Volume 8. Number 2. 127-140 OCTOBER 2019

Jurnal ilmiah pendidikan fisika Al-Biruni
https://ejournal.radenintan.ac.id/index.php/al-biruni/index
DOI: 10.24042/jipfalbiruni.v0i0.3952

e-ISSN: 2503-023X

Specific Open-Ended Assessment: Assessing Students' Critical Thinking Skill on Kinetic Theory of Gases

Riki Perdana[*1], Riwayani[2], Jumadi[3], Dadan Rosana[4], Soeharto Soeharto[5]

[1, 2, 3, 4] Physics Education Department, Yogyakarta State University, Yogyakarta, Indonesia
[5] Doctoral School of Education, University of Szeged, Hungary

*Corresponding Address: rikifisika95@gmail.com

Article Info

Article history:

Received: April 3rd, 2019
Accepted: October 8th, 2019
Published: October 30th, 2019

Keywords:

critical thinking skill;
kinetic theory of gases;
open-ended.

ABSTRACT

The test of critical thinking skills in specific topics in physics is still rarely. This study aimed to develop a specific test in critical thinking skills in the kinetic theory of gases (CTKTG) and also to assess the students' critical thinking skills. This study used the 4D method (Define, Design, Develop, and Disseminate). The CTKTG test was initially tested in four sample groups: interviews with an expert review (N = 3), professional physics teachers (N = 2), and graduate school students (N = 2), students from secondary schools (N = 29). The test was modified based on the revised results in the initial test. After that, the test was given to a group of students in class XI, who were science students (N = 55). The results showed that internal consistency from the CTKTG test was α = .89 (good). The implementation strategies and tactics are the most difficult aspect of critical thinking skill with a mean of 1.37 (very low) and basic classification is easiest with a mean of 2.84 (average). So, the findings showed that the CTKTG test can be used to measure students' critical thinking skills on the topic of the kinetic theory of gases.

INTRODUCTION

Critical thinking skill is owned by active students. This skill is related to others, such as scientific communication and self-confidence (Wismath, Orr, & Zhong, 2014) and students' motivation (Hu, Jia, Plucker, & Shan, 2016). However, Hashim & Samsudin (2019) found that some aspects of students' critical thinking skills were still at the middle level.

Purwati, Hobri, & Fatahillah (2016) also found a similar result, as many as 32.2% of students studied still had low critical thinking skills and 42.8% of the moderate category. Matsun, Sunarno, & Masykuri (2017) found the average value of students in critical thinking skills at a low level with a mean of 65.70. It shows that the level of students' critical thinking skills in Indonesia is very low. This ability has become the main key in policymaking (Szenes, Tilakaratna, & Maton, 2015). Therefore, research on this ability still needs to be done, especially in a specific topic in the learning process.

Measuring critical thinking skills in Physics is found to lack scholars' agreements. In the beginning, scholars suggested that measuring critical thinking skill evaluate the general skill of thinking. However, further studies found that thinking skill is related to the critical point of view on certain issues. Hence, in this current era, the development of critical thinking is considered an important educational goal, with always increasing in now (Kettler, 2014). Nowadays, it takes anyone with many variation skills such as critical thinking, problem-solving, and the

How to cite Perdana. R., Riwayani., Jumadi., Rosana. D., & Soeharto. S. (2019). Specific Open-Ended Assessment: Assessing Students' Critical Thinking Skill on Kinetic Theory of Gases. *Jurnal ilmiah pendidikan fisika Al-Biruni*, *8*(2), 127-140.

application of some way in thinking process (Ghazivakili et al., 2014). In this era of learning environment, the student should at the advance level about critical thinking skills for their success in life (Kong, 2014).

As an important aspect of the learning process, the teacher must have the critical thinking skill to teach the students anything about critical thinking (Fuad, Zubaidah, Mahanal, & Suarsini, 2017). Because, critical thinking skills help the teacher in the teaching process, especially at discussion and debate with the students in difficult objects (Nasution, Harahap, & Manurung, 2017), such as physics and math.

A lot of effort is being made to improve intelligence and general trends towards critical thinking (Huber & Kuncel, 2015). There are many studies about measuring critical thinking test. Liu, Mao, Frankel, & Xu, (2016) have designed an assessment test to measure the students' critical thinking skills in the dimension of analytical and synthetic.

Pascarella et al (2014) use 32 items of Critical Thinking Test (CTT) from the Collegiate Assessment of Academic Proficiency to measure students' critical thinking skills (clarifying, analyzing, evaluating, and expanding arguments).

Rowland, Lovelace, Saunders, Caruso, & Israel (2016) used the California Critical Thinking Ability Test (CCTST). Carter, Creedy, & Sidebotham (2016) designed the Carter Assessment of Critical Thinking in Midwifery Tests (CACTiM). Shin, Jung, & Kim (2015) developed a test of clinical critical thinking skills (CCTS) and then validated the results. The revised of CCTS was declared reliability and sufficient validity.

Lee (2018) used the C-QRAC test (Collaboration Questions and Answers, Reading, Answering, and Checking) on 85 students to facilitate that skills. The results indicate that there is any direct effect of direct instruction in critical thinking on reading literacy. Stupple et al., (2017) developed a Critical Thinking Toolkit (CriTT) to assess students' beliefs and attitudes about these skills. The results indicate that the test can be used to identify students who need help in improving CT skills.

Gelerstein, Río, Nussbaum, Chiuminatto, & López (2016) designed and validated tests to assess students' CT skills in grades 3 and 4 in language arts lessons using graphic novels. The results of the assessment show more detailed and multidimensional student learning. Mapeala & Siew, (2015) developed a Test of Science Critical Thinking (TSCT) to assess three subcritical thinking skills in fifth-grade students which included distinguishing and comparing, identifying and sequencing. Vieira & Tenreiro-Vieira (2016) adapted the Cornell Critical Thinking test (level X) to assess science learning experiences focused on Critical Thinking.

Dawit Tibebu Tiruneh, De Cock, Weldeslassie, Elen, & Janssen, (2017) developed tests to measure students' CT skills on the topic of electricity and magnetism (CTEM). The findings indicate that the test can be used to measure CT skills specifically in Electric & Magnetism, and is a fundamental basic study for future research that focuses on the integration of CT skills in the certain subject matter. This study focuses on CT skills in the kinetic theory of gases (CTKTG).

Open-ended is different from an interview or questionnaire tests because structured questionnaires limit the explanations of the experiences of participants (Tran, Porcher, Falissard, & Ravaud, 2016). The current test more often uses a multiple-choice test in measuring critical thinking skills or interview/questionnaire. In this study, we use the open-ended format. When open-ended questions are used in large-scale assessments, those involved tend to emphasize the skills assessed by these questions, which are useful in real life (Yan, Yamada, Takagaki, & Koizumi, 2019). For, the novelty reason, we decided to use the open-ended format in assessing students' critical thinking skill, through open-ended tests, we can explore, explain or confirm students' knowledge more deeply than any other test.

The importance of open-ended test first and foremost, it can break the opinion with the right solution (Klavir & Hershkovitz, 2014). They allow respondents to write their answers in their own words (Lee & Lutz, 2016; Popping, 2015) and do not limit their answers (Schonlau & Couper,

2016). They can provide new and valuable answers that may not have been thought of by previous researchers (Gurel, Eryilmaz, & McDermott, 2015). In other words, open questions provide a wealth of information to researchers we decided to measure the aspect CT using essay (open-ended). For that reason, this study shows the results of the reliability, validity, and other aspects of developing a test designed to measure CT skills, specifically on the kinetic theory of gases. This study aimed to develop a CT test and assess the level of students' critical thinking skills.

METHODS

This research was the development of research using the 4D model. The 4D model consists of four stages, including define, design, develop, and analyze. The summary of this model as shown in Figure 1.

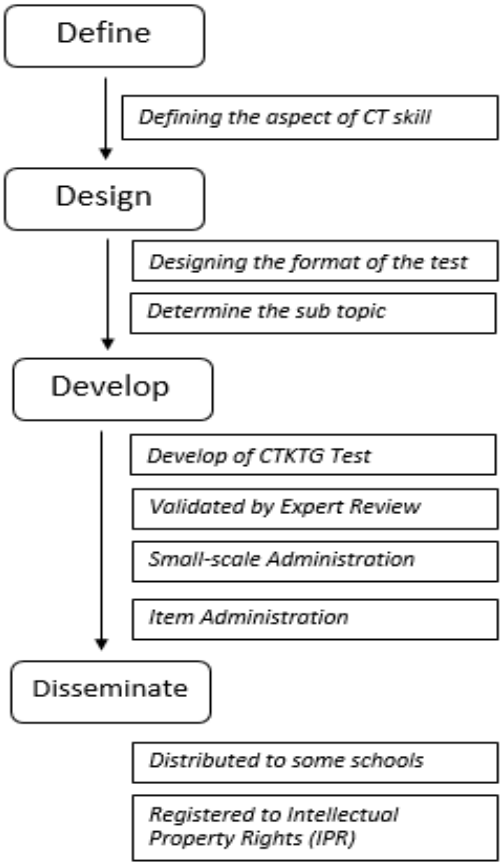

Figure 1. Summary of Research Methodology

Define

The first stage in developing the CTKTG test was defining critical thinking (CT) and selecting the CT skills that should be targeted in the test. Table 1 includes the test from any researchers collected by (Tiruneh et al., 2017),

Table 1 Critical Thinking Test by Researchers

CT Instrument	Targeted CT components
CCTT-Level Z	Analysis, evaluation, deduction, introduction, and overall reasoning skill
CCTT-Level Z	Induction, deduction, credibility, prediction and experimental planning, fallacies, and assumption identification
Ennis-weir CT essay test	Getting the point, identifying reasons and assumption, stating one's point of view, offering a good reason, seeing other possibilities, and responding appropriately to and/or avoiding argument weakness
HCTA	Verbal reasoning, argument analysis, hypothesis testing, likelihood/uncertainty analysis, and problem-solving and decision-making
Watson-Glaser Critical Thinking Appraisal	Inference, recognition of assumptions, deduction, interpretation, and evaluation of arguments

Design

The second stage was to design the format of the items used and the topic in physics. In this study we used open-ended format. We designed the CTKTG test based on the aspect of CT, indicator, and sub-topic. We also designed the criteria of students' CT level on the Kinetic theory of gases.

Develop

The third stage was to develop items with the CT component that is matched with the topic with the kinetic theory of gases and then tested on a small number of students. The CTKTG test was initially tested in four sample groups: interviews with the expert review (N = 3), professional physics teachers (N = 2), and graduate school students (N = 2), students from secondary schools (N = 29).

All items were reviewed by experts with following the criteria by Dawit Tibebu Tiruneh et al., (2017): (a) Are the items suitable for measuring CT skills in the desired domain? (b) Is the item statement clear, complete, and suitable for the participant?

After reviewing the component, the reviewer asked to do the content

validation. Content validation is one of the psychometric methods that aimed to assess the intended to be measured precisely or not (Cheng et al., 2016). This involved subjective opinions of "experts" about items that are judged by three categories: "important," "useful, but not important," or "unnecessary."

In assessing items that were "important", we can calculate it using the following formula (1) using the content validity ratio (CVR). Items that are considered "important" were then inserted into the final instrument, while items that "fail" reach the critical level removed(Ayre & Scally, 2014).

$$CVR = \frac{n_e - \frac{N}{2}}{\frac{N}{2}} \quad (1)$$

n_e is the number of panelists indicating “essential” and N is the total number of expert reviews. The minimum value of CVR, as shown in table 2,

Table 2. Minimum Value of CVR

No. of Expert Review	Minimum value
5	.99
6	.99
7	.99
8	.75
9	.78
10	.62
11	.59
12	.56
13	.54
14	.51
15	.49
20	.42
25	.37
30	.33
35	.31
40	.29

Two physics professors, one doctor, two magister students in the Graduate School Program at Yogyakarta State University, and two professional physics teachers were asked to review the 10 items. The review process of each item based on the accuracy of information and clarity of diagrams, phrases or words.

Small-scale paper-pencil administration

After the review process has been finished based on expert advice, the CTKTG items were administrated to a small group of students (N=29). The main purpose of this test is to determine whether the response can be assessed based on the assessment guide developed, and obtain an estimate of the time needed to complete the test.

Item Administration

The last step was to conduct a large-scale trial after going through the developing stage. The test was modified based on the revised results in the initial test. After that, the CTKTG test was given to a group of students in class XI, science students (N = 55).

The administration of the test lasted in 90 minutes. After incorporating all the revisions, the test was administered to physics students (N= 55) in the science class of Senior High School in Yogyakarta. Item administration was following a step by Tiruneh, De Cock, Weldeslassie, Elen, & Janssen, (2017), before began the test the researcher conveyed to the students the purpose of the test, general direction on how to answer the item, and instructions for taking the test seriously and being told about the time took about one hour to complete.

RESULTS AND DISCUSSION

Define

The result of this stage is the design of critical thinking components. Component of critical thinking skills for the CTKTG test is compiled based on the Ennis-weir CT essay test after reviewing all the tests mentioned above about the criteria by the author. The test focused on the following elements of CT skills: reasoning, argument analysis, hypothesis testing, likelihood and uncertainty analysis, and decision-making.

Design

The result of this stage was to design the format of the items used and the topic in physics. The CTKTG test based on the aspect of CT, indicator, and subtopic as shown in Table 3,

Table 3 Component of CT

Component of CT	Indicator	Sub-topic
Basic classification	Focus on the problems	Pressure in an ideal gas
	Analyze	General

Component of CT	Indicator	Sub-topic
	arguments	equation of ideal gas
Building Basic Skill	Consider the procedure for finding evidence	Boyle-Gay Lussac's Law
	Involves a little guess	Boyle-Gay Lussac's Law
Making the conclusion	Use logical conditions to make conclusions	Boyle-Gay Lussac's Law
	Identify and use distinctive features or patterns in the data to draw conclusions	Charles's Law
Advance clarification	Know the content validity of a definition	The kinetic energy of ideal gas
	Identifying assumptions	Root mean square velocity
Implement strategies and tactics	Understand the total problems and take action	Pressure in an ideal gas
	Choose the criteria for considering possible solutions	The kinetic energy of ideal gas

Students were asked to complete 10 questions according to aspects of CT skills. All of the items were also validated by experts. Assessment of student skills based on the rubric using levels 0 - 4. The table below shows the skill level of students based on their test results,

Table 4 Level of Critical Thinking Skill

Range	Level
x <2.40	Very low
2.40 ≤ x < 2.80	Low
2.80 ≤ x < 3.20	Average
3.20 ≤ x < 3.60	High
3.60 ≤ x ≤ 4.00	Very high

Develop

The results of this stage were content validation by an expert review and the review on small paper administration, CTKTG item with reliability and validity scale, the level of difficulty and discrimination.

The reviewers argued that the CTKTG items were suitable to assess the targeted CT skills on The Kinetic Theory of Gases. Any feedback from them about the items and some revise all of the items.

Analysis of students' responses showed that there were no significant revisions to CTKTG items. Besides, several relevant answers were found, so that revisions to the assessment guidelines were made.

Table 5. Item of CTKTG

The aspect of CT: Basic classification
Indicator of CT: Focus on the problems **Bloom Taxonomy:** C4 Analysis **Question 1:** Every year, the hot air balloon festival is always held in Europe. All hot air balloons are required to meet good flight requirements. One requirement is to use a quality heater. Participants are prohibited from using a bad heater because it can be fatal during flight. Analyze the focus of the problem in the case above! Give reasons for the problem.
Indicator of CT: Analyze arguments **Bloom Taxonomy:** C4 Analysis **Question 2:** Rina wants to be a professional chef. He then enrolled in one of the cooking training institutions. When cooking food, Rina is told by her teacher to close the heated pot. The teacher said that by closing the pan the food would quickly cook. If it is assumed that the gas is ideal, do you agree with the suggestion? Give your reasons.
The aspect of CT: Building Basic Skill
Indicator of CT: Consider the procedure for finding evidence **Bloom Taxonomy:** C5 Prediction **Question 3:** Toni wants to experiment on the concept of an ideal gas. In a closed laboratory, he heated the temperature of the gas so that it changed to 2 times all. If the gas volume is constant, then predict the change in pressure measured by Toni in accordance with the procedure in the ideal gas law concept?
Indicator of CT: Involves a little guess **Bloom Taxonomy:** C4 Analysis **Question 4:** Is it correct or incorrect, if it is said that every two types of ideal gas that are heated will produce the same kinetic energy? Explain your opinion!
The aspect of CT: Making the conclusion
Indicator of CT: Use logical conditions to make conclusions **Bloom Taxonomy:** C4 Analysis **Question 5:** Anto and Budi are conducting simulation experiments on the ideal gas concept. Andi conducted Boyle's legal trial while Budi conducted a legal trial Gay Lussac. The results of these experiments are:

To increase the pressure, Andi must reduce the volume and Budi must increase the temperature.
Analyze the results of the experiment and draw conclusions that show the relationship between pressure, temperature, and volume

Indicator of CT: Identify and use distinctive features or patterns in the data to draw conclusions
Bloom Taxonomy: C5 Conclude
Question 6
Doni experimented with a simulation of Charles's law on ideal gas. The results of the experiments are then copied in the table below:

No	Temperature (K)	Volume (cm^3)	V/T
1	230.15	20.07	11,467
2	215.15	18.76	11.468
3	192.15	16.75	11.471
4	168.15	14.66	11.469
5	140.15	12.22	11.468

Based on these data, determine the right conclusions and write the equation of Charles's law from this experiment!

The aspect of CT: Advance clarification
Indicator of CT: Know the content validity of a definition
Bloom Taxonomy: C5 Validating
Question 7:
Determine which statement is right!
A. If the temperature of the gas in a closed container is bigger than before, so the average velocity of the gas is also bigger than before.
B. If the average velocity of the gas is before than before, the pressure of the gas will be smaller than before.

Indicator of CT: Identifying assumptions
Bloom Taxonomy: C5 Predicting
Question 8:
Rico experimented to determine the relative velocity of the gas. If there are two types of gas assuming the two gases have the same density and pressure. If the volume of container B is twice container A, then determine the relative speed of gas B!

The aspect of CT: Implement strategies and tactics
Indicator of CT:
Bloom Taxonomy: C4 Analysis
Question 9:
Joni wants to join in the hot air balloon race. He plans to buy several supporting devices such as heating machines. However, Joni was confused about how to determine a good heater, whether it produces the most heat or not. He then concluded that there was no need for the most heat-producing machines. This is because it will cause around the balloon to become hot and wasteful of energy. Also, there is the help of wind encouragement so the hot air balloon can float upward. Determine the problem contained in the statement! Is Joni doing the right thing? Explain

Indicator of CT: Choose criteria for considering possible solutions
Bloom Taxonomy: C4 Analysis
Question 10:
A scientist wants to use the ideal gas concept to produce large kinetic energy. Then He calculated to find great energy. If the initial condition of pressure is 100 Pa, the temperature is 300 K and the volume is 1 m^3, determine the appropriate solution chosen by the scientist.
Solution 1: change the pressure to 50 Pa, replace the volume become 0.5 m^3, make the temperature constant
Solution 2: make the pressure constant, replace the volume be 0.5 m3 and reduce the temperature to be 200 K
Solution 3: make the pressure be constant and volume, and raise the temperature to 400 K.
In your opinion, which solution should be chosen by these scientists to produce large kinetic energy? Analyze the case and give your reason.

Internal Consistency/Reliability

Internal consistency is the most basic part of the measurement which refers to the homogeneity of the items on the test (Hajcak, Meyer, & Kotov, 2017). In other words, homogeneity or internal consistency is a level that shows the extent to which an item can measure the same thing (Davenport, Davison, Liou, & Love, 2015). We measured the internal consistency by Cronbach alpha formula:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_i V_i}{V_t}\right] \qquad (2)$$

Where n = number of items, Vt = variance of the total scores and Vi = variance of the item's score. In this test, we found the α =.89 (good) based on Table 6,

Table 6. Internal Consistency Cronbach

Cronbach's alpha	Internal consistency
$\alpha \geq .9$	Excellent
$.9 > \alpha \geq .8$	Good
$.8 > \alpha \geq .7$	Acceptable
$.7 > \alpha \geq .6$	Questionable
$.6 > \alpha \geq .5$	Poor
$.5 > \alpha$	Unacceptable

Validity Test

Validity testing was used to show how accurate the instrument is. In other words, it is the degree of accuracy of a valid test item that precisely measures what you want measured (Siregar, Surya, & Syahputra, 2017). Wang (2017) said that validity indicates whether the test developed is effective, how effective the test is, and how the test characteristics are measured. In this study, we used Pearson's product-moment correlation coefficient r (S) to relationship value between the results. To determine the items are valid or not, we can compare the Pearson product-moment correlation coefficient r(S) with r_{table} (.2241) using SPSS. If r(s) of the item $>r_{table}$, the items are valid.

Table 7. Items Validity

Number of Item	r(S)	Validity result
1	.746	valid
2	.717	valid
3	.657	valid
4	.607	valid
5	.762	valid
6	.827	valid
7	.586	valid
8	.595	valid
9	.776	valid
10	.821	valid

We also determined the validity of the test using Content Validity Ratio (CVR) by expert judgment and compute the index based on Lawshe's formula. The results as shown in table 8,

Table 8. CVR of Item by Expert Review

Exp	Item									
	1	2	3	4	5	6	7	8	9	10
1	3	3	3	3	3	3	3	3	3	3
2	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3
4	3	3	3	3	3	3	3	3	3	3
5	3	3	3	3	3	3	3	3	3	3
6	3	3	3	3	3	3	3	3	3	3
7	3	3	3	3	3	3	3	3	3	3
ne	7	7	7	7	7	7	7	7	7	7
CVR	1	1	1	1	1	1	1	1	1	1

Item Difficulty

The difficulty of items is an important parameter for each new item added to the test (Loukina, Yoon, Sakano, Wei, & Sheehan, 2016). It is very important in education for teachers and item makers (El Masri, Ferrara, Foltz, & Baird, 2017). The difficulty of the question is the measure of the percentage of students who answer the question correctly and the value for the index of difficulty range 0% (very difficult) to 100% (very easy)(Tomak, Bek, & Cengiz, 2016). In other words, the difficulty of the item is the comparison of the number of students who answer right from wrong (X. Bai & Ola, 2017).To compute item difficulty of the test using a program existing now (QUEST). The index range difficulty level and the result of the test, as shown in table 9 and table 10,

Table 9. Index Range of Difficulty Level

Index	Difficulty Scale	Decision
$b \geq 2$	Very Difficult	To be discarded
$1 < b \leq 2$	Difficult	To be revised
$-1 < b \leq 1$	Moderate	Good item
$b < -2$	Easy	To be revised

The statistic for the CTKTG items is shown in table 10,

Table 10. Difficulty of Items

Item	Index	Difficulty Scale
1	.62	Moderate
2	.65	Moderate
3	.37	Moderate
4	.52	Moderate
5	.49	Moderate
6	.59	Moderate
7	.46	Moderate
8	.28	Moderate
9	.66	Moderate
10	.51	Moderate

Item Discriminant

The difficulty of the item is important in maintaining or rejecting the test items given. However, information about item difficulties is not enough, we must also consider discriminatory items (Perkins & Frank, 2018). Item discrimination is very important in determining the quality of the item. This value provides information about the differences in abilities measured by each individual based on the tests made (Khairani & Shamsuddin, 2016).

It is an index that shows how well items can distinguish people with certain levels of ability, especially students in high and low level (Tasca et al., 2016). Ten is used to measure the extent to which an item can predict the overall performance of a test (Xue Bai & Ola, 2017). The following rules of a discriminant level similar to that used by (Quaigrain & Arhin, 2017) as shown in table 11:

Table 11 Index Range of Discriminant Level

Index Range	Discrimination Level
0 < 0.19	The poor item should be eliminated or completely revised
0.20 ≤ x < 0.29	The marginal item needs some revision
0.30 ≤ x < 0.39	Reasonably good item but possibly need little revision for improvement
x ≥ 0.40	Very good item

The discrimination index (ID) is calculated using the following formula(Xue Bai & Ola, 2017),

$$ID = \frac{(\overline{X_C} - \overline{X_w})}{Std}\sqrt{p(1-p)} \qquad (3)$$

Where Xc is the mean total score for students who have responded correctly to the item; Xw is the mean total score for students who have responded incorrectly to the item; p is the item difficulty for the item and Std is the standard deviation of the total exam scores. The discrimination index is shown in Table 12,

Table 12. Items Discrimination

Item	Discriminant Index	Discriminant Level
1	.75	Very good item
2	.72	Very good item
3	.66	Very good item
4	.61	Very good item
5	.76	Good item
6	.83	Very good item
7	.59	Very good item
8	.58	Very good item
9	.78	Very good item
10	.82	Very good item

Disseminate

We measure the difficulty level by the test was given to the participant (N = 55). The difficulty indices for the CTKTG items from 0.58 to 0.82. Most items are at a moderate level and the discriminant level is very good. We know that all goods are good items. The value of the validity of the instrument can be obtained from the relationship or correlation between the instrument that was developed with the instrument that already exists and has previously been considered valid. In this study, we use SPSS to determine r-value to show convergent validity (Pearson correlation) and a Kolmogorov-Smirnov.

Test show that test distribution is normal. The summary of r value from SPSS for all items is shown in table 1. Based on r table, we know that with N = 55 and ∝ = .05, r table is = .2241, so all items are valid.

Based on the test, students were given ten questions according to the aspect of critical thinking skills. The result is revealed in Table 12. The table shows the level of their answers in the test,

Table 13. The Level of Critical Thinking Skill of the Students

Component of CT	Mean	SD	Category
Basic classification	2.84	1.12	Average
Building Basic Skill	1.49	1.54	Very low
Making the conclusion	1.95	1.63	Very low
Advance clarification	1.55	1.66	Very low
Implement strategies and tactics	1.37	1.58	Very low
Overall	**1.84**	**0.32**	**Very low**

Among the ten questions that administrated on the students, answers of the students in basic classification show the highest mean of 2.84 (average). Moreover, the answers to implement strategies and tactics present the lowest mean of 1.37 (very low). It can be gleaned from the table that the students have a very low level of critical thinking skill (mean = 1.84, SD = 0.32). These findings are similar to (Azis, Muhammad, & Yusuf, 2016) which found that the highest aspect possessed by students was basic classification (3,375) and the lowest advance clarification (1,875).

Based on the reliability scale (α = 0.89), the open-ended form was more effective than others, such as multiple choice only

0.78 (Hwang & Chen, 2017). Similar results found by Harjo, Kartowagiran, & Mahmudi (2019), the internal reliability with the open-ended format of their study shows α = 0.94. Besides that, through open-ended tests, we can explore, explain or confirm students' knowledge more deeply than any other test. We also registered all the items to intellectual property rights (IPR).

CONCLUSION AND SUGGESTION

All items are valid and the test distribution is normal. Item difficulty on level moderate and item discrimination on a level very good. So the CTKTG test is a good instrument for measuring CT skill in the kinetic theory of gases. But, to obtain more valid results, it requires a larger number of respondents and varies from several levels of student education. Based on the results and discussion, the level of students in CT skills is very low. It shows that aspects implement strategies and tactics are the most difficult aspect of students' critical thinking skills and basic classification is the easy aspect.

ACKNOWLEDGMENT

We thank Prof. Dr. Mundilarto, Prof. Dr. Herman, and Dr. Supahar as an expert review from Graduate School Program at Yogyakarta State University that has assessed and validated this instrument so that it is suitable for use in learning.

AUTHOR CONTRIBUTIONS

RP develops ideas and designs research. RY collects data. JA performs statistical calculations. DR compiled the research results. SS conducts discussion and article info.

REFERENCES

Ayre, C., & Scally, A. J. (2014). Critical Values for Lawshe's Content Validity Ratio: Revisiting the Original Methods of Calculation. *Measurement and Evaluation in Counseling and Development*. *47*(1), 79-86. https://doi.org/10.1177/0748175613513808

Azis, A., Muhammad Aqil Rusli, A., & Yusuf, M. (2016). Critical Thinking Skill of Student Through Top Down Approach in Physics Learning. *ICMSTEA 2016*.

Bai, X., & Ola, A. (2017). A Tool for Performing Item Analysis to Enhance Teaching and Learning Experiences. *Issues in Information Systems*, *18*(1), 128–136.

Bai, Xue, & Ola, A. (2017). A Tool for Performing Item Analysis to Enhance Teaching and Learning Experiences. *Issues in Information Systems*, *18*(1), 128–136.

Carter, A. G., Creedy, D. K., & Sidebotham, M. (2016). Development and Psychometric Testing of the Carter Assessment of Critical Thinking in Midwifery (Preceptor/Mentor version). *Midwifery*. 34(1), 141-149. https://doi.org/10.1016/j.midw.2015.12.002

Cheng, P. G. F., Ramos, R. M., Bitsch, J. A., Jonas, S. M., Ix, T., See, P. L. Q., & Wehrle, K. (2016). Psychologist in a Pocket: Lexicon Development and Content Validation of a Mobile-Based App for Depression Screening. *JMIR MHealth and UHealth*. *4*(3), e88. https://doi.org/10.2196/mhealth.5284

Davenport, E. C., Davison, M. L., Liou, P. Y., & Love, Q. U. (2015). Reliability, Dimensionality, and Internal Consistency as Defined by Cronbach: Distinct Albeit Related Concepts. *Educational Measurement: Issues and Practice*. *34*(4),4-9. https://doi.org/10.1111/emip.12095

El Masri, Y. H., Ferrara, S., Foltz, P. W., & Baird, J. A. (2017). Predicting Item Difficulty of Science National Curriculum Tests: The Case of Key Stage 2 Assessments. *Curriculum Journal*. *28*(1), 59-82. https://doi.org/10.1080/09585176.2016.1232201

Fuad, N. M., Zubaidah, S., Mahanal, S., & Suarsini, E. (2017). Improving Junior High Schools' Critical Thinking Skills Based on Test Three Different

Models of Learning. *International Journal of Instruction*, *10*(1), 101–116.

Gelerstein, D., Río, R. del, Nussbaum, M., Chiuminatto, P., & López, X. (2016). Designing and Implementing A Test for Measuring Critical Thinking in Primary School. *Thinking Skills and Creativity*. 20, 40-49. https://doi.org/10.1016/j.tsc.2016.02.002

Ghazivakili, Z., Norouzi Nia, R., Panahi, F., Karimi, M., Gholsorkhi, H., & Ahmadi, Z. (2014). The Role of Critical Thinking Skills and Learning Styles of University Students in Their Academic Performance. *Journal of Advances in Medical Education & Professionalism*. *2*(3) 95-102.

Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2015). A Review And Comparison of Diagnostic Instruments to Identify Students' Misconceptions in Science. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(5), 989–1008. https://doi.org/10.12973/eurasia.2015.1369a

Hajcak, G., Meyer, A., & Kotov, R. (2017). Psychometrics and the Neuroscience of Individual Differences: Internal Consistency Limits Between-Subjects Effects. *Journal of Abnormal Psychology*. *126*(6), 823-834. https://doi.org/10.1037/abn0000274

Harjo, B., Kartowagiran, B., & Mahmudi, A. (2019). Development of Critical Thinking Skill Instruments on Mathematical Learning High School. *International Journal of Instruction*, *12*.(4), 149–166.

Hashim, A., & Samsudin, N. S. B. (2019). Level of Critical Thinking Skills Among Students of Tahfiz School. *International Journal of Academic Research in Business and Social Sciences*, *9*(1), 926-931. https://doi.org/10.6007/IJARBSS/v9-i1/5491

Hu, W., Jia, X., Plucker, J. A., & Shan, X. (2016). Effects of a Critical Thinking Skills Program on the Learning Motivation of Primary School Students. *Roeper Review*. *38*(2), 70-83. https://doi.org/10.1080/02783193.2016.1150374

Huber, C. R., & Kuncel, N. R. (2015). Does College Teach Critical Thinking? A Meta-Analysis. *Review of Educational Research*. *86*(2), 431-468. https://doi.org/10.3102/0034654315605917

Hwang, G. J., & Chen, C. H. (2017). Influences of an Inquiry-Based Ubiquitous Gaming Design on Students' Learning Achievements, Motivation, Behavioral Patterns, and Tendency Towards Critical Thinking and Problem Solving. *British Journal of Educational Technology*. *48*(4), 950-971. https://doi.org/10.1111/bjet.12464

Kettler, T. (2014). Critical Thinking Skills Among Elementary School Students: Comparing Identified Gifted and General Education Student Performance. *Gifted Child Quarterly*. *58*(2), 127-138. https://doi.org/10.1177/0016986214522508

Khairani, A. Z., & Shamsuddin, H. (2016). Assessing Item Difficulty and Discrimination Indices of Teacher-Developed Multiple-Choice Tests. In *Assessment for Learning Within and Beyond the Classroom*. https://doi.org/10.1007/978-981-10-0908-2_35

Klavir, R., & Hershkovitz, S. (2014). Teaching and Evaluating 'Open - Ended' Problems. *ResearchGate*.

Kong, S. C. (2014). Developing Information Literacy and Critical Thinking Skills Through Domain

Knowledge Learning in Digital Classrooms: An Experience of Practicing Flipped Classroom Strategy. *Computers and Education*. 78, 160-173. (https://doi.org/10.1016/j.compedu.2014.05.009

Lee, W., & Lutz, B. D. (2016). An Anchored Open-Ended Survey Approach in Multiple Case Study Analysis. *ASEE Annual Conference & Exposition*. New Orleans: Louisiana.

Lee, Y. H. (2018). Scripting to Enhance University Students' Critical Thinking in Flipped Learning: Implications of The Delayed Effect on Science Reading Literacy. *Interactive Learning Environments*. *26*(5), 569-582. https://doi.org/10.1080/10494820.2017.1372483

Liu, O. L., Mao, L., Frankel, L., & Xu, J. (2016). Assessing Critical Thinking in Higher Education: The Heighten^{tm} Approach and Preliminary Validity Evidence. *Assessment and Evaluation in Higher Education*. *41*(5), 677-694. https://doi.org/10.1080/02602938.2016.1168358

Loukina, A., Yoon, S., Sakano, J., Wei, Y., & Sheehan, K. M. (2016). Textual Complexity as a Predictor of Difficulty of Listening Items in Language Proficiency Tests. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*.

Mapeala, R., & Siew, N. M. (2015). The Development and Validation of a Test of Science Critical Thinking for Fifth Graders. *SpringerPlus*. *4*(741). https://doi.org/10.1186/s40064-015-1535-0

Matsun, M., Sunarno, W., & Masykuri, M. (2017). Penggunaan Laboratorium Riil dan Virtuil pada Pembelajaran Fisika dengan Model Inkuiri Terbimbing ditinjau dari Kemampuan Matematis dan Keterampilan Berpikir Kritis. *Jurnal Pendidikan Fisika*. *4*(2), 137-152. https://doi.org/10.24127/jpf.v4i2.541

Nasution, N. E. A., Harahap, F., & Manurung, B. (2017). The Effect of Blended Learning on Student's Critical Thinking Skills in Plant Tissue Culture Course. International Journal of Science and Research. *International Journal of Science and Research (IJSR)*, *6*(11), 1469–1473. https://doi.org/https://doi,org/10.21275/ART20171836

Pascarella, E. T., Martin, G. L., Hanson, J. M., Trolian, T. L., Gillig, B., & Blaich, C. (2014). Effects of Diversity Experiences on Critical Thinking Skills Over 4 Years of College. *Journal of College Student Development*. *55*(1), 86-92. https://doi.org/10.1353/csd.2014.0009

Perkins, K., & Frank, E. (2018). An Item Analysis and a Reliability Estimate of a Classroom Kinesiology Achievement Test. *Online Submission-ERIC*.

Popping, R. (2015). Analyzing Open-ended Questions by Means of Text Analysis Procedures. *BMS Bulletin of Sociological Methodology/ Bulletin de Methodologie Sociologique*. *128*(1), 23-39. https://doi.org/10.1177/0759106315597389

Purwati, R., Hobri, H., & Fatahillah, A. (2016). Analisis Kemampuan Berpikir Kritis Siswa dalam Menyelesaikan Masalah Persamaan Kuadrat pada Pembelajaran Model Creative Problem Solving. *KadikmA*, *7*(1), 84–93.

Quaigrain, K., & Arhin, A. K. (2017). Using Reliability and Item Analysis to Evaluate a Teacher-Developed Test in Educational Measurement and Evaluation. *Cogent Education*. *4*(1), 1-11. https://doi.org/10.1080/2331186X.20

17.1301013

Rowland, K., Lovelace, M. J., Saunders, M. J., Caruso, J. P., & Israel, N. (2016). Citizen Science: The Small World Initiative Improved Lecture Grades and California Critical Thinking Skills Test Scores of Nonscience Major Students at Florida Atlantic University. *Journal of Microbiology & Biology Education*. *17*(1), 156-162. https://doi.org/10.1128/jmbe.v17i1.1011

Schonlau, M., & Couper, M. P. (2016). Semi-Automated Categorization of Open-Ended Questions. *10*(2), 143-152. *Doi.Org*. https://doi.org/10.18148/srm/2016.v10i2.6213

Shin, S., Jung, D., & Kim, S. (2015). Validation of a Clinical Critical Thinking Skills Test in Nursing. *Journal of Educational Evaluation for Health Professions*. *12*(1), 1-6. https://doi.org/10.3352/jeehp.2015.12.1

Siregar, T. P., Surya, E., & Syahputra, E. (2017). Quality Analysis of Multiple Choice Test and Clasical Test at X Grade Students of Senior High School. *Jurnal IJARIIE University of Medan*, *3*(2), 2153–2159.

Stupple, E. J., Maratos, F. A., Elander, J., Hunt, T. E., Cheung, K. Y., & Aubeeluck, A. V. (2017). Development of the Critical Thinking Toolkit (CriTT): A measure of student attitudes and beliefs about critical thinking. *Thinking Skills and Creativity*, *23*, 91–100. https://doi.org/https://doi.org/10.1016/j.tsc.2016.11.007

Szenes, E., Tilakaratna, N., & Maton, K. (2015). The Knowledge Practices of Critical Thinking. In *The Palgrave Handbook of Critical Thinking in Higher Education*. Palgrave Macmillan, New York https://doi.org/10.1057/9781137378057

Tasca, G. A., Cabrera, C., Kristjansson, E., MacNair-Semands, R., Joyce, A. S., & Ogrodniczuk, J. S. (2016). The Therapeutic Factor Inventory-8: Using Item Response Theory to Create a Brief Scale for Continuous Process Monitoring for Group Psychotherapy. *Psychotherapy Research*. *26*(2), 131-145. https://doi.org/10.1080/10503307.2014.963729

Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J., & Janssen, R. (2017). Measuring Critical Thinking in Physics: Development and Validation of a Critical Thinking Test in Electricity and Magnetism. *International Journal of Science and Mathematics Education*. *15*(4), 663-682. https://doi.org/10.1007/s10763-016-9723-0

Tomak, L., Bek, Y., & Cengiz, M. A. (2016). Graphical Modeling for Item Difficulty in Medical Faculty Exams. *Nigerian Journal of Clinical Practice*. https://doi.org/10.4103/1119-3077.173701

Tran, V. T., Porcher, R., Falissard, B., & Ravaud, P. (2016). Point of Data Saturation was Assessed Using Resampling Methods in a Survey with Open-Ended Questions. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2016.07.014

Vieira, R. M., & Tenreiro-Vieira, C. (2016). Fostering Scientific Literacy and Critical Thinking in Elementary Science Education. *International Journal of Science and Mathematics Education*, *14*(4), 659-680. https://doi.org/10.1007/s10763-014-9605-2

Wang, F. (2017). Research on the Construction of Test Questions Bank Based on Educational Measurement Theory. *International Conference on Frontiers in Educational*

Technologies and Management Sciences, (pp. 341–344). Beijing: Research Institute of Management Science and Industrial Engineering.

Wismath, S., Orr, D., & Zhong, M. (2014). Student Perception of Problem Solving Skills. *Transformative Dialogues: Teaching & Learning Journal*, *7*(3), 1-17.

Yan, K., Yamada, H., Takagaki, M., & Koizumi, R. (2019). Status of Mathematics Tests in the National Center Test for University Admissions in Japan and How to Improve It. *Juntendo Medical Journal*, *65*(3), 255–260.

APPENDIX

Indicator of CT	Taxonomy bloom	Question	Key	Guide Score
Implement strategies and tactics	C4 Analysis	A scientist wants to use the ideal gas concept to produce large kinetic energy. Then He calculated to find great energy. If the initial condition of pressure is 100 Pa, the temperature is 300 K and the volume is 1 m^3, determine the appropriate solution chosen by the scientist. Solution 1: change the pressure to 50 Pa, replace the volume become 0.5 m^3, make the temperature constant Solution 2: make the pressure constant, replace the volume be 0.5 m3 and reduce the temperature to be 200 K Solution 3: make the pressure be constant and volume, and raise the temperature to 400 K. In your opinion, which solution should be chosen by these scientists to produce large kinetic energy. Analyze the case and give your reason.	Solution 3 Based on the concept of average kinetic energy, the greater the temperature has the greater the energy.	Score 1: If the answer and the reason are wrong. Score 2: • If the answer is correct, but the reason is wrong or not following the key or the answer key. • If the answer is wrong, but the reason is correct or following the key referred to like the answer key. Score 3: • If the answer is correct, the reason is not in accordance with the key or the answer key. Score 4: • If the correct answer is accompanied by the right reason according to the key or the answer key. • If the answers and reasons can be categorized correctly but not listed in the answer key.