that the differentiating power of question items reflects the differences in answers to question items between groups of students with high ability and those with low ability (Qurrota et al., 2022). This is also supported by a statement from Setyaedhi, which states that intelligent students have a greater chance of answering questions correctly than those who are less intelligent (Setyaedhi et al., 2023). Discriminating power is the ability of the items to differentiate students with high ability (understanding the material) from students with low ability (lacking understanding of the material) (Muluki et al., 2020). This is in line with the statement from Qurrota et al., which states that the differentiating power of question items reflects the differences in answers to question items between groups of students with high ability and those with low ability (Qurrota et al., 2022). This is also supported by a statement from Setyaedhi, which states that intelligent students have a greater chance of answering questions correctly than those who are less intelligent (Setyaedhi et al., 2023).

A distractor is said to be good if all the distractors can function (Arbiatin & Mulabbiyah, 2020). This statement is in accordance with research conducted by Setyaedhi as well as Muluki's research, which stated that distractors were chosen by at least 5% of test participants (Muluki et al., 2020; Setyaedhi et al., 2023). Ideally, the distractor should be chosen only by incompetent or incapable subjects, while capable subjects will not choose the distractor. As with the answer key, of course, in reality, there is still a chance that a competent subject will choose the wrong distractor. If the proportion remains smaller than the proportion of distractor voters from the incompetent subject group, then the distractor can still be considered effective. The more students who choose distractors, the more distractors have carried out their function well (Widiyanto, 2018). The poor quality of the distractor is caused by the distractor being too obvious or misleading (Suci Mitra & Helendra, 2022). This statement is in line with Muluki's statement, which states that a bad distractor indicates that the distractor is too conspicuous and heterogeneous (Muluki et al., 2020). For this reason, the distractor must be as similar as possible to the answer key. A bad distractor is a distractor that is not chosen at all by students because it looks too misleading. So it is not easy for question makers to create distractors so that the answers are not easily guessed correctly by students.

Valid question items mean that the question items can distinguish between students who have achieved the learning objectives compared to students who have not achieved the learning objectives (Amalia et al., 2021). Thus, the test can measure what it should measure and not something else (Muluki et al., 2020; Sijabat et al., 2024; Utomo, 2018). Several things that can affect validity are: 1) insufficient time to complete the questions; 2) cheating in the test; 3) inconsistent scoring; 4) test takers are unable to follow the directions given in the standard test; 5) there are jockeys. Question items can be valid if they are well constructed and include material that represents the measuring target (Suci Mitra & Helendra, 2022). One of the factors that influences validity is the student's answer factor by guessing the answer (Ardhani, 2020; Santoso, 2018).

The relationship between validity and reliability concerns the accuracy of the test in measuring the symptoms to be measured, while reliability refers to the extent to which a measurement can be trusted or consistent (Sijabat et al., 2024; Yusup, 2018). It is important that test instruments have validity and reliability requirements (Arikunto, 2018; Ramadhan et al., 2024; Sarea & Ruslan, 2019). This is in line with Muluki's statement, which states that a test instrument that has good validity for each item will also have a high level of reliability (Muluki et al., 2020). Question items to measure students' abilities must be

considered; among other things, the questions must be valid and reliable. Apart from that, the questions are said to be good if they are not too easy or too difficult; the questions must also be good. The test must be able to differentiate between intelligent and unintelligent students (Friatma & Anhar, 2019; Sarea & Ruslan, 2019). Based on the evaluation, this training was hampered by several obstacles, such as weak signals in carrying out online activities, participants being late due to various needs, but this. does not dampen the enthusiasm of teachers to gain knowledge. Training consisting of high school teachers from various subjects has a huge impact on schools. Many teachers ask about the cases experienced in their schools and how and why the development of tests can improve the quality of teaching in schools. The knowledge and skills gained during the training are used in the evaluation process at their school. Of course, the trainee teachers will share their experiences during the training in order to improve the quality of education.

In this training, there are several limitations, namely: 1) the resource person cannot provide all the material, only an outline of the material required; 2) the calculation of the level of difficulty and differentiating power of the questions uses multiple choice, so it really depends on the sample being analyzed. The research results will be different if the questions are tested on different samples; 3) the analysis of the questions only provides information or examples to the teacher, is not accompanied by creating new questions, and is also not tested again. The implication of this training is to increase teacher competence in analyzing question items. This requires full and continuous support from the Department of Education and local teachers to develop skills in analyzing question items for high school teachers, with training models that are proven to be effective and the use of methods that encourage teacher interaction and involvement. maximally.

4. CONCLUSION

After the service activity in the form of training on developing learning outcomes tests with high school teacher participants, it can be concluded that this activity is very beneficial for high school teachers because it can increase teachers' understanding of carrying out learning evaluations in schools, such as: a) increasing teacher competence regarding test development; b) increasing competence in analyzing the question items; c) teachers can apply fun learning evaluation techniques. This training is in line with the expectations of high school teachers, who have been hampered by learning evaluations, so they really need this material for their daily lives. It is recommended that further training be carried out to analyze the questions in the form of descriptions or essays. Further training is needed to compose HOTS-quality questions.

5. ACKNOWLEDGE

Surabaya State University, Department of Curriculum and Educational Technology, in collaboration with the high school, would like to express their gratitude for providing various facilities during the implementation of Pkm. PKM was held on Tuesday, June 11, 2024. On this occasion, the MOA (memorandum of agreement) was also signed between the Department of Curriculum and Educational Technology and the SMA.

6. REFERENCES

Adom, D., Mensah, J. A., & Dake, D. A. (2020). Test, Measurement, and Evaluation: Understanding and Use of the Concepts in Education. *International Journal of Evaluation and Research in Education*, *9*(1), 109–119. https://doi.org/10.11591/ijere.v9i1.20457.

Aisyah, S. N., Taena, L., & Ili, L. (2021). Pengembangan Tes Hasil Belajar Ekonomi Kelas X SMA melalui Google Formulir. *Jurnal Online Program Studi Pendidikan Ekonomi*, *6*(3), 106--114. http://ojs.uho.ac.id/index.php/JOPSPE/article/view/22761.

Alfiatunnisa, E., Khairunnisa, H. Z., Hayati, S., & Maulida, V. L. (2022). Uji Validitas dan Reliabilitas Terhadap Kemandirian Siswa Sekolah Dasar Kelas 1. *Jurnal Hurriah: Jurnal Evaluasi Pendidikan Dan Penelitian*, *3*(2), 29–36. https://academicareview.com/index.php/jh/article/view/81.

Amalia, N. R., Halik, A., & Mukhlisa, N. (2021). Analisis Butir Soal Matematika Pada Siswa Sekolah Dasar. *Pinisi Journal of Education*, *1*(1), 219–230. https://ojs.unm.ac.id/PJE/article/view/25840.

Anastasi. Anne and Urbina, S. (2017). *Psichological Testing* (Seventh Ed). New Jersey: Prentice-Hall, Inc.

Arbiatin, E., & Mulabbiyah, M. (2020). Analisis Kelayakan Butir Soal Tes Penilaian Akhir Semester Mata Pelajaran Matematika Kelas Vi Di Sdn 19 Ampenan Tahun Pelajaran 2019/2020. *El Midad*, *12*(2), 146–171. https://doi.org/10.20414/elmidad.v12i2.2627.

Ardhani, Y. (2020). Pelajaran Teknologi Dasar Otomotif Kelas X Teknik Kendaraan Ringan Otomotif di SMK

Muhammadiyah Gamping Periode 2018/2019. *Jurnal Pendidikan Vokasi Otomotif*, *3*(1), 85–94. https://doi.org/10.21831/jpvo.v3i1.34917.

Arif, N., Yuanita, P., & Maimunah. (2022). Pengembangan Instrumen Tes Kemampuan Pemecahan Masalah Matematis Berbasis Taksonomi SOLO pada Materi Barisan dan Deret. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, *06*(02), 2318–2335. https://j-cup.org/index.php/cendekia/article/view/1498.

Arifin, Z. (2017). Kriteria Instrumen dalam suatu Penelitian. *THEOREMS (The Original Research of Mathematics)*, *2*(1), 28–36. https://jurnal.unma.ac.id/index.php/th/article/view/571/537.

Arikunto, S. (2018). *Dasar - Dasar Evaluasi Pendidikan*. Bumi Aksara.

Bashooir, K., & Supahar. (2018). Validitas dan reliabilitas instrumen asesmen kinerja literasi sains pelajaran Fisika berbasis STEM. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(2), 168–181. https://doi.org/10.21831/pep.v22i2.20270.

Birch, C., Lichy, J., Mulholland, G., & Kachour, M. (2017). An enquiry into potential graduate entrepreneurship: Is higher education turning off the pipeline of graduate entrepreneurs? *Journal of Management Development*, *36*(6), 743–760. https://doi.org/10.1108/JMD-03-2016-0036.

Budiantoro, T., Kurniawan, B., Negeri, P., Laut, T., Negeri, P., & Laut, T. (2021). *Validitas Dan Reliabilitas Instrumen Keterampilan Komunikasi Dan Keterampilan Kolaborasi Pada Mata Kuliah Bahasa Indonesia*. *7*. https://jht.politala.ac.id/index.php/jht/article/view/89.

Dewi, S. K., & Sudaryanto, A. (2020). Validitas dan Reliabilitas Kuisioner Pengetahuan , Sikap dan Perilaku Pencegahan Demam Berdarah. *Prosiding Seminar Nasional Keperawatan Universitas Muhammadiyah Surakarta*, 73–79. https://publikasiilmiah.ums.ac.id/handle/11617/11916.

Elviana. (2020a). Analisis Butir Soal Evaluasi Pembelajaran Pendidikan Agama Islam Menggunakan Program Anates. *Jurnal Mudarrisuna*, *10*(2), 58–74. http://dx.doi.org/10.22373/jm.v10i2.7839.

Elviana. (2020b). Analisis Butir Soal Evaluasi Pembelajaran Pendidikan Agama Islam Menggunakan Program Anates. *Jurnal MUDARRISUNA*, *10*(2), 58–74. https://jurnal.ar-raniry.ac.id/index.php/mudarrisuna/article/view/7839.

Erawati, N. K. (2018). Analisis Tes Penilaian Pencapaian Kompetensi Pada Mahasiswa Kebidanan. *Jurnal Penjakora*, *5*(2), 111–120. https://ejournal.undiksha.ac.id/index.php/PENJAKORA/article/view/17287/10378.

Fahrurrozi, M., & Laili Rahmawati, S. N. (2021). Pengembangan Model Instrumen Evaluasi Menggunakan Aplikasi Kahoot Pada Pembelajaran Ekonomi. *Jurnal PROFIT Kajian Pendidikan Ekonomi Dan Ilmu Ekonomi*, *8*(1), 1–10. https://doi.org/10.36706/jp.v8i1.13090.

Faiz, A., Permana Putra, N., & Nugraha, F. (2022). Memahami Makna Tes, Pengukuran (Measurement), Penilaian (Assessment), Dan Evaluasi (Evaluation) Dalam Pendidikan. *Jurnal Education and Development*, *10*(3), 492–495. https://journal.ipts.ac.id/index.php/ED/article/view/3861.

Farida, & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. *Al-Mu'arrib: Jurnal Pendidikan Bahasa Arab*, *I*(1), 34–44. https://jurnal.lp2msasbabel.ac.id/index.php/AL-MUARRIB.

Friatma, & Anhar. (2019). Analysis of validity, reliability, discrimination, difficulty and distraction effectiveness in learning assessment. *Journal of Physics: Conference Series*, *1387*(1). https://doi.org/10.1088/1742-6596/1387/1/012063.

Hanifah, U. (2014). Pentingnya Buku Ajar yang Berkualitas dalam Meningkatkan Efektivitas Pembelajaran Bahasa Arab. *Tarbiyah, Jurnal Ilmu*, *3*(1), 99–137. https://diglosiaunmul.com/index.php/diglosia/article/view/125.

Hasyim, A. F., Munawar, B., & Ma'arif, M. (2021). Penggunaan Media Video Untuk Meningkatkan Pemahaman Karakteristik Arus Searah Dan Bolak-Balik Pada Peserta didik MAN 1 Pandeglang. *Jurnal Pendidikan*, *9*(1), 5–24. https://unimuda.e-journal.id/jurnalpendidikan/article/view/545.

Hodyanto, H., & Saputro, M. (2018). Workshop Pembuatan Dan Analisis Butir Soal menggunakan ITEMAN Pada Madrasah Aliyah Miftahul Huda Kecamatan Sungai Ambawang. *Jurnal Transformasi*, *14*(2), 85–90. https://journal.uinmataram.ac.id/index.php/transformasi/article/view/578.

Idrus. (2019). Evaluasi Dalam Proses Pembelajaran. *Evaluasi Dalam Proses Pembelajaran*, *2*, 920–935. http://dx.doi.org/10.35673/ajmpi.v9i2.427.

Iskandar, A., & Rizal, M. (2018). Analisis Kualitas Soal di Perguruan Tinggi Berbasis Aplikasi TAP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(1), 12–23. https://doi.org/10.21831/pep.v22i1.15609

Jumrah, Rukli, & Sulfasyah. (2023). Pengembangan Instrumen Tes Berbasis HOTS dengan Pendekatan Pengukuran Rasch pada Pelajaran Matematika Topik Bangun Ruang untuk Siswa Sekolah Dasar. *Jurnal Basicedu*, *7*(1), 11–27. https://jbasic.org/index.php/basicedu/article/view/4207.

Lia, R. M., Rusilowati, A., & Isnaeni, W. (2020). NGSS-Oriented Chemistry Test Instruments: Validity and Reliability Analysis with the Rasch Model. *Research and Evaluation in Education*, *6*(1), 41–50. https://doi.org/10.21831/reid.v6i1.30112.

Litna, K. ., Mertasari, N. M. ., & Sudhirta, G. (2021). Pengembangan Instrumen Tes Higher Order Thinking Skills (HOTS) Matematika SMA Kelas X. *Jurnal Penelitian Dan Evaluasi Pendidikan Indonesia*, *11*(1), 10–20. https://ejournal2.undiksha.ac.id/index.php/jurnal_ep/article/view/278.

Ma'rifah, U., Algiovan, N., & Sutarsyah, C. (2021). An Item Analysis of English Test During Online Learning. *International Journal of Multicultural and Multireligious Understanding*, *8*(12), 647–654. http://ijmmu.comhttp//dx.doi.org/10.18415/ijmmu.v8i12.3396.

Magdalena, Sundari, Nurkamilah, & Nasrullah. (2020). Analisis bahan ajar. *Jurnal Pendidikan Dan Ilmu Sosial*, *2*, 311–326. https://doi.org/10.36088/nusantara.v2i2.828.

Manfaat, B., & Nurhairiyah, S. (2021). Pengembangan Instrumen Tes untuk Mengukur Kemampuan Penalaran Statistik Mahasiswa Tadris Matematika. *Jurnal Pembelajaran Matematika*, *1*(1), 1–19. https://syekhnurjati.ac.id/jurnal/index.php/eduma/article/view/41.

Mappalesye, N., Sari, S. S., & Arafah, K. (2021). *Pengembangan Intrumen Tes Kemampuan Berpikir Kritis Dalam Pembelajaran Fisika*. *1*, 69–83. https://ojs.unm.ac.id/JSdPF/article/view/19091.

Masitoh, L. F., & Aedi, W. G. (2020). Pengembangan Instrumen Asesmen Higher Order Thinking Skills (HOTS) Matematika di SMP Kelas VII. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *4*(2), 886–897. https://doi.org/10.31004/cendekia.v4i2.328.

Muliyani, T., & Huriaty, D. (2016). Pengembangan Instrumen Tes Geometri dan Pengukuran pada Jenjang SMP. *Math Didactic: Jurnal Pendidikan Matematika*, *2*(2), 91–98. https://doi.org/10.33654/math.v2i2.33.

Muluki, A., Bundu, P., & Sukmawati, I. (2020). Analisis Kualitas Butir Tes Semester Ganjil Mata Pelajaran IPA Kelas IV Mi Radhiatul Adawiyah. *Jurnal Ilmiah Sekolah Dasar*, *4*(1), 86–96. https://doi.org/10.23887/jisd.v4i1.23335.

Muttaqin, M. Z., & Kusaeri. (2017). Pengembangan Instrumen Penilaian Tes Tertulis Bentuk Uraian untuk Pembelajaran PAI Berbasis Masalah Materi Fiqh. *Jurnal Pemikiran Dan Penelitian Pendidikan*, *15*(1), 1–23. https://journal.uinmataram.ac.id/index.php/tatsqif/article/view/11.

Ndiung, S., & Jediut, M. (2020). Pengembangan Instrumen Tes Hasil Belajar Matematika Peserta Didik Sekolah Dasar Berorientasi Pada Berpikir Tingkat Tinggi. *Premiere Educandum : Jurnal Pendidikan Dasar Dan Pembelajaran*, *10*(1), 94. https://doi.org/10.25273/pe.v10i1.6274.

Nengsi, A. R., & Efrina, G. (2019). Optimasi validitas dan reliabilitas tes pilihan ganda buatan guru mata pelajaran ips sd. *Journal Innovation in Islamic Education: Challenges and Readiness in Society 5.0, 4th International Conference on Education*, 43–48. https://ojs.iainbatusangkar.ac.id/ojs/index.php/proceedings/article/view/2138.

Nurfillaili, U., Yusuf, M., & Santih, A. (2016). Pengembangan Instrumen Tes Hasil Belajar Kognitif Mata Pelajaran Fisika pada Pokok Bahasan Usaha dan Energi SMA Negeri Khusus Jeneponto Kelas XI Semester I. *Jurnal Pendidikan Fisika*, *4*(2), 83. http://journal.uin-alauddin.ac.id/indeks.php/PendidikanFisika.

Purnamasari, N. L. (2019). Metode Addie pada Pengembangan Media Interaktif Adobe Flash pada Mata Pelajaran TIK. *Jurnal Pendidikan Dan Pembelajaran Anak Sekolah Dasar*, *5*(1), 23–30. https://jurnal.stkippgritulungagung.ac.id/index.php/pena-sd/article/view/1530.

Purnomo, P., & Maria Sekar Palupi. (2016). Pengembangan Tes Hasil Belajar Matematika Materi Menyelesaikan Masalah yang Berkaitan dengan Waktu, Jarak dan Kecepatan untuk Siswa Kelas V. *Jurnal Penelitian (Edisi Khusus PGSD).*, *20*(2), 151–157. https://e-journal.usd.ac.id/index.php/JP/article/view/872.

Puspasari, H., & Puspita, W. (2022). Uji Validitas dan Reliabilitas Instrumen Penelitian Tingkat Pengetahuan dan Sikap Mahasiswa terhadap Pemilihan Suplemen Kesehatan dalam Menghadapi Covid-19 Validity Test and Reliability Instrument Research Level Knowledge and Attitude of Students Towards. *Jurnal Kesehatan*, *13*, 65–71. http://www.ejurnal.poltekkes-tjk.ac.id/index.php/JK/article/view/2814.

Putri, R. Z., & Sari, R. (2020). Pengembangan dan Validasi Instrumen Tes untuk Mengukur Keterampilan Menyelesaikan Masalah Peserta Didik SMA pada Pelajaran Fisika. *Jurnal Penelitian Pembelajaran Fisika*, *11*(1), 17–25. https://doi.org/10.26877/jp2f.v11i1.3993.

Quansah, F., Amoako, İ., & Ankomah, F. (2019). Teachers' Test Construction Skills in Senior High Schools in Ghana: Document Analysis. *International Journal of Assessment Tools in Education*, *6*(1), 1–8. https://doi.org/10.21449/ijate.481164.

Qurrota, A. A. S., Siskawati, F. S., & Irawati, T. N. (2022). Analisis Kelayakan Butir Soal pada Media INTERMATHLY (Interesting Mathematic Monopoly). *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *6*(1), 634–654. https://doi.org/10.31004/cendekia.v6i1.1181.

Ramadhan, M. F., Siroj, R. A., & Afgani, M. W. (2024). Validitas and Reliabilitas. *Journal on Education*, *6*(2), 10967–10975. https://doi.org/10.31004/joe.v6i2.4885.

Rizky Ananda Setiyawan, & Palupi Sri Wijayanti. (2020). Analisis Kualitas Instrumen Untuk Mengukur Kemampuan Pemecahan Masalah Siswa Selama Pembelajaran Daring Di Masa Pandemi. *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, *1*(2), 130–139. https://doi.org/10.46306/lb.v1i2.26.

Sa'diyyah, F. N., Mania, S., & Suharti. (2021). Pengembangan Instrumen Tes untuk Mengukur Kemampuan Berpikir Komputasi Siswa. *Jurnal Pembelajaran Matematika Inovatif*, *4*(1), 17–26. https://doi.org/10.22460/jpmi.v4i1.17-26.

Santoso, A. (2018). Karakteristik Butir Tes Pengantar Statistika Sosial Berdasarkan Teori Respon Butir. *Jurnal Pendidikan Matematika Dan Sains*, *VI*(2), 1–11. https://journal.uny.ac.id/index.php/jpms/article/view/23959.

Sarea, M. S., & Ruslan, R. (2019). Karateristik Butir Soal: Classical Test Theory VS Item Response Theory? *Didaktika: Jurnal Kependidikan*, *13*(1), 1–16. https://jurnal.iain-bone.ac.id/index.php/didaktika/article/view/296.

Setyaedhi, H. S., Mustaji, & Fitri, C. (2023). Empirical Quality of Final Exam Questions in a Learning Management System-Based Course. *JPI (Jurnal Pendidikan Indonesia)*, *12*(1), 78–88. https://doi.org/10.23887/jpiundiksha.v12i1.52262.

Sholihah, I., Fahrurrozi, & Abdul Gafur, H. (2017). *Pengembangan Modul Pembelajaran Ekonomi Berbasis Guided Inquiry untuk Meningkatkan Hasil Belajar Siswa Kelas XI MA NW Sukamulia*. *1*(2), 104–117. https://ejournal.unsri.ac.id/index.php/jp/article/view/13090/0.

Sijabat, M. P., Hutabarat, K., Sitorus, L., & Syahrial. (2024). Analisis Soal Tes Hasil Belajar Siswa Soal Berstandar Nasional Bahasa Indonesia Kelas 5 Sekolah Dasar. *Jurnal Basicedu*, *8*(5), 1265–1277. https://journalstkippgrisitubondo.ac.id/index.php/PKWU/article/view/67.

Suci Mitra, P., & Helendra. (2022). Analisis Kualitas Butir Soal Ujian Akhir Semester Ganjil Tahun Pelajaran 2020 / 2021 Mata Pelajaran Biologi Kelas X SMA Negeri 1 Teluk Sebong Analysis of Quality The Question Final Exam Odd Semester 2020 / 2021 Biology Class X SMA Negeri 1 Teluk Sebong. *Biodidaktika: Jurnal Biologi Dan Pembelajarannya*, *17*(2). http://dx.doi.org/10.30870/biodidaktika.v17i2.16493.

Sugiono, Noerdjanah, & Wahyu, A. (2020). Uji Validitas dan Reliabilitas alat Ukur SG Posture Evaluation. *Jurnal Keterapian Fisik*, *5* no *1*, 55–61. https://jurnalketerapianfisik.com/index.php/jpt/article/view/167.

Sukiman. (2017). *Sistem Penilaian Pembelajaran*. Media Akademi.

Sulistianingsih. (2020). Pengetahuan Guru tentang Konstruksi Tes , Penguasaan Materi Pelajaran Sains dengan Reliabilitas Tes Buatan Guru. *Jurnal Ilmu Pendidikan (JIP) STKIP Kusuma Negara*, *11*(2), 145–153. https://jurnal.stkipkusumanegara.ac.id/index.php/jip/article/view/120.

Supriyati, Y., & Dudung, A. (2019). *Penilaian Kelas*. Karima (Karya Ilmu Media Aulia).

Surapranata, S. (2004). *Analisis, Validitas, Reliabilitas dan Interpretasi Hasil tes*. Remaja Rosdakarya.

Syaifudin. (2020). Validitas dan Reliabilitas Instrumen Penilaian Pada Mata Pelajaran Bahasa Arab. *Cross-Border:Jurnal Kajian Perbatasan Antarnegara*, *3*(2), 106–118. http://journal.iaisambas.ac.id/index.php/Cross-Border/article/view/553/447.

Tarumasely, Y. (2020). Perbedaan Hasil Belajar Pemahaman Konsep Melalui Penerapan Strategi Pembelajaran Berbasis Self Regulated Learning. *Jurnal Pendidikan Dan Kewirausahaan*, *8*(1), 54–65. https://doi.org/10.47668/pkwu.v8i1.67.

Utomo, B. (2018). Analisis Validitas Isi Butir Soal Sebagai Salah Satu Upaya Peningkatan Kualitas Pembelajaran Di Madrasah Berbasis Nilai - Nilai Islam. *Jurnal Pendidikan Matematika*, *1*(2), 145–159. http://journal.stainkudus.ac.id/index.php/jmtk%0AANALISIS.

Warju, Rizki, S., Soeryanto, & Adi, R. (2020). Analisis Kualitas Butir Soal Tipe HOTS pada Kompetensi Sistem Rem Siswa di Sekolah Menengah Kejuruan. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *17*(1), 1–9.

Wartoni, & Benyamin, P. I. (2020). Strategi Pengembangan Tes Objektif (Pilihan Ganda). *Diegesis : Jurnal Teologi*, *5*(1), 1–8. https://doi.org/10.46933/DGS.vol5i1.

Widayanti, W., Bistari, & Suparjan. (2021). Analisis Butir Soal Pilihan Ganda Penilaian Tengah Semester Pada Pembelajaran Tematik Kelas V Sekolah Dasar Negeri 39 Pontianak Kota. *Jurnal DIDIKA : Wahana Ilmiah Pendidikan Dasar*, *7*(2). https://doi.org/10.29408/didika.v7i2.4370.

Widiyanto, J. (2018). Evaluasi Pembelajaran (Sesuai dengan Kurikulum 2013) : Konsep, Prinsip & Prosedur. *Unipma Press*, 257.

Widiyawati, Y., Nurwahidah, I., & Sari, D. S. (2019). Pengembangan Instrumen Integrated Science Test Tipe Pilihan Ganda Beralasan Untuk Mengukur HOTS Peserta Didik. *Saintifika*, *21*(2), 1–14. issn: 1411 – 5433

Yusup, F. (2018). Uji Validitas Dan Reliabilitas Instrumen Penelitian Kuantutatif. *Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan*, *7*(1), 17–23. https://doi.org/10.21831/jorpres.v13i1.12884.