JPI, Vol. 9 No. 1, March 2020
p-ISSN: 2303-288X, e-ISSN: 2541-7207 DOI: 10.23887/jpi-undiksha.v9i1.20884

Items Quality Analysis Using Rasch Model To Measure Elementary School Students' Critical Thinking Skill On Stem Learning

Ghullam Hamdu*, F N Fuadi, A Yulianto, Y S Akhirani
Elementary School Teachers Education, Universitas Pendidikan Indonesia, Tasikmalaya
e-mail: ghullamh2012@upi.edu, fitrinf@upi.edu, ade.yulianto95@gmail.com, lie.anasta07@gmail.com

Abstract

Critical thinking as one of the 21st century competences required by students needs to be developed and analyzed by employing qualified assessment instrument. Test is a kind of critical thinking assessment instrument which quality is developed and analysed to create a meaningful learning. A total of 10 multiple choices items were developed based on critical thinking indicators. The items were then given to forty two 4th grade students in one of the elementary schools in Tasikmalaya-West Java after obtaining STEM learning. Focus group discussions were conducted to construct and validate the instrument. The result of the test was analyzed using Rasch model with the assistance of Winsteps software version 3.75. The results indicated that the analysis using the Rasch model could explain the critical thinking items' quality based on the level of difficulty and suitability and could categorize students' abilities and their suitability for STEM learning conducted.

Keywords: *Items'quality; The Rasch Model; Critical Thinking.*

1. Introduction

Assessment is considered as an important aspect to measure students' skill. The 21st century skills have become the skills that students must acquire in accordance with the development of science and technology (Kivunja, 2015; Hugerat & Kortam, 2014). In addition, critical thinking is included as one of the skills (Kay & Greenhill, 2011; Sahidah Lisdiani et al., 2019). It is a fundamental skill in assessing and making decision (Fisher, 2011; Kay & Greenhill, 2011). Moreover, it is a competency needed by students for their personal and professional lives in the future (Bezanilla et al., 2019). The critical thinking skill was taught by approaching and solving the problems based on persuasive argument, logic, and rationality which involves verification, evaluation, choosing the right answer for the task given and reasoned rejection of other alternative solutions (Barnhart & van Es, 2015). Various indicators of critical thinking skills that can be developed based on "Assessing 21st Century Skills for Teachers and Students" were analyzing arguments, claiming or proving; making conclusion or reasoning; judging or evaluating; making decisions or problem solving (P21 Partnership for 21st Century Skills: http://www.p21.org/our-work/p21-framework). These indicators can be used as a reference to assess students.

In general, there are limited qualified instrument to measure students' critical thinking skills, especially at the elementary school level. The items provided tend to be conceptual and in remembering level which did not show the learning authenticity and the assessment process. Authentic assessment can be carried out if the learning process was authentic as well (Swaffield, 2011). Moreover, teachers need to master critical thinking skills to understand what must be taught and evaluated. However, teachers often put aside the importance of presenting critical based learning (Kek & Huijser, 2011). Elementary school teachers focused on activities that only develop low-level thinking skills without considering other activities that demand students' critical thinking (Assaraf & Orion, 2010). The remembering activities which commonly implemented by teachers were not relevant to the concept and meaningful learning (ŽivkoviĿ, 2016). There are many concepts in science material taught in elementary schools including force and motion of objects. This concept is

* Corresponding author.

Received 10 September 2019; Accepted 26 March 2020; Available online 31 March 2020

p-ISSN: 2303-288X, e-ISSN: 2541-7207 DOI: 10.23887/jpi-undiksha.v9i1.20884

one of the most important concepts in teaching Science as it is the basic concept required to understand advanced materials. Force and motion are the basic concepts for studying mechanics at a higher level, especially Newton's laws of motion (Panprueksa et al., 2012). Teaching the concepts of force and object motion requires an integrative approach of Science, Mathematics, and Engineering so that the learning is meaningful for students.

Test result in Indonesia showed a discrepancy between the national and international test instruments (Winarti & Patahuddin, 2017). Indonesia generally acquired the lowest score on international tests such as TIMMS, PIRLS, and PISA compared to other countries. This lack of learning outcomes indicated that some distressing trends in Mathematics and Science. Only a small number of students performed well at TIMSS and PIRLS, while at PISA, there were no students as samples who performed well at level 6 (the highest) in Mathematics or Science from 2009 (Pedro & et.al, 2013). PISA data in 2018 demonstrated that Indonesia only reached level 1 for Mathematics, Science and reading skill (OECD, 2019). The tests presented in those international instruments generally explored higher order thinking skills, more specifically on critical thinking. Therefore, the thinking skill needs to be trained through an appropriate learning process so that more authentic results are obtained. One of the recommendations from PISA was to emphasize integrated learning: integrating different subjects, integrating diverse students, and integrating various learning contexts such as real-life contexts with a variety of resources from the community. These various learning processes need to be designed so that students can be successful authentically.

The implementation of learning by exploring authentic abilities in this study was carried out by applying STEM (Science, Technology, Engineering, and Mathematics) learning. It is expected that the integrated approach of STEM education can support the students in the future to solve real-world problems by applying across disciplines concepts and critical thinking, collaboration, and creativity (Burrows & Slater, 2015). The skill capacity resulted from STEM learning overlaps with the skills needed in 21st century education. Therefore, STEM learning can bridge the gap between education and the skills needed in the 21st century, especially critical thinking skill (Putra & Kumano, 2018).

One of the things that influences the success of STEM learning in schools is the curriculum structure and the skill level and teachers' readiness to teach it (Blackley & Howell, 2015). The STEM learning approach is increasingly popular, but remains challenging and difficult to understand for teachers (Wahono & Chang, 2019; Shernoff et al., 2017). For teachers, content knowledge is the basis for applying STEM approach in the classroom (Putra & Kumano, 2018). However, most teachers had acquired training in only one subject (Honey et al., 2014), and most schools and classes at all levels separate STEM as a specific subject. This is a significant challenge for teachers who are interested in implementing integrated STEM. In the context of elementary education in Indonesia, only science and mathematics are included in 2013 curriculum, while technology and engineering subjects are only a minor part or not included in the curriculum. Although STEM education in elementary schools emphasizes more on Science and Mathematics, STEM can assist students to develop their critical thinking skill. Measuring critical thinking skill demands a good test instrument. A good test instrument must meet several criteria, including having various good items' difficulty level and suitability of the items with the indicators being measured. This test instrument can be used as an assessment to measure elementary school students' critical thinking skill. It is essential to determine the instrument quality. Therefore, to find out the instrument quality, a good instrument analysis is needed.

Learning assessment provides good information for teachers to help students learn better. Beside the tools from the Classical Test Theory (CTT) approach that commonly used by teachers, another approach called objective measurement which based on probability is an alternative tool that can provide a more precise measurement. The Rasch model that provides psychometric analysis technique can be used by teachers to develop test items and to present relevant information related to students' learning assessments (Sumintono, 2018). The analysis of this test instrument using the Rasch model is included in the item response measurement theory. This measurement could explain the interaction between the subjects and the test items. This will make the measurement have more precise and objective results

(Sumintono & Widhiarso, 2014). Furthermore, the Rasch model is a well-studied measurement approach that models the relationships between item difficulties, people's abilities, and the probability of responses given (Andrich, 1981). The advantage of the Rasch model compared to classical theory is that this model can identify wrong answers from experts, identify improper judgments, and predict missing data based on systematic response patterns (Goodwin & Leech, 2003; Ratna et al., 2017; Fahmina et al., 2019). Using the Rasch model, this study aims to determine the quality of the items in measuring the critical thinking skill of elementary school students and to measure the level of students' critical thinking skill after the students learn using STEM.

2. Method

The data were collected by distributing 10 critical thinking based written multiple choices test items referred to "Assessing 21st Century Skills for Teachers and Students" to 4th grade students in one of the elementary schools in Tasikmalaya, West Java. The test was constructed by adapting the teaching materials and learning processes carried out previously. The written test items were not constructed independently but always accompanied by the construction of other teaching devices such as lesson plans, media, students' worksheets, and teaching materials. The written test with other teaching devices were reviewed by conducting focus group discussions (FGD) with STEM learning development team to compile and validate the instruments. The acquired suitability of teaching material and the scope of the learning process from the FGD is the final result of the instrument validation process. The test was given to students after they took part in STEM learning. The test results of these students were then analyzed using the Rasch Model. The Rasch Model software application used is Winsteps 3.75. The process of analyzing the data is illustrated in Figure 1. The description of the result of the written test development form is illustrated in Table 1.

Table 1. The Result of Written Test Construction

Test Items Construction Form	Result
Kinds of the written test	Multiple choices (A, B, C, and D)
Items number	10 items
Higher order thinking skill	Critical thinking (based on "*Assessing 21st Century Skills for Teachers and Students*)
The analyzed critical thinking indicator	Making conclusion, analyzing, analyzing and claiming, claiming, evaluating.
Teaching material	Force and motion of objects
Participants as the primary data sources	42 fourth grade students in an elementary school in Tasikmalaya, West Java
Secondary data sources	Observation result of STEM learning and the result of teachers' interview
The duration to finish the task &	20 minutes (08.00-08.20)
Situation during the test	Done in the following day after the STEM-based learning implemented

The Rasch Model was used to analyze the written test results related to students' critical thinking skill indicators after the implementation of STEM-based learning and identify the use of constructed written test items. Other description result was the STEM learning process used to help explaining the result Winsteps 3.75 display. The analysis was done analytically by describing Wright map and table generated and displayed from *Winteps* software 3.75. One of the main components of Rasch's analysis is the Wright map that visually depicts the relationship between people and the question items (Wilson, 2008).

The person-infit and person-outfit statistics from the Rasch model were used as person fit statistics to detect deviant responses (Widhiarso & Sumintono, 2016). Based on the person-fit scores, participants had three categories. They were classified as high, medium and low from their skill analysis. On the other hand, the person-fit score for the quality of the

items was based on the chosen answer to the written test items given to the students. Analysis was then carried out to critical thinking indicators related to students' chosen answers in the written test result.

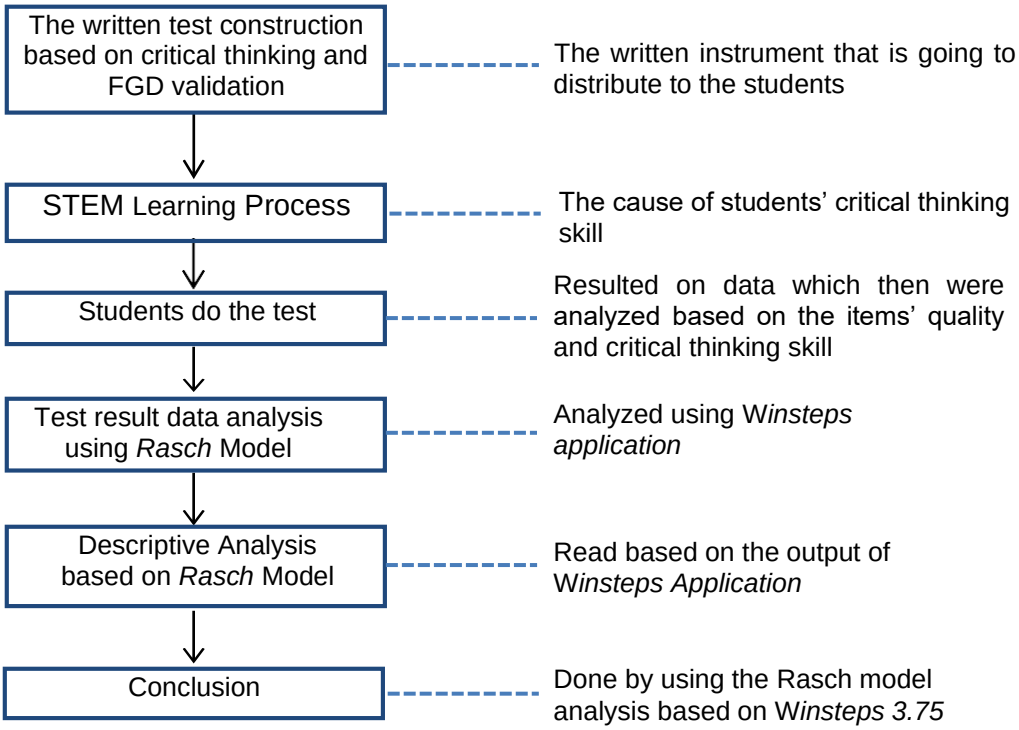

Figure 1. The Data Analysis Process

3. Result and Discussion

Developed critical thinking written test indicators are listed in Table 2. The development of these test items was conducted through an intensive FGD process by deriving from the curriculum at first. Therefore, the materials to be discussed during the learning could be obtained which then became the basis for making test items. The determined critical thinking indicators were then employed to make written test in the form of multiple choices items with the number of items for each critical thinking indicator was different. The unequal item number for every indicator was based on several things, including time needed to construct the item, similarity of the items model developed within the scope, and the depth of material presented as problems. The time allotment to do every item was similar for every item. Thus, students are expected to answer the questions with a more measured time. Question items with similar intention to measure were selected based on the comprehensiveness. This was related to the depth of the materials. Items which were not depth enough and derived from low-level thinking skill were eliminated by integrating them with higher-level thinking items. The illustration of the items in the form of Wright map is presented in Figure 2.

Table 2. The Indicators of The Items

Critical Thinking Indicators	Item Number
Making conclusion	1
Analyzing	2 and
Analyzing and Claiming	3
Claiming	4, 5, and 8
Evaluating	6,9, and 10

In Figure 2, the Wright map illustrates the distribution of students' abilities and the

p-ISSN: 2303-288X, e-ISSN: 2541-7207 DOI: 10.23887/jpi-undiksha.v9i1.20884

items' difficulty level with similar scale. The left side of the Wright map illustrates the distribution of students' abilities. P05 student acquired the highest level of ability with a logit value above the standard deviation (T) which shows different high intelligence (outliers), while P15 student with a logit value below the T limit indicates a very low ability. On the other hand, the right side of Wright map illustrates the distribution of items' difficulty. Item number I8 was categorized in the highest difficult level with a logit value beyond the T limit. This indicated that the probability to work on the item correctly was very small. Item number I8 with the other two items belonged to the claiming indicator. However, based on the Wright's map, the distribution items in this level was varied. Item number I8 was categorized as difficult level, number I5 was medium and number I4 was regarded as an item that could be answered by most of the students. Nevertheless, there were 9 students who could not answer the question. This signified that there were approximately 21% out of 42 students who thought that item number I4 (item with low difficulty based on the Wright's map) was challenging to answer correctly. These 21% of the students based on the distribution in Wright's map generally thought that some of the questions presented were difficult. Wright's map shows that there were questions in evaluating indicator (I6, I9 and I10) possessed almost similar difficulty level. On the other hand, only one student (P05) having good ability was able to answer all the questions presented. The distribution in this Wright map was a general description but it could provide quite clear interpretations regarding the items' difficulty level. Further analysis of the Wright map is explained using a more analytical table based on the distribution of the written test items' difficulty constructed and the distribution of students' abilities from the written test result

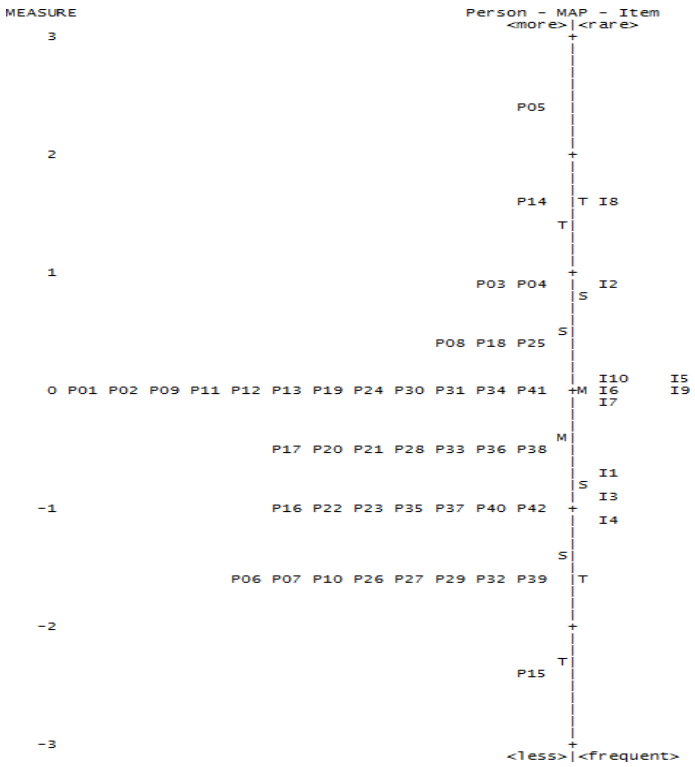

Figure 2. Wright Map: The Distribution of Students' Skill and Items' Difficulty Level with Similar Scale

Items' Difficulty Level Analysis (Item Measure)

Table 3 presents several columns that provide information about each item's difficulty level. The classification of the items' difficulty level was based on the combination of standard deviation (SD) value and the average logit value (Sumintono & Widhiarso, 2015). The categories were tough items with logit value greater than 1SD; difficult items with logit value of 0.0 +1 SD; easy items with logit value of 0.0 -1 SD; and very easy items with logit

p-ISSN: 2303-288X, e-ISSN: 2541-7207 DOI: 10.23887/jpi-undiksha.v9i1.20884

value smaller than -SD.

Table 3. The Items' 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
8	6	42	1.64	.47	1.21	.7	1.29	.7	.12	.33	83.3	86.5	I8
2	10	42	.91	.39	.96	-.1	.92	-.2	.41	.37	76.2	78.5	I2
5	16	42	.12	.34	1.10	.8	1.07	.4	.30	.38	59.5	68.3	I5
10	16	42	.12	.34	.97	-.2	1.14	.8	.38	.38	73.8	68.3	I10
6	17	42	.01	.34	1.09	.8	1.12	.7	.28	.38	71.4	66.7	I6
9	17	42	.01	.34	1.14	1.1	1.20	1.1	.23	.38	71.4	66.7	I9
7	18	42	-.11	.34	.75	-2.3	.68	-2.0	.64	.38	78.6	66.3	I7
1	23	42	-.67	.34	.86	-1.2	.81	-1.1	.52	.37	73.8	66.7	I1
3	25	42	-.90	.34	1.02	.2	1.04	.3	.34	.37	69.0	67.6	I3
4	27	42	-1.13	.35	.91	-.6	.90	-.4	.44	.36	71.4	69.5	I4
MEAN	17.5	42.0	.00	.36	1.00	-.1	1.02	.0			72.9	70.5	
S.D.	6.1	.0	.78	.04	.13	1.0	.18	.9			5.9	6.3	

Based on data in Table 3 presented above, the result of the items analysis can be grouped as follows: 1) tough items group for item no. 8 (I8), and item no. 2 (I2); 2) difficult items group for item no. 5 (I5), item no. 10 (110), item no. 6 (I6), and item no. 9 (I9); 3) easy items group for item no. 7 (I7), and item no. 1 (I1); 3) very easy items group for item no. 3 (I3), and item no. 4 (I4).

From the description above (data in Table 1 and Table 2), it can be implied that there were different difficulty levels for each item in the same indicator.

Table 4. Items' Difficulty Levels

Critical Thinking Indicators	Item Number	Difficulty Level
Making Conclusion	1	Easy
Analyzing	2	Tough Easy
	7	
Analyzing and Claiming	3	Very Easy
Claiming	4	Very Easy
	5	Difficult
	8	Tough
Evaluating	6	Difficult
	9	Difficult
	10	Difficult

The items' difficulty level within an indicator based on the results of the test are different. Similar items' difficulty level for students to answer the questions about critical thinking was in evaluating indicator. These results indicated that the items' type for each number given has equal weight for the students. On the contrary, the items given in claiming indicator possessed different difficulty level within the indicator. This showed that the written test items given had different difficulty weight for each item even though they were in the same critical thinking indicator. The differences among the difficulty level indicated that the items developed did not have difficulty level consistency even though they were in the same indicator. Moreover, it is possible for the items developed after being tested on students to have different perceptions. The difficulty level categories can be obtained after field trials. Therefore, it was very possible that the result of the filed trial had different or similar level of difficulty within the common indicator.

The determination of the items' difficulty level using Rasch analysis was not based on similar percentage distribution as in the case using conventional analysis. In conventional analysis, the categorization of items' difficulty level was conducted by using the percentage of upper, lower and middle limits. The 25% were categorized for difficult and easy items and 50% for medium questions. The division of this percentage was usually done by directly

p-ISSN: 2303-288X, e-ISSN: 2541-7207 DOI: 10.23887/jpi-undiksha.v9i1.20884

arranging students using normal curve. The normal curve showed the ideal condition for the items quality which must meet the criteria for a balanced number of items based on percentages (Arikunto, 2012). For example, 25% each for a number of easy and difficult items and 50% for a number of medium items. However, Rasch's calculation is largely determined by the result of students' responses / answers to the problem. Therefore, it is common that the result of this study indicating no standard calculation of the items' percentage of questions for every difficulty level category. The results of Rasch's analysis were real based on the students' responses or answers. The closer the items to the normal curve distribution, the more proportional the distribution of the items according to the level of difficulty. If we put the data in Table 4 in percentage, the result would be 20% items were tough; 40% items were difficult; 20% items were easy; and 20% items were very easy. If we use this percentage, the distribution of the items was nearly resembling the normal curve with an assumption that there was a proportion of normal difficulty items. If we put them in normal proportion, there would be 20% items were tough, 30% items each were difficult and easy and 20% items were very easy. Further Rasch's analysis to test the suitability of the difficulty level for each item will be presented as followed.

3.2 The Items' Suitability (Item Fit Order)

The item fit level can be seen by using three criteria, namely outfit means-square value (Outfit MNSQ), Outfit Z-Standard (Outfit ZSTD), and Point Measure Correlation (PT-Measure Corr) (Boone et al., 2014; Bond & Fox, 2015; Sumintono & Widhiarso, 2015). The criteria employed to check the fitness of the items that were not appropriate (outlier or misfit)) were 1) the value of outfit means-square (Outfit MNSQ) received: 0.5 <MNSQ <1.5; 2) Outfit Z-Standard (Outfit ZSTD) value received: -2.0 <ZSTD <+2.0; 3) Point Measure Correlation (PT-Measure Corr) Value: 0.4 <PT-Measure Corr <0.85 (Bone et al, 2014).

In Table 5, the MNSQ scores for all items were accepted and the ZSTD values for all questions were also accepted, but only item 1, 2, 4, and 7 are accepted in the PT-Measure Corr values.

Table 5. Items' Statistics: Misfit 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
8	6	42	1.64	.47	1.21	.7	1.29	.7	A .12	.33	83.3	86.5	I8
9	17	42	.01	.34	1.14	1.1	1.20	1.1	B .23	.38	71.4	66.7	I9
10	16	42	.12	.34	.97	-.2	1.14	.8	C .38	.38	73.8	68.3	I10
6	17	42	.01	.34	1.09	.8	1.12	.7	D .28	.38	71.4	66.7	I6
5	16	42	.12	.34	1.10	.8	1.07	.4	E .30	.38	59.5	68.3	I5
3	25	42	-.90	.34	1.02	.2	1.04	.3	e .34	.37	69.0	67.6	I3
2	10	42	.91	.39	.96	-.1	.92	-.2	d .41	.37	76.2	78.5	I2
4	27	42	-1.13	.35	.91	-.6	.90	-.4	c .44	.36	71.4	69.5	I4
1	23	42	-.67	.34	.86	-1.2	.81	-1.1	b .52	.37	73.8	66.7	I1
7	18	42	-.11	.34	.75	-2.3	.68	-2.0	a .64	.38	78.6	66.3	I7
MEAN	17.5	42.0	.00	.36	1.00	-.1	1.02	.0			72.9	70.5	
S.D.	6.1	.0	.78	.04	.13	1.0	.18	.9			5.9	6.3	

From the description and the table above, items number 1, 2, 4, and 7 met the MNSQ, ZSTD, and PT-Measure Corr values. On the other hand, items number 3, 5, 6, 8, 9, and 10 only meet the MNSQ and ZSTD values. If the items did not fulfil all the three criteria (MNSQ, ZSTD, and Pt. Measure Corr), it can be concluded that the questions were not good enough so that they need to be repaired or replaced (Boone et al., 2014; Bond & Fox, 2015). Therefore, referring to that statement, all items analyzed had accepted difficulty level and were worth to be maintained since not all categories were not met. The result of this Rasch analysis showed that there were various levels of difficulty according to Table 4. By using this test, the items constructed have appropriate difficulty level based on data presented in Table 4 which had four levels, namely tough, difficult, easy and very easy. Good test items could identify students' various abilities with diverse levels of difficulty. If the test level of difficulty is high, it can be confirmed authentically that students cannot answer correctly or do not

p-ISSN: 2303-288X, e-ISSN: 2541-7207 DOI: 10.23887/jpi-undiksha.v9i1.20884

understand the questions given. However, if the test difficulty level is low, it can be confirmed authentically that many students can answer correctly or easily. The analysis of the Rasch model based on the above result can determine the validity of a test well (Baghaei & Amrahi, 2011).

The result presented were based on the teaching material related to STEM learning that was carried out. Referring to the data, the test items were appropriate to be used to identify students' critical thinking skills in the learning process carried out previously. However, the difficulty level of this written test items would be different if it was carried out to other students in different schools and conducted by different teachers. The result obtained would be similar if the test was given to students and teachers with similar characteristics. For example, similar school cluster characteristic and the teachers' level of understanding of STEM learning. Therefore, it is important to conduct trials on other research by applying different methods to be more reliable. To see whether any differences or similarities resulted from the implementation of STEM learning and the critical thinking based written test items.

3.3 Students' Critical Thinking Level Analysis (Person Measure)

Beside analyzing critical thinking items and their indicators, an analysis of students' abilities in working on written test questions was conducted. The students' critical thinking skill level can be identified through their work on written test items since it provided information about the effectiveness of STEM-based learning implementation carried out previously. Therefore, the result of this analysis can provide more effective recommendations in helping students during the learning process, especially STEM learning process carried out previously. The data on the students' skill level is presented in Table 6.

Table 6. Person 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%	Person
5	9	10	2.44	1.09	.65	-.2	.30	-.5	.70	.25	90.0	90.0	P05
14	8	10	1.55	.84	1.04	.3	.83	.0	.33	.32	70.0	80.4	P14
3	7	10	.94	.74	1.05	.3	1.19	.6	.24	.34	80.0	73.3	P03
4	7	10	.94	.74	.69	-.9	.62	-.9	.71	.34	80.0	73.3	P04
8	6	10	.44	.69	1.34	1.3	1.51	1.6	-.10	.35	60.0	67.7	P08
18	6	10	.44	.69	1.11	.5	1.19	.7	.20	.35	60.0	67.7	P18
25	6	10	.44	.69	.85	-.5	.84	-.4	.52	.35	80.0	67.7	P25
1	5	10	-.02	.67	.91	-.4	.88	-.4	.46	.34	70.0	62.8	P01
2	5	10	-.02	.67	.94	-.2	.91	-.3	.43	.34	70.0	62.8	P02
9	5	10	-.02	.67	1.11	.6	1.09	.4	.22	.34	70.0	62.8	P09
11	5	10	-.02	.67	.72	-1.4	.67	-1.3	.69	.34	90.0	62.8	P11
12	5	10	-.02	.67	.72	-1.4	.67	-1.3	.69	.34	90.0	62.8	P12
13	5	10	-.02	.67	.89	-.5	.86	-.5	.49	.34	90.0	62.8	P13
19	5	10	-.02	.67	.90	-.4	.84	-.5	.49	.34	50.0	62.8	P19
24	5	10	-.02	.67	.97	-.1	.92	-.2	.40	.34	50.0	62.8	P24
30	5	10	-.02	.67	1.37	1.7	1.52	1.8	-.13	.34	50.0	62.8	P30
31	5	10	-.02	.67	1.05	.3	1.20	.8	.24	.34	70.0	62.8	P31
34	5	10	-.02	.67	1.06	.4	1.03	.2	.28	.34	70.0	62.8	P34
41	5	10	-.02	.67	1.17	.9	1.18	.7	.14	.34	50.0	62.8	P41
17	4	10	-.47	.68	.66	-1.6	.61	-1.3	.73	.33	90.0	66.2	P17
20	4	10	-.47	.68	.95	-.2	1.35	1.1	.28	.33	90.0	66.2	P20
21	4	10	-.47	.68	.69	-1.5	.64	-1.2	.70	.33	90.0	66.2	P21
28	4	10	-.47	.68	1.10	.5	1.02	.2	.23	.33	50.0	66.2	P28
33	4	10	-.47	.68	.69	-1.5	.64	-1.2	.70	.33	90.0	66.2	P33
36	4	10	-.47	.68	.90	-.4	.83	-.4	.47	.33	70.0	66.2	P36
38	4	10	-.47	.68	1.05	.3	.97	.0	.29	.33	50.0	66.2	P38
16	3	10	-.96	.72	1.31	1.0	1.21	.6	-.04	.30	60.0	70.8	P16
22	3	10	-.96	.72	1.31	1.0	2.11	2.0	-.24	.30	60.0	70.8	P22
23	3	10	-.96	.72	1.03	.2	.95	.1	.28	.30	80.0	70.8	P23
35	3	10	-.96	.72	1.09	.4	1.00	.2	.22	.30	60.0	70.8	P35
37	3	10	-.96	.72	1.32	1.1	2.09	1.9	-.24	.30	60.0	70.8	P37
40	3	10	-.96	.72	.91	-.2	.79	-.3	.44	.30	60.0	70.8	P40
42	3	10	-.96	.72	.63	-1.3	.53	-1.1	.75	.30	80.0	70.8	P42
6	2	10	-1.55	.82	.92	.0	.81	.0	.36	.25	80.0	79.9	P06
7	2	10	-1.55	.82	.99	.1	.87	.0	.29	.25	80.0	79.9	P07
10	2	10	-1.55	.82	1.19	.5	1.64	1.0	-.08	.25	80.0	79.9	P10
26	2	10	-1.55	.82	1.24	.7	1.24	.6	-.04	.25	80.0	79.9	P26
27	2	10	-1.55	.82	.94	.0	.88	.1	.32	.25	80.0	79.9	P27
29	2	10	-1.55	.82	.92	.0	.81	.0	.36	.25	80.0	79.9	P29
32	2	10	-1.55	.82	1.22	.6	1.18	.5	.00	.25	80.0	79.9	P32
39	2	10	-1.55	.82	.99	.1	.87	.0	.29	.25	80.0	79.9	P39
15	1	10	-2.41	1.07	1.14	.4	1.36	.7	-.05	.19	90.0	90.0	P15
MEAN	4.2	10.0	-.43	.74	.99	.0	1.02	.1			72.9	70.5	
S.D.	1.8	.0	.93	.09	.20	.8	.37	.8			13.5	7.6	

p-ISSN: 2303-288X, e-ISSN: 2541-7207 DOI: 10.23887/jpi-undiksha.v9i1.20884

The information of students' skill level categories can be seen from the standard deviation (SD) value and the starting point of the average logit person value (Sumintono & Widhiarso, 2015). Based on the results of the Rasch model through the use of Winsteps 3.75 there were three students' skill level categories, namely high, moderate, and low. The result of these categories was based on the SD value (standard deviation = 0.93) and MEAN value (-0.43). Thus, the range of the category values are as followed: if the students' skill> SD (0.93) then they possessed high skill, if the SD (0.93) <students' skill <MEAN (-0.43) then they were categorized as moderate, If the students' skill <MEAN (-0.43) then they had low skill.

From the range aforementioned before, the students 'abilities can be described as follows: 1) high level skill category with 6 students by considering that the person measure with logit value of +2.44 (P05), logit value of +1.55 (P14) , logit value of +0.94 (P03, P04), and logit value +0.44 (P08, P18, P25); 2) moderate level skill category with 19 students for the person measure with logit value of -0.02 (P41, P34, P31, P30, P24, P19, P13, P12, P11, P9, P2, P1) and logit value of -0.47 (P38, P36, P33, P28, P21, P20, P17); 3) low level skill category with 16 students for the person measure with logit value of -0.96 (P42, P40, P37, P16, P22, P23, P25), logit value of -1.55 (P6, P7, P10, P26 , P27, P29, P32, P39) and logit value of -2.41 (P15).

Rasch model analysis was able to identify students' abilities so that they can be categorized into high, moderate or low level. The categories were specific to the results of the written test items given. The result of the test indicated that students' critical thinking skill after STEM-based learning implementation was mostly in moderate and low level. Only seven students who were in high category. Therefore, it can be concluded that the critical thinking based written test items provided optimum information when it was given to students with moderate and low abilities. Similar to the items' category, the determination of student categories using conventional analysis was based on the rank from the highest to the lowest. For example, high level students for 25%, moderate level students for 50% and low level students for 25% so that a normal curve can be obtained. Rasch analysis provided authentic result of students' level without showing percentages like in normal curve. However, if the percentage of these students' categories is close to or equal to the normal curve, it can be concluded that the students' abilities in the class were varied. This is also largely determined by the construction of the items given.

The results of Rasch analysis illustrated that the STEM-based learning carried out was only able to make 7 students out of 42 students (17%) in a high category. On the other hand, there were 19 students (45%) in moderate category and 16 students (38%) in low category. Thus, normal curve could not represent the students with high and low abilities. The data described that students with high skill were not balanced in number or percentage. Therefore, it could be verified that students' skill in learning needed to be improved, especially critical thinking skill.

The result of the Rasch analysis can also provide recommendation that based on the result of the critical thinking based written tests, STEM learning needed to optimize activities which can support and explore students' critical thinking skill. However, the result was only a recommendation and it will only work on STEM-based learning process done with critical thinking based written tests by using situation, time, conditions and respondents with STEM learning criteria.

3.4 The Analysis of Students' Critical Thinking Skill Suitability Level (Person Fit Order)

After mapping the students' skill into high, moderate, and low level, an analysis was carried out to find the students' critical thinking skill suitability level by detecting students' response patterns in their written test. The result of this analysis was able to provide patterns of responses which are not suitable with students' answer based on their skill analyzed previously. Table 6 shows the data about the suitability level of students' skill to work on the critical thinking-based written test.

The criteria used to observe the suitability of students are similar with the criteria for checking the suitability of the item (outlier or misfit) (Bone et al, 2014), namely: 1) means-

p-ISSN: 2303-288X, e-ISSN: 2541-7207 DOI: 10.23887/jpi-undiksha.v9i1.20884

square outfit value (MNSQ Outfit) received: 0.5 <MNSQ <1.5; 2) Outfit Z-Standard (Outfit ZSTD) value received: -2.0 <ZSTD <+2.0; 3) Point Measure Correlation (PT-Measure Corr) value: 0.4 <Pt-Measure Corr <0.85.

Based on the data in Table 7, the MNSQ score of student P22, P37, and P05 was not accepted, the ZSTD value of all students could be accepted, and the Pt-Measure Corr value of student P03, P06, P07, P08, P09, P10, P15, P17, P17, P18, P20, P22, P23, P27, P28, P29, P30, P31, P32, P34, P35, P37, P38, and P41 was not accepted. If the students' skill in the three criteria (MNSQ, ZSTD, and Pt. Measure Corr) was not fulfilled it can be confirmed that the skill was not suitable so it needed to be reviewed or there was a biased skill (Boone et al., 2014; Bond & Fox, 2015). Therefore, according to that statement, all the students' categories (high, moderate and low) analyzed had a suitability level that can be confirmed and was not beyond the reasonable limits of the patterns (high, moderate or low).

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%	Person
22	3	10	-.96	.72	1.31	1.0	2.11	2.0	A-.24	.30	60.0	70.8	P22
37	3	10	-.96	.72	1.32	1.1	2.09	1.9	B-.24	.30	60.0	70.8	P37
10	2	10	-1.55	.82	1.19	.5	1.64	1.0	C-.08	.25	80.0	79.9	P10
30	5	10	-.02	.67	1.37	1.7	1.52	1.8	D-.13	.34	50.0	62.8	P30
8	6	10	.44	.69	1.34	1.3	1.51	1.6	E-.10	.35	60.0	67.7	P08
15	1	10	-2.41	1.07	1.14	.4	1.36	.7	F-.05	.19	90.0	90.0	P15
20	4	10	-.47	.68	.95	-.2	1.35	1.1	G .28	.33	90.0	66.2	P20
16	3	10	-.96	.72	1.31	1.0	1.21	.6	H-.04	.30	60.0	70.8	P16
26	2	10	-1.55	.82	1.24	.7	1.24	.6	I-.04	.25	80.0	79.9	P26
32	2	10	-1.55	.82	1.22	.6	1.18	.5	J .00	.25	80.0	79.9	P32
31	5	10	-.02	.67	1.05	.3	1.20	.8	K .24	.34	70.0	62.8	P31
3	7	10	.94	.74	1.05	.3	1.19	.6	L .24	.34	80.0	73.3	P03
18	6	10	.44	.69	1.11	.5	1.19	.7	M .20	.35	60.0	67.7	P18
41	5	10	-.02	.67	1.17	.9	1.18	.7	N .14	.34	50.0	62.8	P41
9	5	10	-.02	.67	1.11	.6	1.09	.4	O .22	.34	70.0	62.8	P09
28	4	10	-.47	.68	1.10	.5	1.02	.2	P .23	.33	50.0	66.2	P28
35	3	10	-.96	.72	1.09	.4	1.00	.2	Q .22	.30	60.0	70.8	P35
34	5	10	-.02	.67	1.06	.4	1.03	.2	R .28	.34	70.0	62.8	P34
38	4	10	-.47	.68	1.05	.3	.97	.0	S .29	.33	50.0	66.2	P38
14	8	10	1.55	.84	1.04	.3	.83	.0	T .33	.32	70.0	80.4	P14
23	3	10	-.96	.72	1.03	.2	.95	.1	U .28	.30	80.0	70.8	P23
7	2	10	-1.55	.82	.99	.1	.87	.0	u .29	.25	80.0	79.9	P07
39	2	10	-1.55	.82	.99	.1	.87	.0	t .29	.25	80.0	79.9	P39
24	5	10	-.02	.67	.97	-.1	.92	-.2	s .40	.34	50.0	62.8	P24
27	2	10	-1.55	.82	.94	.0	.88	.1	r .32	.25	80.0	79.9	P27
2	5	10	-.02	.67	.94	-.2	.91	-.3	q .43	.34	70.0	62.8	P02
6	2	10	-1.55	.82	.92	.0	.81	.0	p .36	.25	80.0	79.9	P06
29	2	10	-1.55	.82	.92	.0	.81	.0	o .36	.25	80.0	79.9	P29
1	5	10	-.02	.67	.91	-.4	.88	-.4	n .46	.34	70.0	62.8	P01
40	3	10	-.96	.72	.91	-.2	.79	-.3	m .44	.30	60.0	70.8	P40
36	4	10	-.47	.68	.90	-.4	.83	-.4	l .47	.33	70.0	66.2	P36
19	5	10	-.02	.67	.90	-.4	.84	-.5	k .49	.34	50.0	62.8	P19
13	5	10	-.02	.67	.89	-.5	.86	-.5	j .49	.34	90.0	62.8	P13
25	6	10	.44	.69	.85	-.5	.84	-.4	i .52	.35	80.0	67.7	P25
11	5	10	-.02	.67	.72	-1.4	.67	-1.3	h .69	.34	90.0	62.8	P11
12	5	10	-.02	.67	.72	-1.4	.67	-1.3	g .69	.34	90.0	62.8	P12
4	7	10	.94	.74	.69	-.9	.62	-.9	f .71	.34	80.0	73.3	P04
21	4	10	-.47	.68	.69	-1.5	.64	-1.2	e .70	.33	90.0	66.2	P21
33	4	10	-.47	.68	.69	-1.5	.64	-1.2	d .70	.33	90.0	66.2	P33
17	4	10	-.47	.68	.66	-1.6	.61	-1.3	c .73	.33	90.0	66.2	P17
5	9	10	2.44	1.09	.65	-.2	.30	-.5	b .70	.25	90.0	90.0	P05
42	3	10	-.96	.72	.63	-1.3	.53	-1.1	a .75	.30	80.0	70.8	P42
MEAN	4.2	10.0	-.43	.74	.99	.0	1.02	.1			72.9	70.5	
S.D.	1.8	.0	.93	.09	.20	.8	.37	.8			13.5	7.6	

Table 7. Person Statistics: Misfit Order

p-ISSN: 2303-288X, e-ISSN: 2541-7207 DOI: 10.23887/jpi-undiksha.v9i1.20884

An example of the students' responses to the written test analysis can be explained with a scalogram (Guttman Scale) in table 7. Student P22 had an unusual response pattern according to their skill level, where P22 student had a low skill level but was able to answer question no. 8 which had a high level of difficulty (tough). In addition, in answering questions with similar level of difficulty (very easy or item no. 03 & no. 04), student P22 do gave wrong answer to question no. 04 but answered question no.03 correctly. Based on the description, it can be seen that student P22 was guessing and student P22 was inaccurate in answering the questions. Another case was student P05 who had high level of skill and could answer the questions in order from the tough level (item no.02), difficult level (item no. 05, 06, 09, 10), easy level (item no. 01 and 07), and very easy level (item no. 03 and 04). This indicated that student P05 had the suitability to answer the critical thinking based written test.

Table 8. Scalogram (Guttman Matrix)

Person	Item	
	1	
	4317695028	

5	+1111111110	P05
14	+1111110101	P14
3	+1011110110	P03
4	+1111011100	P04
8	+0011101110	P08
18	+0111110010	P18
25	+1101110100	P25
1	+1110100010	P01
2	+1110001010	P02
9	+1011000110	P09
11	+1111001000	P11
12	+1111000100	P12
13	+1111000010	P13
19	+1100111000	P19
24	+1010101100	P24
30	+1001011001	P30
31	+1110001001	P31
34	+1101000110	P34
41	+0011011100	P41
17	+1111000000	P17
20	+1110000001	P20
21	+1110100000	P21
28	+0101101000	P28
33	+1110010000	P33
36	+1100001100	P36
38	+1001101000	P38
16	+0000111000	P16
22	+0100001001	P22
23	+1000100100	P23
35	+0100010100	P35
37	+0011000001	P37
40	+0110010000	P40
42	+1110000000	P42
6	+1000100000	P06
7	+0100010000	P07
10	+0010000010	P10
26	+0000011000	P26
27	+1000000100	P27
29	+1000100000	P29
32	+0000110000	P32
39	+0100010000	P39
15	+0000000100	P15

	1	
	4317695028	

The analysis using the Rasch Model in Table 7 was used to determine the suitability of students' responses of the critical thinking based written test. On the other hand, the result presented in Table 8 was used to identify the direct causes of response patterns that are suitable or not with the students' critical thinking based written tests. For example, student P05 had the best response suitability. This student was able to work on problems ranging from the easiest item (I4) to toughest item (I2) in the correct sequence. This signified that these students had more authentic suitability of the skill. It is different from student P15. This student could only work on item I10 which was more difficult compared to item I4 to I5. Therefore, this student's answer for I10 was accidentally correct as student P15 only predict the answer. The student P15 did not understand the concept. Another example which was commonly found can be seen from student P22. This student did not have a consistent skill. The questions were answered randomly. Student P22 answered correctly item I3 which was in very easy level, item I5 which was in difficult level and item I8 in tough level, while the

p-ISSN: 2303-288X, e-ISSN: 2541-7207 DOI: 10.23887/jpi-undiksha.v9i1.20884

other items were answered incorrectly. It can be assumed that the student with this pattern did not have full understanding of the concepts they have learned. It could be predicted that this student's correct answer on difficult and tough items was only a coincidence.

The result of this data can provide information to the teachers to identify the students' skill and suitability in developing critical thinking skill during STEM-based learning process. Thus, the result provided a recommendation to those who implement the learning to implement STEM-based learning effectively by giving closer attention to students who have inappropriate skill and to improve the students who still have low critical thinking skill. The analysis of this scalogram can further be used by the teachers as the executor of STEM-based learning to describe what the students had acquired from the test results.

The Rasch model analysis had provided comprehensive information about data processing based on students' responses on the critical thinking based written test. Furthermore, the test items' level of difficulty with its indicators for STEM learning could be identified. These results illustrated that the items had been constructed were able to describe the students' skill patterns and their suitability. However, the result of the Rasch Model analysis was more specific to provide a comprehensive picture of STEM learning conducted at that time. The result of the Rasch model analysis could be different or similar depending on the conditions and the learning situations, such as the students' characteristics and the implementation of STEM-based learning in certain classrooms or schools. Nevertheless, the Rasch Model analysis process can be used by teachers in schools to make a comprehensive identification of the learning process connected to the students' responses in the written test

Determining the reliability of tests in classical analysis usually uses raw score intervals which comparison is not clear such as in counting with KR-20 formulation. Therefore, there are extreme scores often included in the reliability test. However, the extreme scores do not have an error variance which make it possible to question the reliability of the test (Boone & Scantlebury, 2006). Thus, classical test theory can have a single standard measurement error. The Rasch measurement for each item and each called as error. If a very large or small percentage of students answered the item correctly, this resulted in a greater error than the items targeted for the average level students (Baghaei & Amrahi, 2011).

4. Conclusion and Suggestions

A comprehensive picture has been presented from the result of the Rasch Model Analysis regarding the items' difficulty and the students' skill measured using critical thinking aspects in STEM-based learning. The distribution of the items from the Rasch Model analysis based on critical thinking indicators resulted on four categories, namely tough, difficult, easy and very easy. The test items indicated that they had various degrees of difficulty for diverse students' abilities (high, moderate and low). Therefore, the test items were suitable to be used in STEM-based learning. On the other hand, the students' critical thinking skill levels in STEM-based learning were generally in the moderate and low categories. Thus, the result of the Rasch Model analysis provided the students' level of suitability which implied that students with low abilities should be assisted more. Moreover, the result of the Rasch model analysis could also be used as a reflection and recommendation for teachers since they are the executor. This is intended to improve the learning process. Last, the Rasch model analysis of the test result can be used by teachers to identify the constructed items quality and the students' abilities resulted from the learning process in school.

Acknowledgement

The authors would like to deliver the greatest gratitude to LPPM Universitas Pendidikan Indonesia for the financial support in publishing the article. The number of research grant funding contract is Number: 5493 / UN40 / KP / 2019 dated on May 28, 2019.

References

Andrich, D. (1981). Book Review : Probabilistic Models for Some Intelligence and Attainment Tests (expanded edition. *Applied Psychological Measurement*, *5*(4), 545–550. https://doi.org/10.1177/014662168100500413

Arikunto, S. (2012). *Dasar-Dasar Evaluasi Pendidikan* (kedua). Bumi Aksara.

Assaraf, O. B. Z., & Orion, N. (2010). System thinking skills at the elementary school level. *Journal of Research in Science Teaching*, *47*(5), 540–563. https://doi.org/10.1002/tea.20351

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling*, *53*(2), 192–211.

Barnhart, T., & van Es, E. (2015). Studying teacher noticing: EXAMINING the relationship among pre-service science teachers' ability to attend, analyze and respond to student thinking. *Teaching and Teacher Education*, *45*, 83–93. https://doi.org/10.1016/j.tate.2014.09.005

Bezanilla, M. J., Fernández-Nogueira, D., Poblete, M., & Galindo-Domínguez, H. (2019). Methodologies for teaching-learning critical thinking in higher education: The teacher's view. *Thinking Skills and Creativity*, *33*(February), 100584. https://doi.org/10.1016/j.tsc.2019.100584

Blackley, S., & Howell, J. (2015). A STEM narrative: 15 years in the making. *Australian Journal of Teacher Education*, *40*(7), 102–112. https://doi.org/10.14221/ajte.2015v40n7.8

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch Model. In *Taylor & Francis*. https://doi.org/10.4324/9781410614575

Boone, W. J., & Scantlebury, K. (2006). The role of rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*(2), 253–269. https://doi.org/10.1002/sce.20106

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

Burrows, A., & Slater, T. (2015). A proposed integrated STEM framework for contemporary teacher preparation. *Teacher Education & Practice*, *28*(2/3), 318–331.

Fahmina, S. S., Masykuri, M., Ramadhani, D. G., & Yamtinah, S. (2019). Content validity uses Rasch model on computerized testlet instrument to measure chemical literacy capabilities. *AIP Conference Proceedings*, *2194*(December). https://doi.org/10.1063/1.5139755

Fisher, A. (2011). *Critical Thinking: An Introduction* (Second). Cambrige University Press.

Goodwin, L. D., & Leech, N. L. (2003). The Meaning of Validity in the New Standards for Educational and Psychological Testing: Implications for Measurement Courses. *Measurement and Evaluation in Counseling and Development*, *36*(3), 181–191. https://doi.org/10.1080/07481756.2003.11909741

Honey, M. A., Pearson, G., & Schweingruber, H. (2014). STEM integration in K-12 education: status, prospects, and an agenda for research. In *STEM Integration in K-12 Education: Status, Prospects, and an Agenda for Research*. https://doi.org/10.17226/18612

Hugerat, M., & Kortam, N. (2014). Improving higher order thinking skills among freshmen by teaching science through inquiry. *Eurasia Journal of Mathematics, Science and Technology Education*, *10*(5), 447–454. https://doi.org/10.12973/eurasia.2014.1107a

Kay, K., & Greenhill, V. (2011). Bringing Schools into the 21st Century. *Bringing Schools into the 21st Century*, 41–65. https://doi.org/10.1007/978-94-007-0268-4

Kek, M. Y. C. A., & Huijser, H. (2011). The power of problem-based learning in developing critical thinking skills: Preparing students for tomorrow's digital futures in today's classrooms. *Higher Education Research and Development*, *30*(3), 329–341. https://doi.org/10.1080/07294360.2010.501074

Kivunja, C. (2015). Exploring the Pedagogical Meaning and Implications of the 4Cs "Super Skills" for the 21st Century through Bruner's 5E Lenses of Knowledge Construction to Improve Pedagogies of the New Learning Paradigm. *Creative Education*, *06*(02), 224–239. https://doi.org/10.4236/ce.2015.62021

OECD. (2019). PISA 2018 insights and interpretations. *OECD Publishing*, 64. https://www.oecd.org/pisa/PISA 2018 Insights and Interpretations FINAL PDF.pdf

Panprueksa, K., Phonphok, N., Boonprakob, M., & Dahsah, C. (2012). Thai Students' Conceptual Understanding on Force and Motion. *International Conference on Education and Management Innovation* , *30*. http://www.ipedr.com

Pedro, C.-I., & et.al. (2013). *Spending more or Spending Better : 76404*. *March*, 1–6.

Putra, P. D. A., & Kumano, Y. (2018). Energy Learning Progression and STEM Conceptualization Among Pre-service Science Teachers in Japan and Indonesia. *New Educational Review*, *53*(3), 153–162. https://doi.org/10.15804/tner.2018.53.3.04

Ratna, I. S., yamtinah, S., Ahadi, Masykuri, M., & Shidiq, A. S. (2017). The Implementation of Testlet Assessment Instrument in Solubility and Solubility Product Material for Measuring Students' Generic Science Skills. *Advances in Social Science, Education and Humanities Research (ASSEHR), Volume 158*, *158*(Ictte), 596–602.

Sahidah Lisdiani, S. A., Setiawan, A., Suhandi, A., Malik, A., Sapriadi, & Safitri, D. (2019). The Implementation of HOT Lab Activity to Improve Students Critical Thinking Skills. *Journal of Physics: Conference Series*, *1204*(1). https://doi.org/10.1088/1742-6596/1204/1/012033

Shernoff, D. J., Sinha, S., Bressler, D. M., & Ginsburg, L. (2017). Assessing teacher education and professional development needs for the implementation of integrated approaches to STEM education. *International Journal of STEM Education*, *4*(1), 1–16. https://doi.org/10.1186/s40594-017-0068-1

Sumintono, B. (2018). Rasch Model Measurements as Tools in Assesment for Learning. *Advances in Social Science, Education and Humanities Research*, *173*(Icei 2017), 38–42. https://doi.org/10.2991/icei-17.2018.11

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi Model Rasch Untuk Penelitian Ilmu-Ilmu Sosial*. TrimKom Publising Home.

Sumintono, B., & Widhiarso, W. (2015). Aplikasi Pemodelan RASCH pada Assessment Pendidikan. In *TrimKom Publising Home*.

Swaffield, S. (2011). Assessment in Education : Principles , Policy & Practice Getting to the heart of authentic Assessment for Learning. *Assessment in Education: Principles, Policy & Practice*, *18*(4), 433–449. https://doi.org/10.1080/0969594X.2011.582838

Wahono, B., & Chang, C. Y. (2019). Assessing Teacher's Attitude, Knowledge, and Application (AKA) on STEM: An Effort to Foster the Sustainable Development of STEM Education. *Sustainability (Switzerland)*, *11*(4). https://doi.org/10.3390/su11040950

Widhiarso, W., & Sumintono, B. (2016). Examining response aberrance as a cause of outliers in statistical analysis. *Personality and Individual Differences*, *98*, 11–15. https://doi.org/10.1016/j.paid.2016.03.099

Wilson, M. (2008). Cognitive diagnosis using item response models. *Journal of Psychology*, *216*(2), 74–88. https://doi.org/10.1027/0044-3409.216.2.74

Winarti, D. W., & Patahuddin, S. M. (2017). Graphic-Rich Items within High-Stakes Tests : Indonesia National Exam (UN), PISA , and TIMSS. *Proceedings of the 40th Annual Conference of the Mathematics Education Research Group of Australasia*, 569–576.

Živkovi Ł, S. (2016). A Model of Critical Thinking as an Important Attribute for Success in the 21st Century. *Procedia - Social and Behavioral Sciences*, *232*(April), 102–108. https://doi.org/10.1016/j.sbspro.2016.10.034