Development of Assessment Instruments In Measuring Critical Thinking Skills of Senior High School Participants of Biology Subject

S. Hidayat¹, E. N. Fadillah²
Universitas Muhammadiyah Palembang, South Sumatra, Indonesia

Corresponding Author: saleh_hidayat@um-palembang.ac.id

Abstract. The purposes of this study are to determine the instrument assessment appropriateness in measuring critical thinking skills and understand the level of critical thinking skills at 10th grade senior high school participants of biology subjects at Palembang. This study is about research and development (R & D). The instrument of development model used is a 4-D development model from Thiagarajan, namely; define, design, develop, and disseminate. Aiken’s V qualitative and quantitative analysis by using the QUEST program was used in this study. The results showed that; 1) Biological assessment instruments used in developing critical thinking skills of senior high school participants of the 10th class of biology subject are worth to be reviewed in terms of test standard characteristics; and 2) the percentage of critical thinking skills mastery level on interpretation indicators: 46.78%; 44.86% analysis; 41.15% evaluation; 49.12% inference; 47.50% explanation; and self-regulation 54.27%.

Keywords: assessment instruments, critical thinking skills, biology

1. Introduction
Learning is an activity that has an important role in manifesting the quality of education. It means that the role of teachers is very important in improving the quality of education. The success of learning is often associated with three things that teachers should do, namely preparation, implementation, and assessment. Curriculum changes from 2006 to the 2013 Curriculum have resulted in changes in several National Education Standards (NES), namely the Graduates Standards Competence (GSC), Content Standards, Process Standards, and Assessment Standards.

Appropriate assessment can produce accurate learning success data and be used as a basis for determining the follow-up. According to Minister of National Education Regulation No. 20 of 2007 about education assessment, assessment system is the process of gathering and processing information to determine the achievement of learning outcomes of students [1]. In line with [2] who stated that assessment is a statement based on a number of facts to explain the characteristics of a person or something. [3] stated that assessment is some procedures that are systematic to collect information and used for making conclusions about the characteristics of a person or object. Furthermore, [4] explained that assessment means evaluating something; take a decision on a matter based on the meaning of good or bad, clever or stupid, etc. [4]. Based on some of these definitions, assessment is all activities carried out by educators and students to assess themselves with the aim that learning and learning activities become better.
Assessment results are not identical with the cognitive learning outcomes, but also, it is the skills which can support learning from students. The 2013 curriculum requires teachers to be able to develop skills possessed by students. One skill that can be assessed from student activities is thinking skills. Critical thinking skills are very important to be mastered by students because they can be used as benchmarks for developing students' skills in carrying out an activity to find a concept. The peak of critical thinking is being able to make decisions which work in the brain called as the ability to think convergently because students must determine what the best one [5]. It means that this skill can train students to be responsive in solving a problem based on their knowledge.

Biology subject is one of the subjects in which there are many concepts that are not just memorized but need to be understood and even found. Finding the concept by itself means that solving a problem while carrying out a series of activities that train critical thinking skills. According to [6], the critical thinking indicator used in this study consists of six indicators, namely interpretation, analysis, evaluation, inference, explanation, and self-regulation. The use of Facione's indicators is because it is the latest indicator to measure critical thinking skills, also it has undergone refinement and rearrangement of the order. This indicator explains each sub-indicator in a simpler way using nouns and examples to make it easy to apply in an educational environment.

Based on the description of the problems that have been described, the assessment instrument product developed, it is expected to give a contribution to the education, especially in the evaluation process of biology learning. The aim is to make it easier for teachers to measure students' critical thinking skills and their mastery. The formulations of the problem of this research are: 1) is the assessment instrument developed worth or qualified according to its assessment aspects? and 2) how big the level of critical thinking skills mastery grade X high school students in biology subjects? This research attempts to answer the problems that have been set.

2. Research Methods
2.1. Types of Research
This is a Research and Development (R & D) study. The purpose of this study is to develop an assessment of the instrument in order to measure the critical thinking skills of the students of High School students in biology subjects in Palembang. The instrument development model used in this study is the 4-D development model proposed by [7], namely define, design, develop, and disseminate.

2.2. Research Subject
The subjects of this study were the class X high school students in even semester in Palembang. The technique of choosing the sample is using purposive sampling technique, namely that the schools are state and private, have an A and B accreditation, have implemented the 2013 curriculum, and has measured critical thinking skills when learning biology. There were 28 high schools with 1780 students as the subjects in this research.

2.3. Procedure
The procedure for developing this assessment instrument includes four stages, namely: define, design, develop, and disseminate.

2.3.1. Define
This stage aims to establish and define an assessment needs of students' skills. This stage is carried out by interviewing and distributing questionnaires to the biology teachers of high schools who teach in class X.

2.3.2. Design
This stage is the design of assessment instrument models and development procedures. At the stage determined KI, KD, the material, describes the indicators of critical thinking skills, designs the question compilation grid, the question script, and the assessment rubric.
2.3.3. Develop
This stage is to develop the instrument as a product. The product developed is in the form of a question compilation lattice, a multiple choice question script accompanied by reasons for giving answers, and an assessment rubric. At this stage, product validation is carried out by the validator, which includes evaluation expert lecturers, material experts, and linguists as well as teachers as implementers of learning. After validation by an expert lecturer, then the product was revised based on the results of validation (revision I). Then validated by the teacher, revised according to the results of the validation (revision II). Then proceed with limited trials, product revisions based on limited trials, and trials on a wide scale.

2.3.4. Disseminate
The disseminate stage is the dissemination of the final product development to the entire population.

2.4. Data, Instruments, and Data Collection Techniques
Data collection techniques in the form of tests and non-tests. Non-test techniques are carried out through interviews and questionnaires, while the technique of test is carried out by using assessment instrument products. The questionnaire used was a question assessment questionnaire by evaluation expert lecturers, material expert lecturers, linguists, and teachers. Questionnaire for evaluation expert lecturers, material expert lecturers, language expert lecturers, and teachers is a closed questionnaire. Interviews and questionnaires are used to find out the information about how teachers assess the critical thinking skills of students conducted so far. The results of this interview and questionnaire have used the assessment as the initial step of the study.

The assessment instrument product is to develop the question script in the form of reasoned multiple choice questions and assessment rubrics and their grids. This product was theoretically validated in advance by three validators and seven biology subject teachers, then empirically validated in a limited trial. The next stage of this problem product is used to measure the mastery of critical thinking skills of high school students.

2.5. Data analysis technique
The initial draft that was validated produced qualitative data. The data is in the form of input and suggestions on the assessment instrument from the validator. Inputs and suggestions from the validator are integrated and used to revise the initial draft that has been given. Assessment results from material experts, evaluation experts, linguists, and teachers were analyzed using the formula from Aiken’s V [8]. The purpose of this analysis is to determine the feasibility of the content of the assessment instrument.

\[ V = \frac{\sum s}{n(c - 1)} \]

Information:
V = validity;
\( s = r - \text{lo} \);
\( \text{lo} = \) lowest validity assessment number;
\( c = \) highest validity assessment number;
\( r = \) number given by an appraiser.

The V value will be interpreted in the range between 0.00—1.00 as the content validity coefficient is good or not good; this value as a measure of whether or not to support the content validity as a whole [9]. To interpret the value of the content validity obtained from these calculations, the classification of validity is used as shown in Table 1.
Table 1. Validity Criteria

| No. | Validity Result | Validity Criteria |
|-----|-----------------|-------------------|
| 1   | 0.81 < V < 1.00 | Very high         |
| 2   | 0.61 < V < 0.80 | High              |
| 3   | 0.41 < V < 0.60 | Enough            |
| 4   | 0.21 < V < 0.40 | Low               |
| 5   | 0.00 < V < 0.20 | Very Low          |

Mastery of critical thinking skills is divided into five categories, namely very low, low, medium, high, and very high. This category as shown in Table 2 [8].

Table 2. Categories of Mastery of Critical Thinking Skills

| No. | Category | Score Range (%) |
|-----|----------|-----------------|
| 1   | Very Low | X ≤ 24.95       |
| 2   | Low      | 24.95 < X ≤ 41.65 |
| 3   | Medium   | 41.65 < X ≤ 58.35 |
| 4   | High     | 58.35 < X ≤ 75.05 |
| 5   | Very High| 75.05 < X       |

Furthermore, the calculation of the percentage of mastery indicators of critical thinking skills was obtained by means of the average score on critical thinking skills that were answered correctly by students with the number of items. If written mathematically are as follows.

\[ NP = \frac{R}{SM} \times 100 \]

Information:
- **NP** = percentage value per indicator of critical thinking skills
- **R** = score obtained on indicators of critical thinking skills
- **SM** = maximum score on indicators of critical thinking skills

This quantitative analysis was done right after the questions were tested by using the QUEST program. Quantitative analysis is used in order to see the internal characteristics such item suitability, problem reliability, and level of difficulty.

2.6. Match Instrument Item

According to [5], in one item or test/case/person is declared as fit with the model with the limitation prediction of INFIT MNSQ from 0.77 to 1.30 [5]. Thus, an item is not fit according to the Rasch model if it has a value of < -2.0 or > +2.0 (probability or opportunity < 0.05).

2.7. Reliability

The results of the QUEST program analysis also show the estimated reliability of tests based on items. From the results of the QUEST program analysis, the estimated reliability value according to the Item Response Theory (IRT) if calculated based on the item is called the item spatial index [5]. If based on the test (case/person) is called the person separation index. The higher the estimation of the item spatial index, the more precisely the whole item is analyzed according to the model used, this is in accordance with the Rasch Model. The higher the person separation index, the more consistent each measuring item is used to measure the test concerned. Reliability estimates based on tests are Cronbach's alpha reliability for polytomous data. According to Wright & Master, the item spatial index (item separation index) is called the sample reliability, while the person separation index is called test reliability.

2.8. Difficulty

The second characteristic is the level of difficulty or difficulty index. Difficulty indexes or difficulty levels are obtained using the QUEST program. The polytomous scale has an x score of 0, 1, 2, 3, ..., m.
Question items are said to be good if the difficulty index is more than -2.00 or less than 2.00. A difficulty is the mean (difficulty) index of difficulty on score 0, score 1, score 2, and score 3 on each item [5].

3. Results and Discussion

3.1. Development Results

The result of the development is the achievements that have been obtained after completing the research procedure. Products developed in the form of assessment instruments that aim to measure the critical thinking skills of high school students in biology subjects. Description of the results of the development of the assessment instrument as follows.

3.1.1. Define

The product of the assessment instrument developed consists of a question grid, a multiple choice question script, and an assessment rubric. The test developed in the form of a multiple choice reasoned with the number of items as many as 22 items.

3.1.2. Design

This stage is the stage of designing an assessment instrument. At this stage the questioning lattice is produced, the initial question text, and the assessment rubric. The formulation of the questions was designed with identity, main competence, basic competence, learning indicators, indicators of critical thinking skills, number questions and cognitive domains.

3.1.3. Develop

The development phase is the stage of developing the assessment of instrument products that have been planned in advance to be prepared as products that were ready to be tested. The steps taken in the development phase of this assessment instrument include validation by expert lecturers and teachers.

Validation of Expert Lecturers

Biological assessment instrument products are developed based on a grid that refers to the main competence, existing basic competence, and indicators, validated by expert lecturers to determine the feasibility of Biological assessment instrument products before being validated by the teacher as the implementer of learning. This biology assessment instrument is in the form of reasoned multiple choice questions. This problem consists of 22 questions. The validation of the biology assessment instrument was carried out by evaluation expert lecturers, material expert lecturers, and language expert lecturers. Evaluator expert validators decide 13 valid items without revisions and 9 valid questions with revisions, validator material experts state 16 valid questions without revisions and 6 valid questions with revisions, while linguist validators state 17 valid questions without revisions and 5 valid questions with revisions. All valid items without revisions can be used directly in limited trials. Valid items that are valid with revisions before being used in limited trials, fixed first. The revision of the three expert validators was then corrected by taking into account the input and suggestions on the validation sheet. All items that have been revised are based on the input and suggestions of the validator, then rearranged so that the products of the assessment instruments are ready to be tested in limited ways.

The results of the validation of the assessment instruments from the expert lecturers indicate that there are several components that require further improvement. After that, it was analyzed using Aiken's V formula with the help of the Microsoft Excel program, it obtained an average value of 0.80 which means it is valid. So, the overall assessment of the validator can be concluded as valid or in other words feasible.

Teacher Validation

Further validation is carried out by seven teachers as the implementer of learning. Validation results are divided into valid without revisions, valid with revisions, and invalid. The question consists of 22 items, representing each indicator of critical thinking skills. Some questions have to input, and suggestions
used to revise assessment instrument products. The results of the validation sheet were analyzed using the Aiken's V formula with the help of the Microsoft Excel program, the average value was 0.92 which means it is valid.

**Limited Trial**

The follow-up to product revision by expert lecturers and teachers is a limited trial. The limited trial was conducted at SMAN 1 Palembang, SMAN 4 Palembang, SMAN 20 Palembang, SMA Muhammadiyah 4, SMA Pembina, SMA Nahdatul Ulama, and SMA Xaverius with the sample of 435 students. This limited trial was conducted to determine the goodness of fit, reliability, and level of difficulty of the assessment instruments developed.

The information about the characteristics of the questions was obtained from the analysis using the QUEST program. The trial was carried out utilizing biology lesson time, namely 2 hours of biology lessons with a total time required of 80 minutes, scoring was done with polytomous.

**A Goodness of Fit in the Limited Trial**

The examination of a test as a whole of the items by using the QUEST program. The overall test fit test was developed by Adam & Khoo based on the mean value of INFIT Mean of Square (Mean INFIT MNSQ) and its default deviation or observing the INFIT t (Mean INFIT t) average value along with its default deviation [5]. If the average INFIT MNSQ is around 1.00 and the standard deviation is 0.00 or the average INFIT t approaches 0.00 and the standard deviation is 1.00, the overall test is fit with the Partial Credited Models 1-Parameter Logic (PCM 1-PL). Based on the calculation, the mean value of INFIT MNSQ is 1.01 and the standard deviation is 0.16 (about 0.00), then the overall fittest with PCM 1-PL.

The fittest for determine of each item on the model follows the rules of Adam & Khoo, namely an item fit to the model if the MNSQ INFIT value is between 0.77 to 1.30. The acceptance threshold of items uses INFIT MNSQ or fit according to the model (between 0.77 to 1.30) and using INFIT t with a limit of -2.00 to 2.00, then the points that match the goodness of fit are obtained. The INFIT MNSQ value has a range from 0.78 to 1.26. Based on the item acceptance limit using MNSQ INFIT or fit according to the model, all 22 items are fit [5].

**Test Reliability in the Trial**

In addition to testing the compatibility, the QUEST program output also displays the estimated reliability of test instrument devices. Based on the results of the analysis with the QUEST program, the reliability of the test was based on the item (item spatial index was 0.84 which was classified as the very high category, while the reliability of the test based on the test (person separation index) was 0.73 classified as high [10].

**Difficulty**

Based on the results of the analysis with a QUEST, the difficulty level of the items lies in the range from -1.60 to 1.16 with an average of 0.00 and a standard deviation of 0.60. The item is said to be good if the difficulty value is more than -2.00 or less than 2.00. The average difficulty value of the critical thinking skills test device used in the trial stage is 0.00 ± 0.60 with the medium category. So, based on the difficulty value of 22 items, there is 1 problem with easy criteria, 16 questions with medium criteria and 5 questions with difficult criteria. The distribution of the level of difficulty of the items can be seen in Table 3.

| No. | Criteria | Total |
|-----|----------|-------|
| 1   | Easy     | 1     |
| 2   | Medium   | 16    |
| 3   | Hard     | 5     |
|     | Number of Question | 22    |
Based on the data in Table 3, it can be obtained that the assessment instrument products received from the trial were limited to 22 items, meaning that all items were received. The questions received are used as the final product in the assessment instrument before being used at the disseminated stage.

3.1.4. Disseminate

After the product of the assessment instrument for critical thinking skills that had been tested in a limited number was revised and obtained multiple choice multiple-choice questions in a limited test of 22 items, meaning that no items were rejected or dropped but the items could be tested in the field entirely.

Disseminate phase involves 21 public and private high schools with accreditation of schools A and B in the city of Palembang. Based on the results of the disseminate stage in the field, the analysis of mastery of critical thinking skills is done by finding the average of percentage value of the correct answers in each type of critical thinking skills that arise. The results of the analysis of mastery of critical thinking skills at the disseminated stage are listed in Table 4.

| No. | Critical Thinking Skills Indicators | Mastery Percentage | Category |
|-----|-------------------------------------|--------------------|----------|
| 1.  | Interpretation                      | 46.78%             | Medium   |
| 2.  | Analysis                            | 44.86%             | Medium   |
| 3.  | Evaluation                          | 41.15%             | Low      |
| 4.  | Inference                           | 49.12%             | Medium   |
| 5.  | Explanation                         | 47.50%             | Medium   |
| 6.  | Self-Regulation                     | 54.27%             | Medium   |

From Table 4 it can be seen that the highest mastery of critical thinking skills in the self-regulation category is 54.27% and the lowest indicator is the evaluation. Low evaluation indicators because the students have not guided to check and re-select the questions that have been made to suit their problems, and students have not guided to assess the quality of the arguments made. Furthermore, most of the teachers stated that evaluation skills were low because students did not understand the material, memorized it more often so that they could easily forget the learning material, the spirit of learning was low, and they were not used to working on questions that involved high-level thinking skills. Most students experience difficulties in remembering the material taught at school, students also tend to be more attached to the method of learning to memorize than understanding which makes them forget the material quickly [11] they memorized it. The low ability is to evaluate it because the students are not accustomed to blaming and justifying the results of problem-solving, so they were still awkward and was not confident in front of the teacher [12]. The low ability of students in evaluating is caused by them, not being able to provide problem-solving properly [13].

In the self-regulation indicator, it is in the high category, namely 54.27%. Based on learning activities, self-regulation has been carried out well. Furthermore, the teacher expressed good self-regulation because they were confident in the concept of learning and were able to apply it with arguments and solutions from themselves. Students who have sufficient self-regulation skills include goal setting, planning, motivation (self-efficacy), attention control, use of learning strategies or learning methods, self-monitoring, seeking help, and self-evaluation [14]. Meanwhile, a student who has achieved a high level of self-regulation is a successful student [15]. Students with a good self-regulation, more often used strategies to regulate their own learning rather than students who are considered incompetent. Students who have good self-regulation cannot be separated from the role of the teacher in regulating the learning process, interactions that occur in the classroom, and teaching material delivered as an external motivation of students [16].
4. Conclusions
From the results of the analysis and discussion can be put forward some conclusions as follows: 1) biological assessment instruments to develop critical thinking skills of class X high school students in biology subjects are said to be feasible in terms of the characteristics of test standards. The assessment from the expert lecturer obtained the Aiken’s V content validity of 0.89. Based on the results of the empirical trial obtained as many as 22 items about the fittest with the value of goodness of fit has a range from 0.78 to 1.26. The reliability of the question obtained a value of 0.84 with a good category and the level of difficulty with the range between -1.60 to 1.16; and 2) Percentage of mastery level of critical thinking skills on interpretation indicators 46.78% (medium), 44.86% (medium) analysis, 41.15% evaluation (low), 49.12% Inference (medium), 47 explorations, 50% (medium), and self-regulation 54.27% (medium).

5. References
[1] Mendiknas, Peraturan Menteri Pendidikan Nasional No. 20 Tahun 2007 tentang Standar Penilaian, Jakarta: Kementerian Pendidikan Nasional, 2007.
[2] E. Kurniawan dan E. Mutaqimah, Penilaian, Jakarta: Depdiknas, 2009.
[3] C. R. Reynolds, R. B. Livingston dan V. Willson, Measurement and assessment in education (2nd ed.), Upper Saddle River: Pearson, 2010.
[4] A. Sudjiono, Pengantar evaluasi pendidikan, Jakarta: Rajawali Pers, 2011.
[5] B. Subali dan P. Suyata, Pengembangan item tes konvergen dan divergen dan penyelidikannya secara empiris, Yogyakarta: Diandra, 2012.
[6] P. A. Facione, “Critical Thinking: What It Is and Why It Counts,” Insight Assessment, pp. 1-28, 2013.
[7] S. Thiagarajan, D. S. Semmel dan M. I. Semmel, Instructional development for training teachers of exceptional children, Minneapolis: University Of Minneasota, 1974.
[8] S. Azwar, Penyusunan skala psikologi, Yogyakarta: Pustaka Pelajar, 2014.
[9] L. R. Aiken, “Three coefficients for analyzing the reliability and validity of ratings,” Educational and Psychological Measurement, pp. 132-142, 1985.
[10] B. Sumintono dan W. Widhiarso, Aplikasi model RASCH untuk penelitian ilmu-ilmu sosial, Cimahi: Trim Komunikata Publishing House, 2014.
[11] Susisolwati, Sajidan dan M. Ramli, “Analisis Keterampilan Berpikir Kritis Siswa Madrasah Aliyah Negeri di Kabupaten Magetan,” dalam Seminar Nasional Pendidikan Sains, Surakarta, 2017.
[12] Pardjono dan Wardaya, “Peningkatan Kemampuan Analisis, Sintesis, dan Evaluasi melalui Pembelajaran Problem Solving,” Cakrawala Pendidikan, pp. 257-269, 2009.
[13] M. Hayudiyan, M. Arif dan M. Risnasari, “Identifikasi Kemampuan Berpikir Kritis Siswa Kelas X TKJ Ditinjau dari Kemampuan Awal dan Jenis Kelamin Siswa di SMKN 1 Kamal,” Jurnal Ilmiah Edutic, vol. 4, no. 1, pp. 20-27, 2017.
[14] D. S. Ahmar, “Hubungan antara Regulasi Diri dengan Kemampuan Berpikir Kreatif dalam Kimia Peserta Didik Kelas XI IPA Se-Kabupaten Takalar,” Jurnal Sainsmat, vol. 5, no. 1, pp. 7-23, 2016.
[15] J. Jakesovaan dan J. Kalendah, “Self-regulated Learning: Critical-realistic Conceptualization,” Procedia - Social and Behavioral Sciences, p. 178–189, 2015.
[16] A. Mahmood, “Testing The Effectiveness of a Critical Thinking Skills Intervention for Initial Teacher Education Students in Pakistan,” University of Southamton, 2017.