New Multiple Choice Questions in Critical Thinking Assessment: As A Way to Distinguish Between Logical Answers and Guess Answers

Aminatuz Zuhriyah *, Agus Suprijono **, Aminuddin Kasdi ***

Post Graduate School
State University of Surabaya

DOI: 10.29322/IJSRP.9.10.2019.p9422
http://dx.doi.org/10.29322/IJSRP.9.10.2019.p9422

Abstract- This study aims to describe a new model of multiple choice questions as an assessment instrument that is able to measure critical thinking skills and be able to distinguish student answers between answers based on logical reasons and speculative answers. This research uses the development research method according to Borg and Gall and only reaches six stages of development. The results showed that the multiple choice question assessment model accompanied by free argumentation was able to measure students' critical thinking skills and was able to distinguish between logical and speculative answers.

Keyword: *Multiple choice questions, free argumentation, critical thinking skills, logical answers, speculative answers.*

I. INTRODUCTION

Until now it was recognized that the multiple choice question model became the test choice that is widely used in schools in various countries. Educational practitioners have long used MCQs in both the National Examination, Semester Exams, and classroom assessments. That is because multiple choice questions are more quickly managed by a scoring machine and are suitable for large groups of students. MCQs are more effective in the assessment process (Douglas, Wilson, & Ennis (2012). These advantages make MCQs an irreplaceable assessment instrument at almost all levels of education in various countries.

Ease of integration with internet technology has also led to increased use multiple choice questions as a practical and fast tool Computer networks allow more flexibility in sending multiple choice questions without any limitations on distance and time, through sophisticated software, can, speed up the examination of test results, also save costs (Cheryl, 2017) On the basis of economic considerations, most schools rely on multiple choice questions (MCQ) that test the knowledge held while competencies such as critical thinking and skills development are usually not assessed.Behind

some of these strengths the MCQ has many weaknesses related to pedagogical objectives, dianta only about MCQ is not able to uncover higher thought processes, even more often raises answers on the basis of guesses. (Henry L. Roediger III, & Elizabeth J. Marsh, 2005). In addition to the tendency to guess, giving feedback causes students to obtain false knowledge, and if the same question is repeated, students only do a recall on the next question, if there is no feedback after the exam then the actual answer is wrong and is thought to be correct by the student will be wrong knowledge henceforth (Henry L. Roediger III, & Elizabeth J. Marsh, 2005). Learning by providing multiple choice tests results in lower learning performance compared to reviewing the previous material (Hoogerheide, et all, 2017).

The results showed that multiple choice questions on the 12th national history subjects (NAEP) in the US cause three constructive processes namely factual recall, reading comprehension, and irrelevant test taking strategies. The findings also reveal that although questions often encourage students to engage in factual withdrawal, they are often not indicators of student knowledge (Smith, Mark D. 2017). Multiple choice questions (MCQ) often overlook some thinking skills such as critical thinking and skill development (Cheryl A. Melovitz Vasan, et al, 2017).

To reduce the negative effects of MCQ questions, several researchers have developed various types of MCQs. Among the results of developing MCQ questions that have been carried out are the Two Tier Multiple Choice or two-level test-based assessment model. The results explained that the two-level multiple choice test reliably and validly was able to diagnose students' understanding of photosynthesis and respiration in plants (Filocha Haslam & David F. Treagust, 1987). However, this two-tier test was later criticized because it was considered unable to distinguish lack of knowledge from misunderstandings as well as from false positive or false negative scientific conceptions (Sreenivasulu and Subramaniam, 2013). Based on these criticisms some researchers developed the Threetier test model so that it provided more accurate results for assessing students' misconceptions due to lack of knowledge. In other words, through conventional or two-tier multiple choice tests, errors due to lack of knowledge are evaluated as misconceptions, while in three tiers errors are not necessarily misconceptions, it could be due to lack of knowledge (Kirbulut, Zubeyde Demet , Omer Geban, 2014).

Based on this research, it shows that although the three-tiered test seems to eliminate many of the disadvantages mentioned in the two-tiered test, other researchers consider the three-tiered test still unable to completely distinguish the choice of beliefs for the main answer (first level) from the choice of beliefs for reasoning (second level). Therefore a three-tier test is considered to overestimate student grades and underestimate their lack of knowledge (Kaltakci-Gurel, Eryılmaz, and McDermott 2015).

Finally Gurel et al (2017) developed a four-level multiple choice test. In the first stage consists of questions and answer choices as in multiple choice tests in general. In the second stage contains the level of confidence regarding the answers in the first stage. In the third stage contains relevant principles that justify the response in the first stage, while in the fourth stage contains the level of confidence regarding the answers in the third stage (Gurel, Derya Kaltakci, Ali Eryilmaz & Lillian Christie McDermott, 2017). The results showed that the four-tier format for diagnostic instruments showed good utility to explore students 'understanding of reaction kinetics and reveal students' understanding of concepts (Gurel, Derya Kaltakci, Ali Eryilmaz & Lillian Christie McDermott. 2017).

Yoonhee & Marshall (2018) develops multiple choice questions with feedback at the end of the test after students work on the questions with the aim of providing long-term memory (Jang, Yoonhee & Elaine Marshall, 2018). The results showed that two days later, participants who received feedback showed an increase in retention or memory, even participants who were only given treatment in the form of feedback by displaying the correct answers outperformed the control group that did not get feedback, regardless of whether the students' answers true or false (Jang, Yoonhee & Elaine Marshall. 2018). Furthermore, there is also the development of a model that combines Multi-tier multiple choice with direct feedback (Maier, Uwe, NicoleWolf, ChristophRandler. 2016). The results showed that the model that combines Multi-tier multiple choice with direct feedback was proven to improve student learning outcomes (Maier, Uwe, Nicole Wolf, Christoph Randler. 2016).

All of the multiple choice problem development models above have not yet limited to how multiple choice questions are constructed as critical thinking assessment instruments that are able to distinguish student answers between logical answers and speculative answers or on the basis of guesses. Hamper is the same as this research, namely the development of multiple choice questions of HTT (Historical Thinking Test) (Smith, Mark D. 2017). This question model was developed because the national exam questions in the field of history studies conducted in the US were considered unable to reveal the ability to think in the history of 12th grade students. To overcome this problem, Smith (2017) developed an HTT multiple choice question consisting of multiple choice questions accompanied by reasons for answers multiple choice selected. Students' reasons are explained in the form of sound recordings In each choice of answers A, B, C, D the student must state the reasons why he chose or did not choose the available options.

If in the field of history, the ability to think history is an important aspect that must be measured in learning outcomes, then in the field of social studies, critical thinking ability is one of the abilities that must be measured in the process of evaluating learning outcomes. Social studies learning in Indonesia refers to the critical education paradigm so that the expected outcome of the learning process is an increase in students' critical thinking skills, so the assessment must also be able to show indicators of critical thinking skills according to existing theoretical footing. in fact there are still many assessments that have not been able to create appropriate assessment procedures for learning purposes (Ennis, Robert H, 1993). One of the main objectives of the assessment of critical thinking is being able to diagnose students' critical thinking levels (Ennis, Robert H, 1993).

Critical thinking is reasoning that is reasonable and reflective by emphasizing decision making about what to believe or do (Ennis Robert H in Fisher, 2008). Robert H. Ennis in Aristika (2015) states that critical thinkers ideally have 5 critical thinking abilities, namely: 1) Elementary clarification (providing basic explanation) which includes, focusing on questions (can identify questions / problems, can identify possible answers, and what is thought does not come out of the problem), Analyze opinions (can identify conclusions from the problem, can identify reasons, can handle things that are not relevant to the problem), try to clarify an explanation through question and answer. 2) The basis for the decision (determining the basis for decision making) which includes, considering whether the source can be trusted or not, observing and considering an observation report. 3) Inference (drawing conclusions) which includes, deducting and considering the results of deduction, inducing and considering the results of induction, making and determining value judgments. 4) Advanced clarification which includes, defines terms and considers definitions, identifies assumptions. 5) Supposition and integration which includes, considers doubtful reasons or assumptions without including them in our thought assumptions, combines abilities and other characters in decision making. Based on the above background this research wants to explain how multiple choice questions accompanied by free argumentation are able to measure critical thinking skills and are able to distinguish between logical answers and speculative answers?

II. RESEARCH METHODS

The research method used in this study is the R&D (Research and Development) method. According to Borgand Gall (1983, p. 772), educational research and development is a process used to develop and validate educational products. This means that educational development research (R&D) is a process used to develop and validate educational products.

From the ten steps of research and development of the Borg & Gall model, the researcher limits the steps used according to the needs of the researcher. Therefore, the researcher did not use the whole stage because of limited time and money. The following are the steps carried out by researchers in carrying out research. First, Research and Information Collectin. At this stage, the initial research and information gathering is done by interviewing, documenting, and observing the field. Interviews were conducted with IPS teachers in MTs. Al-Hasanuddin to analyze problems that occur during the assessment process. Researchers also conducted interviews with class IX students after they were given the USBN 2017-2018 questions to work on. After working on the questions, the researcher asks students questions about the reasons for the answers they choose. Second, Planning. At this stage the researcher prepares a number of things that are prepared before compiling the questions, namely the lesson plan, material, lattice questions, and assessment indicators. Third, Develop Preliminary of Product. At this stage an initial product draft is prepared. This draft is then consulted with expert lecturers to evaluate and correct if there is a discrepancy so that the assessment instruments arranged can be used to measure students' critical thinking skills. Furthermore, the results of evaluations, corrections, and suggestions from expert lecturers are used as a reference to improve the initial draft before conducting field trials. Fourth, Preliminary Field Testing. After the product is declared valid and feasible, then the assessment instruments can be tested in learning activities. This field trial was conducted in two schools, namely 20 students of grade IX MTs. Miftahul Ulum Kalipang Grati Pasuruan and. 24 students of class IX MTs. Al-Hasanuddin Rebalas Grati Pasuruan. This stage is the final stage of the study. The results of this field trial will produce data on the assessment of students' critical thinking abilities, which are then tested for validity, reliability, level of difficulty, difference power, and the effectiveness of fraud.

III. RESULTS AND DISCUSSION

Research and Information Collection

Before researchers determine what products are suitable to be developed to improve the weaknesses of old products, the data collection process is first carried out in the field. As for the results of the data collection process, it can be concluded that the assessment of learning outcomes in the Semester and USBN questions has several weaknesses related to pedagogical elements, namely: Students are very difficult to achieve the target KKM because the material is too much, the majority of exam questions only measure students' memory and understanding levels on material, students are also not given the opportunity to maintain scores if their answers are wrong. Based on the analysis of semester exam questions and USBN questions in the field of social studies in 2018/2019, more than 90% of the questions only measure students' memories and comprehension, not yet at the C4 and C5 levels which are part of the critical thinking ability indicators. Based on the results of interviews with grade IX students after being given assignments to work on USBN questions, some students had difficulty deciphering their arguments and some answered on the basis of guesswork or guesswork. Students assume value is luck so they are forced to surrender to the value they will get tomorrow.

Planning

This study adopted a question model written by Ennis in his journal, theproblem model *Multiple Choice With Written Justification*. As for preparing the product to be developed, several preparations were carried out, namely the Compilation of the RPP, the Compilation of the Questionnaire, the Compilation of the Guidance for Instruction.

Scoring guidelines are based on 12 indicators of critical thinking. each question consisted of 4 indicators that were given a maximum score of 4 for each indicator. So that in each item, the maximum score is 16. Because the number of questions is 10, then the maximum score for all questions is 160. This assessment is applied to the answers of free argumentation. As for scoring for multiple choice answers, it is almost the same as conventional multiple choice questions, namely score 1 for the correct answer and score 0 for the wrong answer.

Example of extraction

In question number 7, Inayatul Maula answers multiple choice with the wrong answer. But he can give a logical reason to the score produced as follows

Multiple choice: 0

Focus on the question: 4

Give a logical reason: 4

Give a rebuttal: 0

State the reason with the correct, concise, and clear sentence: 3

In order to get the choice score double 0 and the argument score 11. As for the weighting calculation per item is calculated by the formula

multiple choice score:

$$\frac{\text{the number of scores obtained by students}}{\text{Maximum score of all questions}} \text{ X } 25$$

So the score of Inayatul Maula in problem number 7 is = 0

Argumentation score:

$$\frac{\text{the number of scores obtained by students}}{\text{Maximum score of all questions}} \text{ X } 75$$

so that the score of argumentation of Inayatul Maula in problem number 7 is

then the total score of Inayatul Maula on question number 7 is 0 + = 5,156

Develop Preliminary of Product

Product developed in this paper is a multiple choice problem model with free argumentation. The form of the problem consists of multiple choice questions and is given a description column below to write the reason why students choose one answer and not choose another answer.

Problems example.

The establishment of factories in the regions has benefited the community in terms of employment opportunities and improving their standard of living. Since the construction of the factory, many people have moved from their work as farmers to factory workers. Their income has also increased, from the average salary of 50,000 per day to 100,000 per day after being rushed to the factory. From these conditions it can be concluded that ...

a. The existence of a factory is the main hope of the community to improve their standard of living
b. Communities can only prosper if in each region factories are built
c. The community must be empowered so that it does not depend on the presence of industry or factory
d. Indonesian people are unskilled and creative

Arguments
……………………………………………………………………………………………………….
……………………………………………………………………………………………………….
……………………………………………………………………………………………………….
……………………………………………………………………………………………………….
……………………………………………………………………………………………………….

Preliminary Field Testing

At this stage the tests on the first two secondary schools that MTs. Miftahul Ulum Kalipang and MTs. Al-Hasanuddin. The first trial was conducted on 20 students of class IX MTs. Miftahul Ulum Kalipang and the 2nd trial were conducted on 24 students of class IX MTs. Al-Hasanuddin. Based on this trial, an analysis of items was produced as explained below.

Difficulty Level

Based on the results of the initial product trial phase, it is stated that the difficulty level index for all questions is moderate. Difficulty test results on each question are explained in the table below.

Table. 1
Analysis of Levels of Problems Early Stage Problem

Number	1st test		2nd test	
	Level Difficulty		Level Difficulty	
	Index	Interpretation	Index	Interpretation
1	0.64	Medium Question	0.62	Medium Question
2	0.43	Medium Question	0.40	Medium Question
3	0.52	Medium Question	0.65	Medium Question
4	0.49	Medium Question	0.55	Medium Question
5	0.54	Medium Question	0.48	Medium Question
6	0.55	Medium Question	0.48	Medium Question
7	0.59	Medium Question	0.63	Medium Question
8	0.53	Medium Question	0.36	Medium Question
9	0.52	Medium Question	0.34	Medium Question
10	0.58	Medium Question	0.42	Medium Question

Based on these data it can be concluded that all items are classified as moderate because the results of the index of the level of difficulty between 0.3 - 0.7.

Distinguishing Power

Based on the results of the initial stage of the test results generated the problem distinguishing as follows.

Table. 2
Analysis of Power Differences Problem Early Stage

Nomor Soal	1st test		2nd test	
	Power Different		Power Different	
	Index	Interpretation	Index	Interpretation
1	0.43	Power Different Good Enough	0.44	Power Different Good Enough
2	0.45	Power Different Good Enough	0.42	Power Different Good Enough
3	0.43	Power Different Good Enough	0.41	Power Different Good Enough
4	0.41	Power Different Good Enough	0.41	Power Different Good Enough
5	0.52	Power Different Good Enough	0.40	Power Different Good Enough
6	0.49	Power Different Good Enough	0.49	Power Different Good Enough
7	0.44	Power Different Good Enough	0.40	Power Different Good Enough
8	0.53	Power Different Good Enough	0.54	Power Different Good Enough
9	0.43	Power Different Good Enough	0.43	Power Different Good Enough
10	0.48	Power Different Good Enough	0.63	Power Different Good Enough

Based on the above data it can be concluded that all the questions have quite a different power because they have an index of $0.40 \leq DP < 0.70$.

The effectiveness of the deception

Based on the results of the initial stage of the trial it was produced that the deception was classified as an effective category because of all the questions, there were students who chose the wrong answers.

Validity

Based on the results of the initial test, the validity of each item is as follows.

Table. 3
Test Item Validity Test Items for Early Stage Trial

No. Soal	1st test		2nd test	
	Validity		**Validity**	
	Index	**Interpretation**	**Index**	**Interpretation**
1	0.8544567	valid	0.79465	valid
2	0.989526	valid	0.8993584	valid
3	0.8850104	valid	0.8767226	valid
4	0.9116114	valid	0.824236	valid
5	0.8961678	valid	0.7679653	valid
6	0.9067636	valid	0.7619441	valid
7	0.8840218	valid	0.7823454	valid
8	0.9195941	valid	0.841298	valid
9	0.9051803	valid	0.7927809	valid
10	0.9010608	valid	0.8851582	valid

Based on the data above, all the questions are declared valid because the results of the validity test show that the r count is greater than the r table. large r table for the number of samples 20 of 0.444.

Reliability

Based on the test results about the early stages of reliability problems generated by 0.97445. That shows that the problem was declared reliable because the reliability coefficient (r11)> 0.6.

From the whole item test, starting from the level of difficulty, distinguishing power, deception effectiveness, validity, and reliability of the questions, all the questions are said to be appropriate because they meet the eligibility criteria of the questions.

IV. CONCLUSION

Based on the results of the research that has been presented previously, there are several important things that can be concluded from the research problem, namely:

MCQs are always an alternative model of questions used in the exam, even though there are some problems in the assessment process related to pedagogical interests, including multiple choice questions it is very possible for children to answer on the basis of speculation or coincidence, multiple choice questions have not been able to measure some of the expected thinking skills, one of which is critical thinking skills.

The construction of questions arranged in exam questions measures more low-level cognitive levels, especially at C1 and C2 levels, namely the level of remembering and understanding. The

development of multiple choice questions accompanied by free arguments becomes alternative questions that can reduce the effect of guessing answers and be able to measure students' critical thinking skills.

Based on the results of the product trial, multiple choice questions accompanied by free arguments have been feasible to be used at a later stage, namely trials on a larger scale, because they meet the criteria of the level of difficulty, differentiation, effectiveness of deception, validity, and reliability of good questions and have gone through validation by two expert lecturers who were declared good.

V. REFERENCES

Henry L. Roediger III,& Elizabeth J. Marsh. (2005). *The Positive and Negative Consequences of Multiple-Choice Testing. Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol 31,5, 1155-1159. http://dx.doi.org/10.1037/0278-7393.31.5.1155

Hoogerheide, Vincent, Justine Staal , Lydia Schaap , Tamara van Gog, Hoogerheide, V. (2017). *Effects of study intention and generating multiple choice questions on expository text retention Learning and Instruction*. https://doi.org/10.1016/j.learninstruc.2017.12.006

Smith, Mark D. (2017). *Cognitive Validity: Can Multiple-Choice Items Tap Historical Thinking Processes?.American Educational Research*. (54): 1256-1287. https://doi.org/10.3102/0002831217717949

Cheryl A. Melovitz Vasan. David O. DeFouw. Bart K. Holland. Nagaswami S. Vasan. (2017). *Analysis of Testing with Multiple Choice Versus Open-Ended Questions: Outcome-Based Observations in an Anatomy Course*.

Anatomical Sciences Education (11): 254–261https://doi.org/10.1002/ase.1739

Filocha Haslam & David F. Treagust (1987): *Diagnosing secondary students' misconceptions of photosynthesis and respiration in plants using a two-tier multiple choice instrument,* Journal of Biological Education, 21:3, 203-211. http://dx.doi.org/10.1080/00219266.1987.9654897

Sreenivasulu, B., and R. Subramaniam. (2013). *"University Students' Understanding of Chemical Thermodynamics."* International Journal of Science Education 35 (4): 601–635

Kirbulut, Zubeyde Demet , Omer Geban. (2014). *Using Three-Tier Diagnostic Test to Assess Students' Misconceptions of States of Matter*. Eurasia Journal of Mathematics, Science & Technology Education. 10(5), 509-521. DOI: https://doi.org/10.12973/eurasia.2014.1128a

Gurel, Derya Kaltakci, Ali Eryilmaz & Lillian Christie McDermott. (2017). *Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics,* Research in Science & Technological Education, DOI: 10.1080/02635143.2017.1310094

Jang, Yoonhee & Elaine Marshall . (2018). *The Effect of Type of Feedback in Multiple-Choice Testing on Long-Term Retention.* The Journal of General Psychology. DOI:10.1080/00221309.2018.1437021

Maier, Uwe, NicoleWolf, ChristophRandler. (2016). *Effects Of A Computer-Assisted Formative Assessment Intervention Based On Multiple-Tier Diagnostic Items And Different Feedback Types*. Computers & Education**. (**95**):** 85-98. https://doi.org/10.1016/j.compedu.2015.12.002

Smith, Mark D. (2017). *Cognitive Validity: Can Multiple-Choice Items Tap Historical Thinking Processes?.American Educational Research*. (54): 1256-1287. https://doi.org/10.3102/0002831217717949

Ennis, Robert H. (1993). *Critical Thinking Assessment.Theory Into Practice*, 32:3, 179-186. DOI: 10.1080/00405849309543594

Fisher, Alec. (2008). *Berpikir Kritis: Sebuah Pengantar*. Jakarta: Erlangga

Aristika, Ayu. (2015). *Tinjauan Tentang Pembelajaran Berbasis Masalah Dalam Meningkatkan Kemampuan Berpikir Kritis Dan Disposisi Matematis*. Seminar Nasional Matematika Dan Pendidikan Matematika UNY.

AUTHORS

First Author – Aminatuz Zuhriyah, Post Graduate Student, State University of Surabaya, Indonesia, amipasuruan@gmail.com
Second Author – Agus Suprijono, Lecturer, Post Graduate School, State University of Surabaya, Indonesia, agussuprijono@unesa.ac.id
Third Author – Aminuddin Kasdi, Lecturer, Post Graduate School, State University of Surabaya, Indonesia, aminuddinkasdi@unesa.ac.id

Corespondence Author – Aminatuz Zuhriyah, State University Of Surabaya, Indonesia, amipasuruan@gmail.com, +6281555880399 725836809