Proceeding Biology Education Conference p-ISSN:2528-5742
Vol. 17, No.1
Page:112-119 July2021

A Survey on the SETS-Based Human Anatomy and Physiology Course: Analysis of Instruments for Assessing Critical Thinking Skills Using Multimedia

Nuril Hidayati[1, a)], Fariha Irmawati[1,b)]

1Department of Biology Education, Faculty of Exact and Sport Science Education, IKIP Budi Utomo, Jl. Simpang Arjuno14B, Malang,East Java, 65119, Indonesia
2Department of Physical Health Education and Recreation, Faculty of Exact and Sport Science Education, IKIP Budi Utomo, Jl. Simpang Arjuno14B, Malang,East Java, 65119, Indonesia

[a)]Corresponding author: hidayatinuril20@gmail.com
[b)]farizha99@gmail.com

Abstract. Learning media must be followed the curriculum, accommodate learning activities, and provide an appropriate evaluation. The research objective was to determine the validity and reliability of the instrument for assessing the effectiveness of multimedia to promote critical thinking skills. A research method is a quantitative approach with a survey design. The study was conducted on 26 Biology Education students of IKIP Budi Utomosamples that were taken at least 20% of the population, 26 participants were taken more than 20% of a total of 82 students. The research instrument was a validation sheet for critical thinking skills assessment and 15 multiple choicesto test the understanding of the cardiovascular system. The validity and reliability data were analyzed using Anates Software. The results showed that the validity of the test instrument met the feasible criteria. The reliability value of the test instrument is 0.65 with a standard deviation of 2.84 which meets the high-reliability criteria. The distinctive power of the test instrument consisted of 20% poor, 26% criteriasufficient, and 53,33% very good. The level of difficulty of the test was 1 question categorized asvery difficult, 7 difficult, 5 moderate, and 2 questions werevery easy. Based on these results, the test instrument to measure students' critical thinking skills can be used as a reliable measuring tool in SETS-based multimedia in the human anatomy course.

Keywords: Virus concept, High school, Biology education.

INTRODUCTION

The development of science and technology will have an impact on education today. Students are required to understand concepts and be able to apply the knowledge they have learned to solve problems. Students are required to find and choose the correct and current sources of knowledge.(Zubaidah, 2016). Current education must be able to focus on helping students learn how to learn so that they can see opportunities in information, technology, work, and social conditions(Barron & Chen, 2008).

Critical thinking skills cannot be developed if learning is still focused on the lecturer. Lecturers must have an understanding that each student can learn actively and find concepts independently and provide flexibility for students to solve problems from various points of view to find a solution (Chee et al., 2009). It needs a special understanding for lecturers to be able to create learning that can build students' critical thinking skills. The results of research on teachers' perceptions of students' critical thinkingshowed that many teachers found it difficult to develop critical thinking skills in the learning that they were carrying (Chee et al., 2009).Critical thinking skills can be developed through innovative learning to understand problems and solve them logically(Walid et al., 2019). The

results of observations in learning the human physiology anatomy subject showed that at the time of evaluation, the research instrument had not been analyzed by the provisions to produce the right measuring instrument.

Innovative learning activities are inseparable from several components such as learning tools and supporting facilities. Learning tools include semester learningplans, evaluation instruments, and teaching materials. Forms of teaching materials that are relevant to education today do not only contain the material being taught but must meet the aspects of teaching as a complete set of learning tools (Hidayati & Irmawati, 2019).However, some teaching materials that have been developed still do not pay attention to the evaluation aspect or teaching materials as a measuring tool for the assessed aspects as seen in the test instruments that have not been tested first. Several teaching materials have been developed to build critical thinking skills, for example in the development of teaching materials to improve critical thinking skills(Mujiyati et al., 2019)and(Susilowati et al., 2018)which is not equipped with an analysis of the test instruments used. Several other studies examined the test instrument and the feasibility of questions as an evaluation tool(Fatimah et al., 2016)and(Hamdi et al., 2018)but were still not equipped with supporting material for learning activities. Based on the description above, this study aims to determine the results of the test instrument analysis to measure critical thinking skills used in multimedia.The instrument developed in this study is following the indicators of critical thinking skills and the needs of the human anatomy and physiology course.

METHOD

The research followed the quantitative approaches with the survey design aims to validate the instruments. The research subjects were 26 students in the Biology Education study program. The sampling technique uses random techniques.The population in this study was 82 students and the sample was taken at least 20% of the total population. 26 students involved in this study more than 20%.The research instrument used was a validation sheet and instrument to measure critical thinking skills consisted of 15 multiple-choice questions.The instrument used to measure critical thinking skills consists of 15 multiple choice questions Validity and reliability data were analyzed using Anatest and SPSS software. Anates is used to analyze multiple-choice and essay tests without calculating or formulating. Anates can produce outputs in the form of validity, reliability, level of difficulty, discriminatory questions, and distractors in multiple-choice tests.The results of the reliability, validity, difficulty level, and different power were analyzed descriptively. The qualityfor the content validity was determined based on five categories as follows: 1 - 1.5 very nonvalid, 1.6 - 2.5 nonvalid, 2.6 - 3.5 less valid, 3.6 - 4.0 quite valid, and 4,1 - 5 valid (Ihsan, 2015).

RESULT AND DISCUSSION

The data obtained from the research were the validation of content and constructs from experts, namely the Educational Evaluation Lecturer from the Biology Education Study Program, and the data from the analysis of the questions that had been tested on 26 Biology Education Study Program students who had taken the Anatomy Physiology course. The research data obtained are in the form of content and construct validity from experts, evaluation lecturers, and data from the analysis of questions that have been tested on 26 Biology Education Study Program students who have taken the Anatomy Physiology course. Human. The results of the construct validity can be seen in Table 1.

Table 1. Results of Test Instrument Construct Validation

Indicators on Subcourse Learning Outcomes	Critical Thinking Skill Indicator	Question Number
Identify the structure and function of the blood, blood vessels, and heart (Subcourse LO 1)	Evaluation (CTS3)	1
	Analysis (CTS 2)	2
	Interpretation (CTS 1)	3
	Explanation (CTS 5)	4
	Analysis (CTS 2)	5
	Inference (CTS 4)	6
	Analysis (CTS 2)	7
Comparing the circulatory mechanisms in humans	Inference (CTS 4)	8

Indicators on Subcourse Learning Outcomes	Critical Thinking Skill Indicator	Question Number
(Subcourse LO 2)	Inference (CTS 4)	9
	Explanation (CTS 5)	10
	Analysis (CTS 2)	11
Identify the type of blood group (Subcourse LO 3)	Explanation (CTS 5)	12
	Inference (CTS 4)	13
Describe the disorders and diseases of the cardiovascular system (Subcourse LO 4)	Explanation (CTS 5)	14
	Analysis (CTS 2)	15

The results of the construct validation show that each indicator in the sub-course learning outcomes shows that each indicator has been included in the question. The number of questions on each sub-indicator of subject learning outcomes is not the same because it is following the breadth and depth of the material being learned. For example, the sub-learning achievement indicator in subject number 1 is presented with 7 questions because the sub course LO1 indicator discusses the structure and function of the blood, blood vessels, and heart wherein this material, if described will discuss blood components and characteristics of blood cells and their functions, the type, and function of the arteries, veins and capillaries as well as the characteristics and mechanisms of action of the heart. The amount of material that students must master in the LO1 sub-course, the number of questions that represent them is also more than the indicators in other LO sub-courses.

Indicators of critical thinking skills used in composing questions use five indicators of critical thinking skills adapted from Fascione. In the developed questions, it can be seen that the distribution of indicators of critical thinking skills, although the numbers are not the same because it follows the suitability of the LO sub-course on the cardiovascular system material.Indicator of critical thinking skills (CTS) 1 on 1 question, CTS 2 is spread over 5 questions, CTS 3 on 1 question, and CTS 4 on 4 questions, and CTS 5 on 4 questions. The dissimilarity of CTS distribution is carried out by prioritizing the conformity aspect of the LO sub-course used. Based on the construct validation, it can be said that the questions developed are appropriate if they are used to measure students' critical thinking skills.

A good question is obtained from the use of indicators developed in this case critical thinking skills that are tailored to the learning outcomes in the curriculum used. Test instruments must be developed following the stated objectives because not all tests meet the appropriate standards, therefore the lecturer must be able to compile an appropriate instrument(Kereh et al., 2015).The similar expressed the importance of the development of the relevant test instrument to measure critical thinking skills following the criteria of reliability and validity as a measurement tool.(Fatimah et al., 2016). To be able to access critical thinking skills, it is necessary to prepare various things related to learning, not only material but also assessment to develop and assess critical thinking skills so thatmust be well arranged (Chee et al., 2009).

The next step is to validate the content assessed by the expert. The summary results of the test instrument validation from the experts are presented in Table 2 below.

Table 2. Summary of Content Validation Results by Experts

No	The aspect of The Assessment	Deskripsi Tiap Aspek	Value
1	Relevance and Representation	Conceptual definition	5
2		Operational definition	5
3		Scoring scale	5
4		Instrument Functions	5
5		Instructions for respondents	4
6		Representation of the number of items	4
7		Answer format	5
8		Drilling	5
9		Population sample	5
10		Time	5
		Average results: 4.8 with valid criteria	
1	The accuracy between the questions and indicators of critical thinking skills	Evaluation (CTS3) with question no 1	5
2		Analysis (CTS 2) with question no 2	5
3		Interpretation (CTS 1) with question no 3	4

No	The aspect of The Assessment	Deskripsi Tiap Aspek	Value
4		Explanation (CTS 5) with question no 4	5
5		Analysis (CTS 2) with question no.5	5
6		Inference (CTS 4) with question no 6	5
7		Analysis (CTS 2) with question no.7	4
8		Inference (CTS 4) with question no. 8	5
9		Inference (CTS 4) with question no.9	4
10		Explanation (CTS 5) with question no 10	5
11		Analysis (CTS 2) with question no.11	5
12		Explanation (CTS 5) with question no.12	5
13		Inference (CTS 4) with question no.13	4
14		Explanation (CTS 5) with question no 14	5
15		Analysis (CTS 2) with question no.15	5
		Average results: 4.7 with valid criteria	
1	Suitability of the question with learning outcomes	Subcourse LO 1 with question no 1	5
2		Subcourse LO 1 with question no 2	5
3		Subcourse LO 1 with question no 3	5
4		Subcourse LO 1 with question no 4	5
5		Subcourse LO 1 with question no 5	5
6		Subcourse LO 1 with question no 6	5
7		Subcourse LO 1 with question no 7	5
8		Subcourse LO 2 with question no. 8	5
9		Subcourse LO 2 with question no 9	5
10		Subcourse LO 2 with question no 10	5
11		Subcourse LO 2 with question no 11	5
12		Subcourse LO 3 with question no 12	5
13		Subcourse LO 3 with question no 13	5
14		Subcourse LO 4 with question no 14	5
15		Subcourse LO 4 with question no 15	5
		Average results: 5 with valid criteria	

Content validations show the average with valid criteria. Valid criteria are obtained from each aspect assessed to the expert regarding the test instrument used.The first aspect that is of relevance and representation discusses the conceptual definition and operational questions that are developed, the scale of assessments, the function of the instrument, the instructions to the user, the representation of the number of items in question with the achievements of learning courses, the format of the answer, engineering scoring, sample and population used to test the questions developed and the timing of the test.The tests carried out on students who have taken human anatomy and physiology courses, this is done to determine the relevance of the questions and retention of student-owned. The second aspect that is measured is the accuracy of the grammar and sentences seen from the use of operational verbs from the indicators of students' critical thinking skills as outlined in the composition of the questions.For example, the suitability of question number 1 with the CTS 3 indicator is evaluated. So that the questions developed are appropriate and can be used to measure students' critical thinking skills.The third aspect is the suitability of the questions to the theory, in this section, an analysis of the suitability of the preparation of questions and the learning outcomes of the course is carried out so that the questions made can be used as a measuring tool to determine students' understanding of concepts.

Content validation is carried out to generate good questions to be used as an appropriate measuring tool. The feasibility of the assessment as a test instrument to measure a process and learning outcomes must go through the validity stage both in content and construct(Walid et al., 2019). The results of the content validity of the experts show good results with valid criteria, so the assessment developed can be used as a test instrument. This test can be used as an appropriate independent measuring tool(Van Lankveld et al., 2017). The next stage is to test the processing of questions by students. Student scores were obtained from the test results shown in Table 3.

Table 3.TheScore Results Obtained by Students at the Time of Test Question Items

No	Student No	Initial Student	Total Score	No	Student No	Initial Student	Total Score
1	22	WHY	11	14	18	ENH	4
2	6	FNT	10	15	19	YHB	4
3	7	LSM	9	16	25	ATP	4
4	10	FZN	9	17	4	DRA	3
5	14	DYS	9	18	20	FSM	3
6	5	MET	6	19	2	FSU	2
7	13	ETP	6	20	8	APM	2
8	26	DLK	6	21	9	RDJ	2
9	12	OTF	5	22	16	MIP	2
10	1	AUT	4	23	21	MNM	2
11	3	LSM	4	24	23	EDM	2
12	11	NIM	4	25	24	HMW	2
13	17	KWK	4	26	15	MLK	1
Average: 4.62, Standard Deviation: 2.84							

The results of the tests conducted by 26 students with 15 questions answered on the cardiovascular system material in the human physiology anatomy subject obtained a mean score of 4.62. Values obtained for most of the students are still low, it shows students still have difficulty in answering the questions presented. Students passed if it gets a value of at least 8. Therefore, it is necessary to test the instrument before it is used in the evaluation. Testing the test instrument was used to determine the accuracy, consistency, and persistence of a question. The test instrument must be tested on another group with the same characteristics before the instrument is given to the actual class(Kereh et al., 2015).The low value obtained when testing the test instrument is usually caused by the reliability of the test instrument with moderate or sufficient criteria(Puspitasari et al., 2019).

Student scores from the trial activity will be analyzed using SPSS to obtain validity data and Anatest to obtain the value of reliability, discriminating power, and the level of difficulty of the items. The results of the analysis are presented in Table 4 below.

Table 4. Test Results of Question Items using Anatest and SPSS

No	The results of the validity of the test (SPSS) r count	Criteria (SPSS)	Distinguishing Power (Anatest)	Degree of difficulty (Anatest)	Correlation (Anatest)	Significance (Anatest)
1	0,357	Valid	28,57	Difficult	0,403	-
2	0,072	Less Valid	28,58	Moderate	0,081	-
3	0,202	Less Valid	28,59	Difficult	0,177	-
4	0,720	Valid	71,43	Difficult	0,698	Very Significant
5	0,451	Valid	57,14	Difficult	0,426	-
6	0,425	Valid	42,86	Difficult	0,426	-
7	0,233	Less Valid	14,29	Difficult	0,208	-
8	0,776	Valid	85,71	Very easy	0,749	Very Significant
9	0,606	Valid	57,14	Very easy	0,600	Significant
10	0,437	Valid	42,86	Difficult	0,436	-
11	0,420	Valid	28,57	Moderate	0,390	-
12	0,293	LessValid	14,29	Moderate	0,341	-
13	0,107	Less Valid	14,29	Very Difficult	0,099	-
14	0,556	Valid	71,43	Moderate	0,607	Very Significant
15	0,553	Valid	85,71	Moderate	0,509	Significant
XY Correlation: 0.48, Results Reliability: 0.65 (Moderate)						

Based on the data obtained from the analysis using SPSS, it is known that the validity value of the questions made is in the valid and less valid categories. Valid and less valid criteria are obtained by comparing r count with r table (the provision of r table with the number of subjects 26 is 0.3172).The item question is valid if the value of r count greater than r table. So the percentage of items that otherwise valid questions amounted to 66.67% (10

questions) and 33.33% (5 questions) with less valid criteria. on items with fewer criteria, the valid question would be improved by changing the grammar and replace with a new question.Validity is very important in the preparation of test instruments because the results are used as a measure of the suitability between the questions and the material. (Mahirah et al., 2016). Good validity results also indicate the good quality of the questions on the test instrument to measure students' abilities following the specified domains or aspects(Mokshein et al., 2019). Validity is one of the requirements that a test instrument must have as an indicator that a test gives almost the same results as a test(Baily et al., 2017).

The reliability value of the final test results is 0.65 which means that the items have medium reliability. a test instrument with a moderate reliability value can be used as a measuring tool in a test. The significance value is obtained by comparing the correlation value of the question items with the total correlation value. The total correlation value in the analysis results is 0.48. The correlation value of each question item is categorized as significant or very significant if the correlation value of each question item is greater than the total correlation value of the instrument. The significance of the question item shows that the 3 questions with the criteria are very significant in questions 3, 8, and 14 and these criteria are significant in questions number 9 and 15.Based on the findings above, the test instruments that have been tested fulfill the validity and reliability aspects. The instrument used to measure thinking skills must have good validity and reliability to be used(Walid et al., 2019).The test instrument developed in the form of a multiple-choice test can be used as a test instrument because it meets the validity and reliability aspects. Multiple-choice questions have high reliability and consistency(Zhongshannvga, 2007). A reliability test is used to determine the reliability or consistency of the instrument seen from the test results which are almost the same in the same conditions(Baily et al., 2017).Even though the reliability value of the test instrument that has been tested is in the sufficient category, the test instrument can be used provided that improvements have been made. The reliability of the instrument can be increased by increasing question items that have high consistency or reducing question items with low consistency(Puspitasari et al., 2019).

The result of the next analysis is the distinguishing power of each item. The criteria for the distinguishing power of the test instrument include, if the diversity index value is ≥ 0,40 then it is considered very good, if it is at a score of 0,30 – 0,39 then it is categorized as good if 0,20 – 0,29 is considered sufficient, and if the value of D ≤ 0,19 is included in the bad criteria(Kereh et al., 2015). The distinguishing power with bad criteria is 20% with 3 questions on numbers 7, 12, and 13. The distinguishing power with criteria is 4 questions sufficient of 26.67% with the distribution of questions number 1, 2, 3, and 11. Distinguishing power with criteria Very good as many as 8 questions with a percentage of 53,33% spread over questions number 4, 5, 6, 8, 9, 10, 14, and 15. Based on these findings, questions with poor and sufficient criteria will be corrected. The test instrument must be known for its distinguishing power to determine the difference between the high and low ability groups (Kusumawati & Hadi, 2018). The distinguishing power of the questions is obtained from the contents of the test instrument itself. The test form used in this instrument is multiple choice. In a multiple-choice test, it must contain one correct answer, have a plausible trick, the alternative answer length must not give a clue to the answer, and the correct answer must appear in every alternative answer(Kumar et al., 2016).

The results of the analysis of the difficulty level value obtained 6,67% with very difficult criteria on question number 13. 46,67% fulfilled the difficult criteria which were spread over 7 questions on numbers 1, 3, 4, 5, 6, 7, and 10 of moderate criteria from 33,33% as many as 5 questions on numbers 2, 11, 12, 14, and 15. The level of difficulty with very easy criteria is 13,33% as many as 2 questions on numbers 8 and 9.The items developed have different levels of difficulty and are scattered throughout the test instruments used. Based on the analysis results, the distribution of the level of difficulty on the test instrument can be said to be almost even. A good test item must have a balance of poor difficulty with a very good difference of 20% each. (Kusumawati & Hadi, 2018). The difference in the difficulty level of each item can be caused by the placement of the types of questions and their order, the placement of the questions that are not appropriate can result in the test results obtained by students.(Debeer & Janssen, 2013).The test instrument that has gone through the trial phase and the value of validity, reliability, differentiation, and difficulty level is known with sufficient to good criteria, it can be said that the instrument can be used as an appropriate measuring tool (Buzi et al., 2019). A test instrument that has gone through many item analysis questions can be stated as an appropriate measuring tool(Lia et al., 2020)and instruments can be used in learning(Hamdi et al., 2018).

CONCLUSION

The test instrument developed for multimedia human physiology anatomy has met the appropriate criteria to be used as a measuring tool for students' critical thinking abilities. This can be seen in the value obtained from the results of the content validity by the expert which shows the valid criteria even though the reliability results are still in the medium category. The differentiation of the distinguishing power of the test instrument is divided into the criteria of poor, sufficient, and very good, while the difficulty level of the test instrument is categorized as very difficult, difficult, moderate, and very easy. The test instrument will be used in the human physiology anatomy course in the next semester.

ACKNOWLEDGEMENTS

Thanks are given to the Ministry of Education and Culture of Higher Education, DRPM, LLDIKTI VII East Java, IKIP Budi Utomo,Department for Research and Community Service of IKIP Budi Utomo, and the Department of Biology Education.

REFERENCES

1. Baily, C., Ryan, Q. X., Astolfi, C., & Pollock, S. J. (2017). Conceptual assessment tool for advanced undergraduate electrodynamics. *Physical Review Physics Education Research*, *13*(2), 1–10. https://doi.org/10.1103/PhysRevPhysEducRes.13.020113
2. Barron, B., & Chen, M. (2008). Teaching for meaningful learning: A review of research on inquiry-based and cooperative learning. In *Powerful Learning: What We Know About Teaching for Understanding* (pp. 11–70). https://doi.org/10.1207/S1532799XSSR0501
3. Buzi, E., Jarani, J., & Isidori, E. (2019). Pro-social and antisocial values in physical education. The validity and reliability of fair play questionnaire in physical education (FPQ-PE) into Albanian language. *Physical Activity Review*, *7*, 143–151. https://doi.org/10.16926/par.2019.07.17
4. Chee, S., Tunku, C., Rahman, A., Phaik, C., Cheah, K., & Rahman College, T. A. (2009). Teacher Perceptions of Critical Thinking Among Students and its Influence on Higher Education. *International Journal of Teaching and Learning in Higher Education*, *20*(2), 198–206. http://www.isetl.org/ijtlhe/
5. Debeer, D., & Janssen, R. (2013). Modeling Item-Position Effects Within an IRT Framework. *Journal of Educational Measurement Summer*, *50*(2), 164–185. https://doi.org/https://doi.org/10.1111/jedm.12009
6. Fatimah, A. W. N., Suryani, N., & Yamtinah, S. (2016). *Development of a Critical Thinking Test Based on Higher-Order Thinking PISA Version : A Tool for Historical Learning in Senior High Schools*. *5*(6), 2013–2016. https://doi.org/http://dx.doi.org/10.18415/ijmmu.v5i2.132
7. Hamdi, S., Suganda, I. A., & Hayati, N. (2018). Developing higher-order thinking skill (HOTS) test instrument using Lombok local cultures as contexts for junior secondary school mathematics. *Research and Evaluation in Education*, *4*(2), 126–135. https://doi.org/10.21831/reid.v4i2.22089
8. Hidayati, N., & Irmawati, F. (2019). Developing digital multimedia of human anatomy and physiology material based on STEM education. *JPBI (Jurnal Pendidikan Biologi Indonesia)*, *5*(3), 497–510. https://doi.org/10.22219/jpbi.v5i3.8584
9. Ihsan, H. (2015). Validitas Isi Alat Uukur Penelitian: Konsep Dan Panduan Penilaiannya. *PEDAGOGIA Jurnal Ilmu Pendidikan*, *13*(3), 173. https://doi.org/10.17509/pedagogia.v13i3.6004
10. Kereh, C. T., Liliasari, Tjiang, P. C., & Subandar, J. (2015). Validitas dan Reliabilitas Instrumen tes Matematika Dasar yang Berkaitan dengan Pendahuluan Fisika Inti. *Jurnal Inovasi Dan Pembelajaran Fisika*, *2*(1), 36–46. https://doi.org/https://doi.org/10.36706/jipf.v2i1.2352
11. Kumar, H., Kumar, S., Dalabh, M., & Ahmad, J. (2016). *Measurement and Evaluation in Education*. http://www.ipesp.ac.th/learning/websatiti/chapter9/unit9_1_4.html
12. Kusumawati, M., & Hadi, S. (2018). An analysis of multiple choice questions (MCQs): Item and test statistics from mathematics assessments in senior high school. *Research and Evaluation in Education*, *4*(1), 70–78. https://doi.org/10.21831/reid.v4i1.20202
13. Lia, R. M., Rusilowati, A., & Isnaeni, W. (2020). NGSS-oriented chemistry test instruments: Validity and reliability analysis with the Rasch model. *Research and Evaluation in Education*, *6*(1), 41–50. https://doi.org/10.21831/reid.v6i1.30112

14. Mahirah, R., Ahmad, D., & , S. (2016). Designing Multiple Choice Test of Vocabulary for the First Semester Students At English Education Department of Alauddin State Islamic University of Makassar. *ETERNAL (English, Teaching, Learning and Research Journal)*, *2*(2), 194–208.
15. Mokshein, S. E., Ishak, H., & Ahmad, H. (2019). The use of rasch measurement model in English testing. *Cakrawala Pendidikan*, *38*(1), 16–32. https://doi.org/10.21831/cp.v38i1.22750
16. Mujiyati, N., Warto, W., & Sutimin, L. A. (2019). Developing a problem-based local history module to improve the critical thinking ability of senior high school students. *Research and Evaluation in Education*, *5*(1), 30–40. https://doi.org/10.21831/reid.v5i1.13334
17. Puspitasari, E. D., Susilo, M. J., & Febrianti, N. (2019). Developing psychomotor evaluation instrument of biochemistry practicum for university students of biology education. *Research and Evaluation in Education*, *5*(1), 1–9. https://doi.org/10.21831/reid.v5i1.22126
18. Susilowati, S., Sajidan, S., & Ramli, M. (2018). Keefektifan perangkat pembelajaran berbasis inquiry lesson untuk meningkatkan keterampilan berpikir kritis siswa. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(1), 49–60. https://doi.org/10.21831/pep.v22i1.17836
19. Van Lankveld, W., Jones, A., Brunnekreef, J. J., Seeger, J. P. H., & Bart Staal, J. (2017). Assessing physical therapist students' self-efficacy: Measurement properties of the Physiotherapist Self-Efficacy (PSE) questionnaire. *BMC Medical Education*, *17*(1), 1–9. https://doi.org/10.1186/s12909-017-1094-x
20. Walid, A., Sajidan, S., Ramli, M., & Kusumah, R. G. T. (2019). Construction of the assessment concept to measure students' high order thinking skills. *Journal for the Education of Gifted Young Scientists*, *7*(2), 237–251. https://doi.org/10.17478/jegys.528180
21. Zhongshannvga, L. L. (2007). *Designing and Revising a Multiple Choice Vocabulary Test*. http://lib.csghs.tp.edu.tw:8080/中山女高學報第7期/004Designing and Revising a Multiple Choice Vocabulary Test-林蕾伊.pdf
22. Zubaidah, S. (2016). Keterampilan Abad Ke-21: Keterampilan yang Diajarkan melalui Pembelajaran. *Prosiding Seminar Nasional Pendidikan Di Pendidikan Biologi STKIP Persada Khatulistiwa Sintang-Kalimantan Barat*, *1*(Desember), 1–17.