Momentum: Physics Education Journal
5 (2), 2021, 194-208

Available at: ejournal.unikama.ac.id/index.php/momentum

Developing authentic assessment instrument based on multiple representations to measure students' critical thinking skills

Muhammad Minan Chusni [1, a] *, Suherman Suherman [2, b]

[1] UIN Sunan Gunung Djati. Kampus 2 Jl. Cimencrang, Kota Bandung, Jawa Barat 40292, Indonesia
[2] University of Szeged. Szeged, Dugonics tér 13, 6720 Hungary
[a] minan.chusni@uinsgd.ac.id; [b] suherman@edu.u-szeged.hu
* Corresponding Author.

Received: 25 June 2021; Revised: 24 July 2021; Accepted: 30 July 2021

Abstract: The purpose of this study was to produce an instruments test used to measure students' critical thinking skills in natural science learning. This research uses a 4-D development model (define, design, develop, and disseminate) involving 118 students at the develop stage and 60 students at the disseminate stage. The instrument developed was an essay test based on multiple representations. Validity was proved by using CVI, and reliability was estimated by using the Item Response Theory. The results showed that the instrument had a very good foreign exchange value. This is reflected in the Aiken V scores on the aspects of substance, construct, language and appearance, respectively, about 1.00, 1.00, 1.00, and 1.00. According to Rasch analysis, the instruments has meet the assumption test for 14 items which is unidimensional, the local independence assumption test, and the parameter invariance assumption test. According to OUTFIT MNSQ Value, the items are fit with PCM 1-PL which functions normally in making measurements. Reliability estimated for items shows a very high consistency of measurement of 0.97 and for person shows a high consistency of measurement of 0.86. The results of the student's CTS measurement showed that the average score was 65.50, with a distribution of high, medium, and low abilities, respectively, about 16.67%, 63.33%, and 20.00%. Thus, according to these results, an authentic assessment based on multiple representations is suitable to measure students' critical thinking skills.

Keywords: authentic assessment, critical thinking skills, multiple representations, Rasch model.

How to Cite: Chusni, M. M., & Suherman, S. (2021). Developing authentic assessment instrument based multiple representations to measure students' critical thinking skills. *Momentum: Physics Education Journal*, *5*(2), 194-208. https://doi.org/10.21067/mpej.v5i2.5790

Introduction

The globalization era has been developed in several fields such as economics, politics, technology, and education. Especially in the field of education, schools are struggling to improve the quality of learning. The quality of learning is determined by, among other things, the quality of the assessment carried out by the teacher in the learning process. Assessment is one of the most important things that must be done in the learning process, (Maba, 2017) stated that assessment and learning are two things that cannot be separated. This is in accordance with the Permendikbud RI Number 23 of 2016 (Menteri Pendidikan dan Kebudayaan Republik Indonesia, 2016) concerning Education Assessment Standards and Permendikbud RI Number 43 of 2019 (Menteri Pendidikan dan Kebudayaan Republik Indonesia, 2019) concerning the Implementation of Examinations organized by the National Education and Examination Unit which states that one of the principles of assessment is integrated, which means that assessment is an integral component of learning activities. Hopfenbeck

et al. (2018) also emphasized that assessment is an integral part of science learning and has a major role in obtaining information related to what is known, what is able to do, and what students learn. Based on the description, it can be concluded that the assessment is carried out in an integrated way with the learning process so that the assessments carried out can be used as feedback, direct learning, and students and learning evaluations. The more quality the learning assessment activities, the more it helps the teacher to understand the strengths and weaknesses of students.

The 2013 curriculum emphasizes students to be more active in the learning process which causes their assessment to also experience adjustments; from assessment of learning outcomes to assessment processes that consider attitudes, behavior, and morals as an inseparable part. Authentic assessment has a strong relevance to the scientific approach, especially in the implementation of learning required by the 2013 Curriculum. Permendikbud RI Number 23 of 2016 requires that the assessment of the learning process is carried out using an authentic assessment approach that assesses the readiness of students, the process, and the learning outcomes as a whole (Menteri Pendidikan dan Kebudayaan Republik Indonesia, 2016). Authentic assessment defines as a meaningful measurement of student learning outcomes in the aspects of attitudes, skills and knowledge. The term authentic is a synonym for genuine, real, or valid. Authentic assessment is often contradicted by assessments that use standard norms-based, namely: multiple choices, true-false, matching, or short answer test standards. These forms are considered less efffective to provide a real picture of the students' thinking processes and the extent to which they understand the material presented.

However, so far, the assessment process in classroom learning is still imbalanced. The assessment process, in general, aims to collect and interpret evidence with the intention of making considerations about the achievement of student learning outcomes or what is known as assessment of learning. In addition, the assessment is generally carried out with the intention of determining the extent to which the learning outcomes of students have been achieved or what is known as assessment for learning. On the other hand, the assessment should be able to have a significant impact on the guidance, management and implementation of education system. Thus, there is a need for a more specific assessment called assessment as learning. This type of assessment is the process of developing and supporting metacognitive of the students. Students are included in an assessment process in which they monitor themselves. This type of assessment aims to prepare feedback descriptions for reflection and to learn self-monitoring. This type of assessment is designed to help students become more independent learners by continuously getting feedback and reflection in order to improve learning (Arends & Kilcher, 2010).

The purpose of authentic-based assessment is that students are able to think through science and establish higher-order thinking skills. In the 2013 curriculum, it is important for students to have higher order thinking skills, including critical thinking, because it can help students develop their intellectual potential, ability to evaluate systematically, and capability of arguing in an organized manner (Chusni et al., 2021)

Permendikbud number 23 of 2016 and number 43 of 2019 state that one of the objectives of learning assessment is to monitor and improve the assessment process in order to improve and balance aspects of attitudes, knowledge and skills to build hard skills and soft skills of students (Menteri Pendidikan dan Kebudayaan Republik Indonesia, 2016, 2019). Soft skills are terms that refer to personality, social skills, language skills, communication skills, negotiation skills, creative thinking and critical thinking skills. In fact, assessment instruments to measure the soft skills of students are still minimal. It can be seen from the National exam questions, especially science lessons, which mostly measure low-level thinking skills and the emphasis on memorization than thinking and problem-solving skills. Study from Mujib et al. (2018) states that the distribution of the science National Exam questions for the elementary school level in the range of Academic Year from 2014/2015 to 2016/2017 is C1 (50%, 50%, 35%), C2 (27.5%, 22, 5%, 40%), C3 (5%, 12.5%, 5%), C4 (17.5%, 15%, 20%), C5 (0%), and C6 (0%). Furthermore, a study conducted by Wijayanti et al. (2019) shows that the distribution of science National Exam preparation questions for the secondary school level is C1 (50%), C2 (22.5%), C3 (5%), C4 (0%), C5 (0%), and C6 (0%). Whereas, in the distribution of the Physics National Exam for High School questions, the teacher's evaluation regarding the distribution

questions was (C1) 0%, understanding (C2) 11.25%, applying (C3) 41.25%, analyzing (C4) 30%, evaluate (C5) 12.5% and create (C6) 5% (Yusrizal, 2016).

Based on Indonesia's rank in the Program for International Students Assessment (PISA) event for several years, it can be used as a standard for the low level of thinking of Indonesian students compared to other countries. Chusni et al. (2020) stated that based on the results of the study, Indonesian students were ranked 38 out of 41 PISA countries in 2000, with a score of 393. PISA in 2003, was ranked 37th out of 44 participating countries with a score of 395. PISA 2006, was ranked 50th out of 57 countries with a score of 393. Most recently, in the PISA assessment released on December 3, 2019, Indonesia was in the 75th position out of 80 countries that participated in the assessment. Indonesian students represented by 12,098 students and carried out using computer-based assessments obtained an average value for reading ability 371, mathematics ability 379 and science ability 396. According to previous research, from the results of the PISA test and evaluation, the performance of Indonesian students is still low (Kartianom & Ndayizeye, 2017; Nugrahanto & Zuchdi, 2019; Suprapto, 2016). Consecutively, the average achievement score of Indonesian students for science is ranked 70 out of 78 evaluated countries. The questions contained in PISA are in the form of questions with a high level of thinking that not only require the ability to memorize, but also critical thinking skills.

Critical thinking is an important skill for problem solving, education and learning. This skill provides a systematic approach in identifying problems, so that the most rapid solutions will be obtained in structured problem solving (Amalia & Wuryandani, 2020; Doleck et al., 2017). Critical thinking skills are valuable skills that can be generated from the learning process in schools. Teachers can make critical thinking skills as an outcome to be achieved during learning. However, learning in school does not always result in the competence of students who have critical thinking skills. Research by Paul et al. (1997) reveals that in the learning process, many teachers direct students to have the critical thinking as a result of learning. Therefore, they encourage students to practice critical thinking during the learning process. However, during its implementation, many teachers cannot define or distinguish the results of critical thinking from conceptual mastery. Especially in science learning, there are still a lot of evaluation questions on formative assessments that tend to require memory skills and solve problems through mathematical solutions in the form of multiple-choice questions, which are often contradicted with authentic assessments. The form of multiple-choice questions was chosen because this model makes it easier for teachers to distribute questions related to the material being studied. Besides, that it also easier for teachers to make assessments because in there is only one correct answer. However, the multiple-choice form of test requires close supervision to avoid the tendency of cheating. Besides, the multiple-choice form tends to be ineffective when applied as an evaluation tool to measure students' higher order thinking skills. Therefore, in this study, it is intended to conduct research on the development of authentic assessment instruments in the form of written essay questions that can be used to measure critical thinking skills in environmental change material. The written test in the form of an essay was chosen as explained by Brookhart and Nitko (2019), who states that the essay test provides more opportunities for students to display the ability to write, organize, express, and explain the relationship between ideas so that they are able to assess students' higher order thinking skills. Environmental change material was chosen because this material is one of the essential materials that must be mastered by grade VII students in science learning.

Encouragement to achieve critical thinking skills is needed so that students better understand the concepts being learned, and can apply them in various situations. Achieving high-order thinking skills such as assessing, interpreting, analyzing, expressing opinions, evaluating requires a means of communication in verbal or written form. The means of communication can be in multiple forms arranged in language which in the form of expression or communication is in the form of writing through graphics, tables, pictures, or other forms (Gebre, 2018; Kristidhika et al., 2020; Moore et al., 2020; Namdar & Shen, 2016; Ngin, 2018; Rachman et al., 2020).

Based on the characteristics of science material, the use of multiple representation is highly regarded by teachers in the learning process. The teacher can design multiple representation tests,

so that these tests can communicate representations to model and interpret them. Study from Abdurrahman et al. (2019); Susilaningsih et al. (2019); and Yanti et al. (2019) showed that multiple representation can improve the implementation of the learning process starting from planning, student activities during learning, responding to student's very good interest, and providing positive results on students' social skills. The results of the study show that multiple representation can improve student's problem solving ability (Prahani et al., 2016), increase creativity in analyzing tests (Mutia & Prasetyo, 2018), help students stimulate the development of thinking skills with various perspectives and approaches (Fonna & Mursalin, 2018).

Multiple representation tests to measure critical thinking skills need to be developed in each material. However, only a few materials in science subjects can accommodate the needs of the test. One of the materials that can support the needs of this type of assessment is environmental change material. Observations have been made at one of the junior high schools in Sleman district, namely MTS Negeri 2 Sleman. Based on observations on science learning, it was found that formative assessment for the cognitive aspects rarely used the high-level thinking instruments. Additional information obtained was that research had never been carried out related to the development of an authentic assessment instrument based on multiple representations in science learning to measure students' critical thinking skills before. Therefore, this study aims to produce an instruments test used to measure students' critical thinking skills in natural science learning.

Method

This study focuses on measuring the utility of a multiple representation-based critical thinking skills test instrument. This instrument consists of 14 questions, where 1 indicator is represented by 2 questions which have been developed based on the modified critical thinking skills indicators. This research uses a 4-D development model (define, design, develop, and disseminate) involving 118 students at the develop stage and 60 students at the disseminate stage. The instrument developed was an essay test based on multiple representations. The analysis of the content validity of the test instruments was analyzed using the Content Validity Index, and the empirical validity of the test instrument reliability was estimated by using the Item Response Theory with Winsteps program. There were 308 students as subject divided into 10 groups, five groups used MSLAM as the experimental class and five groups used the HOT Lab as a control class. All groups did experiments of series-parallel circuit on electrical and elasticity.

Instruments and Procedures

The data obtained are as follows: 1. Qualitative Data that comes from expert validators in the form of comments and suggestions for authentic assessment instruments for critical thinking skills. 2. Quantitative Data, namely: a. The assessment score of the expert validator of the authentic assessment instrument of critical thinking skills. b. Score of student learning outcomes on environmental change material using authentic assessment instruments for critical thinking skills. Data was collected from July to August 2020. The school as the place of research conducted was MTs Negeri 2 Sleman which located at Jalan Magelang km 17, Margorejo, Tempel, Sleman, Daerah Istimewa Yogyakarta. Data collection was carried out during the process of preparing the assessment instrument as well as in the learning assessment process in the classroom, including: (1) Testing the appropriateness of the authentic assessment instrument for critical thinking skills developed through validation by expert validators; (2) Taking data on students' cognitive learning outcomes using authentic assessment instruments for critical thinking skills after learning environmental change material is carried out; and (3) Observing the achievement of students' critical thinking skills in the cognitive aspects after using the developed authentic assessment instruments.

Data Analysis

The data collected from the instruments were analysed as follow:

Analysis of Validity

Content validity is the validity that is estimated through testing the test content with rational analysis or through professional judgment (Gardner & Dunkin, 2018). The data from the results of the assessment by expert validators from the validation sheet of the assessment instrument were analyzed to determine the validity of the content of the developed authentic assessment instrument. In this study, the content validity of the critical thinking skills assessment instrument was analyzed using the Content Validity Ratio (CVR) and the Content Validity Index (CVI). According to Lawshe (Lawshe, 1975), CVR is a content validity approach to determine the suitability of items with domains that are measured based on expert judgment. Content Validity Ratio (CVR) was obtained from a number of experts (panels) who were asked to examine each component of the measurement instrument. The technique of analyzing it is as follows: (1) Validator assessment criteria on the assessment data obtained from the validation is a score. The table is used to convert the score given by the validator into the assessment index value; (2) Calculating the CVR value; (3) Calculating Content Validity Index value (CVI-value); (4) Categorization of CVR and CVI result is in range $-1 < 0 < 1$ (Lawshe, 1975).

Empirical validity Analysis

According to Kvale (1989) empirical validity is the validity obtained based on experience by means of testing. Empirical validity is obtained through the results of test trials to respondents. In this study, the empirical validity of the critical thinking skills instrument was analyzed using the Winsteps 3.37 program with Rasch Modeling, which is the development of the analytical model by Georg Rasch from the response theory item 1 LP (one Logistic Parameter). The item fit with the Rasch model can explain whether the item of the instrument functions normally in making measurements or not. Item fit analysis provides a technique to control the quality needed to assess the validation of test items and person responses (Wright & Stone, 1988). Boone et al. (2014) added that the criteria used to check the suitability of instrument items to be considered in accordance with the model were by looking at the value of OUTFIT Mean Square (MNSQ), OUTFIT Z-standard (ZSTD), and Point Measure Correction (Pt Mean Corr).

Analysis of reliability

According to Lester et al. (2014) reliability means the extent to which the results of a measurement can be trusted. A measurement result can be trusted if the results obtained are relatively the same for several times. The reliability analysis of the test instruments was carried out with the help of the Winsteps 3.37 program. The Winsteps program can provide instrument reliability information, namely person spacing index and item spacing index, and Cronbach's Alpha value, namely the interaction between person and item (Fariña et al., 2019). Yanto (2019) state that the higher the item reliability, the more precise the overall item is analyzed according to the model used. The instrument can be said to be reliable if it has a Cronbach Alpha value> 0.7.

Analysis of level difficulty

The level of difficulty of the instrument items can be obtained in the analysis using the Winsteps program. Hambleton and Swaminathan (1985, p. 36) states that an item is said to be good if the level of difficulty is more than -2.0 or less than +2.0 (-2.0 <difficulty <+2.0).

Analysis of student's achievement

The level of student's ability in answering the instruments can be seen with the help of the Winsteps program with Rasch modeling. Fariña et al. (2019) state that the ability level of these students is indicated by the logit value on the person measure.

Analysis of e-learning achievement of Critical Thinking Skills

The score of the results obtained by students from the critical thinking instruments in the form of numbers is then converted into three categories (Permatasari et al., 2019).

Results and Discussion

Feasibility of authentic instruments of critical thinking skills

Expert judgment trials aim to produce valid instruments in terms of content. A qualitative review has been carried out prior to testing the instruments and measurements involving experts. The aspects assessed at this stage include substance, construction, language and appearance. The validation results from the experts were then analyzed using the Aiken V equation to determine the value of each item. The results of the Aiken V scores on the aspects of substance, construct, language and appearance, respectively, are about 0.90, 0.85, 0.80, and 0.92. Aiken stated that the statistical significance of Aiken V can be determined by correlating the scale used with the number of experts. This study involved seven experts and five category scales with a significance level of 0.05 so that the Aiken V limit for each item was 0.75. Thus, it can be stated that, at the significance level of 0.05, all items are included in the content valid category.

The results in content validation for the written test assessment tool were analyzed using the Lawshe content validity where the CVR validity standard depends on the number of SMEs. The CVR value must meet 0.99 for the items to be declared valid. This applies to content validation using 7 SMEs (Lawshe, 1975). The CVR value obtained from each item is 1 and is fully presented in the attachment. The CVI value obtained from the average CVR is 1. Based on the CVR value that exceeds 0.99, all items are declared valid and fit for use for further research.

Based on the analysis of the instrument items using the Winsteps 3.37 program, the quality of the assessment instrument items can be seen. First, a test of the fulfilled assumptions, namely indicators, is carried out. Unidimensional means that each test item measures only one ability (Fu & Feng, 2018). The results of the analysis with Winsteps obtained Eigenvalues or raw variance data of 48.3% with Unexplained variance in 1st contrast of 7.0% and Unexplained variance in 2nd contrast of 6.5% for authentic assessment instruments of critical thinking skills.

According to Chan et al. (2014), the minimum requirement for unidimensionality is 20%. These results indicate that the unidirectionality of the instrument with a minimum requirement of 20% raw variance has been met. Student's CTS was measured by giving 7 questions, 1 question represent 1 indicator, which showed that students have a low level of CTS. RASCH analysis done included person reliability, item reliability, and fit item measure. Person reliability or score that shows how consistent the students are in answering correctly is 0.87 with a separation index of 2.57. In other words, students may answer questions, with the correct answers, in the "good" category (Boone et al., 2011). Separation index, or score that shows how well a sample of people is able to separate the items, of 2.14 indicated that students' score of CTS has good enough distribution (Boone & Noltemeyer, 2017). Next, determining a separation index that obtains 3.76 is good. These results indicate that respondents can be divided into four large groups, namely groups that have very high, high, low, and very low critical thinking skills scores. This classification indicates that the instrument made is able to distinguish students' abilities quite accurately.

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	18.8	9.9	-.02	.40	.96	-.1	.97	.0
S.D.	7.1	1.7	1.17	.09	.52	1.1	.64	1.0
MAX.	33.0	14.0	2.01	.79	3.34	3.1	4.10	2.9
MIN.	2.0	6.0	-4.79	.30	.18	-2.8	.15	-2.6
REAL RMSE	.44	TRUE SD	1.09	SEPARATION	2.46	Person RELIABILITY	.86	
MODEL RMSE	.41	TRUE SD	1.10	SEPARATION	2.69	Person RELIABILITY	.88	
S.E. OF Person MEAN = .11								

VALID RESPONSES: 70.4% (APPROXIMATE)
Person RAW SCORE-TO-MEASURE CORRELATION = .86 (approximate due to missing data)
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .71 (approximate due to missing data)

Figure 1. Pearson Measured

Figure 1 shows that the value of the person measure shows the score of 0.02, which means that the student's average ability is almost the same as the difficulty level of the question (set by default 0.00). The OUTFIT-MNSQ column shows the score of 0.99 which indicates that the questions that have been given are in the acceptable category, or excellent in measuring. Furthermore, the OUTFIT-ZSTD score of -.1 has a high level of reliability. Referring to Boone et al. (2014); and Linacre (2012), the range of values for OUTFIT-MNSQ is from 0.5 to 1.5 and the range of values for OUTFIT-ZSTD is from -2 to +2. Next is The Cronbach alpha value that shows how the students' internal consistency is in answering questions. Therefore, this score is not fully part of the statistical analysis, but to show the reliability between students (who answer) and the questions (which are asked). In this study, the value of The Cronbach alpha value was 0.71 which means "acceptable", as shown in Figure 1.

Figure 2 shows the results of the measured item test which shows the item's reliability value reaches 0.97. This means that the item is considered very good in measuring the ability of students with a separation index of 5.92, which is in the very good category. This result is supported by Fit Item Order (Figure 3) and variable map between item difficulties and student ability (see Figure 4).

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	155.5	81.7	.00	.14	.95	-.2	1.00	-.1
S.D.	76.3	18.8	.78	.03	.23	1.4	.28	1.2
MAX.	288.0	108.0	1.50	.20	1.48	3.1	1.72	2.3
MIN.	49.0	44.0	-1.14	.11	.53	-2.3	.63	-1.8

REAL RMSE	.14	TRUE SD	.76	SEPARATION	5.34	Item	RELIABILITY	.97
MODEL RMSE	.14	TRUE SD	.76	SEPARATION	5.48	Item	RELIABILITY	.97

S.E. OF Item MEAN = .22

UMEAN=.0000 USCALE=1.0000
Item RAW SCORE-TO-MEASURE CORRELATION = -.85 (approximate due to missing data)
1144 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 1838.04 with 1012 d.f. p=.0000
Global Root-Mean-Square Residual (excluding extreme scores): .8358

Figure 2. Item Measured

Figure 3 shows the item analysis based on the fit order item. Based on the table, the item fit criteria for the model if the OUTFIT MNSQ is between 0.5 to 1.5, the ZSTD value is between -2.0 to +2.0, and the Pt Mean Corr value is 0.4 to 0.8 (Fariña et al., 2019). Based on this description, it can be concluded that 14 instrument items (7 indicators) of authentic assessment of critical thinking skills fit the Rasch model.

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
9	75	88	1.50	.17	.53	-2.3	.63	-1.7	.59	.47	85.2	72.3	I0009
10	90	80	.80	.15	.78	-1.1	.89	-.4	.49	.51	67.5	62.2	I0010
11	112	94	.79	.13	.99	.0	1.13	.7	.50	.51	54.3	56.6	I0011
8	49	44	.76	.20	1.05	.3	1.72	2.0	.13	.55	52.3	60.5	I0008
12	144	96	.45	.12	1.10	.8	1.00	.0	.56	.56	47.9	49.3	I0012
13	91	66	.31	.14	.78	-1.3	.75	-1.1	.56	.55	51.5	52.1	I0013
14	89	55	.16	.15	1.02	.1	1.19	.9	.64	.61	38.2	48.2	I0014
2	172	86	-.18	.12	.88	-.8	.85	-.9	.70	.64	46.5	42.9	I0002
1	128	64	-.37	.14	.80	-1.3	.83	-.9	.60	.65	51.6	45.8	I0001
3	204	89	-.52	.12	1.00	.0	.94	-.3	.68	.66	46.1	44.4	I0003
4	258	105	-.59	.11	.77	-1.8	.73	-1.8	.71	.65	55.2	44.1	I0004
5	281	108	-.86	.11	1.48	3.1	1.41	2.3	.68	.68	25.9	45.1	I0005
6	288	101	-1.10	.12	1.29	1.8	1.17	.9	.72	.66	38.6	47.3	I0006
7	196	68	-1.14	.15	.86	-.7	.82	-.8	.75	.72	50.0	48.4	I0007
MEAN	155.5	81.7	.00	.14	.95	-.2	1.00	-.1			50.8	51.4	
S.D.	76.3	18.8	.78	.03	.23	1.4	.28	1.2			13.3	8.2	

Figure 3. Item Fit Order

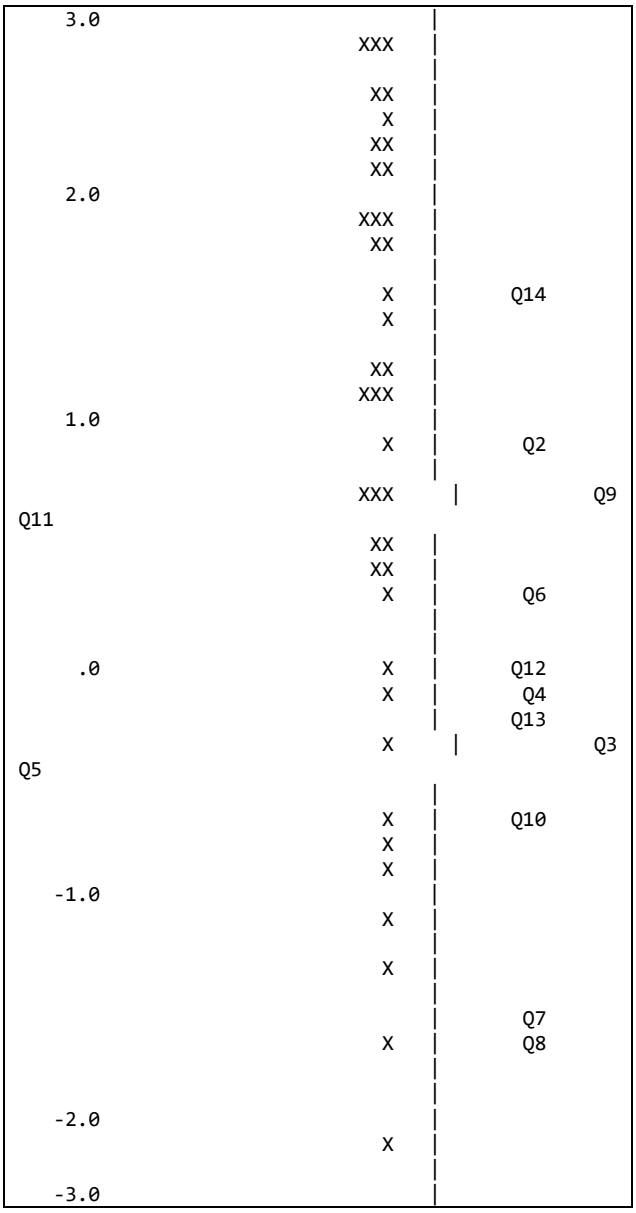

Figure 4. Variable Map

Figure 4 shows how the distribution of responses is associated with the question items answered. The question with the highest difficulty level is in the top position and the item with the lowest difficulty level is the one at the bottom. Labels X and Y indicate the type of question (X = Type 1 tested first; Y = Type 2 tested after the first one) and numbers 1-7 show indicators of critical thinking skills. Based on the data n Figure 3 and Figure 4, it can be seen that all the questions have good criteria for measuring students' critical thinking skills.

Achievement of students' critical thinking skills (Implementation)

Descriptive analysis

In the implementation stage, the data were collected in two stages, namely pre-test and post-test. The data on the results of the implementation carried out are shown in Figure 5.

Based on Figure 5, there are two students who have very extreme scores both in the pre-test and post-test. Based on the distribution of data, it can be concluded that the pre-test items are more spread out symmetrically (normal with sig. 0.017 <0.05) than the post-test which tends to the right (not normal with sig. 0.200> 0.05). More detailed details are shown in Figure 6.

Value Stem-and-Leaf Plot for Group= Pre Test			Value Stem-and-Leaf Plot for Group= Post Test		
Frequency	Stem &	Leaf	Frequency	Stem &	Leaf
2,00	2 .	11	1,00	2 .	8
5,00	2 .	55588	4,00	3 .	5999
8,00	3 .	22222222	3,00	4 .	222
16,00	3 .	5555555599999999	18,00	5 .	000003333377777777
12,00	4 .	222222222222	12,00	6 .	004444477777
4,00	4 .	6666	13,00	7 .	1111555555588
8,00	5 .	00000333	5,00	8 .	22255
Stem width:		10,00	Stem width:		10,00
Each leaf:		1 case(s)	Each leaf:		1 case(s)

Figure 5. Student Score

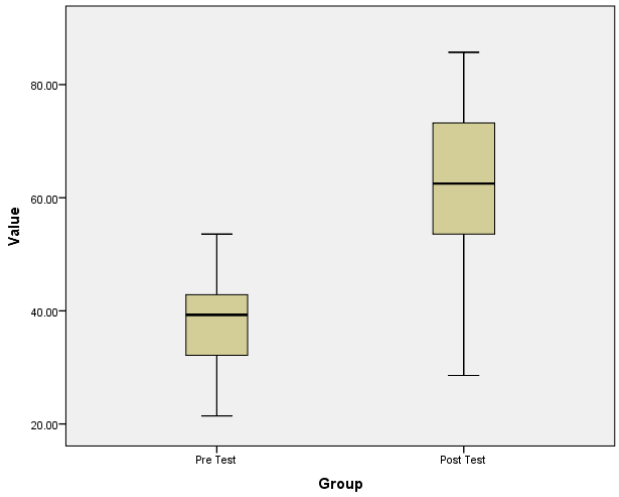

Figure 5. Summary Score Test (Box Plot)

Based on Figure 6, it can be seen that on average, students' score with type 2 questions are higher than those type 1 questions. This means that with the same indicator, type 2 questions are declared easier than type 1 questions. Furthermore, Figure 6 also describes the range of data on the two types of questions which in type 1 students have the ability to answer questions that tend to be the same as the data set which is more centralized, whereas in type 2 the data is spread over a larger cluster. Then, students can be grouped according to their abilities as in Table 1.

Table 1. The distribution of student's ablity in each item

	The number of students of each level		
No. Item	High	Medium	Low
Item 1	13.00	33.00	14.00
Item 2	5.00	38.00	17.00
Item 3	17.00	33.00	10.00
Item 4	0.00	51.00	9.00
Item 5	5.00	49.00	6.00
Item 6	7.00	40.00	13.00
Item 7	6.00	48.00	6.00
Item 8	8.00	34.00	18.00
Item 9	14.00	39.00	7.00
Item 10	21.00	17.00	22.00
Item 11	9.00	48.00	3.00
Item 12	22.00	24.00	14.00
Item 13	12.00	31.00	17.00
Item 14	9.00	46.00	5.00

Instrument's analysis

The results of the instrument testing at the implementation stage were given to sixty samples that were spread into two classes. They indicate that the resulting instrument had a fairly good reliability, as shown in Figure 7.

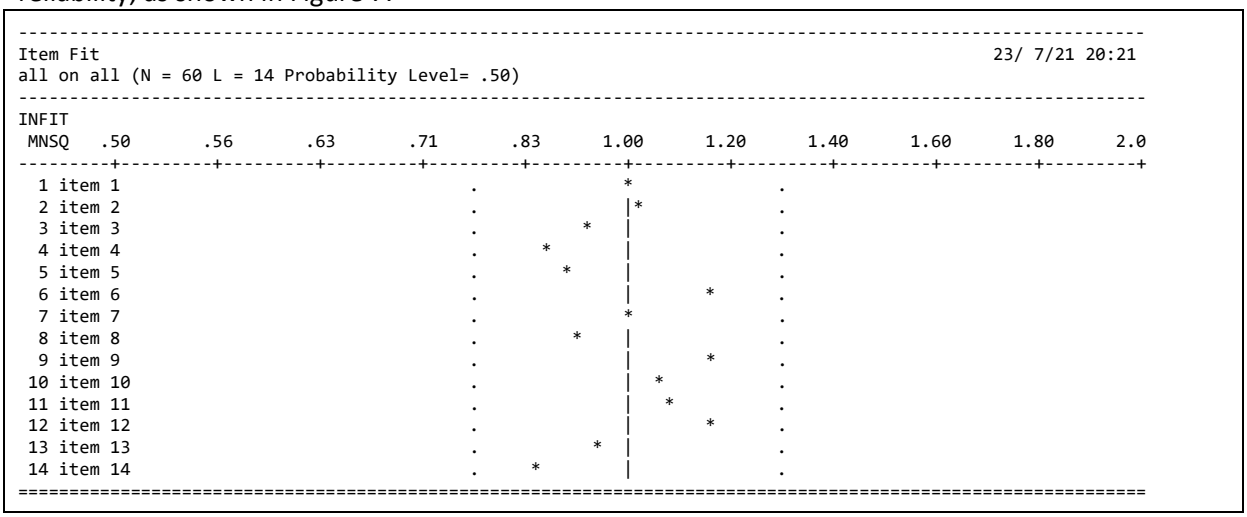

Figure 7. Critical Thinking Skill Test Results on the Implementation

Figure 7 shows the total average value of the three aspects. Most of which are in the good category because they are between -2 to +2.

Case Estimates	
Summary of case Estimates	
Mean	0.94
SD	0.66
Reliability	0.95
Item Separated	3.89
Strata Separation	5.52
Cronbach Alpha (Kr-20)	0.81

Figure 8. Result of Reliability

Figure 8 shows that the questions made have good quality, which is indicated by the reliability value of 0.95. Furthermore, the separation index of 3.89 shows that each item is distributed in good categories. This is reinforced by the separation index of 3.89 (rounded to 4) which shows that the question instrument made is able to differentiate students' abilities into four groups (very high, high, low, very low). Finally, with reference to the Cronbach alpha value of 0.81, it indicates that the resulting instrument is acceptable for measuring critical thinking skills.

Referring to the variable map, the data in Figure 6 shows that the post-test score is better than the pre-test, as evidenced by Figure 9 which indicates that the post-test (denoted by Y) has a lower level of difficulty than the pre-test. When viewed from the FIT model (Figure 10), it can be seen that the Y type (post-test) questions have a pattern that is more similar to the RASCH model compared to the X type (pre-test) questions.

The results of the study showed that the authentic assessment instrument has good quality and is suitable to measure student's CTS. Based on the results of the analysis obtained, it can be seen that the CTS authentic assessment instrument developed is able to measure the ability of students from the highest to the lowest ability. Based on the results of the expert's assessment, it appears that all the questions are of good quality and are feasible to be tested on students. The results of the feasibility test on 118 students showed that the instrument developed was able to measure students' abilities well. This can be seen in the group of students who were divided into six groups from the group with the highest ability to the lowest one which was shown by a separation index of 5.92 (rounded to 6). At the dissemination stage, the separation index slightly decreased to 5.52 (rounded

to 5), which meant that students were grouped into five groups. Thus, it can be concluded that the instrument has good quality in order to measure student ability correctly.

```
                          |       3.3
                          |      13.4
 3.0                      |       1.3
                          |
                          |       2.4
                          |
                          |
                          |
 2.0                      |       5.3    8.4
                        X |       6.4    9.4
                          |       7.3   11.3
                          |      14.4
                        X |      13.3
                      XXX |
                        X |
 1.0                   XX |       2.3    3.2    9.3
                    XXXXX |       7.2
                       XX |       5.2
                  XXXXXXX |      10.4   12.4
                      XXX |       4.3    9.2
                  XXXXXXX |
  .0                  XXX |      14.3
                    XXXXX |       6.3   10.3
                   XXXXXX |          8.3    10.2    11.2
12.3
                     XXXX |       4.2    8.2
                     XXXX |
                          |       6.2   12.2   13.2
                        X |       2.2
-1.0                      |
                        X |       1.2    4.1
                          |
                        X |       3.1
                          |
                          |       6.1
                          |       7.1    8.1    9.1
-2.0                      |       5.1   14.2
                        X |      13.1
                        X |      10.1   12.1
                          |
                          |      11.1
                          |
                        X |
-3.0                      |
                          |      14.1
                          |       2.1
```

Figure 9. Variable Map

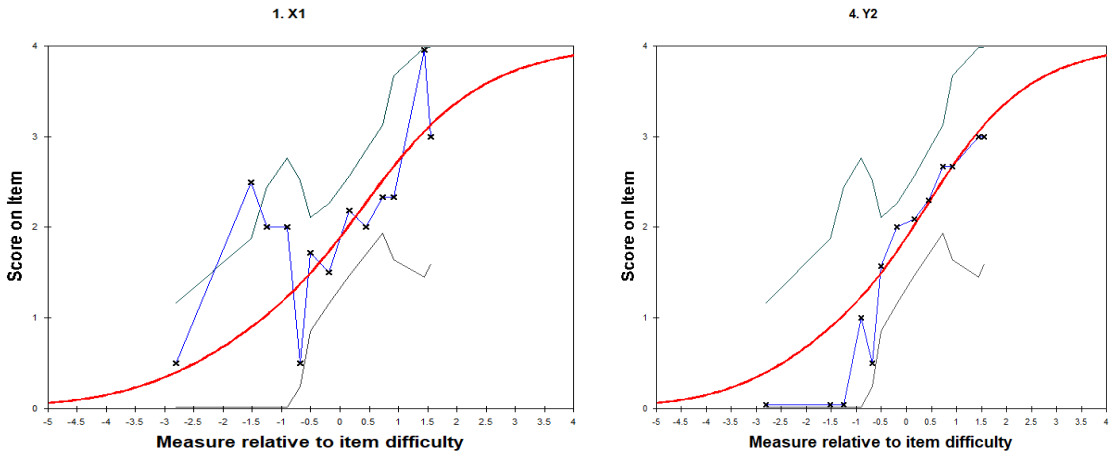

Figure 10. Expected Score ICC

Based on the results of the analysis on each item and the indicators tested, in all aspects, it meets the OUTFIT MNSQ value limit standard with a score range of 0.5 to 1.5. The clarity assumption aspect (Item 1 and item 8) has an OUTFIT MNSQ value of 0.80 (type A) and 0.72 (Type B), an interpretation aspect of 0.7 (type A) and 0.93 (type B), an analysis aspect of 0.74 (type A) and 1.13 (type B), an evaluation aspect of 1.26 (type A) and 1.63 (type B), a reason aspect of 0.73 (type A) and 1.36 (type B), and a self-regulation aspect of 0.96 (type A) and 0.8 (type B). Based on the OUTFIT MNSQ value, it can be seen that there is only 1 question that needs to be revised, namely the type B evaluation aspect (item number 12), and after that it can be used in dissemination test. Furthermore, referring to the variable Map, the results of the analysis show that there are some students with very high critical thinking skills, but on the other hand there are also students with very low critical thinking skills. This of course is influenced by several factors, one of which is the class characteristics factor. The class used as the test subject is a class with input from students with various abilities. This means that students in the class do not have the same abilities from the start, but there are students with very high abilities and students with sufficient abilities. However, the instrument of critical thinking skills developed was able to measure the ability of students from the highest to the lowest abilities. These results are consistent with research conducted by Burhanuddin (2015) which states that authentic assessment by written tests is suitable for assessing the cognitive aspects of students.

Conclusion

Based on the results of the analysis and discussion, the following conclusions were obtained: (1) this study produced an authentic assessment instrument which was suitable for measuring students' critical thinking skills on environmental change material. The feasibility of this authentic assessment instrument is based on the results of the analysis as follows: the instrument has met the content validity requirements by the expert judgment in the very good category and about 14 items of critical thinking skills have obtained empirical evidence of fit with the Rasch Model based on three parameters, namely OUTFIT MNSQ, ZSTD, and Pt Mean Corr; based on the item separation index, the developed critical thinking skills authentic assessment instrument was classified as reliable, and the developed authentic instrument can measure students' critical thinking skills in level high of 16.67%, middle of 63.33%, and low of 20.00% with an average of 10 students belong to the high level, 38 students to the middle level, and 12 students to the low level.

References

Abdurrahman, A., Setyaningsih, C. A., & Jalmo, T. (2019). Implementating multiple representation-based worksheet to develop critical thinking skills. *Journal of Turkish Science Education*, *16*(1), 138–155. https://doi.org/10.12973/tused.10271a

Amalia, S., & Wuryandani, W. (2020). Socio-cultural based learning module for critical thinking ability in elementary school: systematic search. *Acta Educationis Generalis*, *10*(2), 180–187. https://doi.org/10.2478/atd-2020-0017

Arends, R. I., & Kilcher, A. (2010). Teaching for student learning: Becoming an accomplished teacher. In *Teaching for Student Learning: Becoming an Accomplished Teacher*. https://doi.org/10.4324/9780203866771

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, *4*(1), 1–13. https://doi.org/10.1080/2331186X.2017.1416898

Boone, W. J., Townsend, J. S., & Staver, J. (2011). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education*, *95*(2), 258–280. https://doi.org/10.1002/sce.20413

Boone, W. J., Yale, M. S., & Staver, J. R. (2014). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). Pearson.

Burhanuddin. (2015). *Evaluasi keterlaksanaan penilaian otentik (authentic assessment) pada pembelajaran IPA SMP Negeri di Pasangkayu Kabupaten Mamuju Utara* [Universitas Negeri Yogyakarta]. https://eprints.uny.ac.id/25593/

Chan, S. W., Ismail, Z., & Sumintono, B. (2014). A Rasch model analysis on secondary students' statistical reasoning ability in descriptive statistics. *Procedia - Social and Behavioral Sciences*, *129*, 133–139. https://doi.org/10.1016/j.sbspro.2014.03.658

Chusni, M. M., Saputro, S., Raharjo, S. B., & Suranto, S. (2021). Student's critical thinking skills through discovery learning model using e-learning on environmental change subject matter. *European Journal of Educational Research*, *10*(3), 1123–1135. https://doi.org/10.12973/eu-jer.10.3.1123

Chusni, M. M., Saputro, S., Suranto, S., & Rahardjo, S. B. (2020). Review of critical thinking skill in indonesia: Preparation of the 21st century learner. *Journal of Critical Reviews*, *7*(09), 1230–1235. https://doi.org/10.31838/jcr.07.09.223

Doleck, T., Bazelais, P., Lemay, D. J., Saxena, A., & Basnet, R. B. (2017). Algorithmic thinking, cooperativity, creativity, critical thinking, and problem solving: exploring the relationship between computational thinking skills and academic performance. *Journal of Computers in Education*, *4*(4), 355–369. https://doi.org/10.1007/s40692-017-0090-9

Fariña, P., González, J., & San Martín, E. (2019). The use of an identifiability-based strategy for the interpretation of parameters in the 1PL-G and rasch models. *Psychometrika*, *84*(2), 511–528. https://doi.org/10.1007/s11336-018-09659-w

Fonna, M., & Mursalin, M. (2018). Role of self-efficacy toward students' achievement in mathematical multiple representation ability (MMRA). *Jurnal Ilmiah Peuradeun*, *6*(1), 31. https://doi.org/10.26811/peuradeun.v6i1.174

Fu, J., & Feng, Y. (2018). A comparison of score aggregation methods for unidimensional tests on different dimensions. *ETS Research Report Series*, *2018*(1), 1–16. https://doi.org/10.1002/ets2.12194

Gardner, A. K., & Dunkin, B. J. (2018). Evaluation of validity evidence for personality, emotional intelligence, and situational judgment tests to identify successful residents. *JAMA Surgery*, *153*(5), 409–415. https://doi.org/10.1001/jamasurg.2017.5013

Gebre, E. (2018). Learning with multiple representations: Infographics as cognitive tools for authentic learning in science literacy. *Canadian Journal of Learning and Technology*, *44*(1), 1–24. https://doi.org/10.21432/cjlt27572

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Springer Netherlands. https://doi.org/10.1007/978-94-017-1988-9

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research*, *62*(3), 333–353. https://doi.org/10.1080/00313831.2016.1258726

Kartianom, K., & Ndayizeye, O. (2017). What's wrong with the Asian and African Students' mathematics learning achievement? The multilevel PISA 2015 data analysis for Indonesia, Japan, and Algeria. *Jurnal Riset Pendidikan Matematika*, *4*(2), 200–210. https://doi.org/10.21831/jrpm.v4i2.16931

Kristidhika, D. C., Cendana, W., Felix-Otuorimuo, I., & Müller, C. (2020). Contextual teaching and learning to improve conceptual understanding of primary students. *Teacher in Educational Research*, *2*(2), 71–78. https://doi.org/10.33292/ter.v2i2.84

Kvale, S. E. (1989). *Issues of validity in qualitative research.* Studentlitteratur.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–

575. http://caepnet.org/~/media/Files/caep/knowledge-center/lawshe-content-validity.pdf

Lester, P. E., Inman, D., & Bishop, L. K. (2014). *Handbook of tests and measurement in education and the social sciences*. Rowman & Littlefield.

Linacre, J. M. (2012). *A user guide to Winsteps Ministep Rasch model computer programs: Program manual 3.75.0.*

Maba, W. (2017). Teacher's perception on the implementation of the assessment process in 2013 curriculum. *International Journal of Social Sciences and Humanities*, *1*(2), 1–9. https://doi.org/10.29332/ijssh.v1n2.26

Menteri Pendidikan dan Kebudayaan Republik Indonesia. (2016). *Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 23 Tahun 2016 tentang standar penilaian pendidikan*. Kementerian Pendidikan dan Kebudayaan Republik Indonesia.

Menteri Pendidikan dan Kebudayaan Republik Indonesia. (2019). *Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 43 Tahun 2019 tentang penyelenggaraan ujian yang diselenggarakan satuan pendidikan dan ujian nasional*. Kementerian Pendidikan dan Kebudayaan Republik Indonesia.

Moore, T. J., Brophy, S. P., Tank, K. M., Lopez, R. D., Johnston, A. C., Hynes, M. M., & Gajdzik, E. (2020). Multiple representations in computational thinking tasks: A clinical study of second-grade students. *Journal of Science Education and Technology*, *29*(1), 19–34. https://doi.org/10.1007/s10956-020-09812-0

Mujib, N. R., Toenlioe, A. J. ., & Praherdhiono, H. (2018). Analisis butir soal ujian nasional IPA SD/MI tahun 2015 sampai 2017 berdasarkan taksonomi Bloom. *Jktp*, *1*(2), 149–158. http://journal2.um.ac.id/index.php/jktp/article/view/3731

Mutia, N. B., & Prasetyo, Z. K. (2018). The effectiveness of students' worksheet based on multiple representations to increase creative thinking skills. *Journal of Education and Learning (EduLearn)*, *12*(4), 631–637. https://doi.org/10.11591/edulearn.v12i4.8487

Namdar, B., & Shen, J. (2016). Intersection of argumentation and the use of multiple representations in the context of socioscientific issues. *International Journal of Science Education*, *38*(7), 1100–1132. https://doi.org/10.1080/09500693.2016.1183265

Ngin, C. S. (2018). Examining a teacher's use of multiple representations in the teaching of percentages : A commognitive perspective. *Proceedings of the 41st Annual Conference of the Mathematics Education Research Group of Australasia*, 591–598.

Nugrahanto, S., & Zuchdi, D. (2019). Indonesia PISA result and impact on the reading learning program in Indonesia. *Proceedings of the International Conference on Interdisciplinary Language, Literature and Education (ICILLE 2018)*, *297*(Icille 2018), 373–377. https://doi.org/10.2991/icille-18.2019.77

Paul, R., Elder, L., & Bartell, T. (1997). *A brief history of the idea of critical thinking*. California Commission on Teacher Credentialing. http://www.criticalthinking.org/pages/a-brief-history-of-the-idea-of-critical-thinking/408

Permatasari, A. K., Istiyono, E., & Kuswanto, H. (2019). Developing assessment instrument to measure physics problem solving skills for mirror topic. *International Journal of Educational Research Review*, 358–366. https://doi.org/10.24331/ijere.573872

Prahani, B. K., Limatahu, I., W.W, S., Yuanita, L., & Nur, M. (2016). Effectiveness of physics learning material through guided inquiry model to improve student's problem solving skills based on multiple representation. *International Journal of Education and Research*, *4*(12), 231–242. https://www.ijern.com/journal/2016/December-2016/17.pdf

Rachman, D., Soviyah, S., Fajaruddin, S., & Pratama, R. A. (2020). Reading engagement, achievement and learning experiences through kahoot. *LingTera*, *7*(2). https://doi.org/10.21831/lt.v7i2.38457

Suprapto, N. (2016). What should educational reform in Indonesia look like? - Learning from the PISA science scores of East-Asian countries and Singapore. *Asia-Pacific Forum on Science Learning and Teaching*, *17*(2), 1–22.

Susilaningsih, E., Supartono, S., Kristanto, T., Sariana, E., Azizah, P., & Natasukma, M. (2019). The effectiveness of multiple representation oriented learning material with project based learning to improve students' chemistry learning outcomes. *Proceedings of the 6th International Conference on Educational Research and Innovation (ICERI 2018)*, *330*(Iceri 2018), 87–90. https://doi.org/10.2991/iceri-18.2019.18

Wijayanti, M. D., Rahardjo, S. B., Saputro, S., & Mulyani, S. (2019). Item analysis of critical thinking skills instrument to measure effectiveness of scientific group inquiry learning (SGIL) model. *Jurnal Pendidikan IPA Indonesia*, *8*(4), 538–546. https://doi.org/10.15294/jpii.v8i4.20794

Wright, B. D., & Stone, M. H. (1988). Reliability in Rasch measurement. In *Research Memorandum No. 53*. MESA.

Yanti, H., Distrik, I. W., & Rosidin, U. (2019). The effectiveness of students' worksheets based on multi-representation in improving students' metacognition skills in static electricity. *Journal of Physics: Conference Series*, *1155*(1), 012083. https://doi.org/10.1088/1742-6596/1155/1/012083

Yanto, F. (2019). Development of problem-based student worksheet with authentic assessment to improve student's physics problem solving ability. *Journal of Physics: Conference Series*, *1185*, 012075. https://doi.org/10.1088/1742-6596/1185/1/012075

Yusrizal, Y. (2016). Analysis of difficulty level of physics national examination's questions. *Jurnal Pendidikan IPA Indonesia*, *5*(1), 140–149. https://doi.org/10.15294/jpii.v5i1.5803