USEJ 8 (3) (2019)

Unnes Science Education Journal

http://journal.unnes.ac.id/sju/index.php/usej

DEVELOPMENT OF INTEGRATED SCIENCE-BASED SCIENCE LITERACY SKILLS INSTRUMENTS USING THE RASCH MODEL

P. Susongko [1,✉], H. Widiatmo [2], M. Kusuma [3], Y.Afiani [4]

[1] Science Education Department University of Pancasakti Tegal, Indonesia
[2] American College Testing, United States of America
[3] Science Education Department, University of Pancasakti Tegal, Indonesia
[4] Science Education Department, University of Pancasakti Tegal, Indonesia

Article Info

Article History:
Received April 2019
Accepted July 2019
Published December 2019

Keywords:
Instruments, scientific literacy, integrated IPA, Rasch model

Abstract

The objectives of this study are: (1) To compile a test construction to measure scientific literacy skills of integrated science-based high school MIPA program students by referring to the achievements of scientific literacy according to the 2015 PISA standard, (2) Conduct validation of test items using Rasch modeling. The research design uses ADDIE procedural models (Analysis, Design, Development, Implementation, Evaluation). There are three types of validation in the development of instruments namely content validation, psychometric aspects validation and extract validation with Rasch modeling. The instruments were tested on class XII students of the MIPA Program from Tegal City 2 High School and Tegal City 3 High School involving 112 students. The construction of the test consisted of 17 integrated science cases presented in the form of tests each consisting of three questions referring to scientific literacy competencies according to PISA standard 2015. Test items have fulfilled the validity of the content aspects and psychometric aspects. Extract validation using Rasch modeling gives the following results: (1) The level of item difficulty is in the range of -3 to 3, (2) There are 16 items that are compatible with modeling, (3) There are 97.32% of student responses suitable with modeling, (4) There are 2 items containing DIF. Based on consideration of all aspects of validity, there are 14 items worthy of being used as items for scientific literacy tests.

p-ISSN 2252-6617
e-ISSN 2252-6232

✉Correspondence author:
P. Susongko
Science Education Department University of Pancasakti Tegal, Indonesia
Email: purwosusongko@upstegal.ac.id

INTRODUCTION

High scientific literacy of society has a very significant effect on the progress of a Nation. This is because public science literacy has a positive effect on the quality of economic development, democracy, culture and the quality of one's personality (Hanushek, & Woessmann, 2016; Rudolph, & Horibe, 2016; Bereiter, 2002). Therefore in many developed countries, achieving student scientific literacy is the main goal in science education (Hanson, 2016). The main objectives of science education in high school (SMA) mathematics and natural sciences (MIPA) programs include: (1) building and applying knowledge and technology information and demonstrating the ability to think logically, critically, creatively and innovatively, (2) demonstrate ability think logically, critically, creatively and innovatively independently, (3) demonstrate the ability to analyze and solve complex problems, (4) demonstrate the ability to analyze natural phenomena, utilize the environment productively and responsibly and master the knowledge needed for higher levels of education (Republic of Indonesia Ministry of National Education, 2006). This is in line with the achievements of scientific literacy developed by PISA (Program for International Science Student Assessment) which includes, (1) Explaining phenomena scientifically, (2) Evaluating and designing scientific investigations, (3) Interpreting data and evidence scientifically - analyzing and evaluate data, claims and arguments in various representations and draw appropriate scientific conclusions (OECD, 2016).

The competency standards that have been made by the Government have been measured through the National Examination (UN). However, there are some weaknesses in the implementation of the National Examination. First, not using the results of the National Examination (UN) as a determinant of graduation so that there is no guarantee of compliance with competency standards for high school students who graduate. Second, not all subjects that build science competency are tested, students may choose one subject only. This causes the ability of students who graduate not to be comprehensive in accordance with competency standards that should be mastered by students. In this regard, there needs to be a comprehensive exam that ensures that the competencies of high school students are in accordance with the specified competency standards. This exam is expected to take the form of a standard test from three aspects which include the contents, achievements of scientific literacy and measurement models.

Several studies show that science learning presented in an integrated manner has a stronger influence on improving students' scientific literacy (Tamassia, & Frans, 2014; Maria, 2008). This has the consequence of the need to make a comprehensive final examination covering integrated Mathematics, Physics, Chemistry and Biology competencies through integrated science cases. The achievements of the scientific literacy aspects of high school students also need to be considered by looking at the comparison of standards in several developed countries and by looking at the studies carried out by PISA and TIMSS.

Educational measurement model with classical test theory which has been used so far, based on the number of correct numbers so that it only reaches the ordinal level. Ordinal scores cannot be applied to basic arithmetic operations such as added, less, times and therefore need improvement with Rasch modeling which results in scores at the interval level (Mari, et al, 2012).

Classical measurement theory has limitations, namely: (1) test item statistics are very dependent on the characteristics of the subject being tested; (2) the estimation of the competency of the examinee is very dependent on the test items being tested; (3) the standard error of assessing the score applies to all examinees, so there is no standard error in measuring each participant and item; (4) information presented is limited to the number of correct answers; and (5) the parallel test assumption is difficult to fulfill. The weakness that is quite serious and has an impact according to Steven (Mari, et al, 2012) is that the type of data generated from the learning achievement test as well as from the attitude

scale is ordinal rather than interval so that analytical tools that can be used are limited. Even basic arithmetic operations such as added, less, times and divisions cannot be done because the numbers obtained are not integers but the scores are in the form of ordinal data.

The concept of objective measurement in the social sciences and the assessment of education according to Mok and Wright must have five criteria, namely: (1) Producing linear measurements with equal intervals, (2) exact estimation process, (3) identifying items that are not right (misfits) or not general (outliers), (4) Able to cope with lost data, (5) Produce measurements that are independent of the parameters under study (Mok, and Wright, 2004). Of the five conditions, so far only the Rasch model can fulfill the five conditions. The quality of measurements in the assessment of education conducted with the Rasch model will have the same quality as the measurements made in the physical dimension in the field of physics (Sumintono, & Widhiarso, 2014). In measuring modern test theory, the Rasch model is seen as the most objective measurement model. The use of the Rasch model in measuring education has advantages in specific objectivity and stability in the estimation of high grain parameters (Wu, & Adams, R, 2007).

The Rasch model connects the opportunity to correctly answer each item (P (θ)) as a function of ability (θ) with the constant level of difficulty of item (b) through a relationship as in equation 1.

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

The Rasch model is used for dichotomous responses or two categories such as multiple choice forms. Whereas for polytomous responses or more than two categories, the Rasch Model is developed more broadly as a Partial credit model (PCM) or a partial credit model. Opportunities generally in PCM are expressed by equation 2.

$$P(X_{ni} = x) = \frac{exp \sum_{k=0}^{x} (\emptyset_n - \delta_n)}{\sum_{h=0}^{nj} exp \sum_{k=0}^{h} (\emptyset_n - \delta_n)} \quad (2)$$

The Rasch model has been further developed separately from the IRT even the Rasch model has also been developed more widely in the Polytomous scoring. The application of the Rasch Model in academic achievement since its introduction by inventor Georg Rasch in 1960, now extends not only in the world of education even in the world of medicine and public health (Lu, et al, 2013; Smith, et al, 2010; Ayele, et al, 2014).

The Rasch model has long been used in the assessment of science education both in learning achievement tests and psychological test tests as well as interest and motivation to learn science. The basic and practical concept of using the Rasch model in science education assessment is explained quite comprehensively by several experts (Liu, 2010; Sjaastad, 2014). Likewise, the Rasch model is widely applied to surveys of psychological aspects relating to learning Science (Lamb et al., 2012) and some aspects of scientific literacy as well as the nature of science (Neumann et al, 2011).

Science literacy was first used by Hurd in 1958 and by James Bryant Conant in 1952 (Hanson, 2016). This term has become popular and the achievement of scientific literacy is one of the main goals of science education (Hanson, 2016; Holbrook & Rannikmae, 2009: NSTA, 2014: UNESCO, 2010). According to NSES (National Science Education Standard), students' scientific literacy abilities are the result of participating in inquiry-oriented activities thereby developing a fundamental understanding of the basic concepts of science and technology as a provision for them to relate to individuals and society (NRC, 1996).

Bybee (Bybee, 2012) defines scientific literacy as an understanding of science and its application into a social experience and proposes four levels of scientific literacy namely: (1) nominal science literacy, (2) functional science literacy, (3) conceptual and procedural scientific literacy, (4) Multidimensional scientific literacy. Wenning and Vierya (2015) compiled a fairly comprehensive framework relating the relationship of intellectual process skills and scientific practices categorized into increasing

levels of intellectual level achievement and related to the level of inquiry (Wenning, 2007).

PISA defines scientific literacy as the ability to engage with issues related to science, and with scientific ideas, as reflective manifestations. Educated people are scientifically willing to engage in reasoned discourse about science and technology, which requires competence to:

1. Explain phenomena scientifically: recognize, offer and evaluate explanations for various natural phenomena and technology.
2. Evaluate and design scientific investigations: describe and assess scientific investigations and propose ways to answer questions scientifically.
3. Interpret data and evidence scientifically - analyze and evaluate data, claims, and arguments in various representations and draw appropriate scientific conclusions (OECD, 2016). From some definitions of scientific literacy, the definition used by PISA is more operational and easy to apply to the science learning achievement test.

The study conducted by Maria Astrom (Maria, 2008) on the PISA results in 2006 showed that there were differences in scientific literacy skills in students who studied science in an integrated manner and who learned science separately even for female students, this difference was very significant. There is a tendency for countries that carry out integrated science learning to have higher scientific literacy than countries that present science learning separately (Physics, Chemistry, and Biology). In Belgium, countries that provide integrated science learning have higher scientific literacy than other countries (Tamassia, & Frans, 2014). In Indonesia, presenting science in an integrated manner also provides higher scientific literacy capability than presented separately (Yenni, et al, 2017). Some studies also prove that the effect of integrated science presentation provides an increase in science literacy that is better than presenting science with a separate conceptual approach. (Cervetti et al, 2012; Greenleaf et al, 2011)

From several studies, it shows that integrated science competencies support more towards improving student scientific literacy. This will improve the analytical skills of high school students in the MIPA program in reviewing real case cases seen from a science perspective holistically. To achieve this, in the assessment of scientific literacy competence students are made in an integrated science approach. In connection with the above matters, it is necessary to further study how to develop a test to assess science literacy competencies of high school students of the MIPA program based on integrated science.

To develop the test, several problems that must be answered in this study are as follows: (1) How is the construction of integrated science-based science literacy tests for high school students in the MIPA program ?, (2) How is the validity of the content aspects of integrated science-based science literacy tests for MIPA program high school students ?, (3) How is the quality of the psychometric aspects of the integrated science-based science literacy tests for high school students in the MIPA program ?, (4) What is the validity of the construct of integrated science-based science literacy tests for high school students in the MIPA program?

METHODS

This research was conducted at the Laboratory of Science Education Study Program at FKIP Pancasakti University and in the state high school in Tegal City. The form of this research is Research and Development (Research and Development) (Gall, et al, 1999; Haryati, 2012; Richey, & Klein, 2014). The object of this research is the science literacy assessment instrument of high school students of integrated science-based MIPA program which was compiled, revised, and validated by Rasch modeling. In the research design, instrument development uses ADDIE procedural models (Analysis, Design, Development, Implementation, Evaluation) (Molenda, 2003; Wahyuni, 2015).

In the design stage, researchers begin to collect, compile and design products to be

developed. There are three things that are considered in compiling the grid and test items, namely the thematic case of science, the achievement of scientific literacy and the model validation of the test items. The form of the test is given in the test (collection of items), each one thematic case of IPA is presented in one testlet consisting of 3 test items. The test points pay attention to the achievements of scientific literacy developed by PISA 2015. Grain validation using PCM modeling with four categories (0,1,2, and 3). In addition to the aspects of achievement of scientific literacy that is considered, in this test also pay attention to aspects of the content consisting of Physics, Chemistry, Biology, and Mathematics. The casting of each item in one testlet is dichotomous (1 or 0), while the scoring of each testlet is polytomous with four categories of 0.1.2 and 3. For the subject matter obtained from scientific news as well as www. sciencenews.org, www. sciencenewsforstudents.org,www.readworks.org, a collection of integrated science questions about college entrance exams.

In the development stage, the researcher began to validate the instruments he developed. There are three types of validation, namely content aspect validation, psychometric aspects validation and extract validation with Rasch modeling. Content validation is carried out with the consideration of 2 experts related to the test material and the achievements of scientific literacy to be measured. Psychometric aspects of validation involving 2 psychometric experts related to testing construction. For the sake of construct validity, the instrument was tested in the XI class of SMA MIPA program in Tegal 3 and SMA 2 high schools involving 112 students so that the grain parameter estimation became stable.

The validity of the construct used in this study refers to the concept of Messick Extract validity (Messick, 1996; Baghaei, & Amrahi, 2011;), where construct validity is divided into six aspects, namely content, substantive, structural, external, consequential and generalization. Susongko (2016) provides quantitative criteria relating to the indicators of the validity of the constructs according to the Rasch model as described in Table 1

Table 1. Valid test criteria seen from various aspects of validity and the criteria for applying the Rasch Model

The aspect of construct validity	Indicator	Criteria
Content	the item compatibility test (itemfit)	P > 0.05 0,5 <MNSQ<1,5 -2,0 < ZSTD<2,0
	Person-item Map	All item difficulty levels are in the testee ability domain
	Person/Item Map	Testee ability equals or approaches the difficulty level of an item
	Test Information Function	Test information function has a maximum value on the testee ability domain
Substantif	Person fit statistic	P > 0.05 0,5 <MNSQ<1,5 -2,0 < ZSTD<2,0
	Collapsed Deviance / Casewise Deviance /Hosmer-Lemeshow	P<0,05
	accuracy, sensitivity, dan specificity	close to 1.0
Structural	Unidimensional Test	there is one main factor illustrated through Scree Plot the result of factor analysis
	Invariance Test (LRtest)	P< 0,05
Eksternal	The external Person strata separation value	Close to 1,0
Consequential	DIF	The DIF does

	not have a significant

In this study, the software used in analyzing Rasch modeling uses Program R version 3.5.0 with the eRm package version 0.16-2. This software is used because it is open source so that it is easy to access and develop for observers of educational assessment research.

RESULTS AND DISCUSSION

A measuring instrument is considered to have content validity if the measuring instrument contains it can measure the overall content of what will be measured. Validation of the content aspect tests the quality of the test items qualitatively in terms of the validity of the data presented and the achievements of the level of scientific literacy and the involvement of integrated science principles. From the results of the two experts, it can be stated that the instrument of Science Literacy Measurement for MIPA Program High School Students has been made feasible from the aspect of the content or in accordance with the measurement objectives.

The psychometric aspect validation aims to ensure that the test items meet psychometric rules in the preparation of items. Psychometric aspects that need to be considered are material aspects, construction, language, and scientific news narratives. From the results of the assessment of two experts in the psychometric field, it can be concluded that the instrument of Science Literacy Measurement for MIPA Program High School Students that has been made feasible from psychometric aspects and can be followed up with empirical trials.

Construct Validity of Content Aspects

As explained in Table 1 about the criteria for construct validity in the Content aspect, the following will explain some of the results of the analysis data using Rasch modeling for polytomous data (PCM). Table 2 contains the results of the analysis of item compatibility with the model (Item Fit). The item fit basically explains whether an item functions to measure normally or not. Quantitatively the test items that are declared fit or able to function properly are if the MSQ Outfit value is between 0.5 to 1.5 while the outfit value of t is between -2 to 2.0 and the probability of acceptance of Ho (model match) is greater than 0.05 (p> 0.05). The outfit is an outlier-sensitive fit, which is a measure of the sensitivity of the response pattern to an item with a certain level of difficulty from the respondents (students) or vice versa. Outfit t is a t-test for the data compatibility hypothesis with the model.

The value of the MSQ Outfit is calculated from the chi-square value divided by the degree of freedom (Df). From Table 16, it appears that all items, in general, can be accepted as good items except point 16. Point number 16 has MSQ outfit of 1,286, t outfit is 2.12 and p-value is <0.05. This means that item number 16 is seen from out fit-t more than 2.0, which means that the data appears unpredictable while the probability of a model match is also less than 0.05. All criteria reject item number 16 so it can be concluded that at the level of significance 0.05 item number 16 cannot be accepted by the model. The magnitude of the level of difficulty in each category (threshold) can be seen in Table 3.

Table 2. Results of Item Fit Analysis Instruments for Literacy Science Measurement for MIPA Program High School Students

Item Number	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t
1	101.885	111	0.720	0.910	0.904	-0.63	-0.67
2	106.407	111	0.606	0.950	0.946	-0.39	-0.43
3	114.512	111	0.391	1.022	1.011	0.23	0.13
4	100.688	111	0.748	0.899	0.915	-0.79	-0.66
5	103.550	111	0.680	0.925	0.922	-0.65	-0.67
6	93.025	111	0.891	0.831	0.851	-1.53	-1.39
7	111.176	111	0.477	0.993	0.985	-0.03	-0.09
8	100.618	111	0.750	0.898	0.883	-0.78	-0.92
9	103.471	111	0.682	0.924	0.910	-0.66	-0.79
10	100.752	111	0.747	0.900	0.897	-0.79	-0.82

11	102.762	111	0.699	0.918	0.930	-0.72	-0.62
12	117.843	111	0.310	1.052	1.043	0.40	0.34
13	109.316	111	0.527	0.976	0.966	-0.17	-0.27
14	116.516	111	0.341	1.040	1.024	0.38	0.25
15	100.021	111	0.763	0.893	0.908	-0.90	-0.79
16	144.040	111	0.019	1.286	1.192	2.12	1.54
17	100.269	111	0.758	0.895	0.916	-0.61	-0.48

The value of this outfit describes the deviation of the test participant's response from the ideal model. With the outfit value exceeding the fairness limit, it can be stated that the item has a significant deviation from the Rasch model. Deviations, in this case, are some test takers who have the ability lower than the level of difficulty of the item successfully answer the item correctly or some test participants who have the ability above the level of difficulty but did not succeed in correctly answering the item. The incompatibility of responses with the model can be caused by many factors such as the existence of carelessness, misconception or the success of guessing (Sumintono & Widhiarso, 2015). Thus the Rasch model can be used to identify misconceptions.

Many studies show that the Rasch Model can be used to identify the occurrence of misconceptions on large scale tests. This is especially true of mastery tests in physics, chemistry, and science (Herrmann-Abell, & DeBoer, 2011; Wind, & Gale, 2015; Romine et al, 2015; Morris et al, 2012; Edwards, & Alcock, 2010; Sheu et al, 2013; Planinic et al, 2010). Testlet number 16 contains scientific news about solar activity accompanied by three questions that refer to the scientific news. In the first point only measures students' knowledge of electromagnetic waves, but in the second item measures students' ability to interpret readings while the third item measures students' ability to connect physics and mathematical concepts in wave equations. This second and third item is very vulnerable to student misconception.

Table 3. Value of the Difficulty Level of the Items of Science Literacy Measurement Instruments for MIPA Program High School Students

Item	*Threshold*	value	Item	*Threshold*	value	Item	*Threshold*	*Value*
1	C1	-1.810	7	C1	-0.752	13	C1	-0.431
	C2	-3.169		C2	-0.338		C2	0.131
	C3	-2.415		C3	1.322		C3	2.382
2	C1	-1.787	8	C1	-0.947	14	C1	0.127
	C2	-1.459		C2	-0.039		C2	0.598
	C3	-0.350		C3	1.841		C3	2.758
3	C1	-1.071	9	C1	-0.579	15	C1	-0.272
	C2	-0.547		C2	-0.231		C2	0.329
	C3	0.635		C3	1.753		C3	1.913
4	C1	-0.288	10	C1	-1.660	16	C1	-0.223
	C2	0.783		C2	-1.753		C2	0.681
	C3	2.938		C3	0.281		C3	1.886
5	C1	-0.745	11	C1	-0.521	17	C1	-1.563
	C2	-0.786		C2	-0.099		C2	0.032
	C3	1.276		C3	0.868		C3	1.718
6	C1	-1.146	12	C1	-0.606			
	C2	-1.445		C2	1.191			
	C3	-1.258		C3	2.843			

PCM does not require steps to complete the test items in sequence and does not have to have the same difficulties. PCM developed in this instrument has four categories, so PCM analysis produces three thresholds (difficulty level) for each item. From Table 3 it can be seen that the lowest level of difficulty in item number one for threshold 2 is -3,169 while the difficulty level is highest in item number four for Threshold 3 of 2. 938. The level of difficulty of 2,938 means that participants are expected to work on items correctly if they have a minimum capability of 2,938. The level of difficulty of the item is a location parameter that shows the position of the grain characteristic curve in relation to the scale of ability. The parameter level of difficulty of the item is described by a point on the scale of ability where the opportunity to answer correctly is 0.5. The greater the parameter level of difficulty, the greater the ability needed by respondents to get the opportunity to answer the questions correctly as much as 0.5. For more details, Figure 1 and Figure 2 describe the characteristic curves of item number 1 and number 4.

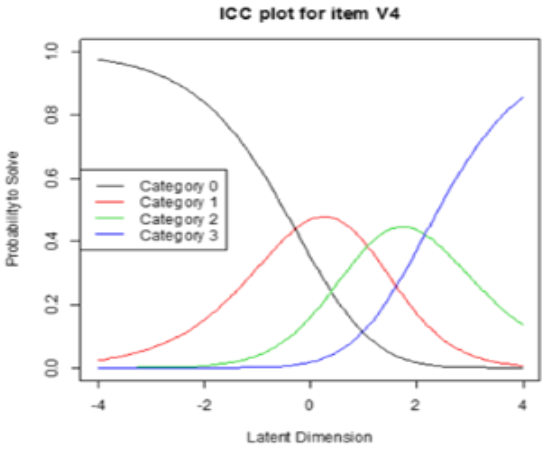


Figure 1. Item Characteristic Curve 1

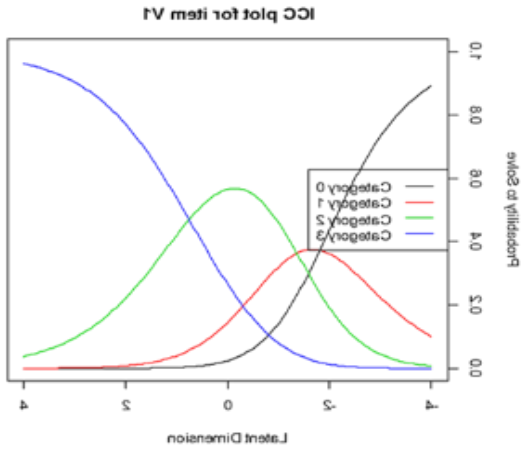


Figure 2. Item Characteristic Curve 2

From Figure 1 and Figure 2, it can be seen that for category 0, the higher the respondent's ability, the lower the chance, on the contrary for category 3 the higher the respondent's ability, the higher the chance to answer the truth. Whereas for categories 1 and 2, this is not the case but the opportunity to answer correctly increases with the increase in ability and will reach a peak in certain abilities then the opportunity will decline again in line with the increase in the ability of respondents.

From Table 3, it can be seen that the difficulty level of the item moves from -3,169 to 2,938. Effective tests have a degree of difficulty between grains of -2.00 to 2.00 (Wright, & Stone, 1979; Hambleton, et al, 1991; Wu & Adam, 2007). However, tests built to measure competencies as well as scientific literacy measurement instruments for MIPA Program High School Students should be able to measure the abilities of all test participants so that the distribution of the level of difficulty is broader than the tests built in the selection test paradigm or tests that use the norm reference. If it is assumed that as developed by response theory / normal distribution items, then the level of difficulty of items for competent measurement can be started from -3.00 to 3.00, because at that interval it can measure around 99.98% of test participants. Thus from the results of the analysis of all the items in the test of scientific literacy measurement instruments for students who have been compiled, are in the interval -3.00 to 3.00 so that it is effective as a competency test. This is made clear by Figure 3 which describes the item map and Figure 4 which describe the person map system where all grain difficulty levels are at predetermined intervals. Figure 5 connects the ability of the test taker and the level of difficulty of the item

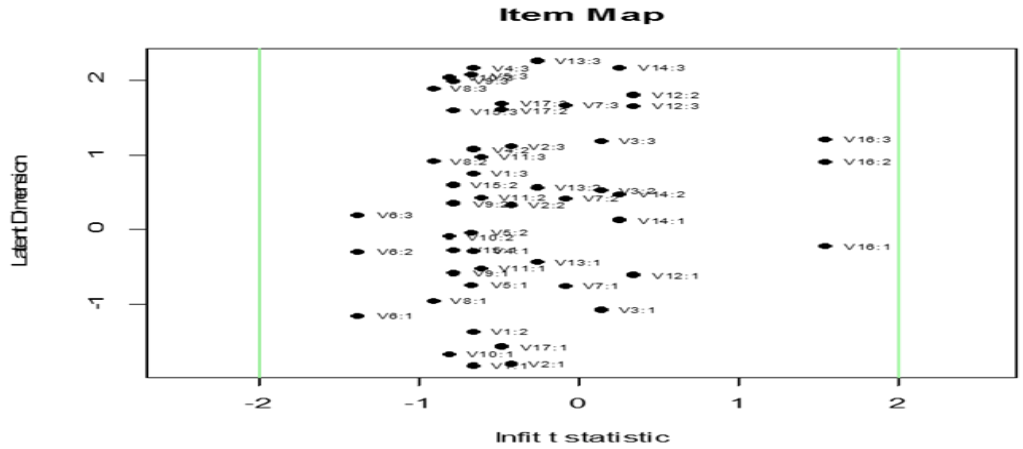


Figure 3. Item Map Items of Science Literacy Measurement Instruments for MIPA Program High School Students

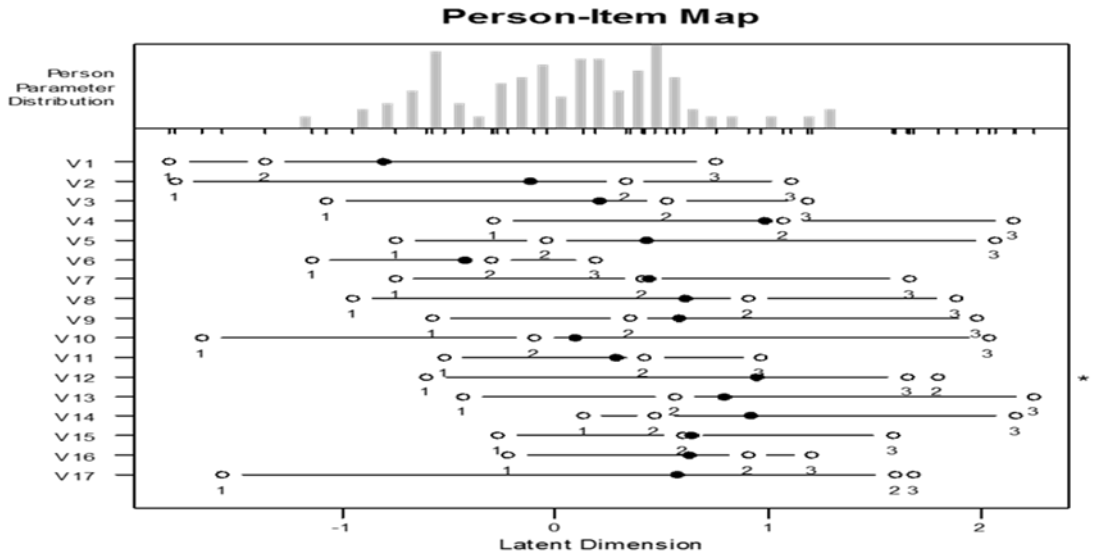


Figure 4. Person-Map It ems of Instruments for Science Literacy Measurement for MIPA Program High School Students

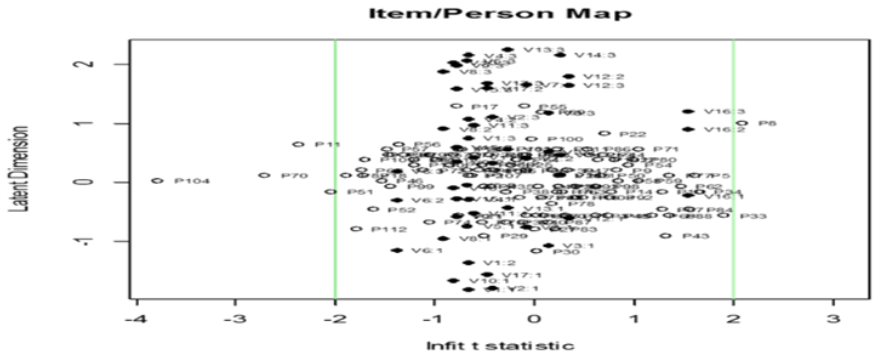


Figure 5. Item / Person Map Items of Science Literacy Measurement Instruments for MIPA Program High School Students

Evidence that the items of scientific literacy measurement instruments for high school students of the MIPA program are effectively used for the ability of test participants between -3.00 to 3.00 is explained by the item information and test functions (Figure 6). The picture explains the information function will be maximal at the interval of students' abilities between 0 to 1.0 and effective between -3.0 to 3.00.

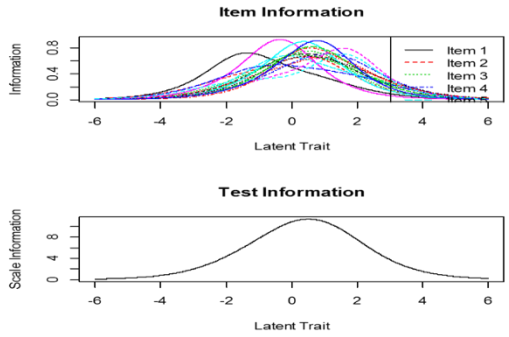


Figure 6. Information Function of Items of Science Literacy Measurement Instruments for MIPA Program High School Students

Construct Validity of Substantive Aspects

To see the quality of construct validity from substantive aspects, a match test of the ability of the test participants to the model was used. This test basically is to test the consistency of the response or the different response patterns of the participants towards the test items based on their level of difficulty. A different response pattern is the incompatibility of the response given based on its ability compared to the ideal model. A test participant who has the ability (Ø) of 1.5 should be able to answer all items that have difficulty levels below 1.5, but in the field, there are certainly some students who are inconsistent or give rise to an aberrant response. How many students experience the aberrant response is a measure of the substantive type of construct validity.

This deviant response can be caused by inaccuracies, cheating or even misconceptions. A person's response test experiences irregularities or is not called a person fit. Criteria for receiving test taker's response are deemed to have deviations or are not the same as the item fit criteria. Quantitatively the response of test participants who were declared fit or not experiencing deviation is if the MSQ Outfit value is between 0.5 and 1.5 while the outfit value of t is between -2 to 2.0 and the chance of acceptance of Ho (model match) is greater than 0.05 (p> 0.05). Table 11-14 contains the results of the person fittest from 112 responses to the science literacy test for high school students in the MIPA program. Of the 112 test participants, there were five test participants who experienced a defiant response from the model. It is seen that the five test participants did not fulfill as many as two p-values and MSQ outfit) from three criteria of person fit. Even one participant (P33) did not meet all the criteria of person fit. The list of test takers is described in Table 4.

Table 4. Test participants who have an aberrant response

Peserta	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t
P8	28.740	16	0.026	1.691	1.731	1.96	2.09
P33	*30.388*	*16*	*0.016*	*1.788*	*1.708*	*2.03*	*1.89*
P34	*29.441*	*16*	*0.021*	*1.732*	*1.563*	*1.98*	*1.65*
P43	*28.521*	*16*	*0.027*	*1.678*	*1.481*	*1.72*	*1.32*
P84	*27.584*	*16*	*0.035*	*1.623*	*1.545*	*1.70*	*1.56*

From this explanation (Table 4) it can be concluded that there are 95.5% responses of reasonable test participants according to the model or not experiencing deviations while there are 4.5% responses experiencing irregularities. The percentage of test takers who have a reasonable response according to this model can be the basis that the test adequately meets substantive validity. Even if you use a 0.01 level of confidence, then all test takers' responses are in accordance with the model.

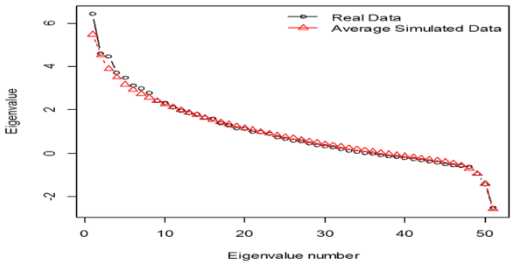
Student responses that deviate from the Rasch model indicate an indication of students doing careless or lucky guesses or even cheating (Sumintono & Widhiarso, 2015). Several studies have shown that person fit can be used as preliminary data for cheating, careless or lucky guesses in conducting tests by students (Shu et al, 2013; Wagner-Menghin et al, 2103; Meyer, & Zhu, 2013; Hohensinn, & Kubinger, 2011 ; Magis et al, 2012; Elhan et al, 2010; Lamprianou, 2010; Liu, & Yu, 2011)

Construct Validity of Structural Aspects

There are two test indicators that have structural aspects of construct validity, that is, the tests are unidimensional and have stability in estimating the parameters of the items and test takers. Tests built on a one-dimensional paradigm must have one dimension so that the measurement results obtained can have meaning. The principle of unidimensional testing is stated by the null hypothesis which states that the second eigenvalue value is not greater than the first eigenvalue value with the alternative hypothesis that the second eigenvalue value is greater than the first eigenvalue value. The results of the unidimensional test analysis with the R program using the ltm package can be seen in Table 5 while the results of the curvature analysis can be seen in Figure 7.

Table 5. Unidimensional Test Results Instruments Items of Science Literacy Measurement for MIPA

Alternative hypothesis: the second eigenvalue of the observed data is substantially larger than the second eigenvalue of data under the assumed IRT model
Second eigenvalue in the observed data: 4.599 Average of second eigenvalues in Monte Carlo samples: 4.5197 Monte Carlo samples: 100 p-value: 0.396



Program High School Students

Figure 7. Graph of the Dimensionality Test Analysis of the Instruments for Science Literacy Measurement Instruments for MIPA Program High School Students

From Table 5, it can be seen that the probability of the resulting test is equal to 0.396, a value greater than 0.05 so that it can be stated that Ho is accepted. If Ho is accepted it means the second eigenvalue and so on is smaller than the first eigenvalue. Such conditions can be stated that the test contains only one dimension. Thus it can be concluded that the scientific literacy test for high school students in the MIPA program can be declared to be unidimensional.

Next to do the measurement invariance test using the LR Anderson test. This test is used to determine the consistency of Rasch modeling parameter estimates. The ideal condition for Rasch modeling occurs when the item difficulty level estimation is consistent (invariant) even though it is obtained from a sample consisting of any subgroup of the population while applying Rasch modeling, in this case using PCM. The results of the Anderson LR test can be seen in Table 6. From the results of the analysis, the p-value of 0.188 means that it accepts Ho so that it can be concluded that the parameter estimation is invariant.

Table 6. Measurement Invariance Test Using the LR Anderson Test

Andersen LR-test: LR-value: 45.489 Chi-square df: 38 p-value: 0.188

Construct Validity of External Aspects

The validity of the external aspect construct is used to determine the extent to which the test results are supported by other measurements (which measure the same or similar domain) so that it can be seen whether it has a strong relationship or not. Ideally, researchers have other, more accurate test data such as standardized scientific literacy tests, general intelligence tests or special talents that support scientific literacy, or could be standardized science learning achievement tests. It can be interpreted that the test of external construct validity is basically an evaluation of an instrument that has been developed. In this regard, researchers will do this in the second year.

One approach to determine the construct validity of external aspects in this first-year study is to use Person Separation reliability or information separation. Separation of Persons is used to classify people based on information

obtained from tests. Low person separation (less than 2) with a relevant sample of people implies that the instrument may not be sensitive enough to distinguish between high and low performance. This means that more items are needed to measure it. The results of Person separation analysis using eRm packages can be seen in Table 7.

Table 7. Person Separation Reliability Tests on the Instruments of Science Literacy Measurement

Separation Reliability: 0.6016 Observed Variance: 0.2396 (Squared Standard Deviation) Mean Square Measurement Error: 0.0955 (Model Error Variance)

Instruments for MIPA Program High School Students

From Table 7 it can be seen that the value of Person Separation reliability is 0.6016. Thus the person separation value for the test is 1,133. From the separation value of the person, it can be seen that the classification of test participants obtained more than one or close to 2. This means that the instruments that have been made can distinguish test participants in two categories namely literate and non-literate. Consequently, the results of this test only distinguish test participants into two groups, namely test takers who have had a minimum of scientific literacy and who do not yet have a minimum of scientific literacy. This information can be followed up in determining the graduation limit for science literacy tests for MIPA Program High School students

Construct Validity Aspects of Consequences

Consequential aspects in the validity of the construct on the implications of the value of the score interpretation as a source of action. Evidence regarding a spects of consequential validity also addresses the actual and potential consequences of testing and using scores, especially in terms of sources of invalidity such as bias, justice, and distributive justice. In this regard, scientific literacy measurements for MIPA Program High School students need to detect test bias.

In Rasch modeling with the eRm package, the detection of grain bias can be approached by determining the items that have a differential item functioning (DIF) using the Waldt Test. DIF is related to the estimation of different grain parameters in different subpopulations, in this case, the test participant is differentiated based on the type of darkness. If an item is considered more difficult or easier by male test takers than women or vice versa, then the item contains DIF. DIF or also called external item bias is not the justification for the occurrence of item bias because to know whether there is a bias or not, a more in-depth qualitative study must be carried out regarding the cause of the emergence of DIF. However, the emergence of DIF can be a clue to the possibility of bias. The list of test items detected by DIF can be seen in Table 8 while the description of DIF can be seen in Figure 8. Statistical criteria with Wald test, items that have DIF are those who have a p-value of less than 0.05 (if using a significance level of 0.05). From Table 8 it is known that there are 4 items indicated to have DIF, namely points 6,7,13 and 17.

Table 8. List of DIF Indicated Test Items by Gender at a Significance Level of 0.05

Butir	Threshold	Z statistic	p-value	Butir	Threshold	Z statistic	p-value
6	C1	2.267	0.023	13	C1	-2.756	0.006
	C2	2.490	0.013		C2	-1.704	0.088
	C3	1.690	0.091		C3	-1.690	0.091
7	C1	0.627	0.531	17	C1	2.022	0.043
	C2	1.937	0.053		C2	1.472	0.141
	C3	2.768	0.006		C3	-0.786	0.432

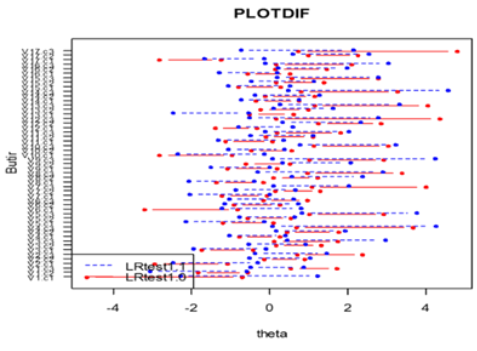


Figure 8. Description of DIF in Science Literacy Test Items for MIPA High School Programs

From Table 8, there are 4 points where the opportunity to correctly answer each item in one testlet is DIF. When using a significance level of 0.05, item number 6 has two thresholds that have DIF, while the other items each have only one threshold. When using a significance level of 0.01, only 7 and 13 only experience DIF. In accordance with the data of test takers, where the proportion of men is only 34.8%, far from the ideal proportion of course researchers must be more careful in determining the level of significance when testing the presence of DIF on items caused by sex. If at the 0.05 significance level it means that the probability of rejecting Ho is correct as much as 0.05, then at the significance level of 0.01 it means the opportunity to reject Ho is correct as much as 0.01. Ho here states that students' responses to tests do not experience DIF. Regarding this in determining the DIF, the researcher chose a significance level of 0.01 so that there were two items that were considered detected by DIF.

Point number 7 contains material about quarks that form exotic particles. Point number 13 contains material about the situation on the Moon. Both of these materials discuss many abstract things. For item number 7, the proportion of students with male sex who answered correctly was 0.521 while for women there were 0.356. In item number 13, the proportion of students with male sex who answered correctly was 0.410 while for women as many as 0.324. Both of these grains benefit men and significantly contain DIF that benefits men. This phenomenon supports several previous studies where it was found that men are easier to think abstractly while women have superiority in concrete thinking (Wilson et al, 2016; Madsen et al, 2013; Dietz et al, 2012; Bates et al, 2013)

From the results of the study, it was found that there were three items that were not suitable to be used as scientific literacy measurement instruments, namely items that did not match the model (item number 16) and items detected by DIF at the significance level of 0.01, namely number 7 and item number 13. While Item others by analyzing validity which includes content, psychometrics, and extracts (content, substantive, structural, external, consequences) fulfill the requirements as a good item. The weakness of this study is that the validity of the criteria for the test instrument has not been carried out. Criteria validity test is needed in order to ensure that the test results are in line with other standard tests that have similar constructs. The validity of this criterion can be tested by comparing the results of this student's literacy test with the results of other tests such as intelligence tests, aptitude tests or national examination results.

CONCLUSION

Integrated science-based science literacy tests for high school students in the MIPA program consist of 17 testlets containing scientific news where each testlet consists of 3 items that refer to the level of achievement of scientific literacy according to the PISA 2015 standard. All items of integrated science-based science literacy tests for high school students the MIPA program has fulfilled the content aspect validity. All items of integrated science-based science literacy tests for high school students of the MIPA program have fulfilled the validity of psychometric aspects. Extract validation using Rasch modeling gives the following results: (1) The level of item difficulty is in the range of -3 to 3, (2) There are 16 items that are compatible with modeling, (3) There are 97.32% of student responses suitable with modeling, (4) There are 2 items containing DIF. Based on consideration of all aspects of validity, there are 14 items from 17 items that are worthy of being used as items for scientific literacy tests.

ACKNOWLEDGMENT

The author thanks the Republic of Indonesia Ministry of Research, Technology and Higher Education for providing funding for this research. Likewise, the author would like to thank all parties involved, especially the principal of SMA 2 and SMA 3 of Tegal City who has supported and given permission for research

REFERENCES

Ayele, Dawit G; Zewotir, Temesgen; Mwambi, Henry (2014). Using Rasch Modeling to Re-Evaluate Rapid Malaria Diagnosis Test Analyses. *International Journal of Environmental Research and Public Health*). *11*(7) 6681-91.

Baghaei, P., & Amrahi, N. (2011). Validation of a Multiple Choice English Vocabulary Test with the Rasch Model. *Journal of Language Teaching & Research*, *2*(5).

Bates, S., Donnelly, R., MacPhee, C., Sands, D., Birch, M., & Walet, N. R. (2013). Gender differences in conceptual understanding of Newtonian mechanics: a UK cross-institution comparison. *European Journal of Physics*, *34*(2), 421.

Bereiter, C. (2002). Design research for sustained innovation. *Cognitive Studies*, *9*(3), 321-327.

Bybee RW. Scientific Literacy in Environmental and Health Education. In Zeyer & Kyburz-Graber, R. (Eds.) Science/ Environment/Health. Towards a Renewed Pedagogy for Science Education. New York: Springer, 2012; p.49-68.

Cervetti, G. N., Barber, J., Dorph, R., Pearson, P. D., & Goldschmidt, P. G. (2012). The impact of an integrated approach to science and literacy in elementary school classrooms. Journal of research in science teaching, *49*(5), 631-658.

Dietz, R. D., Pearson, R. H., Semak, M. R., & Willis, C. W. (2012, February). Gender bias in the force concept inventory?. In *AIP Conference Proceedings* (Vol. 1413, No. 1, pp. 171-174). American Institute of Physics.

Edwards, A., & Alcock, L. (2010). Using Rasch analysis to identify uncharacteristic responses to undergraduate assessments. *Teaching Mathematics and its Applications: An International Journal of the IMA*, *29*(4), 165-175.

Elhan, A. H., Küçükdeveci, A. A., & Tennant, A. L. A. N. (2010). The Rasch measurement model. *Research Issues in physical and rehabilitation Medicine*, 89-102.

Gall, J. P., Gall, M. D., & Borg, W. R. (1999). *Applying educational research: A practical guide*. Longman Publishing Group.

Greenleaf, C. L., Litman, C., Hanson, T. L., Rosen, R., Boscardin, C. K., Herman, J., & Jones, B. (2011). Integrating literacy and science in biology: Teaching and learning impacts of reading apprenticeship professional development. American Educational Research Journal, *48*(3), 647-717.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hanson, S. (2016). The assessment of scientific reasoning skills of high school science students: A standardized assessment instrument.

Hanushek, E. A., & Woessmann, L. (2016). Knowledge capital, growth, and the East Asian miracle. *Science*, *351*(6271), 344-345.

Haryati, S. (2012). Research and Development (R&D) sebagai salah satu model penelitian dalam bidang pendidikan. *Majalah Ilmiah Dinamika*, *37*(1), 15.

Herrmann-Abell, C. F., & DeBoer, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate the hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice*, *12*(2), 184-192.

Hohensinn, C., & Kubinger, K. D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling*, *53*(3), 380.

Holbrook, J., & Rannikmae, M. (2009). The meaning of scientific literacy. *International Journal of Environmental and Science Education*, *4*(3), 275-288.

Kementerian Pendidikan Nasional Republik Indonesia (2006). Peraturan Menteri Pendidikan Nasional Republik Indonesia Nomor 23 Tahun 2006 tentang Standar Kompetensi Lulusan Untuk Satuan Pendidikan Dasar dan Menengah

Lamb, R. L., Annetta, L., Meldrum, J., & Vallett, D. (2012). Measuring science interest: Rasch validation of the science interest survey. *International Journal of Science and Mathematics Education*, *10*(3), 643-668.

Lamprianou, I. (2010). The practical application of Optimal Appropriateness Measurement on empirical data using Rasch Models. *Journal of applied measurement*, *11*(4), 409.

Liu, M. T., & Yu, P. T. (2011). Aberrant learning achievement detection based on person-fit statistics in personalized e-learning systems. *Journal of Educational Technology & Society*, *14*(1), 107-120.

Liu, X. (2010). Using and developing measurement instruments in science education: A Rasch modeling approach. Iap.

Lu, Y. M., Wu, Y. Y., Hsieh, C. L., Lin, C. L., Hwang, S. L., Cheng, K. I., & Lue, Y. J. (2013). Measurement precision of the disability for back pain scale-by applying Rasch analysis. *Health and quality of life outcomes*, *11*(1), 119.

Madsen, A., McKagan, S. B., & Sayre, E. C. (2013). The gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?. *Physical Review Special Topics-Physics Education Research*, *9*(2), 020121.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's LZ* index of person fit with an emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral statistics*, *37*(1), 57-81.

Mari, L., Carbone, P., and Petri, D. (2012). Measurement fundamentals: A pragmatic view. IEEE *Transactions on Instrumentation and Measurement, 61*(8), 2107-2114

Maria, A (2008). Defining Integrated Science Education and Putting It toTest.The Swedish National Graduate. Dissertation. Unpublished

Messick, S. (1996). Validity and washback in language testing. *Language testing*, *13*(3), 241-256.

Meyer, J. P., & Zhu, S. (2013). Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating. *Research & Practice in Assessment*, *8*, 26-39.

Mok, M. and Wright, B. (2004). Overview of Rasch Model Families. In *Introduction to Rasch Measurement: Theory, Models and Applications* (has 1-24). Minnesota: Jam Press.

Molenda, M. (2003). In search of the elusive ADDIE model. *Performance improvement*, *42*(5), 34-37.

Morris, G. A., Harshman, N., Branum-Martin, L., Mazur, E., Mzoughi, T., & Baker, S. D. (2012). An item response curves analysis of the Force Concept Inventory. *American Journal of Physics*, *80*(9), 825-831.

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch- based analyses of a nature of science test. *International Journal of Science Education*, *33*(10), 1373-1405.

NRC. (1996). *National Science Education Standards.* Washington, DC: National Research Council

NSTA. Teaching Science and Technology in the Context of Societal and Personal Issues. Retrieved,2010-2014.

OECD, P. (2015). Assessment and Analytical Framework: Science. *Reading, Mathematic and Financial Literacy,(Interscience: Paris, 2016)*, 24-25.

Planinic, M., Ivanjek, L., & Susac, A. (2010). Rasch model-based analysis of the Force Concept Inventory. *Physical Review Special*

Topics-Physics Education Research, *6*(1), 010103.

Richey, R. C., & Klein, J. D. (2014). *Design and development research: Methods, strategies, and issues*. Routledge.

Romine, W. L., Schaffer, D. L., & Barrow, L. (2015). Development and application of a novel Rasch-based methodology for evaluating multi-tiered assessment instruments: Validation and utilization of an undergraduate diagnostic test of the water cycle. *International Journal of Science Education*, *37*(16), 2740-2768.

Rudolph, J. L., & Horibe, S. (2016). What do we mean by science education for civic engagement?. *Journal of Research in Science Teaching*, *53*(6), 805-820.

Sheu, T. W., Tsai, C. P., Tzeng, J. W., Chen, T. L., & Nagai, M. (2013). An Algorithm of the Misconception Order. In *Applied Mechanics and Materials* (Vol. 284, pp. 3010-3014). Trans Tech Publications.

Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika*, *78*(3), 481-497.

Sjaastad, J. (2014). Enhancing measurement in science education research through Rasch analysis: Rationale and properties. Nordic Studies in Science Education, 10(2), 212-230.

Smith, Adam B, et al (2010). A Rasch and confirmatory factor analysis of the General Health Questionnaire (GHQ) – 12. Journal Health and Quality of Life Outcomes, 8, pages: 45. http://search.proquest.com/docview/902252382?accountid=62691

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model Rasch Untuk penelitian ilmu-ilmu social (edisi revisi)*. Trim Komunikata Publishing House

Susongko, P. (2016). Validation of science achievement test with the Rasch model. *Journal Pendidikan IPA Indonesia*, *5*(2), 268-277.

Tamassia, L., & Frans, R. (2014). Does integrated science education improve scientific literacy?. *Journal of the European Teacher Education Network*, *9*, 131-141

UNESCO. Current Challenges in Basic Science Education. Paris, UNESCO, 2010

Wagner-Menghin, M., Preusche, I., & Schmidts, M. (2013). The effects of reusing written test items: A study using the Rasch model. *ISRN Education*, *2013*.

Wahyuni, S. (2015). Developing science learning instruments based on local wisdom to improve students critical thinking skills.

Wenning, C. J. (2007). Assessing inquiry skills as a component of scientific literacy. J*ournal of Physics Teacher Education Online, 4*(2), 21-24.

Wenning, C. J., & Vieyra, R. E. (2015). *Teaching High School Physics Volume I*. Rebecca Vieyra.

Wilson, K., Low, D., Verdon, M., & Verdon, A. (2016). Differences in gender performance on competitive physics selection tests. *Physical Review Physics Education Research*, *12*(2), 020111.

Wind, S. A., & Gale, J. D. (2015). Diagnostic Opportunities Using Rasch Measurement in the Context of a Misconceptions-Based Physical Science Assessment. *Science Education*, *99*(4), 721-741.

Wright, B. D., & Stone, M. H. (1979). Best test design.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.

Yenni, R., Hernani, & Widodo, A. (2017, May). The implementation of integrated science teaching materials based socio-scientific issues to improve students scientific literacy for environmental pollution theme. In *AIP Conference Proceedings* (Vol. 1848, No. 1, p. 060002). AIP Publishing