Analysis of multiple-choice question (MCQ) of physics final examination in senior high school

N Rahmah¹, Yusrizal² and M Syukri²
¹Department of Science Education, Universitas Syiah Kuala, Banda Aceh, 23111, Indonesia
²Department of Physics Education, Universitas Syiah Kuala, Banda Aceh, 23111, Indonesia
E-mail: syukri.physics@unsyiah.ac.id

Abstract. This study aims to determine the quality of questions in multiple-choice test for School Final Exam of Physics subject at senior high school Pidie academic year 2018/2019, reviewed based on validity, reliability, level of difficulty, differentiating question level, and function of distracting questions. This non-experimental quantitative descriptive study involved year 12 students of three senior high school in Pidie academic year 2018/2019, namely senior high school Unggul Sigli, senior high school 1 Mila, and senior high school 1 Muara Tiga which were selected by stratified random sampling. The data source comes from the entire school final exam question sheet physics subjects for the 2018/2019 school year, answer keys and student answer sheets. The results of the package A and B final exam questions showed that not all of the questions were valid, the reliability of the questions was high, the differentiating question level was not the same, the difficulty level of the questions was high, the differentiating question level was not the same, the difficulty level of the questions was different, and the effectiveness of the question was generally good and some of distracting questions need to be revised.

1. Introduction

Education is learning from the knowledge, skill, and attitude or habit of individuals or groups. Learning is essentially a complex (complicated) process, but with the same intention; to provide learning experiences for students in accordance with the objectives. The ultimate goal of education is a reference towards the implementation of the learning process [1]. The learning process starts from the learning plan designed by the teacher, followed by the implementation of learning based on the initial plan. Then an evaluation is carried out in the form of assessment because one of the main tasks of the teacher as a professional is to assess and evaluate students. The learning evaluation process must be mastered by an educator. Evaluation is an assessment from planning to implementation in measuring students' abilities or making a decision on the results of the assessment. Evaluation of learning outcomes is inseparable from the activity of collecting data, so a tool called evaluation or instrument is needed. Evaluation refers to actions or processes to determine a value whilst in doing evaluation, not only value is related, but also meaning [2].

The test is one of the evaluation tools that serves to measure student learning progress [3]. Test is also used to measure the amount of knowledge of an individual acquires from a subject matter limited to a certain level [4]. Test is one way to interpret the magnitude of a person's ability measured indirectly through a person's response to a stimulus or question [5]. The questions used as material for this test are an
ingredient in assessing student learning to achieve goals in certain competencies. Tests in the world of education when viewed from their forms consist of written tests (multiple-choice, description, matchmaking, short answering, filling, and correct-incorrect), oral tests, and practice tests. While those classified as non-test techniques are observation (usually using tools in the form of checklists, and rating scales), questionnaires, and documentation. There are two forms of learning outcomes test: description form and multiple-choice test [6]. The multiple-choice test is used to measure student learning outcomes in a complex manner. Multiple-choice test is used so that the test can cover all the material that has been taught. Evaluation tools (test questions) can be made directly by the teacher or by the designated team. Good tests are prepared in accordance with the procedures and principles of preparing the test. A good test can be used repeatedly with a few changes.

Based on preparation, the test is divided into two types such as standardized tests and teacher-made tests [7]. Standardized tests were developed to follow procedures and meet strict requirements. In particular learning, teacher-made tests are widely used, both in relation to language skills and language components [8]. Multiple choice questions test can be used to measure students' cognitive learning outcomes (memory, understanding, application, analysis, synthesis, and evaluation). Multiple-choice questions are questions whose answers must be chosen from several possible answers provided. In general, each multiple choice question consists of a subject matter (stem) and choice of answers (option)[9]. The final school exam questions consist of two questions, namely type A and type B with the type of questions in the form of multiple choice and essays. Multiple-choice tests are also called objective tests because the choice of answers is already available, one of which is the key answer to the question being asked. In the multiple-choice test, each item uses several tricky answers (distractor)2.

Analysis of question items is a systematic procedure to assess the quality of questions in the test from student answers which include empirical validity, the level of difficulty of the questions, the level of different questions, the effectiveness of distractor, and the reliability of exam questions [10]. The research conducted by Anita et al. (2018) concluded that the analysis of items of the qualitative analysis is based on constructs, material aspects and language as a whole are appropriate, but there are several questions that need to be corrected in terms of constructs, material and language. Quantitative analysis based on the level of difficulty obtained by the difficult category by 26 questions, moderate category by 9 questions, and easy category by 5 questions. Its differentiating level is 30% poorly categorized, 37.5% is enough, 20% is good, and 12.5% is very bad. The effectiveness of the 80% distracting question were functioning, the level of validity is 65%, the reliability is 0.65 with the medium category so the questions can be trusted to evaluate students. The ability of students to work on problems is still at the level of knowledge and understanding [11].

Based on the problem, the purpose of this study was to determine the quality of the questions in the Multiple-Choice Test (MCT) for the Final Exam of Physics subjects at the Pidie District High School in year 12 MIPA academic year 2018/2019 reviewed based on the validity of the question, reliability of the questions, difficulty level, differentiating level, and function of tricky question. Therefore, from the results of this study we can find out which questions can be maintained, revised or not used in the future. The results of this study are expected to be useful for developing insights for all parties in analyzing student learning outcomes, especially for physics teachers as a reflection and evaluation material in implementing learning outcomes assessment. In addition, the results of this study are also expected to be considered by schools and the government in developing the quality of physics teachers related to the analysis of student learning outcomes.

2. Method
This study used descriptive quantitative non-experimental research. The population in this study was Pidie District Senior High School academic year 2018/2019. This study used three schools as a place of research chosen by stratified random sampling, namely senior high school Unggul Sigli with a total sample of 60 students, senior high school 1 Mila with a total sample of 27 students, and senior high school 1 Muara
Tiga with a total sample of 96 students. The type A questions were chosen by 93 students from all three schools, while the type B questions were chosen by 90 students from the three sample schools. Data collection was gathered from the documentation of all the answers of the students of year 12 high school who took the final exam of the school year 2018/2019. Data analysis included validity test, reliability test, difficulty test, differentiation test, and effectiveness of distractor. The results obtained are then interpreted with the table of each test as follows.

| Table 1. Classification of validity test |
|----------------------------------------|
| **No.** | **Values** | **Criterias** |
| 1       | 0.81 – 1.00 | Very high (very good) |
| 2       | 0.61 – 0.80 | High (good) |
| 3       | 0.41 – 0.60 | Medium (enough) |
| 4       | 0.21 – 0.40 | Low (less) |
| 5       | 0.00 – 0.20 | Very Low (very less) |
| 6       | < 0.00      | Invalid |

| Table 2. Classification of reliability test |
|--------------------------------------------|
| **No.** | **Value of r_{ii}** | **Criterias** |
| 1       | 0.81 – 1.00 | Very high |
| 2       | 0.61 – 0.80 | High |
| 3       | 0.41 – 0.60 | Enough |
| 4       | 0.21 – 0.40 | Low |
| 5       | 0.00 – 0.20 | Very low |

| Table 3. Classification of differentiating level |
|------------------------------------------------|
| **No.** | **Index** | **Criterias** |
| 1       | 0.71 – 1.00 | Very good |
| 2       | 0.41 – 0.70 | Good |
| 3       | 0.21 – 0.40 | Medium |
| 4       | 0.01 – 0.20 | Bad |
| 5       | < 0.0       | Very bad |

| Table 4. Classification of the difficulty level of the questions |
|---------------------------------------------------------------|
| **No.** | **Index** | **Criterias** |
| 1       | 0.00 – 0.30 | Hard |
| 2       | 0.31 – 0.70 | Medium |
| 3       | 0.71 – 1.00 | Easy |

| Table 5. Distracting questions index |
|-------------------------------------|
| **Deceitful index** | **Quality** |
| 76% - 125%           | Very good |
| 51% - 75% or 126% - 150% | Good |
| 26% - 50% or 151% - 175% | Medium |
| 0% - 25% or 176% - 200% | Bad |
| More than 200%       | Very bad |
3. Result and Discussion

3.1. Validity

The validity of a question can be measured using biserial point correlation. The biserial point correlation index ($r_{bis}$) obtained from the calculation results was consulted with $r_{table}$ at a significant level of 5% according to the number of students studied. If $r_{bis} > r_{table}$, the question item is categorized as valid. The validity of the type A question test in the school of which the research was conducted can be explained as follows. The number of students from senior high school Unggul Sigli, senior high school 1 Mila, and senior high school 1 Muara Tiga who took the final school exam with type A question test was 93 students. If consulted with $r_{table}$, the value is 0.202. Question items are said to be valid if the value of $r_{bis}$ is $\geq 0.202$. The results of the analysis show that the type A question test of Physics school final exam academic year 2018/2019 has poor validity. Multiple-Choice Test (MCT) form of school final exam show valid questions totaling 12 questions (34%) and invalid questions by 23 questions (66%).

The validity of the type B question test of research conducted school can be explained as follows. The number of students from senior high school Unggul Sigli, senior high school 1 Mila, and senior high school 1 Muara Tiga who took the final school exam with type B question test was 90 students. If consulted with $r_{table}$, the value is 0.207. Question item is said to be valid if $r_{bis} \geq 0.207$. The results of the analysis show that the type B question test of Physics school final exam academic year 2018/2019 has good validity. Type B question test form show valid questions totaling 21 questions (60%) and invalid questions by 14 questions (40%).

Some of the results of previous studies on the analysis of test validity showed different results that not all test questions given to students were categorized as valid or the validity of the data was not a hundred percent. The research conducted by Anita et al. concluded that 65% of the questions were valid [11]. However, Lili and Raden concluded that the validity of the items showed 28.57% of invalidity, 28.57% were not valid and 42.86% question items were very valid [1]. The results of another study by Alpusari which numbered forty questions item showed that the results of the item validity analysis at 1% coefficient level, valid questions were 16 questions and 24 questions were not valid, while at the 5% coefficient level, 26 questions were valid and 14 questions were not [6].

3.2. Reliability

Reliability means reliable or trustworthy. Reliability of question is calculated using the KR-20 formula for Multiple-Choices Questions. Interpretation of the reliability coefficient ($r_{11}$), that is, if $r_{11} \geq 0.61$ then the questions tested are high and reliable, but if $r_{11} <0.40$ then the questions tested have low reliability and are not reliable. The results of the analysis indicate that the MCQ type A of Final School Exam of Physics subject has $r_{11}$ greater than 0.61 which is equal to 0.703 so that the question has a high reliability or reliable. The results of type B's final school exam analysis also show that the final exam MCQ Physics subject has $r_{11}$ greater than 0.61 which is equal to 0.633 so that the question has a high reliability or reliable.

Some of the results of previous studies have shown different results. The results of the research conducted by Lili and Raden concluded that the analysis of the question items resulted in a test reliability coefficient of 0.79, which means that the questions met the criteria of high reliability [1]. Meanwhile, the results of research conducted by Anita, et al. concluded that the reliability of test questions is 0.65 with a medium category so that questions can be trusted to evaluate students [11].

3.3. Differentiating Level

The criteria for differentiating level are negative (-) means there is no differentiating level, 0.0 - 0.20 categorized as weak differentiating level, 0.20 - 0.39 sufficient differentiating, 0.40 - 0.69 well differentiating level, 0.70 - 1.00 very good differentiating level. Based on the results of the analysis of final school exam question type A, it can be seen that the questions that have negative criteria (-) are 2
questions (5.7%), the questions with weak differentiating level are 32 questions (91.4%), and the question with medium differentiating level merely 1 question (2.9%). Whilst based on the results of the analysis of final exam questions type B, it can be seen that the questions have negative criteria (-) by 3 questions (8.6%), 27 questions of weak differentiating level (77.1%), and 5 questions of enough differentiating level (14.3%).

Some of the results of previous studies showed different results. The research conducted by Anita et al. concluded that the differentiating level has not functioned properly. Where the differentiating level of 30% is categorized as bad, 37.5% is categorized sufficient, 20% is categorized good, 12.5% is categorized as very bad [11]. Whereas according to Lili and Raden concluded that the percentage of differentiating level is known that most items have medium differentiating level [1]. Another study conducted by Alpusari concluded that the results of the analysis of the distinguishing power of questions that get very poor category is one question in question number 20, questions that are categorized as bad were 15, 15 questions were enough, and 9 questions were categorized as good [6].

3.4. Level of difficulty
The level of difficulty of the question is to examine the test questions in terms of difficulties so that questions can be categorized as easy, medium, and difficult. The results showed that there were 30 questions in Type A categorized as difficult (85.7%), 4 questions as medium category (11.4%), and 1 question as easy category (2.9%). While the type B exam test questions from the results of the study showed that difficult category questions were 25 questions (71.4%) and 10 questions categorized as medium (28.6%).

The results of previous studies show different research results. The research conducted by Lili and Raden concluded that the majority (more than 50%) of General Physics Questions items were difficult [1]. While the research conducted by Anita, et al. concluded that quantitatively based on the level of difficulty is not balanced [11]. Meanwhile, the research from Alpusari showed that the difficulty of the questions that were very easy by 17 questions, items that were easy were 9 questions, items that categorized as medium were 11 questions, items that had difficult categories was one question, and items that were very difficult were two questions [6]. Anizam, et al. suggested that the degree of difficulty is not significantly correlated with differentiation at 1% significance level. The level of difficulty is mostly significantly correlated with item validity at a significance level of 1%. Differentiating level is significantly correlated with item validity at 1% significance level. The reliability of a question is higher if the question has more valid items [12].

The difficult questions make students overwhelm to solve the problem. While the easy questions make students do not have the effort to solve or underestimate the matter. These difficult and easy questions need to be followed up, whether they will be revised or will be discarded so that they can be reused in the upcoming exam. While the medium question items can be stored in the question bank so that they can be used on future exams.

3.5. Effectiveness of Tricky Answers
The tricky answers from good items will be chosen equally by students who answer incorrectly. Conversely, the items that are not good enough will be chosen by the students unevenly. The results showed that there were tricky answers functioning well and there were also tricky answers which were less functional in deceiving students so they needed to be revised or replaced. The effectiveness of tricky answers of the final school exam type A for Physics subject academic year 2018/2019 in senior high school Pidie district can be explained as follows. Regarding number 1 and 2, all of the point answers are well. The question number 3 module A, did not work so it needs to be replaced. The question number 4 module B, is less functioning so it needs to be revised. Questions number 5, 6, 7, 8, 9, all of the point was functioning properly. The question number 10, point A, did not work so it needs to be replaced. Questions number 11 and 12 which E as distractor are less functioning so they need to be revised. The question
number 13 for point B and E did not work so they need to be revised. About number 14, all of the distractor points were functioning properly. The question number 15, C distractor did not work so it needs to be revised and point D did not work so it needs to be changed. Question number 16 all of the distractor points were functioning properly. The question number 17 point A did not work so it needs to be changed. The question number 18 point A was less functioning so it needs to be revised. The questions number 19 and 20 all of the distractor points were functioning properly.

The question number 21 point A and E were not functioning so that it needs to be revised. The question number 22 point B and E were less functioning so it needs to be changed. The question number 23 point B was less functioning so it needs to be revised. The number 24, all of points are functioning properly. The questions number 25 and 26 point B and E were less functioning so they need to be revised. Questions number 27 and 28 all of points were functioning properly. The question number 29 point B and D did not work so it needs to be revised. Number 30, all of the points were functioning properly. The question number 31 point C and E were less functioning so it needs to be revised. The questions number 32 and 33 all of the points were functioning properly. The questions number 34 points B and D were less functioning so it needs to be replaced. Problem number 35, all of the points were functioning properly.

The results of the study indicate that the multiple-choice module B has distractors that work well and there was also less functioning that need to be revised or replaced. Question number 1, all of the points did not function properly, so they need to be revised or replaced. The questions number 2 and 3, all of the points were functioning properly. The question number 4 points B and D were not functioning so that it needs to be revised. The questions number 5 and 6 all the distractors function properly. The question number 7 points C and D are less functioning so it needs to be revised. The number question 8, distractor point E did not work so it needs to be revised. The question number 9 points B and E were not functioning so they need to be changed. Questions number 10 and 11 all of the points were functioning properly. The question number 12 point A was less functioning so it needs to be revised. The number 13 question E did not work so it needs to be replaced. The question number 14 points C and E were not working so it needs to be replaced. Problem number 15 all of the points were functioning properly. The question number 16 point E was less functioning so it needs to be revised. The question number 17 point A did not work so it needs to be changed. The number 18 all of the points were functioning properly. The question number 17 point A did not work so it needs to be changed. The question number 19 points A and C did not work properly so they need to be replaced. The questions number 20 and 21 points A and E did not function properly so they need to be replaced.

The question number 22 point B was less functioning so it needs to be revised. The question number 23 point E did not work so it needs to be replaced. The question number 24 point B was less functioning so it needs to be revised. The questions number 25 and 26 point E were less functioning so they need to be revised. Question number 27, all of the points were functioning properly. The number 28 point A was less functioning so it needs to be revised. The number 29 points C and D did not work so it needs to be revised. Number 30, all of the points were functioning properly. The question number 31 point C did not work so it needs to be revised. The questions number 32 and 33 all of the points were functioning properly. The question number 34 points B and E were less functioning so it needs to be revised. The number 35 point D did not work so it needs to be revised.

The results of this study show different from several studies that have been done before. The research conducted by Anita et al. concluded that the effectiveness of 80% distractor functioned properly so that the tricky question effectiveness could be accepted and used [6]. While the research by Septiana suggested that the effectiveness of the class X of biology questions were 40 questions, there were 2 questions including good criteria, 10 questions enough criteria, 18 questions were bad criteria, and 10 questions were very bad criteria, in class XI there were 3 very good criteria, 6 good criteria, 12 criteria were medium, 14 criteria were bad criteria, and 5 were very bad criteria [13]. Whereas the research by Kurniawan found that the effectiveness aspects of the distractor were 11 questions (44%) were effective categories and 14 questions (56%) were ineffective categories [14].
According to Khaerudin, the spread of answer choices is used as the basis for question review. This is intended to find out whether the answer is available or not. An answer choice (which is tricky) can be said to function if the distractor is at least chosen by 5% of the test participants/students, and more is chosen by a group of students who do not understand the material [15].

4. Conclusion

Based on the results of data analysis, it can be concluded that the Multiple-Choice Questions (MCQ) module A has 12 valid questions and 23 invalid questions. While the MCQ module B has 21 valid questions and 14 invalid questions. The reliability of type A is 0.703 and the question has a high level of reliability. While the reliability of type B is 0.633 and the question has a high level of reliability. Differentiating level in module A questions are known that the questions do not have a differentiating level because the problem has negative criteria (-) by 2 questions, questions with weak differentiating level by 32 questions, and 1 question with medium differentiating level. While the differentiating level of module B questions is known that the questions that do not have a differentiating level because the question has negative criteria (-) by 3 questions, 27 questions with weak differentiating level, and 5 questions with enough differentiating level. The module A questions which are included in the difficult category have 30 questions, the questions which are included in the medium category have 4 questions, and 1 question categorized as easy. While the type B questions which are included in the difficult category, there are 25 questions, 10 questions as medium category. The tricky effect of type A and B questions are generally good and a small number (some questions) are problematic, the questions need to be revised or replaced.

Based on the results of this study, it can be suggested a number of suggestions, the teacher needs to improve the quality of the questions based on validity, reliability, level of difficulty, and the power of different test questions so that in making the items that will come better.

References

[1] Maenani Lili O and Raden 2015 Analisis butir fisika ulangan umum kenaikan kelas x madrasah aliyah se-kabupaten bajaranegara, jawa tengah tahun pelajaran 2011/2012 Jurnal Ilmiah Fisika, Pembelajaran dan Aplikasinya 1 5
[2] Arifin Z 2014 Evaluasi Pembelajaran (Bandung:Remaja Rosda Karya)
[3] Puspitasari D 2014 Analisis instrumen tes ulangan akhir semester ganjil mata pelajaran fisika kelas X RSBI kabupaten slamen tahun pelajaran 2012/2013 Jurnal Pendidikan Fisika 2 1
[4] Djaali and Pudji M 2008 Pengukuran Dalam Bidang Pendidikan (Jakarta:Gramedia Wijayaarana Indonesia)
[5] Djemari M 1999 Pengukuran, penilaian, dan evaluasi Makalah disampaikan pada penataran evaluasi pembelajaran matematika SLTP untuk guru inti matematika di MGMP SLTP (Yogyakarta : PPPG Matematika Yogyakarta)
[6] Alpusari M 2014 Analisis butir soal konsep dasar IPA 1 melalui penggunaan program komputer anates versi 4.0 for Windows Jurnal Primary PGSD FKIP Universitas Riau 2 106-115
[7] Nurgiyantoro B 2011 Penilaian Dalam Pembelajaran Bahasa dan Sastra (Yogyakarta: BPFE)
[8] Djiwandono S 2009 Tes Bahasa Pegangan Pengajar Bahasa (Jakarta: Indeks)
[9] Kunandar 2014 Penilaian Autentik (Penilaian Hasil Belajar Peserta Didik Berdasarkan Kurikulum 2013) (Jakarta: Raja Grafindo Persada)
[10] Suyono 2009 Prinsip dan Dasar Evaluasi Pendidikan (Bogor: Ghialia Indonesia)
[11] Anita S T and Zuldafril 2018 Analisis kualitas butir soal fisika kelas x sekolah menengah atas Edukasi:Jurnal Pendidikan 1 35-47
[12] Anizam Z, Muhyiatul F and Rahma N 2013 Hubungan antara validitas butir, reliabilitas, tingkat kesukaran dan daya pembeda soal ujian semester genap bidang studi biologi kelas XI SMA/MA Negeri di kota padang tahun pelajaran 2010/2011 Prosiding Semirata FMIPA Universitas Lampung (Lampung: FMIPA Universitas Lampung) pp 39-47

[13] Septiana N 2016 Analisis butir soal ulangan akhir semester (UAS) biologi tahun pelajaran 2015/2016 kelas X dan XI pada MAN Sampit Jurnal EduSains 2 115-121

[14] Kurniawan T 2015 Analisis butir soal ulangan akhir semester gasal mata pelajaran IPS Sekolah Dasar Journal of Elementary Education 1 1-6

[15] Khaerudin 2017 Administrasi, analisis butir dan kaidah penulisan tes Jurnal Madaniyah 1 97-128