IJERE International Journal of Educational Research Review

Analysis of Students Critical Thinking Skills Using Partial Credit Models (PCM) in Physics Learning

Diena Shulhu Asysyifa[1], Jumadi[2], Insih Wilujeng[3], Heru Kuswanto[4]

ARTICLE INFO

Article History:
Received 15.10.2018
Received in revised form
24.01.2019
Accepted
Available online 01.04.2019

ABSTRACT

The ability to think is divided into low-order thinking skills (LOTS) and high-order thinking skills (HOTS). The abilities expected by the learners through physics learning is the ability to think critically. Therefore, in the assessment of physics learning outcomes of learners, should contain items that are intended to measure the ability. This study aims to analyze the critical thinking skills of learners using the response theory item (Item Response Theory / IRT), with Partial Credit Models (PCM) approach. The form of test used is two tier multiple choice (TTMC) according to the scoring polytomous. TTMC was chosen because have reasoning option to analyze critical thinking skills. PCM was chosen because it corresponds to the characteristics of the test response, namely the form of a polytomous. The subjects were sciences senior high school students. The result of parameter estimation of critical thinking ability of learners shows that there are no students who have highest critical thinking ability, 1.67% of students have high critical thinking ability, 60% of students have average critical thinking ability, 1.67% learners have low critical thinking skills, and 3.33% of learners who have lowest critical thinking skills. Thus, the critical thinking skills of learners in physics lessons still need to be developed.

Keywords:
Critical Thinking Skill, PCM, Two Tier Multiple Choice, Physics Education, İtem Response Theory.

INTRODUCTION

Education can change a person's mindset to always make changes and improvements in all aspects of life. Education can provide supplies for social life. Learners are expected to have the thinking skills to be able to solve problems and face the challenges of globalization. In typical lists of skills needed for the 21st century, critical thinking appears at or near the top (Bailin & Siegel, *2003;* Trilling & Fadel, 2009; Walser, 2008). Effective communication, curiosity, and critical thinking skills are no longer only desirable outcomes of elite liberal arts education, but the essential competencies for life in the 21st century (Wagner, 2008).

The ability to think is divided into two namely low-order thinking (LOT) and high-order thinking (HOT). Based on Bloom's taxonomy, memorization and understanding are classified as low-level thinking while analyzing, synthesizing, and evaluating are classified as high-level thinking (Zohar & Dori, 2003). High-level thinking skills is needed by learners associated with the needs of learners to solve problems faced in everyday life, one of them is critical thinking skills (Noer, 2009). Critical thinking is to think reasonably and reflectively on making decisions about what to do. In other words, decision-making is taken after reflection and evaluation on what is believed. Think critically as a systematic process that gives students the opportunity to formulate and evaluate their own beliefs and opinions. The purpose of critical thinking is to examine an opinion or idea, including to consider or think based on the proposed opinion (Sapriya, 2011). Such considerations are usually supported by acceptable criteria. Ideally critical individuals have 12 critical thinking skills grouped into five aspects of critical thinking skills, such as: 1) Elementary clarification, 2) The basis for the decision, 3) Inference, 4) Advanced clarification and 5) Supposition and integration (Ennis, 2002).

Tests can be classified in several kinds depending on the shape, type, and variety (Zainul & Nasution, 2001). Function of learning result test as tool for placement, formative function, diagnostic function, and summative function. Based on the form, the test of learning outcomes can be grouped into three types: two-tier multiple choices, multiple-tier multiple choice and essay (Gronlund & Linn, 1990). Improving the quality of education can't be separated from the application of an assessment that can accurately measure the

[1] Corresponding e-mail: dienasyifa06@gmail.com, https://orcid.org/0000-0002-1434-7406
[2]; https://orcid.org/0000-0002-4055-5065
[3]; https://orcid.org/0000-0003-1900-7985
[4]; https://orcid.org/0000-0002-2693-8078
Universitas Negeri Yogyakarta[1,2,3,4]

outcome of a learning process. This means that to assess the final outcome in the learning required a quality assessment tool. The fact that the multiple-choice test used in high school for Physics measures ability: remembering, understanding, and applying. The multiple-choice test that used still to measure low-level thinking skills, has not measured the high-order thinking skills of Physics (Istiyono, 2014). Students who experience multiple choice tests tend to achieve lower science scores than those who experienced constructed responses (Tjalla, 2010).

The fact that multiple choice tests are more widely used than other forms of testing. This is because the multiple-choice test used to have advantages, among others: (1) the material being tested can cover most of the learning materials, (2) the students' answers can be corrected easily and quickly, and (3) the answers to each question is certainly true or wrong, so that an objective assessment (Sudjana, 2013). Research on the use of multiple choice items to obtain a level of weakness understanding the concept of Basic Physics has been carried out (Obaidat & Malkawi, 2009). It is said that traditional analysis of multiple choice items with a focus on scores and correlations of true answers cannot optimally provide the information needed by the teacher. The scoring of multiple-choice is dichotomy score, that is correct answer given score 1 and wrong answer given score 0. Meanwhile, to measure critical thinking ability require consideration of reason of student answer that question. Multiple choice test formats can be modified to assess how learners reach the conclusions of their answers.

The two-tier multiple-choice test was introduced as a multiple-choice test modification (Treagust, 1998). This two-tier multiple-choice instrument is an objective test consisting of two levels, the first level being the first-tier and the second-tier. The first part contains the questions of the knowledge aspect. The second section contains a set of possible reasons for the answer in the first section. The second level is to improve the high-level capability and ability to reveal the reasons (Adodo, 2013). Two-tier multiple choice has an advantage because in this test other than students working on a test item that expresses a particular concept the student must also reveal the reason why choose the answer (Suwarto, 2012). Therefore, instead of multiple-choice tests, two-tier tests can be easily used by teachers to increase students' knowledge level and prevent students' alternative conceptions (Tuysuz, 2009; Kubiszyn & Borich, 2013). This test is developed from a multiple-choice item designed in proportion to the format in reasoning test.

The scoring of two-tier multiple choice is usually done by the score of polytomous where the scores of more than two categories are given according to certain criteria. Decision-making can improve measurement, because it is a development of the polytomous scoring system, using many categories (Baker, Rounds & Zeron, 2000). Among a number of polytomous scoring models, the scoring of the partial credit model has a scoring characteristic that is in accordance with the crime of physics. In this case the authors are interested to reveal empirically critical thinking skills using the model Partial Credit Models (PCM).

PCM is the development of the IRT 1 parameter of logistic (1-PL) model and includes the Rasch model. Use of more than two categories is sorted to record the results of an individual's interaction with an item (usually to recognize the truth or completion level) (Master, 1999). The item of polytomous is the item that has the response of more than two answers so it is easier to correct the students' answers as well as detect the ability of the learners (Wardani, Yamtinah & Mulyani, 2015). When it is assumed that an item follows a partial credit pattern then higher individual abilities are expected to have higher scores than low-ability individuals (Widhiarso, 2010). PCM is also appropriate for analyzing responses to the measurement of critical thinking and conceptual understanding in science (Linden & Hambelton, 1997). PCM is an analytical model of the IRT form (Item Response Theory) in which the learners' response to the problem can illustrate a particular ability. This model was developed to describe the relationship between the characteristics of the items with the characteristics or the nature of the respondents.

The estimation of testers' ability is based on the result of response analysis or student's answer in the test. The theory used in the analysis of test results is usually the classical test theory (CTT) that is widely developed and a mainstream among psychologists and education experts, as well as other areas of behavioral studies (Embreston & Reise, 2000). However, it was found that CTT has a weakness because it is examinee sample dependent and item dependent sample (Fan, 2008; Hambleton & Swaminathan, 1985).

That weakness triggers a new theory that is more adequate, namely Item Response Theory / IRT. If CTT focuses on information at the test level, IRT mainly focuses on information at the grain level. Implementation of the IRT model is based on several assumptions: (1) the results of a participant's test of an item can be predicted by a set of factors called traits or capabilities; and (2) the relationship between the test results of

participants on a grain and a set of traits is described by a monotonically rising function called the characteristic curve (ICC) item (Hambleton, Swaminathan & Rogers, 1991; Harvey & Hammer, 1999; Suryabrata, 2000). Then the ICC curve explains the relationship between traits and test results of participants on each item.

Item response theory is a modern method of measurement commonly used in item analysis. Item response theory seen from the item characteristics of the question is determined by the response of the test participants (both high and low ability). Based on the IRT model, the relationship between examinee's responses and test items can be explained by so-called item characteristic curve (ICC) (Wang, 2006). The development of the IRT is based on two features, namely latent trait or abilities and Item Characteristics Curve (ICC). Latent trait is the ability of test participants on a question item can be expected by a set of factors. The ICC shows the relationship between the testers' ability on a question item and the underlying latent ability device (Hambleton & Swaminathan, 1985).

The first and most important step in the IRT application is parameter estimation, both the testers' ability parameter (θ) and the item characteristic parameter. The approach that can be used to estimate the item parameter is the maximum likelihood estimation (MLE) method (Matthew, 2007). The basic principle of the MLE method is if there are random instances $x_1, x_2, \dots, x_n$ of the distribution having a functionality of chance probability $f(x_n;\theta), \theta \epsilon \Omega$. This function of the coexistence of opportunity is seen as a function of θ (Hogg & Craig, 1978). The Categorical Response Function (CRF) graph is the relationship between the true answering opportunities obtaining category k score on the -j item with the ability of test participants (θ) (Toit, 2003). The higher the ability of the test participants, the chances of answering a correct item correctly will increase.

Based on some of these descriptions, the researcher will guess the critical thinking ability of the participants of multiple choice test with the reason of physics subject with the approach of politico theory granular theorist with Partial Credit Models (PCM) model.

METHOD

Type of research used in this research is qualitative research with descriptive approach. Determination of subjects in this study, using sampling technique purposive sampling where sampling of data sources with certain considerations (Sugiyono, 2012). The sample of the test is constituted from 61 students of SHS 1 Depok Yogyakarta, who were took science as concentration of their studies, in the XI MIA class of academic year of 2017-2018. The participants were selected via "random sampling" method. In the other side, the instrument test consists of 12 multiple choice questions compiled with reference to the indicator of critical thinking skills in physics learning.

Material

The data collection techniques were used two-tier multiple-choice test to assess the student's critical thinking skills. The items of "Two-tier multiple-choice test" were scored as (4) "Question and reason answers are correct", (3) "The answer to the question is wrong but the reason is right", (2) "The answer to the question is correct but the reason is wrong", and (1) "Question answers and wrong reasons". The results of the test were analyzed by item parameter estimated and participant parameter estimated.

Data Analyses

The empirical validity of the test instrument was counted using Partial Credit Model (PCM). PCM is a polytomous scoring model derived from the Rasch model in the dichotomous data (Retnawati, 2016). PCM was used to analyze the test items which have several steps to solve them. The synchronization of the test item and the PCM model was interpreted based on the average means of INFIT Mean of Square (Mean INFIT MNSQ) and the standard deviation (Hambleton & Swaminathan, 1985). If the average mean of INFIT MNSQ was 1.0 and the standard deviation was 0.0 or the mean of INFIT t approached 0.0 and the standard deviation was 1.0, the entire test items were synchronized with the model. An item or testee/case/person is declared to be suitable to the model in the range of INFIT MNSQ of 0.77 to 1.30. In addition, the item is declared to be good when the index of difficulty was more than -2.0 or less than 2.0.

For data analyses, PARSCALE and QUEST Program were used. Each item can be obtained from the item-item characteristic parameter, i.e the difficulty level of the $-j$ (B_j) and the degree of difficulty k between

the categories of items about $-j$ (B_{jk}). Next calculate the capability parameters (θ) and estimate the PCM through the Categorical Response Function (CRF) graph for each item.

FINDINGS

Item test reliability

In the main field testing phase, the test instrument on critical thinking skills was tried out and the result can be seen in Table 1.

Tabel 1. The result of reliability test on critical thinking skills

Parameter	Estimate Item	Estimated Testee
INFIT MNSQ	0,99 ± 0,33	0,98 ± 0,33
OUTFIT MNSQ	1,57 ± 2,23	1,57 ± 3,92
Reliability of estimate	0,55	0,66
Average difficulty	0,63 ± 1,60	

Based on the analysis, the reliability of the instrument (test) is qualified as good instrument.

Item Parameter Estimates

The estimation results of the characteristic parameters of critical thinking test items using the PCM model show that the item has a level of difficulty of various questions. The results of grain parameter estimation using the PARSCALE program are presented in Table 2.

Table 2. Summary Statistics of Parameter Estimates

PARAMETER	MEAN	STN DEV	N
SLOPE	1,876	0,000	12
LOG (SLOPE)	0,629	0,000	12
THRESHOLD	0,620	0,000	12
GUESSING	0,000	0,000	0

In each item in Table 3., the estimation of power different parameters (a_i) can be seen in the SLOPE parameter, while the difficulty level parameter estimation results (B_{ij}) can be seen in the THRESHOLD parameter and can also be seen GUESSING parameters. In this case the estimation result of the estimation parameter for all items is 0 (zero), this indicates the analysis guessing by students was unconsidered. The level of difficulty in the item as a whole is classified as medium with a value of 0.629. Furthermore, to adjust grain parameters in PCM can be done by reducing the parameters β items with each category parameter. The category parameters can be seen in the STEP PARAMETER in Table 3.

Table 3. Parameters Category

Scoring Function	1,000	2,000	3,000	4,000
Step Parameter	0,000	0,917	1,516	-2,433
S.E.	0,000	0,306	0,316	0,553

$(\bar{\beta})$	$(\overline{\beta_1})$	$(\overline{\beta_2})$	$(\overline{\beta_3})$	$(\overline{\beta_4})$
0,629	0,629	-1,689	-0,288	3,062

The estimation results of the level of difficulty level of the above points show that an increasingly high ability is needed to obtain higher value categories. The level of difficulty to reach the category of value 2 (correct answer but wrong reason) is -1,689 which means it needs average (medium) ability; The level of difficulty to reach the value category 3 (wrong answer but the correct reason) is -0.288 which means that it is needed ability above average (high); and the level of difficulty to reach the value category 4 (answer and correct reason) is 3,062 which means that it requires very high abilities. While the value category 1 (answers and wrong reasons) can be obtained by the ability of low to very high because this category gives a score

even though for the wrong answer. A higher category threshold in PCM scoring is not always greater than the threshold of the previous category [36].

Estimated Ability of Participants

The results of the estimation of the ability of the test participants are presented in the following histogram.

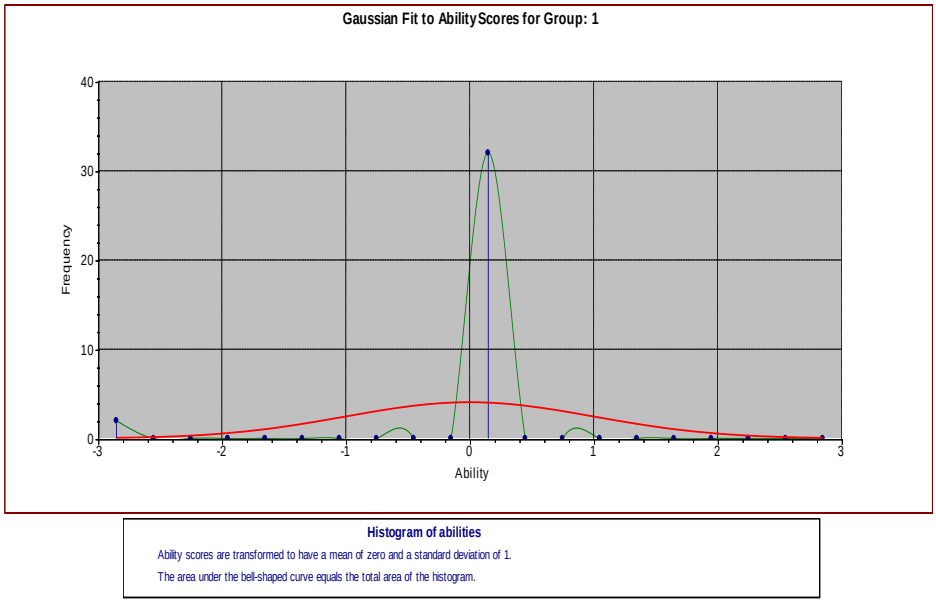

Figure 1. Graph of CRF Estimation of Students' Ability

The histogram above shows that students' critical thinking skills are not spread normally by interpreting them into criteria in Table 4.

Table 4. Critical thinking ability categories

Sample	Ability value	Interpretation
0	2,00 - 3,00	Very high (above average)
1	1,00 - 2,00	High
36	-1,00 - 1,00	Average
1	-2,00 - 1,00	Low
2	-3,00 – (-2,00)	Very low

By using the criteria, the results of the estimation of the parameters of critical thinking skills of students showed that no students with critical thinking abilities were very high, while 1.67% had high critical thinking skills, 60% had average critical thinking skills, 1.67% have critical thinking skills below average (low) and 3.33% have very low critical thinking skills. This shows that students' critical thinking skills in physics are still at the level of low order thinking ability.

PCM Model Estimation

PCM assessments are presented in the Categorical Response Function (CRF) graph for each item. The ICC chart shows the probability of answers to each student's ability. From this graph shows that the greater the value of students' abilities, the greater the likelihood that students will answer correctly with the right reason (category 4). The average competent student will answer in the medium category, that is, only the answer or reason is correct (categories 2 and 3). Low-ability students will answer in the low category, either the answer or the reason is wrong (category 1). The picture below is a PCM assessment for item 1.

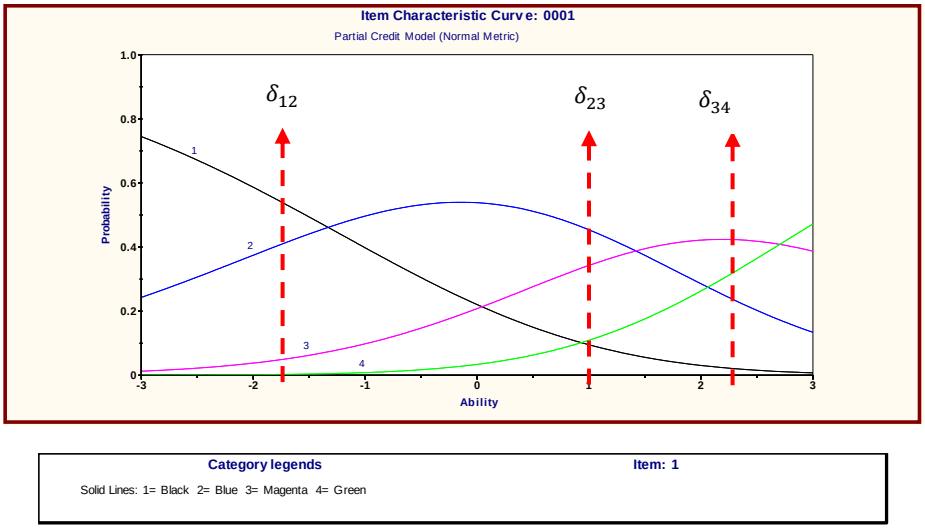

Figure 2. Graph of CRF for item number 1

Based on the CRF graph for item 1, the probability of the level of difficulty relative to answer is equal to B_1= 0,75, B_2 = 0,55 , B_3 = 0,43 and B_4 = 0,50. Graph CRF on item number 1, there are intersection points between categories. The level of difficulty of the category related to the transition of one category to the next category is considered as the point where two categories have the same probability of being selected for the level of relevance. Item 1 has (δ_{12}) = -1,3, (δ_{23}) = 1,4, (δ_{34}) = 2,7 on the range of capability scales $-3 > \theta > 3$. The δ_{12} means that individuals who have the ability level θ below -0.25 have the probability to answer wrong questions for the wrong reasons too (category 1). While individuals who have the ability level above -0.25 have the probability to give the right reasons even though the answer is wrong.

RESULT, DISCUSSION AND SUGGESTIONS

The instrument test used to measure students' critical thinking skills consisted of 12 test items. Test reliability calculated based on measurement error was calculated based on the estimation according to the test (Wright & Master, 1999: 96) reached 0.66 so the reliability was high.The results of item analysis tests of mastery of critical thinking skills PCM 1-PL used the QUEST program with the INFIT MNSQ limit 0.98 means that all items meet the requirements.The results of item analysis indicate that all items have an INFIT MNSQ value in the lower and upper limit of the range, 0.77 to 1.30. Thus, all items can function as items measuring the critical thinking skills (Istiyono, 2016). In addition, analysis of test items shows the average value of INFIT t is 0.0 with a standard deviation of 0.98. An item serves as a measuring item of critical thinking skills if the item is declared fit with the model because it meets the fit requirements the statistics required in the QUEST program, which is fit with the model when the analyzed item has an average value average INFIT t approaches 0.0 with standard deviation 1.0 (Adam & Kho, 1996). Thus, overall the items analyzed were fit according to PCM 1-PL.

The difficulty level parameter estimation results (B_{ij}) shown by PARSCALE output that an increasingly high ability is needed to obtain higher value categories. The level of difficulty in the item as a whole is classified as medium with a value of 0.629. The level of difficulty to reach the category of value 2 (correct answer but wrong reason) is -1,689 which means it needs average (medium) ability; The level of difficulty to reach the value category 3 (wrong answer but the correct reason) is -0.288 which means that it is needed ability above average (high); and the level of difficulty to reach the value category 4 (answer and correct reason) is 3,062 which means that it requires very high abilities.

Assessing tests is based on the steps that the examinee can complete. Even though they have only just completed the initial stage, the exam participants have already received scores. The highest score is of course obtained when the examinee has completed all phases of the exam questions in that clause. This assumption was then developed into PCM. When it is assumed that a point follows a partial credit pattern, the higher the ability of individuals is expected to have a higher score than individuals who have the ability to decrease (Widhiarso, 2010). According to Wright & Masters, PCM is also appropriate for analyzing responses to the

measurement of critical thinking and conceptual understanding in science (Linden & Hambleton, 1997). The physics achievement test is a test that is done with the right steps.

Table 4 shows the results of estimating the parameters of critical thinking skills that showed that no students with critical thinking abilities were very high, while 1.67% had high critical thinking skills, 60% had average critical thinking skills, 1.67% had critical skills below average (low) and 3.33% thinking skills have very low critical thinking skills. This shows that students' critical thinking skills in physics are still at the level of low order thinking abilities. This illustrates that critical thinking skills are less developed by teachers in schools. There are two possibilities behind it. The first possibility is that teachers do not develop critical thinking skills. While critical thinking skills are characteristic of science learning. As a result, students lack mastery of critical thinking skills. It is possible that both teachers have trained critical thinking skills, but are less oriented to divergent patterns as a basis for skill development. It may be because one of the dominant factors is the habit of teachers taking measurements with multiple-choice forms that are clearly oriented to the development of convergent thinking patterns.

Findings of the study indicates that students' critical thinking skills in physics are still at the level of low order thinking ability. Some researches support this finding, as several studies point out (Istiyono, 2017; Oliveira and Rodrigues, 2004; Rivard, 2004 and Newton, 1999), science classrooms are still strongly teacher directed, that is, the teaching and learning model used is mainly the transmission model that does not foster critical thinking. The lower critical thinking skills find on students have difficulty in analyzing the arguments presented in the problem and have difficulty considering the definition. The ability of students to answer questions is still in the stage of memorizing and understanding. The same thing is shown by other studies that the difficulties and lack of understanding of students is partly because the critical thinking skills possessed are low (Chee, 2010; Khol & Noah, 2008; Sari, Parno & Taufiq, 2016). This indicates that there is still a lack of learning efforts, both strategies and assessment systems that accommodate students to achieve higher-order thinking skills.

The choice test by the teacher in high school cannot be released from the form tests commonly used in high-level examinations, such as the tests used in the national exam (UN) and college selection. The teacher is more complicated to discuss the test questions used in the National Examination and college selection. The further impact is the use of multiple choice tests in the National Examination and college selection (Subali & Surastuti, 1991). There are many educators who had failed to give questions regarding the knowledge content of thingking skills of the students, the educators are only able to give questions regarding the aspect of students' that are memorized and understood by the concept (Jensen, 2014). Students are almost never drilled to apply critical thinking methods to solve problems.

In the context of assessment for learning, simple practice tests can affect tests that require complex thinking if the test is associated with learning experiences. The test instrument with a one-size-fits-all spirit requires educators to organize learning programs that are more oriented to being able to understand the test or termed teaching for the test (Jehlen, 2007). The results of studies in the US compiled since 1990 show that high-risk tests have a positive and negative effect. The positive effect is that schools are motivated to achieve better performance, there are also teachers who change learning strategies in a better direction, which is more oriented to problem solving. However, negative effects arise in the form of the emergence of stress and fatigue, some even have an impact on the decline of the morale of teachers and students (Abrams, 2007: 80-86).

Further studies that used the two-tier multiple-choice test to analyze critical thinking skills are necessary. Thus, the ability to think critically as part of high thinking skills still needs to be developed. In addition, quantitative and qualitative studies which investigate why students critical thinking skills and the right test should be carried out. Discussions about the necessity and quality of the instrument of test frequently come to the fore.

REFERENCES

Adams, R.J dan Khoo, S.T. (1996). *Quest* : *The interactive test analysis system version 2.1. Victoria*. The Australian Council for Educational Research.

Adodo, S. O. (2013). Effects of two-tier multiple choice diagnostic assessment ıtems on students' learning outcome in basic science technology (BST). *Academic Journal of Interdisciplinary Studies, 2*(2),201–210. doi:org/10.5901/ajis.2013.v2n2p201.

Bailin, S., & Siegel, H. (2003). *Critical thinking*. In N. Blake, P. Smeyers, R. Smith, & P. Standish (Eds.), The Blackwell guide to the philosophy of education (pp. 181-193). Blackwell Publishing.

Baker, J. G., Rounds, J. B., & Zeron, M. A. (2000). A comparison of graded response and rasch partial credit models with subjective wellbeing. *Journal of Educational and Behavioral Statistic, 25*(3), 253- 270.

Chee, T.C. (2010). Common misconceptions in frictional force among university physics students. *Journal on Teaching and Learning, 16*(2), 107-116.

Du Toit M. (2003). *IRT from SSi: Bilog-MG, Multilog, Parscale, Testfact*. Lincolnwood: Scientific Software International. Inc.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologist*. NJ: Lawrence Erlbaum Associates Inc.

Ennis, R.H. (2002). *Goals for a critical thinking curriculum and its assessment*. In Arthur L. Costa (Ed.), Developing minds (3rd Edition). Alexandria, VA: ASCD. Pp. 44-46.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their ıtem/response person statistics. *Educational and Psychological Measurement, 58* (3), 357-381.;

Gronlund NE., & Linn RL. (1990). *Measurement and evaluation in teaching* (6th ed). New York: Collier Macmillan Publishers.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and application*. Boston, MA: Kluwer Inc.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of ıtem response theory*. CA: Sage Publication Inc.

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist, 27* (3), 353-383.

Hogg RV, Craig AT. (1978). *Introduction to mathematical statistics*. New York: Macmillan Publishing Co Inc.

Istiyono, E. (2014). Pengukuran kemampuan berpikir tingkat tinggi fisika peserta didik SMA di DIY. Disertasi doktor, tidak diterbitkan, Universitas Negeri Yogyakarta, Yogyakarta.

Istiyono, E. (2017). *The analysis of senior high school students' physics HOTS in bantul district measured using PhysReMChoTHOTS*. AIP Conference Proceedings 1868, 070008 (2017); https://doi.org/10.1063/1.4995184.

Jensen, J. L, et al. (2014). Teaching to the Test or Testing to Teach: Exams Requiring Higher Order Thinking Skills Encourage Greater Conceptual Understanding. *Educational Psychology Review*, 1-24.

Khol, P.B. & Noah, D.F. (2008). Patterns of multiple representation use by experts and novices during physics problem solving. *Physical Review Special Topic- Physics Education Research* 4.010111.

Kubiszyn, T., & Borich, G. D. (2013) *Educational testing and measurement: Classroom application and practice*. Hoboken, NJ: Willey.

Masters, G.N. (1999). *Partial credit model*. Dalam J.P. Keeves & G.N. Masters (Eds.). Advances in Measurement in Educational Research and Assessment. Amsterdam: Pergamon.

Matthew, SJ. (2007). Marginal maximum likelihood estimation of ıtem response model in R. *Journal of Statistical Software,20(10).* http://www.jstatsoft.org/ [27 April 2012].

Newton, P. G. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education, 21*(5), 553-576.

Noer, S.H. (2009). Peningkatan Kemampuan Berpikir Kritis Matematis Siswa SMP Melalui Pembelajaran Berbasis Masalah. Prosiding.

Obaidat, I. & Malkawi, E., (2009). The grasp of physics concepts of motion: ıdentifying particular patterns in students' thinking‖. Georgia Southern University: *International Journal for the Scholarship of Teaching and Learning, 3*, (1), Januari, 11-12.

Oliveira; M. and Rodrigues, A. (2004). *Portfolio as a strategy to interrelate research in education and physics teachers practices*. In M. Michelini (Ed.), Quality development in teacher education and training: Second International GIREP Seminar 2003 selected contributions, Forum, Udine, Italy.

Rivard, L. P. (2004). Are language-based activities in science effective for all students, including low achievers? *Science Education, 88*(3), 420-442.

Sapriya. (2011). Pendidikan IPS: Konsep dan Pembelajaran. Bandung: PT Remaja Rosdakarya.

Sari, ALR., Parno & Taufiq, A. (2016). Kemampuan Berpikir Kritis dan Pemahaman Konsep Fisika Siswa SMA pada Materi Hukum Newton. Prosiding Seminar Nasional Pend.IPA Pascasarjana Universitas Negeri Malang.

Subali, Bambang & Surastuti, Etty. (1991). Persepsi Siswa Kelas III SMA terhadap Lembaga Bimbingan Tes. *Jurnal Kependidikan, 21*(2): 1-9.

Sudjana, N. (2013). Penilaian Hasil Belajar Mengajar. Bandung: PT Remaja Rosdakarya.

Sugiyono. (2012). Metode Penelitian Kuantitatif Kualitatif dan R&D. Bandung: Alfabeta.

Suryabrata, S. (2000). Pengembangan Alat Ukur Psikologi. Yogyakarta: Andi.

Suwarto (2012). Pengembangan The Two-Tier Diagnostic Tests Pada Bidang Biologi. Proceeding Seminar Nasional: Profesionalisme Guru Dalam Perspektif Global.

Tjalla, A. (2010). *Potret Mutu Pendidikan Indonesia Ditinjau dari Hasil-Hasil Studi Internasional*. In: Temu Ilmiah Nasional Guru II: Membangun Profesionalitas Insan Pendidikan Yang Berkarakter dan Berbasis Budaya, 24–25 November 2010, Tangerang Selatan.

Tognolini, J., & Davidson, M. (2003). *How do we operationalise what we value? Some technical chalenges in assessing higher order thinking skills*. Paper presented in the National Roundtable on Assessment Conference, Darwin, Australia.

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2),159–169. https://doi.org/10.1080/0950069880100204

Trilling, B. & Fadel, C (2009). *Learning and innovation skills. 21st century skills learning for life in our times*. (pp45-60). San Francisco: Jossey-Bass

Tuysuz, C. (2009). Development of two-tier diagnostic ınstrument and assess student's misunderstanding in chemistry. Scientific Research and Essay 4.

Van der Linden, W. J & Hambleton, R. K. (1997). *Handbook of modern ıtem response theory*. New York: Springer-Verlag New York, Inc

Wagner, T. (2008). *The global achievement gap: Why even our best schools don't teach the new survival skills our children need--and what we can do about it*. New York: Basic Books.

Walser, N. (2008). Teaching 21st century skills. *Harvard Education Letter, 24*(5), 1-3.

Wang, F.-H. (2006*). Application of componential IRT model for diagnostic test in a standard conformant elearning system.* In Sixth international conference on advanced learning technologies (ICALT'06).

Wardani, R. K., Yamtinah, S., & Mulyani, B. (2015). Instrumen Penilaian Two-Tier Test Aspek Pengetahuan Untuk Mengukur Keterampilan Proses Sains (KPS) Pada Pembelajaran Kimia Untuk Siswa SMA/MA Kelas X, 4(4),156–162.

Widhiarso W., (2010). Model Politomi dalam Teori Respon Butir. Fakultas Psikologi UGM, Yogyakarta.

Widhiarso, W. (2010). Model politomi dalam teori respons butir. Yogyakarta: Psikologi UGM.

Wright, B.D. & Masters, G.N. (1982). Rating scale analysis. Chicago: Mesa Press.

Zainul, A., & Nasution, N. (2001). Penilaian hasil belajar. Jakarta: PAU-PPAI, UT.

Zohar, A., & Dori, Y. J. (2003). Higher order thinking skills and low-achieving students: Are they mutually exclusive. *Journal of the Learning Sciences, 12*(2), 145–181.