REiD (Research and Evaluation in Education), 3(2), 2017, 114-123
Available online at: http://journal.uny.ac.id/index.php/reid
Research Article

Developing physics problem-solving skill test for grade X students of senior high school

[1]Amipa Tri Yanti Nadapdap; *[2]Edi Istiyono
*Graduate School of Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
*Email: edi_istiyono@uny.ac.id

Submitted: 24 July 2017 | Revised: 29December 2017 | Accepted: 29 December 2017

Abstract

This research aimed to develop a physics problem-solving skill (PSS) test for grade X students of senior high school which met test instrument characteristics and feasibility. The development stages included: (a) test designing, (b) test trial, and (c) test revision and preparation. The designing stage included: (1) needs analysis, (2) mapping, (3) drawing conclusion, (4) determining test purpose, (5) determining competencies, (6) determining materials, (7) preparing answers, (8) writing items, (9) validating content, (10) improving and preparing the test, and (11) preparing the scoring guide with PCM. The trial stage consisted of: (1) determining trial subjects, (2) performing trial, and (3) analyzing trial result data based on IRT. The study was performed in Kulonprogo involving 281 students. The result shows that the instrument fulfills content validity with Aiken's V of 0.95 to 0.98. Based on INFIT MNSQ criteria, 52 items fit PCM, item difficulty index ranges from -1.47 to 0.88, meaning that all items are good, and information function analysis and SEM show that the test fits the ability between -1.3 and 2.7. Therefore, the test instrument meets the characteristics and feasibility to measure physics PSS in high school.

Keywords: *problem-solving skill, testing, physics, assessment*

How to cite item:

Nadapdap, A., & Istiyono, E. (2017). Developing physics problem-solving skill test for grade X students of senior high school. *REiD (Research and Evaluation in Education), 3*(2), 114-123. doi:http://dx.doi.org/10.21831/reid.v3i2.14982

Introduction

Assessment in education must be performed in order to measure student's cognitive skills. It is expected to increase the success of learning process. Thus, a series of test assessment instruments should be developed.

A test is a planned measurement instrument used by educators to give an opportunity to students to show their achievement and it is related to predetermined objectives (Cangelosi, 1995). A test can show the success rate of teaching based on target aspects. Its preparation is adjusted to its purpose, e.g. a summative test is used to measure student's achievement, formative test is to measure the success of learning process and a diagnostic test is to examine student's difficulty before a teaching and learning process.

There are other test types used to measure certain skills, such as cognitive, affective and psychomotor skills. A test has many variations in its preparation, i.e. multiple choice, sentence completion, listing, true-false, essay, matching, and modified form (Tonidandel, Quiñones, & Adams, 2002) Therefore, a test should be developed consistently, adjusted to its form and measurement purpose.

Problem solving is a skill which should be improved in the 21st century. Indonesia is a

developing country in terms of education, so problem-solving skill (PSS) is a skill which must be mastered by students in Curriculum 2013 (K-13). Rating in K-13 is done in the form of authentic assessment that assesses the start of the input, process and results (outputs) of learning, including attitudes, knowledge and skills. An assessment technique is relevant with the scientific learning process and able to assess the students' ability in the teaching and learning process and results. Regulation of Minister of Education and Culture No. 59 of 2014 states that problem-solving skill is required to achieve the objectives of K-13 to give students the life skills to be an individual and citizen who is faithful, productive, creative, innovative, and affective, as well as able to contribute to social life, nation, country, and world civilization. This skill is expected to produce scientific students (Nadapdap & Lede, 2016). Therefore, problem-solving skill test should be developed.

Problem-solving consists of four parts: (1) understanding a problem; (2) preparing a plan for solution; (3) performing a plan; (4) reexamining (Pólya, 1957). The indicators of problem-solving development according to Helaiya (2010) are including: (1) the ability to identify problem and problem-solving process; (2) the ability to define problem by thinking about different situations from the reality; (3) the ability to think of many possible alternatives of some solutions; (4) the ability to verify result of solution; and (5) the ability to verify in a solution acquisition process. Therefore, the aspects of a problem-solving test can be developed, including: (1) understanding; (2) planning a solution in problem solving; (3) describing a problem; (4) finding a way to solve a problem according to the planned solution; (5) bringing about a problem; and (6) evaluating the problem solving result assessment (Helaiya, 2010).

In physics teaching, PSS is the main topic in physics education research (PER) because it has long-term benefits. Further, physics PSS can help students understand the concepts of physics in real terms.

The most important part in teaching physics is students are expected to understand the real world. The theory of learning is based on one's process with its various interactions to gain experience which makes one have changes in cognitive, affective and psychomotor skills (Slameto, 2010).

According to Bloom, cognitive process thinking consists of Lower Order Thinking which consists of abilities to memorize, understand and apply, and Higher Order Thinking which consists of the ability to analyze, evaluate and create. PSS is a part of higher order thinking (Carvalho et al., 2015). Higher order thinking skills (HOTS) are: (1) higher order thinking at the upper part of Bloom's cognitive taxonomy, (2) teaching purpose behind cognitive taxonomy which can prepare students to perform knowledge transfer, (3) ablility to think, which means that students can apply the knowledge and skills that they develop during the learning process in a new context (Brookhart, 2010).

PSS can be measured by using a test which is consistent with the purpose of student's higher order thinking. Besides, the test which is used has to require the use of knowledge and skills in the new situation. In order to assess the HOTS, something new should be used. One of the ways to do that is using a test which is in the valid category — a test which is aimed to measure the HOTS.

One of the modern measurement theories is called Generalized Partial Credit Model (GPCM). GPCM is the improved Partial Credit Model (PCM). The PCM discriminant items are constant or 1, while the value GPCM discriminant varies. PCM is also appropriate for analyzing the response to the measurement of critical thinking and conceptual understanding in science (Istiyono, 2016). PCM was developed to analyze the test items that require several steps to resolve.

GPCM can be applied to tests, which is done with the steps that are clear for the testee. A physics achievement test is a test administered following the exact steps. Therefore, GPCM is expected to be applied properly.

Multiple-choice test has advantages, including: (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 obviously right or wrong, so it is an objec-

tive assessment (Istiyono, 2016). Therefore, using a multiple-choice item test to measure the problem-solving skills is good to do.

Assessment in education uses two kinds of measurement theories: classical measurement theory and modern measurement theory or item response theory (IRT). The classical test theory (CTT) is also called the True-Score Classical Theory. The CTT is so named for the elements of this theory have been developed and applied for a long time, but still survive today (Suryabrata, 2002). According to the classical theory of measurement, measuring by using measurement score result is usually conducted partially based on the steps that must be taken in order to correct an answer items. Scoring is conducted at every step and score each item participant adds a score obtained by the students of each step, and the ability is estimated by the raw scores.

A scoring model is not necessarily right, because the level of difficulty of each step is not taken into account. Since a test is an instrument that provides stimulus in the form of a command or a question which requires a response from the test participants, the response which is given by the test participants stated in a score is easy to interpret.

In addition, the scoring results of a multiple choice test is gained by the use of a dichotomous model, which means that if the item response is correct, it is given a score of 1 and if the response is wrong, it is given a score of 0. Teachers do not use polytomous scoring models that would be more equitable because it considers item response measures. These dichotomous scoring models have yet to appreciate the steps of problem solving, because different error rates will result in the same score of 0. Dichotomous scoring models are certainly less fair. One of the scoring guidelines that can be selected is the provision of each category, as presented in Table 1.

HOTS is interdependent with students' problem-solving skill. Physics PSS can really help students solve physics problems in learning. With that skill, students are expected to solve a given problem with an effective solution. An accurate solution is seen based on the aspects to be measured, the aspects which measure students' problem-solving skill consistent with a students' operational stage of formal thinking. High school students are 17 years old in average, an age when they can think abstractly and logically which is categorized as problem solving stage.

Table 1. Scoring category & description

Category	Guidelines
Category -1	The students are wrong in writing the concepts used and the results are wrong. This is indicated by the students that answer question one and also one of the reasons
Category -2	The students are wrong in writing the the concepts used but the results can be correct. This is indicated by the students' correct answer to questions wrong basis.
Category -3	The students are correct in writing the concepts used but the end result is wrong. This is indicated by students' wrong answer to the question and correct reason.
Category -4	The students are correct in writing the concepts used and the results are correct. This is indicated by the students' correct answer to questions and correct reason.

(Istiyono, 2016)

Thinking skill is required in scientific thinking. Further, scientific thinking is involved in hipothetico-dedutive and inductive types (Piaget, 2005). Scientific thinking is working effectively and systematically, as well as proportionally. In terms of PSS, at that age, students can draw conclusions and interpret and develop hypotheses.

However, the existing test did not describe the skill which demands thinking consistent with the optimization of the characteristics of student's ability (Eraikhuemen & Ogumogu, 2014). Therefore, the higher the characteristics of the cognitive development stage, the more orderly and abstract the students' thinking.

The appropriate assessment to get information on student's thinking skill based on characteristics is by giving an appropriate test for measuring the thinking competence level. However, the current development of assessment is only based on the Classical Theory assumption in which scoring is performed step by step and student's score per item is gained

by adding the student's score in every step, and the skill is estimated by raw scores. Thus, an assessment which can cover the thinking skill level such as problem identification to assessment should be developed (Gok, 2010). Therefore, a physics problem-solving skill test instrument was developed for grade X students of high school. The purpose of the study was to produce an instrument to measure physics PSS in grade X students in their even semester and to get the characteristics of the physics PSS assessment instrument.

Method

This study is a developmental study with quantitative approach. The instrument development used in this study was the modified Orindo and Antonio model (Oriondo & Dallo-Antonio, 1998). The developed assessment instrument was a physics PSS test for grade X students in their even semester of 2016/2017 academic year.

Population

The study was performed in public high schools (or *Sekolah Menengah Atas Negeri* – *SMA Negeri*) in Kulonprogo Regency, Yogyakarta, i.e. *SMA Negeri* 1 Wates, *SMA Negeri* 2 Wates and *SMA Negeri* 1 Pengasih. The trial subjects were 281 students. The sample consisted of the students who had received similar tested materials in the three schools and they were selected not based on ranking.

The valid instrument was used in the form of a PSS test instrument packed in two packages of materials, each containing 30 questions with 8 anchor items of multiple-choice type reasonably ready for use in empirical testing. Testing is done by testing the instrument to 281 students.

The respondents were chosen from the class which had studied the materials of elasticity, static fluid, temperature and heat and optical equipment. They were classes X of *SMA Negeri* 2 Wates, SMA Negeri 1 Wates, and *SMA Negeri* 1 Pengasih Kulon Progo. The test results were analyzed by reference of the test using the criteria of acceptance of instrument suitability with Rash model, seen from the mean value of INFIT MNSQ (Mean Of Square) which ranged from 0.77 to 1.33 (Adams & Khoo, 1996, p. 30).

The trial sample in the analysis by IRT consisted of 281 students, who were required in IRT model research. Some experts consider that the bigger the sample size, the better the measurement result will be. One of the bases for using 281 students as the trial sample was Shin, who was using 200 to 1000 (Shin, 2009). Therefore, the 281 students used in this measurement was considered adequate.

Data Collection Technique

The instrument development was based on the aspects and sub-aspects of PSS test, including: (a) test design, (b) test trial, and (c) test revision and preparation. Meanwhile, the instrument designing stage consisted of: (1) needs analysis, (2) mapping, (3) drawing conclusion, (4) determining test purpose, (5) determining tested competencies, (6) determining tested materials, (7) preparing test answers, (8) writing items, (9) validating content by expert, (10) improving and preparing test, and (11) preparing scoring guide with Partial Credit Model (PCM). The trial stage consisted of: (1) determining trial subjects, (2) performing trial, and also (3) analyzing the trial result data based on IRT.

Figure 1 shows the test development stage. The test developed was a physical test used in high school with problem-solving aspect. The test was developed in the form of a multiple choice item consisting of 60 items including 8 anchor items. The test developed yielded 2 sets of problems with package A of 30 questions and package B of 30 questions. Each package has 8 anchor items.

The data analysis employed in this study was Partial Credit Model 1 PL (PCM 1-PL) for the testing item fitness of the physics PSS test for grade X students of high school. Based on IRT, the sample was adequate and good according to PCM 1-PL model (Adams & Khoo, 1996). The content validity analysis was performed qualitatively by material experts using Aiken index. The content validity analysis was performed qualitatively by material experts using Aiken's V index. Based on the index, the item was valid if the minimal Aiken's V is 0.87 (Aiken, 1980).

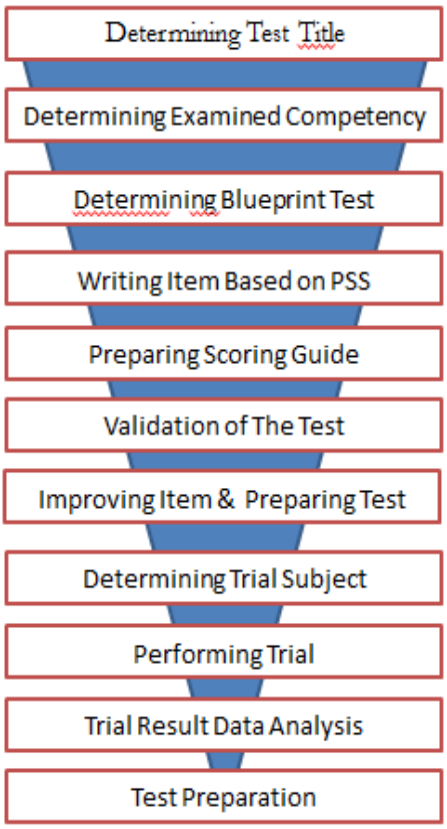

Figure 1. Phases of test development

The data analysis was performed on several aspects, including (1) the fitness of instrument items, (2) the reliability, (3) the item characteristic curve (ICC), (4) the difficulty index, and also (5) the total information function and standard error measurement (SEM). The goodness of the fit test for the overall test and testees (case/person) was based on the average INFIT Mean of Square (Mean INFIT MNSQ) and its standard deviation, or by the observation of the average INFIT *t* (Mean INFIT *t*) and its standard deviation. If the average INFIT MNSQ was approximately 1.0 and its standard deviation was 0.0 or the average INFIT *t* was approaching 0.0 and its standard deviation was 1.0, then the whole test fits the model. An item or testee/case/person fits a model in the INFIT MNSQ ranging from 0.77 to 1.30. An item was good if the difficulty index was over -2.0 or less than 2.0 (Hambleton & Swaminathan, 1985). The test reliability was tested by testing the information function and the following criteria presented in Table 2.

Table 2. Criteria of ideal score

Score Criteria Realibility	Category
>0.94	Excellent
0.91 – 0.94	Very Good
0.81 – 0.90	Good
0.67 – 0.80	Acceptable
<0.67	Questionable

Findings and Discussion

The development resulted in a problem-solving skill test with two sets of problems, coded A and B, each consisting of the materials of: elasticity, static fluid, temperature and heat, and also optical instruments. Table 3 shows the item distribution with eight items as the anchor items with the aspects of identification, planning, application, and also assessment.

Table 3. Distribution test

Subject		Elasticity	Static Fluid	Temperature and Heat	Optic
Aspect/ Sub aspect					
Identify	Distinguish	1a* 1b*	8a 8b	17a 17b	24a 24b
	Identify	2a 2b			25a *25b*
Plan	Formulate	3a 3b	9a 9b, 10a 10b	19a 19b	
	Devise	4a 4b			26a 26b
Apply and Execute	Connect	5a 5b	12a 12b, 11a 11b, 16a *16b*		28a 28b
	Apply		13a 13b	21a *21b*	29a 29b
				20a, 20b, 18a 18b	
	Analyze	6a 6b	14a* 14b*	23a* 23b*	27a*, 27b*
Evaluation	Investigate	7a *7b*		22a 22b	30a 30b
	Conclude		15a 15b		

The research product was validated by two assessment experts and five practitioners to assess the feasibility. Aiken index is in the range of 0.8 to 1.00. It can be interpreted that all of the items have good content validity and have supported overall content validity.

The fit goodness was tested for overall test items. The fitness of the overall test items used the principle developed by Adams and Khoo (1996, p. 30) based on INFIT Mean of Square (Mean INFITMNSQ) and its standard deviation or observing the average INFIT *t* (Mean INFIT t) and its standard deviation.

If the average INFITMNSQ was approximately 1.0 and its standard deviation 0.0 or the average INFIT *t* approached 0.0 and its standard deviation 1.0, the overall test fits PCM 1-PL model. Table 4 shows the average INFITMNSQ is 1.00 and its standard deviation 0.02, so the overall test fits PCM 1 PL model.

The fitness determination of each item followed the principle of Adams and Khoo (1996, p. 30) in which an item fits the model if INFIT MNSQ ranges from 0.77 to 1.30. With INFIT MNSQ as the item acceptance limit or fit according to the model (ranging from 0.77 to 1.30) and by using the INFIT *t* from -2.0 to 2.0, the items which met the goodness of fit were found. The INFIT MNSQ value ranged from 0.99 to 1.03. With INFIT MNSQ as the item acceptance limit or fit according to the model (ranging from 0.77 to 1.30), all of the 52 items fit the PCM.

Table 4. Testing the statistic fit parameter level

No	Test Parameter	Item estimation	Case Estimation
1	Average and std.deviation	-0.25 ± 0.28	0.17 ± 0.02
2	INFIT MNSQ	1.00 ± 0.02	1.00 ± 0.12
3	Outfit MNSQ	1.00 ± 0.02	1.00 ± 0.12
4	INFIT ZSTD	0.09 ± 0.75	0.06 ± 1.84
5	Average difficulty	1.00 ± 0.95	
6	Estimate Reliability	0.8	

The result of the reliability testing shows that the value of the reliability of the instrument is 0.28. Based on the relative value, the whole item is reliable as it corresponds to the reliability of the interpretation data of the Rasch model sufficiently categorized.

Figure 2 shows the goodness of item with an analysis by quest. Based on results of the analysis, it can be concluded that the entire test items are in accordance with the PCM model with the whole item being within the range of INFIT MNSQ PSS from 0.77 to 1.33 and using INFIT t with the limit of -2.0 to 2.0 in accordance with Figure 2 that no item exceeds the acceptance limit. In conclusion, 52 items fit the PCM model.

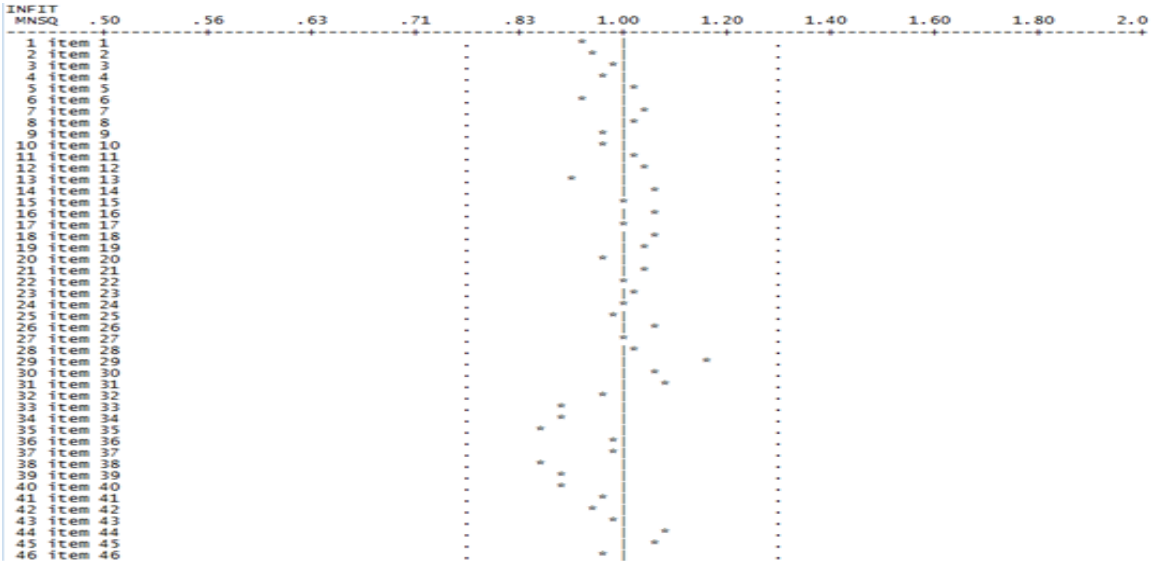

Figure 2. Goodness of fit instrument

Based on the result of analysis, the reliability of the instrument is 0.80. The reliability is adequate. The instrument has adequate strength and reliability because it consists of the items which have high information function (Hambleton & Swaminathan, 1985, p. 94). It may be because the test fits the skill of the tested students.

An item is categorized as good if the difficulty index is higher than -2.0 or less than 2.0 (Hambleton & Swaminathan, 1985, p. 36). Based on the analysis result, the items difficulty is between -0.95 and 1.0 with an average of 0 and standard deviation of 0.32. Therefore, based on the difficulty level, 52 items are good. The average difficulty of the aspect of problem-solving skills are shown by Table 5.

Table 5. Average difficulty of the aspect of problem-solving skills

Aspect	Difficulty
Identify	-0.13
Plan	-0.16
Apply and Execute	0.20
Evaluate	0.54

Construct validity is empirically proven by goodness of fit in the partial credit model (PCM). Table 4 shows the average value and standard deviation of INFIT MNSQ are 1.00 and 0.02, respectively, so the test fits PCM 1 PL. This means that the test is empirically valid. The test contains valid aspects of the PSS. This is because: (1) the items were developed consistently with the appropriate instrument item development procedure, (2) the items were developed from indicators derived from the aspects of the problem-solving skill and physics materials, (3) the test consisted of 52 items whose content validity was examined through expert judgment, and (4) the tryout respondents (students) worked on the test seriously (Istiyono, Mardapi, & Suparno, 2014). The difficulty level b for good item varies between -2.00 and 2.00. An item with the difficulty level of -2.00 is very easy, while that with the difficulty level of 2.00 is very difficult. Based on the test characteristics, the problem-solving skill test had the reliability coefficient, test information function, and estimation parameter which were reliable and had high stability.

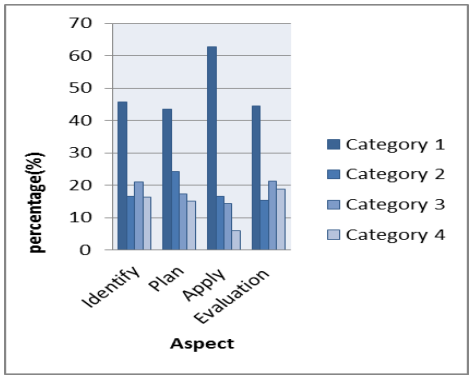

Figure 3. The percentage of difficulty level of aspect

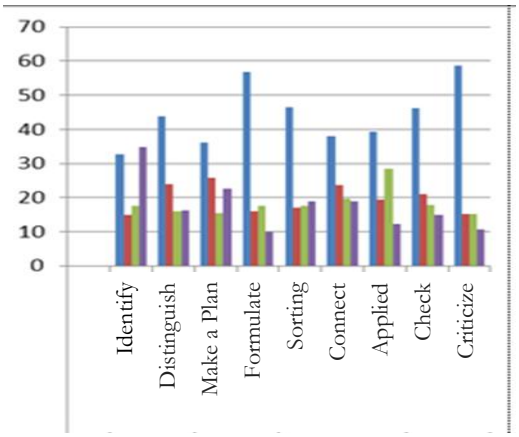

Figure 4. The percentage of difficulty level of sub-aspect

Figures 3 and 4 show the percentage of the aspects and sub-aspects that have been tested. The percentage of the results indicates that the frequency of students' responses to the per item categories of each aspect and the sub-category is put into category one, two, three and four. The first category states that the frequent answers are with a score of one whereas a score of four is expressed by the fourth category.

The percentage of each difficulty level of each item is shown in Figure 3. It shows that the highest difficulty level is in the application aspect. Category 1 percentage shows that most students answer correctly in score 1, so the item is difficult. Figure 3 shows that the percentage of the application in category 1 is 64 and that in category 4 is 6. Figure 4 shows the level of difficulty of each aspect of the problem-solving skill.

The differences between the classical theory and the modern theory in educational assessment can be illustrated by five students

A, B, C, D, and E taking the test as many as 5 items with five alternatives type. The wrong item was given a score of 0 and a maximum of four is given to the correct answer.

The most difficult aspect is the evaluation and implementation aspect. This shows that the students' problem-solving skill in evaluation and implementation aspects is still low.

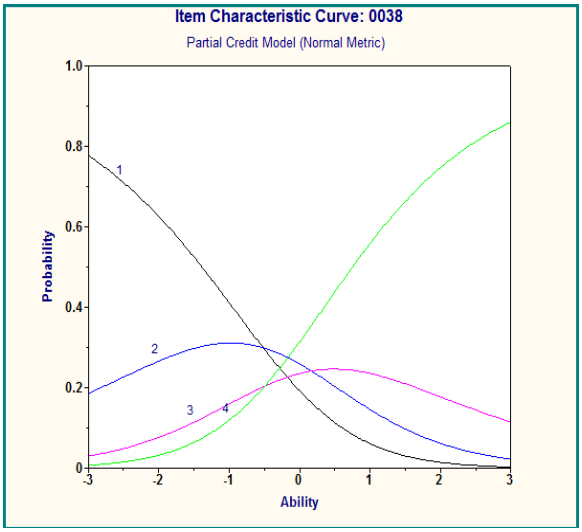

Figure 5. ICC of item no. 38

The characteristic of the item is indicated by the item characteristic curve (ICC) and the difficulty index. Based on the result of the ICC analysis, 52 items are equivalent to the number of the questionnaire items developed. Figure 5 shows an example of ICC for item 38. It shows that in Category 1, the ability of most of the students is very low (θ = -3), in Category 2, the ability of most of the students is low (θ = -1), in Category 3, the ability of some students is high (θ = 1), in Category 4, the ability of most of the students is very high (θ = 3). The difficulty level ranges from small to large sequential categories 1, 2, 3, and 4.

Based on Figure 6, the measurement information is in the range of the ability of -1.3 to 2.7. Therefore, the test instrument is suitable to be used for the students with -1.3 to 2.7 so that in that range, information function shows the ability level estimated by the test (Thorpe et al., 2007, p. 179). The assessment of learning achievement in physics is an assessment of the results of the physics learning process which is a number that describes the characteristics of individual students.

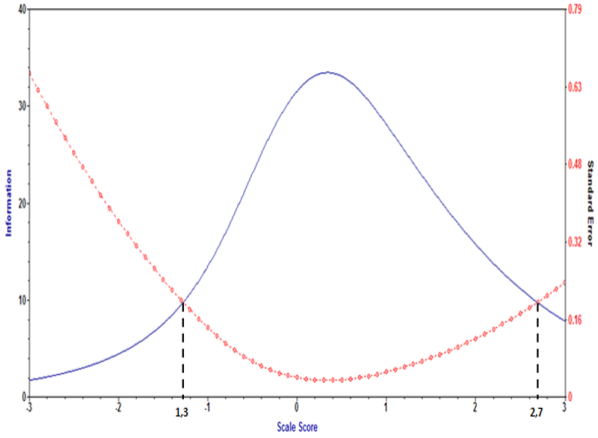

Figure 6. Information function & Standard Error Measurement (SEM)

The relationship between the information function and SEM shows the grand contribution of the test to expressing the latent ability as measured by the test. The greater the value of IF given by the item on the test, the fewer the measurement errors. Therefore, the test is suitable to be used in measuring students' problem-solving skill in the ability categories of medium, low, and high.

Based on the discussion, the test is feasible to use in measuring students' PSS, because: (1) the developed items were consistent with the appropriate instrument item development procedure, (2) the items were developed from problem-solving indicators, (3) the test consists of 52 items whose content validity was examined through expert judgment, and (4) the tested respondents (students) did the test seriously because they were observed by their teachers. This was consistent with the finding of Istiyono et al. (2014). Therefore, the instrument is expected to be able to be used to measure problem-solving skill appropriately. Problem-solving assessment can help students understand a problem quickly (Gok, 2010). Thus, this instrument can be used to measure the exact problem-solving skills.

Conclusion

The problem-solving skill instrument developed in the form of a multiple choice test is based on the problem-solving skills in the physics materials of elasticity, static fluid, temperature and heat and optics consisting of set A and set B each wih 8 anchor items has 52 items.

The problem-solving skill test fulfills the content validity by expert judgment and has empirical evidence of construct validity which fits Partial Credit Model (PCM) based on polytomous data of four categories. The reliability PSS test has met the requirement (reliability coefficient of 0.79). In terms of difficulty level of 52 test items, it is good, between -2 and +2. Thus the test is suitable for measuring the problem-solving ability of students in medium, low and high category of tray.

Based on the information function, the PSS test is appropriate for measuring students' problem-solving skill from -1.3 to 2.7 with a good item difficulty level. Therefore, the test is qualified and so it can be used to measure the physics problem-solving skill of grade X students of high school.

References

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

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, *40*(4), 955–959. https://doi.org/10.1177/001316448004000419

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

Cangelosi, J. (1995). *Merancang tes untuk menilai prestasi siswa*. (D. Tedjasudhana, Ed.). Bandung: Institut Teknologi Bandung.

Carvalho, C., Fíuza, E., Conboy, J., Fonseca, J., Santos, J., Gama, A. P., & Salema, M. H. (2015). Critical thinking, real life problems and feedback in the sciences classroom. *Journal of Turkish Science Education*, *12*(2), 21–31.

Eraikhuemen, L., & Ogumogu, A. E. (2014). An assessment of secondary school physics teachers conceptual understanding of force and motion in Edo South Senatorial District. *Academic Research International*, *5*(1), 253–262.

Gok, T. (2010). The general assessment of problem solving processes and metacognition in physics education. *Eurasian Journal of Physics and Chemistry Education*, *2*(2), 110–122. Retrieved from http://www.eurasianjournals.com/index.php/ejpce

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory : principles and applications*. Boston, MA: Kluwer Nijhoff.

Helaiya, S. (2010). *Development and implementation of life skills programme for student teachers*. Vadodara: Maharaja Sayaji Rao University of Baroda.

Istiyono, E. (2016). The application of GPCM on MMC test as a fair alternative assessment model in physics learning. In *Proceeding of the 3rd International Conference on Research, Implementation and Education of Mathematics and Science (ICRIEMS), 16-17 May 2017* (pp. 25–30). Yogyakarta: Universitas Negeri Yogyakarta. Retrieved from http://seminar.uny.ac.id/icriems/sites/seminar.uny.ac.id.icriems/files/prosiding/PE-04.pdf

Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (Pys-THOTS) peserta didik SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(1), 1–12. https://doi.org/10.21831/pep.v18i1.2120

Nadapdap, A. T. Y., & Lede, Y. (2016). Authentic assessment of problem solving and critical thinking skill for improvement in learning physics. In *Proceeding of International Seminar on Science Education (ISSE), 29 October 2016* (pp. 37–42). Yogyakarta: Universitas Negeri Yogyakarta.

Oriondo, L. L., & Dallo-Antonio, E. M. (1998). *Evaluation educational outcomes*. Manila: Rex Printing Compagny.

Piaget, J. (2005). *The psychology of intelligence* (Electronic version). Taylor & Francis.

Pólya, G. (1957). *How to solve it: A new aspect of mathematical method*. Doubleday: Garden City.

Regulation of Minister of Education and Culture No. 59 of 2014 on the curriculum 2013 of senior high school/Madrasah Aliyah (2014). Republic of Indonesia.

Shin, S.-H. (2009). How to treat omitted responses in Rasch model-based equating. *Practical Assessment, Research & Evaluation*, *14*(1), 1–8. Retrieved from http://pareonline.net/getvn.asp?v=14&n=1

Slameto. (2010). *Belajar dan faktor-faktor yang mempengaruhi*. Jakarta: Rineka Cipta.

Suryabrata, S. (2002). *Pengembangan alat ukur psikologis*. Yogyakarta: Andi Offset.

Thorpe, G. L., McMillan, E., Sigmon, S. T., Owings, L. R., Dawson, R., & Bouman, P. (2007). Latent trait modeling with the Common Beliefs Survey III: Using item response theory to evaluate an irrational beliefs inventory. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *25*(3), 175–189. https://doi.org/10.1007/s10942-006-0039-9

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, *87*(2), 320–32.