Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram
http://ojs.ikipmataram.ac.id/index.php/prismasains/index
e-mail: prismasains.pkpsm@gmail.com

June 2020. Vol. 8, No. 1
p-ISSN: 2338-4530
e-ISSN: 2540-7899
pp. 46-56

Development of the Test Instrument for Measuring Students' Critical Thinking skills on Fluid Material

[1]Ni Nyoman Sri Putu Verawati, [2*]Saiful Prayogi, [2]Muhammad Yusril Yusup, [3]Hafsah Taha

[1]Physics Education Department, Universitas Mataram. Jl. Majapahit No. 62, Mataram, 83125, Indonesia

[2]Physics Education Department, Universitas Pendidikan Mandalika. Jl. Pemuda No. 59A, Mataram, 83126, Indonesia

[3]Faculty of Sciences and Mathematics, Sultan Idris Education University. Proton City, 35900, Malaysia

*Corresponding Author e-mail: saifulprayogi@ikipmataram.ac.id

Received: April 2020; Revised: June 2020; Published: June 2020

Abstract

This study aimed to develop a test instrument to measure students' critical thinking skills on fluid material. Characteristics of the instrument and its validity estimation are also described. This research and development study employed five stages of research, namely information collecting (literature review and preparation of the subject matter), planning (defining and formulating objectives), developing preliminary form of the test instrument, preliminary field testing (expert validation), and main product revision (in accordance with the recommendations in the preliminary field testing). The content and construct validity were estimated by expert validation. Results of the instrument validation showed average scores for each component (content validity index of 4.63 and construct validity of 4.75), both of which are in the *very valid* category. The final result of the instrument validation is 4.69 (very valid if: Va > 4.21), with 98.7% reliability. Description of the study result is presented further in this article.

Keywords: Test instrument, Critical thinking skills, Fluid material

How to Cite: Verawati, N., N., S., P., Prayogi, S., Yusup, M., Y., & Taha, H. (2020). Development of the Test Instrument for Measuring Students' Critical Thinking Abilities on Fluid Material. *Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram,* 8(1), 46-56. doi:https://doi.org/10.33394/j-ps.v8i1.2487

https://doi.org/10.33394/j-ps.v8i1.2487

INTRODUCTION

One measure of progress for a nation is the quality of education and one of the important aspects of quality education, namely students who can compete globally and are able to solve problems in daily life (Megawati et al., 2020). In terms of global competitiveness, the quality of Indonesian education seems to be still low. The results of the Program for International Student Assessment (PISA) study conducted by the Organization for Economic Cooperation and Development (OECD) in 2015, Indonesia ranked 62 out of 70 countries. Indicators of PISA assessment included the ability of students to solve problems and higher-order thinking skills (OECD, 2017). The results of the 2015 Trends in International Mathematics and Science Study (TIMSS) study showed that Indonesia ranked 46 out of 51 countries in science achievements (Mullis et al., 2015). In addition, studies other than PISA and TIMSS have long reported that in order to face the challenges and developments of the modern age, it is necessary not just conceptual knowledge, but the skill to apply knowledge and various thinking skills called 21st Century Skills (Partnership for 21st Century Skills, 2002). One of the skills contained in 21st Century skills, namely critical thinking (Prayogi et al., 2018). Critical thinking is basically a

detailed description of several characteristics which include the process of interpretation, analysis, evaluation, inference, explanation and self-regulation (Facione, 2011). One famous contributor in the tradition of critical thinking is Robert Ennis. Ennis (1996) provide the definition about the concept of critical thinking, namely critical thinking as sensible and reflective thinking that focuses on deciding what to believe or do.

Improving the quality of education delivery is marked by curriculum reform. For example, Finland is a developed country in the field of education that has long placed critical thinking as one of the goals of learning in its curriculum content (Horn & Veermans, 2019). In Indonesia, curriculum content which is oriented towards the development of a variety of thinking skills, especially critical thinking skills, has begun to be noticed with the implementation of the Curriculum-2013 (K-13) (Prayogi et al., 2019). In several other developing countries critical thinking is one of the high-level skills most often discussed and linked to educational goals, because it is believed to play a central role in logical thinking, problem solving, an in decision making (Butler, 2012). Various factors can contribute to achieving learning goals towards increasing critical thinking skills, one of which is the ability of teachers to carry out and utilize the critical thinking assessment process itself, because it can stimulate students to develop critical thinking skills (Herpiana & Rosidin, 2018).

The learning objectives towards training students' critical thinking are contained in K-13 in Indonesia. However, learning in reality there are still many that are oriented solely on efforts to develop and test students 'memory so that students' thinking abilities are reduced and simply understood as the ability to remember (Herpiana & Rosidin, 2018), whereas they should be oriented towards achieving thinking and solving skills the problem is in accordance with the contents of the 2013 Curriculum. The impact is an evaluation tool used to measure aspects of ordinary cognitive abilities. In other cases, it was also found that although the learning process aimed at training students' critical thinking, in the assessment process the teacher did not use the instrument of critical thinking to measure learning outcomes according to the learning process intended (Sudrajat et al., 2018). Whereas learning assessment is one of the elements in learning used to determine the extent of the achievement of student competencies and the effectiveness of the learning process undertaken to achieve learning objectives (William, 2013). Assessment is an inseparable part of the learning process. Learning assessment is important so that students know what they have learned or show what they have not learned (Jabbarifar, 2009).

Test instruments to measure critical thinking skills have been developed by experts and researchers before, including; California Critical Thinking Skill Test (Facione, 1990), California Critical Thinking Disposition Inventory (Facione & Facione, 1992), Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980), Ennis-Weir Critical Thinking Essay Test (Ennis & Weir, 1985), Cornell Critical Thinking Test (Ennis, Millman, & Tomko, 1985), Halper Critical Thinking Assessment Manual (Halpern, 2010), and many others. However, to date the multivariate nature of the definitions offered for critical thinking, the assessment of critical thinking, their level of generality or specificity, and their practical impact on broader academic achievement is still being debated (Liu et al., 2014). According to Lai (2011), these instruments vary greatly in purpose and format, and critical thinking assessments tend to be general in nature. Norris (1989) has long argued that the facts about the level of uniqueness in critical thinking have not been resolved because of the many theories in different views, making measurement and evaluation of critical thinking difficult. First, the type of a person's conclusions is still unclear to the extent that researchers cannot agree whether critical thinking is a general or special subject. Second, it is difficult to assess the transfer of critical thinking with other subjects, because transfer to other contexts may be different from the specificity of knowledge in critical thinking. Thus, despite the fact that previous researchers or experts have identified critical thinking skills or abilities and dispositions that differ from one another, it turns out to describe separate effects and the use of difficult judgments in practice (Prayogi, 2019). Therefore, the measurement of critical thinking in a more specific context needs to be

done and the instrument used as an important measurement tool to be developed, in this study is the instrument for measuring critical thinking skills in fluid material. The instrument for measuring critical thinking skills developed was tested for its validity (content and construct), and its reliability.

The purpose of this study, namely: a) describe the characteristics of test instruments for the measurement of students 'critical thinking skills on fluid material, and b) describe the validity and reliability of instruments measuring students' critical thinking skills on fluid material. The test instrument was developed based on critical thinking indicators that have been used by researchers previously, namely in the aspects of analysis, inference, evaluation, and decision making (Prayogi et al., 2017; Prayogi et al., 2018; Prayogi et al., 2019; Prayogi, 2019; Wahyudi et al., 2019; Verawati et al., 2019). The results of this study can be a reference in the development of test instruments to measure critical thinking skills.

METHOD

This research and development (R & D) study aimed to develop an instrument to measure students' critical thinking skills. According to Borg and Gall (1983), a R & D study involves 10 stages of research, that are research and information collecting, planning, develop preliminary form of product, preliminary field testing, main product revision, main field testing, operational product revision, operational field testing, final product revisions, and dissemination. In accordance with the needs of this research, adaptation and modification are carried out into 5 stages of research, namely: 1) information collecting, includes literature review and preparation of the subject matter; 2) planning, includes defining and formulating objectives; 3) develop preliminary form of instrumen test product; 4) preliminary field testing (expert validation); and 5) main product revision in accordance with the recommendations in the preliminary field testing. The research phase is summarized in Figure 1.

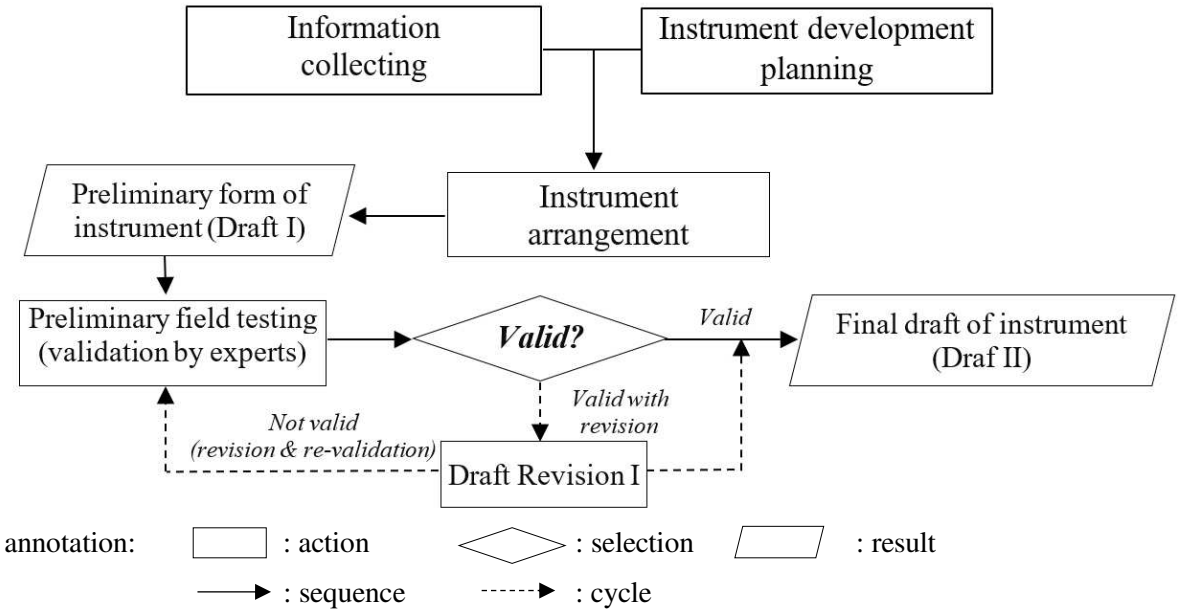


Figure 1. The research stage is the development of the test instrument

Information collecting includes reviewing the literature, preliminary observations, and determination of the subject matter, this is done to determine the learning needs associated with the plan to develop instruments for measuring students' critical thinking abilities. The instrument development planning includes defining and formulating objectives. Next, the researcher designed the draft of test instrument in Bahasa Indonesia to measure students' critical thinking skills and then validated it. Validation contains two elements of validity, namely content validity and construct validity. The instrument developed was validated by 2 validators, with the validator criteria experts in the field of physics and at least had conducted studies on

critical thinking. Suggestions and input from the validator in the next will be followed up to improve the instruments developed. The validity of the instrument by experts shows the quality of the instrument in terms of content and construct validity. Content validity refers to the extent to which the test measures the domain of content to be measured. There are three aspects of content validity: domain definition (operational definition of the content domain), domain representation (a test match between content and cognitive specifications), and domain relevance (relevance of test items to content domains) (Sireci, 1998). While the construct validity refers to the extent to which the operationalization of the construct is defined by a theory (Cronbach & Meehl, 1955). Reliability tests were also conducted and calculated using the equation of percentage of agreement by Emmer and Millett in Borich (1994), the instrument is said to be reliable if it has a match percentage of ≥ 75%. If the match percentage is high, then the instrument is said to be reliable. Reliability shows the level of consistency of the instrument based on observers' judgment in this context is the validator.

$$Percentage\ of\ Agreement = 100\left(1 - \frac{A-B}{A+B}\right)$$

annotation:

A = The frequency of the aspects observed by the observer by giving a high frequency

B = The frequency of the aspects observed by other observers by giving a low frequency

The data of validation result were analyzed descriptively qualitatively, that is, by averaging the scores obtained from the validators. Validity assessment uses a scale of 5 (highest score 5, lowest score 1). Scores obtained from expert judgment are then converted into qualitative data and categorized as in Table 1.

Table 1. Conversion of quantitative data into categories (Prayogi, 2019)

Interval Score		Category
$X > X_i + 1{,}8\ Sbi$	$X > 4{,}21$	very good
$X_i + 0{,}6\ Sbi < X \leq X_i + 1{,}8\ Sbi$	$3{,}40 < X \leq 4{,}21$	good
$X_i - 0{,}6\ Sbi < X \leq X_i + 0{,}6\ Sbi$	$2{,}60 < X \leq 3{,}40$	enought
$X_i - 1{,}8\ Sbi < X \leq X_i - 0{,}6\ Sbi$	$1{,}79 < X \leq 2{,}60$	less
$X \leq X_i - 1{,}8\ Sbi$	$X \leq 1{,}79$	not good

annotation: X (empirical score), X_i (mean ideal), *Sbi* (ideal deviation standard)

The level of validity (Va) is determined by calculating the average score of indicators and aspects for each expert, by adapting the interval of values in Table 1, then the level of instrument validity as in Table 2.

Table 2. Level of instrumen validity (Prayogi, 2019)

Interval Score	Category
$Va > 4{,}21$	very valid
$3{,}40 < Va \leq 4{,}21$	valid
$2{,}60 < Va \leq 3{,}40$	enought valid
$1{,}79 < Va \leq 2{,}60$	less valid
$Va \leq 1{,}79$	not valid

The instrument for measuring critical thinking skills is said to have a good degree of validity, if the minimum level of validity achieved is valid. If the achievement level of validity is under valid, a revision is necessary.

RESULTS AND DISCUSSION

Characteristics of test instruments for the measurement of critical thinking skills

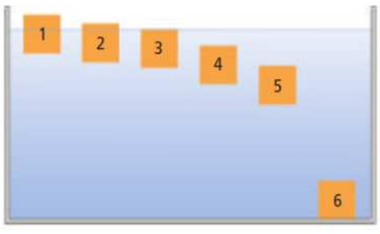
Characteristics of test instruments for the measurement of critical thinking skills in this study are based on theoretical studies from experts who develop high-level thinking instruments. In general, the preparation of assessment instruments generally involves three

things, namely: a) clearly determining what will be assessed; b) arrange test assignments or questions; and c) determine the criteria for mastery of the thing being assessed. In this study the assessed context, test questions, and mastery criteria were judged to focus on critical thinking skills. Furthermore, according to Brookhart (2010), in the preparation of high-level thinking assessment instruments, there are three things that need to be considered, namely: a) using stimulus; b) use a new context; and c) distinguish between the level of difficulty and complexity of the thought process. Test instruments to measure high-level thinking according to Scully (2017) can be arranged by taking into account the following matters: a) manipulation of target verbs in accordance with the aspects measured, b) item flipping, meaning items simple questions can be modified (reversed) for example by presenting specific examples in the problem bar, and asking respondents to identify the rules or concepts underlying them, c) the use of high quality distractors, and d) tapping multipple neurons, meaning presenting several items content where the instrument can be a stimulus that allows the interconnection between knowledge. Elaboration of existing theories, the determination of the characteristics of the critical thinking skills test instrument in this study refers to the following things.

a. The suitability of the content domain to be measured and assessed. This research has determined the critical thinking skill as an aspect to be measured in fluid material, namely on the indicators of analysis, inference, evaluation, and decision making. Operationally the skill to think critically on the aspect of analyzing is the process of actually identifying the relationship between statements, concepts, and descriptions of a situation or phenomenon, and describing reasons as a form of representation to express opinions and beliefs. Referencing is the process of concluding information correctly using logical reasoning from data, statements, principles, concepts, or other forms of representation. Evaluating is the process of evaluating the credibility of a statement, description, experience, situation, or opinion, as well as describing the reasons for the evaluation. Making a decision is the process of choosing choices or actions between a set of alternatives based on criteria or strategy.
b. Item compatibility with critical thinking indicators. 14 questions were developed for fluid matter to accommodate the critical thinking indicators that are used as a reference for the preparation of items. One example problem to measure the critical thinking skill on the indicator analysis as in Figure 2.
c. Give a stimulus in which the problem presented is closely related to the real-world context. On the subject of fluid matter, for example, in rivers, objects are often found in a floating position on the water, partially submerged, floating in water, and even submerged entirely in the riverbed. This case is then elaborated into the idea of preparing critical thinking skills items that measure analytical skills in the sub-subject matter of density, where students are asked to analyze the state of objects immersed in water with different conditions based on their density and determine the position of each object immersed. This problem requires the critical thinking skill in this context of analysis, because it is not possible for students to determine the position of each object if it cannot analyze the state of an object based on its density. The item construction of the critical thinking skill instrument in the aspect of analysis as in Figure 2.

Berikut ditampilkan lima densitas atau massa jenis benda yang dicelupkan ke dalam bak berisi air (perhatikan gambar di samping!)

a. 0,85 g/cm^3
b. 0.95 g/cm^3
c. 1,05 g/cm^3
d. 1,15 g/cm^3
e. 1,25 g/cm^3



Massa jenis air adalah 1,0 g/cm^3. Diagram pada gambar menunjukkan enam kemungkinan posisi kelima benda jika dimasukkan ke dalam bak berisi air tersebut. Analisislah keadaan benda dengan memilih posisi dari 1-6 untuk kelima benda, berikan alasanmu!

Figure 2. The example of test instrumen in analysis aspect

The validity and reliability of the test instruments measure critical thinking skills

The critical thinking skills test instrument that has been compiled is tested for its validity and reliability. Validity test includes content validity and construct validity. Content validity according to Sireci (1998), i.e. "…refers to the degree to which a test measures the content domain it purports to measure". There are three aspects of content validity: domain definition (operational definition of content domain), domain representation (how well a test matches its content and cognitive specipication), and domain relevance (relevance of each test item to the content domain). Content validity refers to the extent to which the test measures the domain of content to be measured. While the construct validity according to Cronbach and Meehl (1955), i.e "…refers to the extent to which operationalizations of a construct are defined by a theory," its means that the construct validity refers to the extent to which the operationalization of the construct is defined by a theory. Reliability shows the level of consistency of the instrument based on observers' judgment in this context is the validator. Reliability is calculated using the percentage of agreement equation. The validation sheets are developed as Figure 3, Figure 4, Figure 5, and Figure 6 (examples for analysis indicators). Figure 3 explains the critical thinking skills test instrument, Figure 4 and Figure 5 explain aspects of content validity, and Figure 6 explain of construct validity.

1. Here are shown the density of five objects dipped in a tub of water (see the figure on the side!)
   a. 0,85 g/cm³
   b. 0.95 g/cm³
   c. 1,05 g/cm³
   d. 1,15 g/cm³
   e. 1,25 g/cm³

   The density of water is 1.0 g / cm³. The diagram in the figure shows the six possible positions of the five objects if inserted into the tub of water. Analyze the state of objects by selecting positions from 1-6 for the five objects, give your reasons!

2. According to research data, atmospheric pressure is proportional to 100,000 Pa, this means that the weight of the earth's atmosphere is proportional to 100,000 N per meter- square on earth and the pressure is proportional to the weight of a minibus. Our body is surrounded by this pressure every day. Why is our body not destroyed by this pressure?

3. The large cross section of a hydraulic pump has a surface area of 50 times the small cross section of a hydraulic pump. A stocky man wants to provide enough force to lift an object placed on a small cross section. If the mass of the object is 10 kg, according to your analysis can the man give a strong force to lift the object at a small cross section?

4. The mass of an aluminum beam is 25.0 gr. (a) Analyze how much is the volume? (b) how much the tension of the rope that holds the beam when the beam is fully submerged in water? The density of alumanium is 2700. kg/m³.

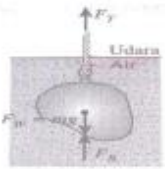


Figure 3. Critical thinking skills test instrument for analysis aspect

The aspects to be assessed: Content validity (domain definition)

The test instrument presents statements, data, phenomena, facts, or laws that enable students to conduct an analysis (the process of actually identifying the relationship between statements, concepts, and descriptions of a situation or phenomenon, and describing reasons as a form of representation to express opinions and beliefs).

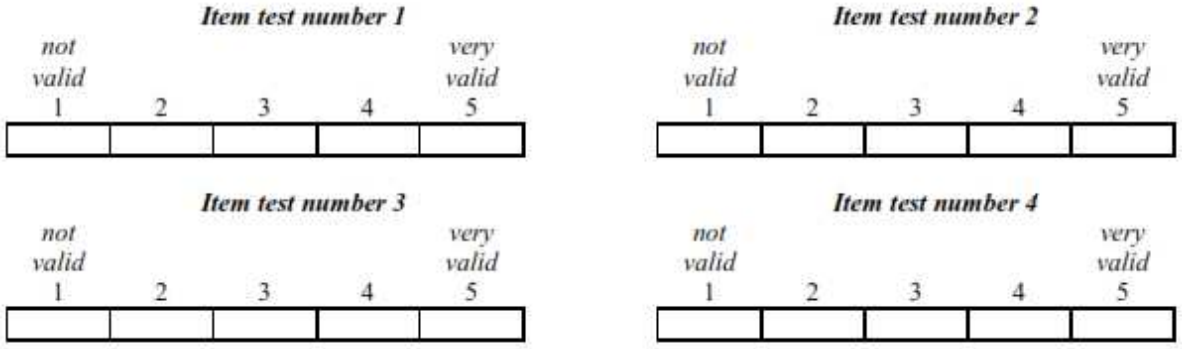


Figure 4. The validation sheet in the aspect of content validity (domain definition)

The aspects to be assessed: Content validity (domain relevance and representation)

The test instruments that was developed are relevant and represent the content of the analysis.

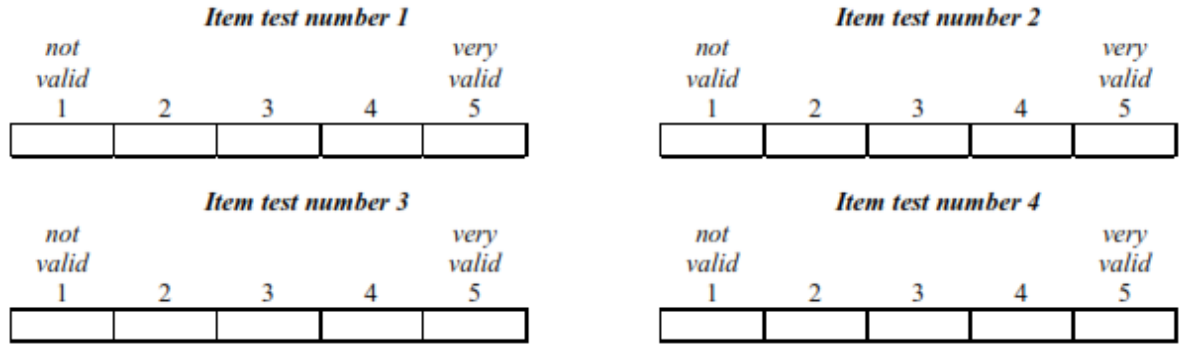


Figure 5. The validation sheet in the aspect of content validity (domain relevance and representation)

The aspects to be assessed: Construct validity

The test instrument that was developed factually contains the correctness of the concept in terms of theory.

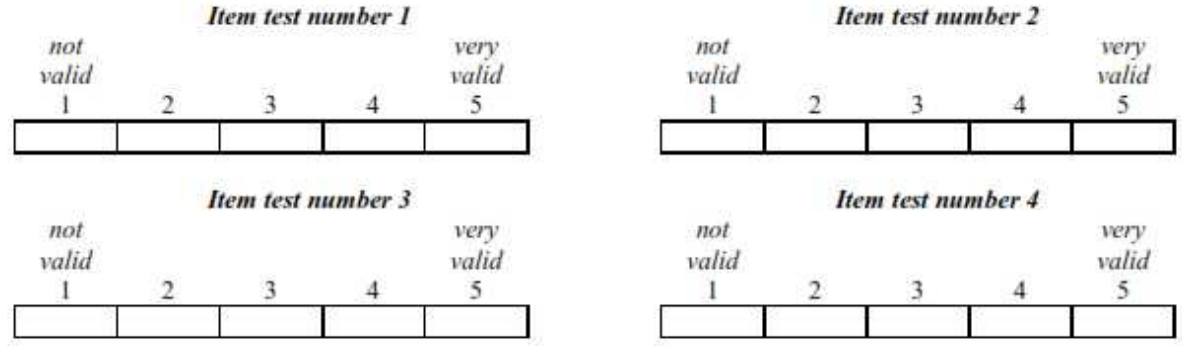


Figure 6. The validation sheet in the aspect of construct validity

The results of the validity and reliability tests of the instrumen that was developed are presented in Table 3.

Table 3. The results of the validity and reliability tests of the instrumen

Indicator	Items	Content Validity				Mean	Construct Validity		Mean
		DD		DRR					
		V_1	V_2	V_1	V_2		V_1	V_2	
Analysis	1	4	2	4	5	3.75	5	5	5
	2	4	4	5	5	4.5	5	5	5

Indicator	Items	Content Validity				Mean	Construct Validity		Mean
		DD		DRR					
		V_1	V_2	V_1	V_2		V_1	V_2	
	3	5	5	5	5	5	5	5	5
	4	5	5	5	5	5	5	5	5
mean						4.56			5
Inference	5	5	5	5	5	5	5	5	5
	6	5	5	5	5	5	5	4	4.5
	7	4	5	4	5	4.5	4	5	4.5
	8	4	5	4	5	4.5	4	5	4.5
mean						4.75			4.63
Evaluation	9	5	5	5	5	5	5	5	5
	10	3	5	3	5	4	3	5	4
	11	5	5	5	5	5	5	5	5
	12	3	5	4	4	4	4	5	4.5
mean						4.5			4.63
Decision making	13	5	5	5	5	5	5	5	5
	14	4	5	4	5	4.5	4	5	4.5
mean						4.75			4.75
mean score						4.75			4.75
mean score of each item validity						4.63			4.75
mean score of validity (Va)						4.69 (very valid)			
Percentage of agreement						98,7% (reliable)			

annotation: DD (domain definition), DRR (domain representation and relevance), V (validator)

In recent years, there has been increased recognition of the role of measuring instruments in the assessment process in education. That is, in addition to providing evaluative information about a student, assessment can and must also function as a mechanism to assist learning to achieve desired learning goals (Wiliam, 2013). As specific objectives are established in learning, specific instruments are developed for the intended purpose. For example, learning objectives towards learning outcomes of critical thinking skills, the measurement and assessment instruments must also be oriented towards critical thinking skills. In compiling an instrument, it is not enough just to determine the topic or material to be assessed, but it is also necessary to determine more specifically what thought processes will be assessed for certain materials. Criteria for measuring critical thinking instruments have been established in this study, namely the suitability of the content domain to be measured and assessed in critical thinking, the suitability of the item items with indicators of critical thinking, and the instrument items can provide a stimulus in which the questions presented are closely related to the real world context . Critical thinking skills criteria are set in this research, namely on aspects of analysis, inference, evaluation, and making decisions. Furthermore, the instrument item construction format is in the form of essay questions, this is in accordance with the suggestion from Brookhart (2010) that the measurement of high-level thinking including critical thinking can use two item item formats, namely essay questions or performance assessment. Explanation in students' answers can provide more valid information about their critical thinking abilities, so that the evaluator in this case the teacher can give an interpretation of the level of critical thinking students.

Accommodating the criteria for the preparation of instruments for measuring critical thinking skills as explained earlier, the instrument for thinking critical thinking skills has been

developed and compiled in the form of essay questions. A total of 14 test items were developed to accommodate the four indicators of critical thinking set in this study. The instruments compiled were then validated (content and construct) by two validators. Validation results show the average score for each component, namely the content validity of 4.63 and construct validity of 4.75. The average score which is the final result of the instrument validation is 4.69 categorized as very valid (very valid if: Va> 4.21), a percentage of agreement rating of the validators is 98.7% with reliable criteria.

The average score of content validity on aspects of analysis, inference, evaluation, and making decisions in a row is 4.56, 4.75, 4.5, and 4.75, all with very valid criteria. These results indicate that the instrument is able to measure the content domain to be measured, in this context is the content of critical thinking. A more specific meaning is that the test instruments developed are in accordance with the operational definitions of the aspects of critical thinking that are measured as well as, are relevant and represent questions of critical thinking. The average construct validity score on aspects of analysis, inference, evaluation, and making decisions in a row of 5, 4.63, 4.63, and 4.75, all with very valid criteria. That is, the operationalization of the instrument item developed construct is based on theory. The developed instrument has also been reliable, this is indicated by the percentage of compatibility by two validators of 98.7%. In accordance with the objectives to be achieved, the instrument has been developed in measuring the critical thinking skill and has been declared valid and reliable.

CONCLUSION

The instrument for measuring critical thinking skills was developed with criteria; a) the suitability of the content domain to be measured and assessed in critical thinking, b) the suitability of the item items with the critical thinking indicator, and c) the instrument items can provide a stimulus where the questions presented are closely related to the real world context. The measurement instruments of critical thinking skills developed have been valid in content and construct, as well as reliable based on the validator's instrument rating. Therefore, it can be used to measure students' critical thinking skills on fluid material.

RECOMMENDATION

The instruments developed need to be empirically tested in class in high school students and their empirical validity is calculated, the results will be a strong support for the results of the content and construct validity tests that have been obtained previously.

ACNOWLEDGEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Borg, W. R., & Gall, M. D. (1983). *Educational research: An introduction (4th ed)*. White Plains, NY: Longman Inc.

Borich, G. D. (1994). *Observation skills for effective teaching*. Columbus, OH: Merrill

Brookhart, S. M. (2010). *How to assess higher-order thinking in your classroom*. Virginia: ASCD Publication.

Butler, H. A. (2012). Halpern critical thinking assessment predicts real-world outcomes of critical thinking. *Applied Cognitive Psychology, 25*(5), 721-729.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Ennis, R. H., & Weir, E. (1985). *The Ennis -Weir Critical Thinking Essay Test*. Pacific Grove, CA: Midwest Publications.

Ennis, R. H., Millman, J., & Tomko, T. N. (1985). *Cornell Critical Thinking Tests*. Pacific Grove, CA: Midwest Publications.

Facione, P. A. (1990). *The California Critical Thinking Skills Test-college level. Technical report #2. Factors predictive of CT skills*. Millbrae, CA: California Academic Press.

Facione, P. A., & Facione, N. C. (1992). *The California Critical Thinking Dispositions Inventory*. Millbrae, CA: California Academic Press.

Facione, P. (2011). *Critical thinking. What it is and why its counts.* Millbrae, CA: The California Academic Press.

Halpern, D. F. (2010). *Halpern Critical Thinking Assessment manual*. Vienna, Austria: Schuhfried GmbH.

Herpiana, R. & Rosidin, U. (2018). Development of instrument for assessing students' critical and creative thinking ability. *J. Phys.: Conf. Ser. 948*, 1-7.

Horn, S. & Veermans, K. (2019). Critical thinking efficacy and transfer skills defend against 'fake news' at an international school in Finland. *Journal of Research in International Education, 18*(1), 1-19.

Jabbarifar, T. (2009). The importance of classroom assessment and evaluation in educational system. In *Proceedings of the 2nd International Conference of Teaching and Learning* (pp. 1–9). Malaysia: INTI University College.

Lai, E. (2011). *Critical thinking: A literatur review*. Pearson Research Reports. Retrieved from http://images.pearsonassessments.com/CriticalThinking ReviewFINAL.pdf

Liu, O. L., Frankel, L., & Roohr, K.C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report*, *14*(10), 1-23.

Megawati, Wardani, A. K., & Hartatiana. (2020). Kemampuan berpikir tingkat tinggi siswa SMP dalam menyelesaikan soal matematika model PISA. *Jurnal Pendidikan Matematika, 14*(1): 15-24.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2015). *TIMSS 2015 international results*. Boston: TIMSS & PIRLS International Study Center.

Norris, S. (1989). Can we test validly for critical thinking? *Educational Researcher, 18*(9), 21-26.

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 results: Collaborative problem solving* (Volume V). Paris: OECD Publishing.

Partnership for 21st Century Skills. (2002). *Learning for the 21st century: A report and mile guide for 21st century skills*. Tucson, AZ: Author.

Prayogi, S., Yuanita, L. & Wasis. (2017). Critical-Inquiry-Based-Learning: Model of learning to promote critical thinking ability of pre-service teachers. *Journal of Physics: Conference Series 947,* 1-6, doi:10.1088/1742-6596/947/1/012013

Prayogi, S., Yuanita, L. & Wasis. (2018). Critical-Inquiry-Based-Learning: A model of learning to promote critical thinking among prospective teachers of physic. *Journal of Turkish Science Education, 15*(1), 43-56.

Prayogi, S., Muhali, Yuliyanti, S., Asy'ari, M., Azmi, I. & Verawati, N. N. S. P. (2019). The effect of presenting anomalous data on improving student's critical thinking ability. *International Journal of Emerging Technologies in Learning, 14*(6), 133-137.

Prayogi, S. (2019). Pengembangan model pembelajaran Critical-Inquiry-Based-Learning untuk meningkatkan kemampuan berpikir kritis mahasiswa. *Disertasi*. Surabaya: Universitas Negeri Surabaya.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research & Evaluation, 22*(4), 1-13.

Sireci, S. G. (1998). Gathering and analyzing content validity. *Educational Assessment 5*(4), 299-321.

Sudrajat, A.K., Saptasari, M. & Tenzer, A. (2018). Developing formative assessment instruments to measure students' critical thinking ability in human circulatory system

material for 11th grade students of UM High School Laboratorium. *Jurnal Penelitian Pendidikan, 18*(3), 243-251.

Verawati, N. N. S. P., Prayogi, S., Gummah, S., Muliadi, A. & Yusup, M. Y. (2019). The effect of conflict-cognitive strategy in inquiry learning towards pre-service teachers' critical thinking ability. *Jurnal Pendidikan IPA Indonesia, 8*(4), 529-537

Wahyudi, Verawati, N. N. S. P., Ayub, S., & Prayogi, S. (2019). The effect of scientific creativity in inquiry learning to promote critical thinking ability of prospective teachers. *International Journal of Emerging Technologies in Learning, 14*(14), 122-131.

Watson, G., & Glaser, E. M. (1980). *Watson-Glaser Critical Thinking Appraisal, forms A and B manual*. San Antonio, TX: The Psychological Corporation.

Wiliam, D. (2013). Assessment: The bridge between teaching and learning. *Voices from the Middle, 21*(2), 15-20.