HISTORIA: Jurnal Program Studi Pendidikan Sejarah Volume 12 (1) 2024
ISSN 2337-4713 (E-ISSN 2442-8728)

Analysis of Multiple Choice Questions as an Evaluation Tool for the end of Year Assessment (PAT) in the Subject of Indonesian History Class XI SMA

Muhammad Syahroni[1*], Ana Nurhasanah[2], Arif Permana Putra[3]
[1,2,3]History Education, Faculty of Teacher Training and Education, Sultan Ageng Tirtayasa University, Indonesia
*correspondence email: 2288160036@untirta.ac.id

Received 23 August 2023; Received in revised form 27 September 2023; Accepted 29 September 2023

Abstrak

Penelitian ini bertujuan untuk mengetahui validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektifitas pengecoh soal pilihan ganda sebagai alat evaluasi mata pelajaran sejarah di kelas XI SMAN 2 Pandeglang. Penelitian ini bersifat kuantitatif dan deskriptif. Siswa kelas XI B dan D IPS di SMAN 2 Pandeglang sebagai sampel penelitian. Berdasarkan hasil penelitian pada 30 butir soal Penilaian Akhir Tahun (PAT) mata pelajaran sejarah Indonesia kelas XI IPS SMAN 2 Pandeglang tahun ajaran 2021/2022 terdapat 13 butir soal valid (43,3%) dan 17 butir soal (56,7%) yang tidak valid. Angka reliabilitas 0,49 menunjukkan bahwa baik soal maupun hasil tidak reliabel. Pada daya pembeda, 20 butir soal (66,7%) masuk dalam kategori lemah, 7 butir soal (23,3%) masuk dalam kategori cukup, 1 (satu) butir soal (3,3%) masuk dalam kategori baik, 2 (dua) butir soal (6,7%) masuk dalam kategori sangat lemah, dan tidak ada terdapat butir soal dengan kategori sangat baik. Indeks kesukaran menunjukan bahwa 4 (empat) butir soal atau 13,3% termasuk dalam kategori mudah, 26 butir soal atau 86,7% yang tergolong sukar, dan tidak terdapat butir soal yang termasuk dalam kategori sangat sukar. Pada efektifitas pengecoh seluruh butir soal (30) memiliki keefektifan pengecoh dengan kriteria baik. Hasil validasi oleh dosen ahli terhadap penelaahan soal menunjukkan bahwa 27 dari 30 butir soal telah memenuhi aspek kesesuaian butir soal dengan indikator, 29 dari 30 butir soal sesuai dengan kaidah keilmuan, seluruh butir soal memiliki satu kunci jawaban. Sementara itu, hasil validasi terhadap konstruksi soal menunjukan bahwa seluruh soal telah memenuhi aspek kaidah penulisan butir soal. Dan, sebanyak 27 dari 30 butir soal dianyatakan memenuhi kaidah bahasa dalam penulisan butir soal.

Kata Kunci: analisis butir soal, pilihan ganda, sejarah Indonesia.

Abstract

This study aims to determine the validity, reliability, level of difficulty, discriminating power, and the effectiveness of the multiple choice distractor as an evaluation tool for history subjects in class XI SMAN 2 Pandeglang. This research is quantitative and descriptive. Students of class XI B and D IPS at SMAN 2 Pandeglang were used as the research sample. Based on the results of the research on 30 year-end assessment questions (PAT) for Indonesian history class XI IPS SMAN 2 Pandeglang for the 2021/2022 academic year, there were 13 valid questions (43.3%) and 17 questions (56.7%) which invalid. The reliability score of 0.49 indicates that neither the questions nor the results are reliable. On discriminating power, 20 items (66.7%) fall into the weak category, 7 items (23.3%) fall into the sufficient category, 1 (one) item (3.3%) falls into the good category, 2 (two) items (6.7%) fall into the very weak category, and there are no items in the very good category. The index of difficulty shows that 4 (four) items or 13.3% are included in the easy category, 26 items or 86.7% are classified as difficult, and there are no items included in the very difficult category. On the effectiveness of the distractor, all items (30) have the effectiveness of the distractor with good criteria. The results of the validation by the expert lecturer on the study of the questions showed that 27 out of 30 items had fulfilled the suitability aspect of the items with indicators, 29 out of 30 items were in accordance with scientific principles, all items (30) had one answer key. Meanwhile, the results of the validation of the construction of the questions showed that all questions met the

aspects of the rules for writing items. And, as many as 27 of the 30 items were stated to fulfill the language rules in writing the items.
***Keywords**: analysis of question items, multipe choice, indonesian history.*

INTRODUCTION

Ideally, instruments or questions used to measure student learning outcomes must be carefully designed, prepared using question writing rules, tested, and evaluated for quality. Make a grid of questions first, develop scoring guidelines and apply them to avoid biased assessment results on question descriptions, arrange questions and distract the correct answer key, pay attention to the quality of the instruments or questions that must be prepared so that they are in accordance with the criteria for good questions and rules. preparing questions, testing questions and analyzing them to find out the quality of each question prepared are things that educators need to pay attention to in order to avoid these problems. It is hoped that teachers will develop quality questions and be able to measure student learning outcomes accurately to ensure they are aligned with learning objectives (Kusnandar, 2014: 65).

The problem with the current test is that the collection of questions used as an evaluation tool has not been thoroughly analyzed so the quality of the questions is still unclear. The quality of students can be affected by these significant issues. According to Jandaghi & Shaterian (2018), who published their research in the journal Validity, Reliability, and Difficulty Indies for Instructor-Built Exam Questions, student quality is one of the most significant problems in the teaching and learning system. In order for exam questions to have a higher quality than the expected passing standard, there needs to be a standard. The questions on an exam have a big impact on how well a student does. To get good results, exam questions must be checked for level of difficulty, distinguishing power, validity and reliability (Jandaghi & Shaterian, 2018: 36).

The questions that have been prepared must be tested first. This is expected to determine the nature of the questions in order to measure student abilities. Students' final exam scores or results will be affected if the teacher's questions are not of renowned quality (Kunandar, 2014: 167). Knowing the characteristics of a good question can help determine its quality. Validity, dependability, level of difficulty, differentiating power, and effectiveness of distractors for multiple choice questions are important components of a good question (Arikunto, 2013: 222). The government is trying to overcome this problem by publishing a guidebook for preparing good questions, compiling HOTS (Higher Order Thinking Skill) type questions, and assessing elementary and middle school learning outcomes, in

addition to revising the 2013 curriculum to improve the quality of education. . However, not all educators carry out assessments in accordance with the requirements set by the government. Looking at the statements above, it can be emphasized that the truth of most of the tests used as learning assessment tools has not been detailed from top to bottom. Educators must carry out analytical activities to ensure that the test questions given to students are of high quality and in line with the educators' measurement objectives (Arikunto, 2013: 222).

Educators have not carried out a comprehensive analysis of the end-of-year assessment questions given to students, according to observations made on January 22 2021 at SMA Negeri 2 Pandeglang. This is because teachers do not have much time, making it impossible to carry out a thorough analysis. So far, educators have been able to determine whether a question is good or not based on whether students pass it or not and how difficult the question is based on the operational verbs in the question. Of course, this will result in the unknown quality of the test equipment as an evaluation tool in terms of validity, dependability, distinguishing power, level of difficulty and effectiveness of distractors. By analyzing the question items, it can be determined which ones should be revised and which ones should be eliminated, so as to identify questions that are quality and useful as measurement instruments. Teachers will be able to obtain accurate evaluation results thanks to the high quality of the questions. Evaluation findings will reveal what needs to be done next, as well as information about student learning feedback, learning progress, and learning programs. As a learning evaluation tool, recommendations from relevant studies to answer the problems mentioned above must be analyzed.

In subjects such as history, there is also the problem of question quality. The aim of this research is to improve the quality of history learning in schools and learning outcomes for students. The question to be analyzed is what differentiates the current research from previous research. This research was conducted on end-of-year multiple choice questions (PAT) because this summative exam has a significant impact on student learning outcomes and determines whether students can continue to the next material or receive a grade increase. Besides that, it is still a challenge for researchers to find studies on the subject of historical analysis. The majority of analysis is carried out in the fields of science and social sciences, such as accounting and economics. This research is needed because of the many problems and lack of research on problem analysis in history courses. Research will be conducted on the analysis of multiple

choice questions related to the level of difficulty of the questions, validity, reliability, distinguishing power, level of difficulty, and effectiveness of delusions based on explanations of expert opinion, reality, and observation results.

METHOD

This research uses evaluative research as its method. Research that requires the existence of criteria, benchmarks, or standards that are used as a comparison between the data obtained and data that has been processed to represent the actual condition of the object is called evaluation research. researched. In addition, the gap between actual conditions and the expected conditions outlined in the criteria is the aim of getting a general idea of whether the research subject meets the criteria or not. According to Arikunto, the aim of this research is to compare the data or information that has been collected with the criteria and draw conclusions (Arikunto, 2014:36).

This research uses a combination of quantitative and qualitative methods (Creswell, 2016: 5). A mixed approach is an approach in which different designs and quantitative and qualitative data sets are combined. When compared with a single approach, this combination may offer a more comprehensive understanding. The quality of the even semester summative questions in class XI history subjects at SMA Negeri 2 Pandeglang is measured using a quantitative approach, which will be shown by calculating the numbers. Meanwhile, subjective methodology is used to understand the exploration information as a whole. so that numerical data resulting from quantitative analysis activities can be understood thoroughly and easily.

This research was carried out in class XI IPS at SMA Negeri 2 Pandeglang using UKBM (Independent Learning Activity Unit) learning combined with a semester credit system program.

Information sorting is planned to obtain data that is relevant, precise and can be used appropriately in accordance with learning objectives. Interviews and documentation are methods used in research to collect data. Interviews are a method of gathering information by asking relevant questions. Historical records are called documentation. Documents can be written, visual or monumental works carried out by an individual (Sugiyono, 2013: 82). As stated by Arikunto (2014: 158). The word “document” refers to written material and is the origin of documentation. Research uses documentation methods to look at written things such as books, magazines, documents, regulations, meeting minutes, and so on.

Open interviews were used to collect information about the preparation of

questions and the use of formative tests in this research. Multiple choice questions are also used as a documentation method for class XI B and D IPS students at SMAN 2 Pandeglang. The documentation approach begins with collecting data on student learning achievements, learning syllabus, question papers, summative papers, and answer keys to even semester questions from history teachers, as well as responses from all class XI IPS students to the summative questions tested.

The data analysis technique used is Item Response Theory (IRT), also known as question answering theory, used in the process of analyzing the quality of questions. Mathematical functions are used in this theory to establish a relationship between a student's ability and the likelihood that they will answer a question correctly. Item Response Theory (IRT) is the relationship between a student's ability or level of achievement and the probability of answering a question item correctly. The following formula will be used to carry out quantitative analysis with the help of computer programs, especially Microsoft Office Excel 2010.

The analysis carried out includes Validity test, Reliability test, Difficulty Level Test, Differentiating Power test, and Distractor Effectiveness test. As well as drawing conclusions at the end.

RESULTS AND DISCUSSION

Qualitative Analysis

This analysis is carried out while studying the questions before the exam. This analysis, also known as logical analysis, requires preparing questions, discussing these questions, and determining whether the function of the problem is based on material aspects (Sudarsono, 2012: 138). The study can utilize the study sheet that has been prepared. The purpose of this study sheet is to simplify the implementation process (Kusaeri and Suprananto, 2012: 166).

The results of the qualitative analysis are in accordance with the question grid, but three questions—numbers 13, 23, and 24—do not match the indicators. In terms of content, the material is scientifically correct, one question is not correct, number 17. There are two aspects of the assessment that are not correct in terms of construction. The first answer choice is clearly stated in questions 13 and 28. Numbers 8 and 14 are the second choice in terms of time and number. There are two aspects of the assessment that are less precise in terms of language; the first is the communicative aspect of language, which is listed in number 29. In numbers 6 and 14 there are two grammatical sentences.

Quantitative Analysis

a. Validity

Indicators of instrument accuracy or validity are valid measures or standards. Testing the validity of objects in the End of Year Evaluation Questions (PAT) in the Indonesian History subject class XI IPS at SMAN 2 Pandeglang for the 2021/2022 academic year uses the second item relationship equation (rxy).

The product moment (rxy) correlation formula can be used to determine the validity of a question. The rxy correlation index is obtained from consultation calculations with rtable at a significance level of 5%, depending on the number of students taking the test. If rcount is greater than rtable, then the condition for seeing the question is true. At SMAN 2 Pandeglang there are 70 students taking classes XI IPS B and D. After looking at the r table, we found 68 participants with a score of 0.2199. If rcount is less than 0.2199, then the item is considered valid.

The results of the analysis show that in the End of Year Assessment Questions (PAT) on Indonesian History for class declared invalid or 56.7%. Valid questions (43.3%) indicate that the items have been able to carry out their function, namely measuring what should be measured. Then, there are various reasons why 56.7% of the questions are invalid. This is in accordance with Grounlund's theory in Zainal Arifin (2014: 247) which states that administration and assessment factors, student answer factors, and instrument factors all have an influence on the validity of test results. Students' tendency to answer quickly and incorrectly can have an impact on the End of Year Assessment Questions (PAT) for the Indonesian History subject for the 2021/2022 academic year.

The description above shows that the End of Year Assessment Questions (PAT) for the Indonesian History Subject Class XI IPS for the 2021/2022 Academic Year have low validity quality. The question bank can be used to store valid question items. On the other hand, invalid items must be corrected by adjusting indicators and improving technical capabilities in compiling items.

b. Reliability

The calculation of reliability on the question of the End of Year Assessment (PAT) of Indonesian History Subject class XI IPS at SMAN 2 Pandeglang in the 2021/2022 school year was carried out manually using the help of the Microsoft Excel program using the KR 20 formula. Reliability is the level or degree of consistency of an instrument. A measurement has high reliability if the measurement is able to produce reliable data. The reliability of the score is measured by K-R 2, at $r11 \geq 0.70$ then the question tested has high reliability, but if $r11 \leq 0.70$ then the question tested has low reliability or unreliable.

Based on the results of research conducted manually using the help of the Microsoft Excel program, it shows that the question of the End of Year Assessment (PAT) of the Indonesian History Subject class XI IPS at SMAN 2 Pandeglang in the 2021/2022 school year on multiple choice questions has a reliability of r11 = 0.49.

The results of these calculations indicate that the question of the End of Year Assessment (PAT) of Indonesian History Subject XI IPS class at SMAN 2 Pandeglang in the 2021/2022 academic year on multiple choice questions has a low level of reliability because it has a reliability coefficient (r11) ≤ 0.70. The low reliability coefficient of the question is due to the limited number of items made by the teacher, so that the teacher can add the number of valid questions.

A test instrument that has good validity on each item will also have a high level of reliability as well. This is in line with the theory of Suharsimi Arikunto (2013: 101) which states that a test consisting of many items will be more valid than a test consisting of only a few items. The high and low level of validity can indicate the high and low reliability coefficient, so the longer the test, the higher the reliability. Based on this description, it can be concluded that the question of the End of Year Assessment (PAT) of Indonesian History Subject class XI IPS at SMAN 2 Pandeglang in the 2021/2022 school year multiple choice questions are questions that are not good in terms of reliability.

c. Distinguishing Power

The ability of an item to distinguish between participants with high scores and those with low scores is known as distinguishing power. If the question is answered correctly only by intelligent students, the differentiating power is useful. By using Microsoft Excel program, the differentiating power of Indonesian History End of Year Assessment (PAT) questions of class XI IPS SMAN 2 Pandeglang in 2021/22 academic year was calculated manually. Translation of the side effects of the power splitting calculation using the accompanying rules.

The results of the analysis of multiple choice questions of Indonesian History End-of-Year Assessment (PAT) class XI IPS B and D at SMAN 2 Pandeglang showed that 20 questions (or 66.7%) were in the poor category, 7 questions (or 23.3%) were in the poor category. in the moderate category, 1 item (or 3.3%) was in the good category, 2 items (or 6.7%) were in the very weak category, and 0 questions or 0% met the very high criteria. Questions with low differentiating power do not allow to distinguish student abilities. so that low-ability students can answer questions correctly and high-ability students can answer incorrectly.

Zainal Arifin (2014: 273), the calculation of differentiating power is a

measure of the extent to which a question is able to distinguish students who master the material from students who do not/do not master the material based on certain criteria. Anas Sudijono provides support stating that question compilers must realize that students' abilities vary and knowing the differentiating power of each item is very important because it is one of the guidelines for compiling question items (Anas Sudijono, 2011: 386). The question of the end-of-year assessment (PAT) of Indonesian History class XI IPS SMAN 2 Pandeglang in the 2021/2022 academic year has poor discrimination power because the majority of items or 66.7% are in the Bad category. This conclusion can be drawn from the discussion above. The next learning outcome test question bank should contain items with sufficient, good, and excellent discriminating power; items with poor discriminating power can be improved. It should not be used again for future tests, especially for items with negative discriminating power because these items are of very poor quality.

d. Difficulty Level

The difficulty level of each question can be used to determine its quality. There are three levels of difficulty: easy, medium, and difficult. Based on the findings of the analysis of multiple choice questions of the End of Year Assessment (PAT) of Indonesian History subject at SMAN 2 Pandeglang, there are four questions that fall into the easy category (13.3%), 26 questions that fall into the medium category (86.7%), and zero questions that fall into the difficult category (0%). The following is a follow-up that can be done after checking the level of difficulty of the questions (Anas Sudijono, 2011: 376-378):

a) Items that have a difficulty level in the good category (medium difficulty level) should be stored in the question bank so that they can be issued again in the future.
b) Question items that are classified as difficult questions, there are 3 possible follow-ups, namely: 1) The item is discarded and will not be used again in the next test. 2) Re-examined the factors that cause the item concerned to be difficult to answer by the testee. Improvements can be made by simplifying the question sentence so that it does not cause multiple interpretations. Furthermore, these items can be issued again in future learning outcomes tests. 3) Items are retained to be used again in tests that are very strict in nature, in the sense that most of the testees will not be passed in the selection test.
c) Question items that fall into the easy category, there are 3 possible follow-up actions, namely: 1) The item is discarded and will not be

reused in the next test. 2) Re-examine the factors that cause the item concerned to be difficult for the testee to answer. Improvements can be made by fixing the options and making the question sentence more complex. Furthermore, the item can be issued again in the next learning outcomes test. 3) Items are retained to be used again on tests that are loose in nature. In this condition, the test is only a formality.

e. Excerpt Effectiveness

The use of Microsoft Excel program was used to manually calculate the effectiveness of the distractors on the End of Year Assessment Question (PAT) of Indonesian History subject in class XII IPS SMAN 2 Pandeglang in the 2021-2022 academic year. The following criteria were used to interpret the distractors of each item: Then, using a standard adapted from the Likert scale, interpret the effectiveness of each item as follows: a) If all four answers to the question's exceptions can function properly, then the question can be said to have very good exception effectiveness. b) If there are three working answers to the question, then the question is said to have good effectiveness. c) If there are two working answers, the question is said to have fairly good effectiveness. d) If there is only one working answer, the question is said to have poor effectiveness. e) If all the answers to the triggers do not work, the question is said to have poor tracer effectiveness.

Analysis of the End of Year Assessment (PAT) questions of the Indonesian History course for class XII IPS SMAN 2 Pandeglang in the 2021-2022 academic year found that thirty questions (100%) met the required standards. Because the distractors meet the good criteria of 100 percent, the value of the effectiveness of the distractors is in the good category. The following can be followed (Anas Sudijono, 2011: 417). a) Excerpts that work well can be used again in future learning outcome tests. b) Checkers that have not functioned properly are repaired or replaced with other checkers. The way to make a good exception is as follows (Sumarna Surapranata. 2005: 136). 1) Use answer choices that learners can understand. 2) Use words that sound the same. 3) Use ones that are roughly related. 4) Use the language of the book that is undoubtedly true.

CONCLUSION

Item analysis of year-end assessment (PAT) of Indonesian history subject in class XI IPS SMAN 2 Pandeglang in 2021/2022 academic year resulted in the following conclusions: Validity, reliability, distinguishing power, difficulty level, and the effectiveness of the checkers. 1)

Based on the research findings, multiple choice questions of Indonesian history End of Year Assessment (PAT) class XI IPS SMAN 2 Pandeglang in the 2021/2022 academic year there are 13 valid questions (43.3%) and 17 invalid questions (56.7%). 2) ased on the findings of the reliability research, it shows that the multiple choice questions of the end-of-year assessment (PAT) of Indonesian History subject class XI IPS at SMAN 2 Pandeglang in the 2021/2022 academic year have a reliability of 0.49, this indicates that both the questions and the results are inconsistent. 3) Based on the research findings, the Discrimination Power reveals that the Indonesian History End of Year Assessment (PAT) question of XI social studies class at SMAN 2 Pandeglang in the 2021/2022 academic year contains multiple choice questions. Twenty questions (or 66.7%) are in the weak category, seven questions (or 23.3%) are in the sufficient category, one item (or 3.3%) is in the good category, two items (or 6.7%) are in the good category. 4) Based on the research findings, the End of Year Assessment (PAT) questions of Indonesian history subject in class XI IPS SMAN 2 Pandeglang in the 2021/2022 academic year, there are four multiple choice questions that fall into the easy category, or 13.3%, and four questions that fall into the medium category. 5) Based on the research, the Year-End Assessment Question (PAT) of Indonesian history class XI IPS SMAN 2 Pandeglang in the 2021/2022 academic year has 30 multiple choice questions (100%) with good criteria, this shows the effectiveness of learning. Because the effectiveness of the triggers meets the good criteria of 100%, the effectiveness value of the triggers is in the good category. 6) It is known that based on the overall analysis of the questions, the multiple choice End of Year Assessment (PAT) questions of Indonesian history class XI IPS SMAN 2 Pandeglang in the 2021/22 academic year amounted to 5 questions (16.7%). with quality category, 23 questions (76.7%) with low quality category, and 2 questions (6.6%) with low quality category. 7) Thanks to the Principal of SMAN 2 Pandeglang Dra. Hj. Lilis Lismunah, M.Pd who has given permission as a place of research in this paper. And thanks to Desi Tresnasari, S.Pd. as a teacher of Indonesian history who has helped a lot in the research process.

REFERENCE

Aiken, L.R. (1995). *Psychological Testing and Asessment*. Boston: Alliyn and Bacon.

Anastasi, Urbina, S. (1997). *Psycological Testing*. New Jersey: Prentice Hall, Inc.

Azwar, S. (2013). *Sikap Manusia: Teori dan Pengukurannya*. Yogyakarta: Pustaka Pelajar.

Arifin, Zainal. (2012). *Evaluasi Pembelajaran (Edisi Revisi)*. Jakarta. Direktorat Jendral Pendidikan Islam Kementerian Agama Islam.

Arifin, Zainal. (2014). *Evaluasi Pembelajaran*. Bandung: PT Remaja Rosdakarya.

_____.(2013). *Dasar-Dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.

_____.(2014). *Prosedur Penelitian: Suatu Pendekatan Praktik*. Jakarta.: PT Rineka Cipta.

Brinkerhoff, R. O. et al. (1986). *Program Evaluation: A Practitioner's Guide for. Trainers and Educators. Fourth Printing*. Massachusetts: Kluwer-Nijhoff Publishing.

Creswell, J.W. (2016). *Research Design: Pendekatan Metode Kualitatif, Kuantitatif, dan Campuran (Terjemahan Edisi 4)*. Yogyakarta: Pustaka Pelajar.

Darmawan, D. (2013). *Metode Penelitian Kuantitatif*. Bandung: Remaja Rosdakarya.

Depdiknas. (2007). *Panduan Penulisan Soal Pilihan Ganda*. Jakarta: Pusat Penilaian Pendidikan Balitbang Depdiknas.

Djamarah, S. (2005). *Guru dan Anak didik dalam interaksi edukatif suatu pendekatan teoritis psikologis*. Jakarta: PT. Rineka Cipta.

Djiwandono, S. (2008). *Tes Bahasa dalam Pengajaran*. Bandung: ITB.

Gronlund, E. Norman.(1982). *Constructing Achievement Test*. London: Prentice Hall.

Kusaeri dan Suprananto. (2012). *Pengukuran dan Penilaian Pendidikan*. Yogyakarta: Graha Ilmu.

Majid, A. (2014). *Penilaian Autentik Proses dan Hasil Belajar*. Bandung: PT. Remaja Rosdakarya Offset.

Mardapi, D. (2008). *Teknik Penyusunan Instrumen Tes dan Nontes*. Yogyakarta: Mitra Cendekia.

Muhson, A. (2015). *Panduan Penggunaan Anbuso Versi 6.1*. Yogyakarta: Universitas Negeri Yogyakarta.

Mulyasa, E. (2005). *Kurikulum Tingkat Satuan Pendidikan*. Bandung: PT. Remaja Rosdakarya.

Nurgiyantoro, B. (2011). *Penilaian Pembelajaran Bahasa*. Yogyakarta: BPFE.

Popham, W. J. (1995). *Classroom Assesment: What Teacher Need to Know*. Boston: Allyn and Bacon.

Pramana, Y.A. (2013). *Aplikasi Microsoft Excel 2010 untik Menganalisis Butir Soal Pilihan Ganda*. Skripsi. Semarang: UNS

Purwanto, Ngalim. (2013). *Prinsip-Prinsip dan Teknik Evaluasi Pengajaran*. Bandung: PT. Remaja Rosdakarya.

Purwanto. (2011). *Evaluasi Hasil Belajar*. Yogyakarta: Pustaka Pelajar.

Ratnawulan, E & Rusdiana. (2014). *Evaluasi Pembelajaran*. Bandung: Pustaka Pelajar.

Slameto. (2003). *Belajar dan Faktor-Faktor yang Mempengaruhi Edisi Revisi*. Jakarta: Rineka Cipta.

Subali, B. (2014). *Evaluasi Pembelejaran (Proses dan Produk)*. Purwokerto. Universitas Muhammadiyah Purwokerto.

Sudijono, A. (2009). *Pengantar Evaluasi Pendidikan*. Yogyakarta: PT Raja

Wijayanti, H. (2014). *Analisis Butir Soal Objektif UAS Semester Genap Kelas VII Mata Pelajaran Ilmu Pengetahuan Sosial (IPS) Tahun Pelajaran 2013/2014 Di SMP 3 Balung*. Skripsi. Jember. Universitas Jember Press.

Yulista H., Zulfan, & M. Arifin (2018). *Analisis Tingkat Kesukaran Soal dan Daya Pembeda Soal Mata Pelajaran Sejarah Kelas Xi Semester Ganjil di SMA Negeri 5 Banda Aceh Tahun Pelajaran 2015-2016*. Jurnal. Banda Aceh: Universitas Syiah Kuala.

Yuniarti, A.D. (2013). *Analisis Butir Soal Olimpiade Ekonomi VI pada Prodi Pendidikan Ekonomi FKIP UNEJ*

Tingkat SMA Sederajat. Skripsi. Jember. Universitas Jember Press.

Werdiningsih, G. (2015). *Analisis Kualitas Butir Soal Ulangan Akhir Semester Gasal Mata Pelajaran Ekonomi Kelas XII IPS SMAN 2 Banguntapan Tahun Ajaran 2014/2015*. Skripsi. Yogyakarta: Universitas Negeri Yogyakarta.

Tjalla, A.(2020) *Potret Mutu Pendidikan Indonesia Ditinjau dari Hasil-Hasil Studi Internasioanal*. Jurnal. Jakarta: Universitas Negeri Jakarta.

Toksosz, S & Ertunc, A. (2017). *Item Analysis of a Multiple Choice Exam*. Jurnal. Turkey. University, İstiklal Yerleşkesi Turkey.

Umamah, N. (2014). *Bahan Ajar: Perencanaan Pembelajaran Bidang Studi*. Jember: Universitas Jember.

Utami, I. (2016). *Analisis Butir Soal Pilihan Ganda Ulangan Akhir Semester Genap Tahun Pelajaran 2014/2015 Mata Pelajaran PKn Kelas IV SD di Kecamatan Depok, Sleman, Yogyakarta*. Skripsi. Yogyakarta: Universitas Sanata Yogyakarta.

Wahyudi, D. (2011). *Analisis Kualitas Butir Soal Mata Pelajaran Pendidikan Agama Islam dalam Pencapaian Kompetensi Siswa SMA Negeri 2 Kebumen*. Skripsi. Yogyakarta: Universitas Islam Negeri Sunan Kalijaga Yogyakarta.