Journal for the Education of Gifted Young Scientists, 8(4), 1583-1602, Dec 2020
e-ISSN: 2149- 360X
jegys.org

youngwisepub.com
© 2020

Research Article

Revising of the Integrating Scientific Literacy Skills Scale (ISLS) with Rasch Model Analysis

Purwo Susongko[1]*, Mobinta Kusuma[2], Yuni Arfiani[3], Achmad Samsudin[4], Adam Aminudin[5]

Department of Science Education, University of Pancasakti Tegal, Indonesia,

Article Info

Received: 22 August 2020
Revised: 11 November 2020
Accepted: 01 December 2020
Available online: 15 Dec 2020

Keywords:
Integrating Scientific Literacy Skills Scale (ISLS)
Rasch Model
Scale development
Science program students

Abstract

In this study it was aimed to develop and analyze instruments of integrating scientific literacy skills scale (ISLS) for science program students of senior high school with a Rasch Model Analysis. In developing and analyzing instruments we use the Messick's validity (1996) approach which consists of five aspects including content, substantive, structural, external, and consequential. ISLS consisted of 14 cases of integrated science presented in the form of a testlet. Each case consists of three questions given to scientific literacy competencies according to PISA 2015 standards. The research design uses the ADDIE procedural model (Analysis, Design, Development, Implementation, Evaluation). Participants consisted of 310 grade XII students of the science program from two senior high schools in Tegal City, Indonesia. Constructive validation with Rasch modelling gives the following results. The level of conformity of the items is in the range of -3 to 4. All the items that are suitable for modelling. As many as 95.16 % of student responses match modelling. Has no items containing DIF. It can be said that ISLS, which consists of 14 items, is suitable for measuring Integrating Scientific Literacy Skills for science program students of senior high school.

To cite this article:

Susongko, P, Kusuma, M, Arfiani, Y., Samsudin, A., & Aminudin, A. (2020). Revising of the Integrating Scientific Literacy Skills Scale (ISLS) with Rasch Model Analysis. *Journal for the Education of Gifted Young Scientists*, *8*(4), 1583-1602. DOI: http://dx.doi.org/10.17478/jegys.781583

Introduction

Several studies have been carried out in developing instruments to identify scientific literacy skills. Noted less than the last 10 years, there are various studies such as: “Instrument Development in Measuring the Scientific Literacy Integrated Character Levels of Junior High School Students” (Jufri et al. 2019); “The development of scientific literacy assessments to measure students' scientific literacy skills in energy themes” (Rusilowati et al. 2018); “Development and validation of scientific literacy scale for colleges preparedness in STEM with freshmen from diverse institutions” (Benjamin et al. 2017); “Developing an Instrument of Scientific Literacy Assessment on the Cycle Theme” (Rusilowati et al. 2016); “Development and Validation of Scientific Literacy Achievement Tests to Assess Senior Secondary School Students' Literacy Acquisition in Physics” (Adeleke & Joshua, 2015); “Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments” (Gormally et al. 2012), and; “Assessing Student's Level of Scientific Literacy Using Interdisciplinary Scenarios” (Soobard & Rannikmäe, 2011). The development of these instruments shows that the measurement of scientific literacy skills is very important, both in the world of education and in life in society.

[1] Corresponding Author: Associate Professor of Science Education Assessment, University of Pancasakti Tegal, Indonesia (purwosusongko@upstegal.ac.id), Orcid no: 0000-0001-9126-1027
[2] Department of Science Education, University of Pancasakti Tegal, Indonesia (mobintakusuma@upstegal.ac.i), Orcid no: 0000-0002-5924-5075
[3] Department of Science Education, University of Pancasakti Tegal, Indonesia (yuniarfiani@upstegal.ac.id), Orcid no: 0000-0002-9557-2102
[4] Assist. Prof. of Physics Education, Indonesia University of Education, Indonesia (achmadsamsudin@upi.edu), Orcid no: 0000-0003-3564-6031
[5] Department of Physics Education, Indonesia University of Education, Indonesia (adamhadiana@upi.edu), Orcid no: 0000-0001-7409-9195

In the world of education, the term scientific literacy was first used by Hurd in 1958 and James Bryant Conant in 1952 (Hanson, 2016). This term has become popular because the achievement of scientific literacy is one of the main goals of science education (Hanson, 2016; Holbrook & Rannikmae, 2009). Bybee, (2012) defines scientific literacy as an understanding of science and its application in social experience and proposes four levels of scientific literacy, including:

- Nominal scientific literacy
- Functional science literacy
- Conceptual and procedural scientific literacy
- Multidimensional scientific literacy

In addition, the National Science Education Standard (NSES) revealed that students' scientific literacy skills are the result of developing a fundamental understanding of the basic concepts of science and technology as their provisions relating to individuals and society.

In the community, high scientific literacy skills will significantly influence the progress of a Nation. This is due to the scientific literacy of the community has a positive effect on the quality of economic development, democracy, culture, and the quality of one's personality (Abdul Rahim & Chun, 2017; Hanushek & Woessmann, 2016; Md-Ali et al. 2016; Rudolph & Horibe, 2016). Therefore, in many developed countries, the achievement of student scientific literacy is the aim of science education (Hanson, 2016). This is in line with the achievements of scientific literacy developed by Programme for International Student Assessment (PISA) which includes;

- Explaining phenomena scientifically
- Evaluating and designing scientific investigations
- Interpret scientific data and evidence-analyse and evaluate data, claims and arguments in various representations and draw appropriate scientific conclusions (OECD, 2015). However, some of the instruments that have been developed by researchers in the previous discussion have not shown integration with science learning.

Several studies show that science learning presented in an integrated manner has a stronger influence on increasing student scientific literacy (Heng et al. 2015; Suhandi & Samsudin, 2019; Suryana et al. 2020; Tamassia & Frans, 2014; Yenni et al. 2017). This gives the consequence the need for a comprehensive final examination covering Mathematics, Physics, Chemistry, and Biology competencies in an integrated manner through integrated science cases. This is one of the challenges for teachers in the 21st century (Nordin & Ariffin, 2016). The achievement of aspects of scientific literacy of students also needs to be considered by looking at the standard comparison in several developed countries and by looking at studies that have been carried out by the Program for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS).

- Scientific literacy is defined by PISA as a reflective form of the ability to be in problems related to science or with the scientific ideas. Competencies needed by people who are involved in science problems include competence for: Explain phenomena scientifically, such as recognizing, offering, and evaluating explanations for various natural and technological phenomena,
- Evaluating and designing scientific investigations, such as describing and evaluating scientific investigations and proposing ways to answer questions scientifically
- Interpret scientific data and evidence - analyses and evaluate data, claims and arguments in various representations and draw appropriate scientific conclusions (OECD, 2015). Of the several definitions of scientific literacy, the definitions used by PISA are more operational and easier to apply to the science achievement test. Measurements made on these tests must also be carried out objectively.

Objective measurements in social sciences and educational assessment according to (Wright & Mok, 2004) must have five criteria, namely:

- Producing linear measures with the same interval,
- Appropriate estimation process,
- Identifying misfits or outliers,
- Able to overcome lost data and produce measurements that are sovereign of the limitations considered. 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 the specific objectivity and stability of the estimation of high grain parameters (Wu & Adams, 2007).

Whereas in the measurement of classical tests there are some shortcomings, consists of; statistics of test items highly depend on the characteristics of the subject being tested, the estimated the examinees ability is very reliant on the items being tested, standard errors in estimating scores apply to all examinees, so there is no standard measurement error for each participant and the items are absent, the information presented is limited to the number of correct answers; and assumptions of parallel tests are difficult to fulfill. Even the types of data generated from achievement tests and attitude scales are ordinal rather than intervals so the analytical tools that can be used are limited (Mari et al. 2012). Thus, that the Rasch model is used that the measurements made are objective.

The Rasch model also has other advantages such as linking the chance of correctly answering for item (P (*θ*)) as the ability function (*θ*) with the item difficulty near constant (*b*). This connection can be shown in Equation 1.

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

In addition, the Rasch model can be used for dichotomous responses or two categories as well as multiple choice questions. Whereas for responses that are political in nature or more than two categories, the Rasch Model is developed more broadly as a Partial Credit Model (PCM) or partial credit model. General opportunities in PCM are expressed by Equation 2.

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

The Rasch model has been further developed separately from IRT (Pratiwi et al. 2020; Samsudin, 2020; Sumintono, 2018; Susongko, 2016). Even the Rasch model has also been developed more broadly in scoring polytomous. The implementation of the Rasch modelling in learning since it was presented by its discoverer Georg Rasch, is now widespread not only in the education world but also in the medicine world and the health (Smith et al. 2010). Likewise, for surveys on psychological aspects, related to science learning (Lamb et al. 2012) and some aspects of scientific literacy as well as the nature of science (Neumann et al. 2010). But instruments of science-based integrated science literacy capabilities are still rarely found. To develop these instruments, the following problems must be answered: These problems are, for example, how the test is constructed, the validity of the content and psychometric aspects and the validity of the constructs. Interestingly, this research was conducted in Tegal City, even though Indonesia is a country with diverse cultures and ethnicities. Likewise, in Tegal, which is dominated by Javanese tribes. This research can be developed if applied in other cities with different ethnicities. This is because each tribe has different characteristics and culture.

Problem of Study

Susongko et al. (2019) have developed an integrated science-based literacy scale consisting of 17 items and were tested on 112 senior high school students in the city of Tegal. Validation with the Messick model (1996) with the Rasch approach to the scale shows that 14 items meet all aspects of the validity of Messick (1996). The weakness of this study is that it does not involve many respondents, so it is possible that the parameter estimates of the resulting items are not stable. In Rasch modeling, the use of large sample size will increase the stability of the item parameter estimation (O'Neill et al. 2020). Another weakness in this study is that it uses student respondents at grade XI, even though it is by the purpose of the assessment for students at grade XII. Therefore it is necessary to revise the ISLS scale by re-validating the 14 items that were valid in the preliminary research and involving target respondents in a larger number.

This study, it was aimed to revise ISLS that scale developed by Susongko et al. (2019) for science program students at senior high school with a Rasch Model Analysis. In developing and analyzing instruments used the Messick's validity (1996) approach which consists of five aspects including content, substantive, structural, external, and consequential. Research problem is that,

- Is the ISLS scale revised/developed according to the Rasch Model Analysis suitable for senior high school students?

Method

Research Design

This research includes Research and Development with ADDIE (Analysis, Design, Development, Implementation, Evaluation) design (Branch, 2009). In the analysis phase, researchers determine the needs and objectives of the product be developed. The product of this study is an instrument that measures the science literacy competency of high school students of science programs through thematic study of science problems by involving students' abilities in mathematics, physics, chemistry, and biology. To build objectivity this instrument is validated by Rasch modeling (Bond et al. 2020).

In the design phase, researchers begin to collect, arrange and design products to be developed. At the development phase, researchers begin to validate the instruments they are developing. At the implementation phase, the researcher makes observations by providing integrated science-based science literacy capabilities. At the evaluation phase, an external validity test is carried out using external criteria such as the intelligence test or the National Examination test results. For this research, it is limited to the analysis, design and devolution stages (Haladyna, & Rodriguez, 2013).

Participants

Participants in this research were 310 grade XII students of the Sciences Program from two senior high schools in Tegal, each with 102 male and 208 female students. Their age range is 16-18 years. All students come from the city of Tegal and surrounding areas. The initial abilities and family background of students are very diverse as a result of the application of zoning policies in the acceptance of high school students. The city of Tegal is located in the province of Central Java, with a distance of 165 km from Semarang City (the capital of Central Java) as shown in Figure 1. People in Central Java are dominated by Javanese, in contrast to West Java which is dominated by Sundanese (Aminudin et al. 2019; Saddhono & Rohmadi, 2014).

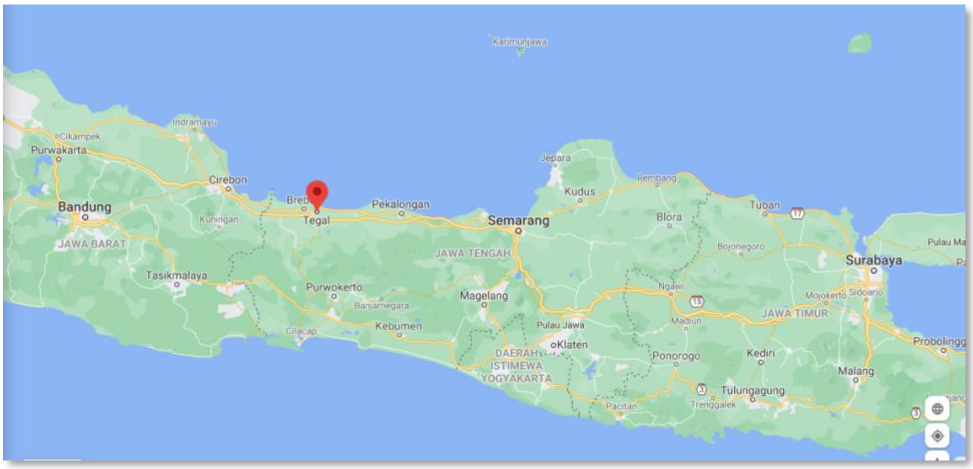


Figure 1.
The Map of Tegal Seen from Semarang (Source by Google Maps)

Data Collection Tools

Integrating Scientific Literacy Skills Scale (ISLS)

This scale was firstly developed by Susongko (2019) using 17 testlets with validation using Messick (1996) based on the Rasch model resulting in 14 valid items. These items need to be re-validated by involving respondents according to the target assessment objectives using a larger sample size.

Whereas the integrated science-based science literacy skills instrument is presented in the form of a testlet, which consists of 14 integrated science cases. Each case consists of three questions that refer to scientific literacy competencies according to the 2015 PISA standard. Questions were prepared by researchers and six science teachers from SMAN 2 Tegal and SMAN 3 Tegal. ISLS item was reviewed by two science education experts from Sebelas Maret University and one psychometric professor from Universitas Negeri Yogyakarta. Examples of ISLS items can be seen Figure 2.

Data Analysis

Data analysis was performed by dividing validity into three types, namely content validation, psychometric aspect validation, and construct validation with Rasch modelling. Content validation is carried out by an assessment involving two experts relating to the test material and the achievement of scientific literacy to be measured. Assessors are asked to answer whether the test items have met several criteria such as: the truth of the scientific news presented, the validity of the data presented, the suitability of the questions with the indicators of scientific literacy, the correctness of the answer keys and the involvement of integrated science capabilities.

Validation of psychometric aspects involves two psychometrics experts related to testing construction. Aspects of the test construction assessed include aspects of the material, construction, language and narrative of the testlet.

Meanwhile, for construct validity, which refers to the concept of the validity of the construct of (Messick, 1996), where the construct validity is divided into six aspects namely content, substantive, structural, external, consequential and generalization (Baghaei & Amrahi, 2011). Susongko (2016) provides quantitative criteria relating to indicators of construct validity according to Rasch modelling as described in Table 1.

Table 1.
Valid Test Criteria are Seen from Various Aspects of Validity and Criteria by Applying the Rasch Model

The construct validity aspect	Indicator	Criteria
Content	Test item compatibility (itemfit)	P > 0.05 0,5 <MNSQ<1,5 -2,0 < ZSTD<2,0
	Person-item Map	All item difficulty levels are in the ability tester domain
	Person/Item Map	The ability of the tester is equal to or near the level of difficulty of the item
	Test Info Meaning	Function Test information has a maximum value in the domain of the ability of the tester
Substantive	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 that is described through the Scree Plot results of factor analysis
	Invariance Test (LRtest)	P< 0,05
External	Strata Person separation value	close to 1,0
Consequential	DIF	There is no significant DIF

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

Results and Discussion

Development of ISSL scale consists of 6 stages: analysis, design, development, implementation, evaluation stages.

Analysis Phase

In the analysis phase, researchers determine the needs and objectives of the product be developed. The development of an integrated science-based science literacy assessment instrument for high school students in the science program aims to meet the need for a comprehensive exam that can ensure that the competencies of high school students in the Mathematics and Natural Sciences program are in accordance with predetermined competency standards. The National Examination that has been conducted is not enough to measure the competency standards that have been determined due to several things including: The National Examination is not a determinant of graduation so there are no guarantee of compliance with competency standards for high school students who graduate, students only choose one of the points lessons from three science subjects, there is no measurement of the ability of integrated science as a basis for competence in scientific literacy. This instrument is expected to take the form of a standardized test with regard to three aspects which include content, scientific literacy achievements and measurement models.

Design Phase

In the design phase, researchers begin to collect, arrange and design products to be developed. There are three things that must be considered in compiling the lattice and test items, namely the thematic cases of natural sciences, scientific literacy achievements and the validation models of the test items. The form of the test is given in the *testlet* (collection of items) for each one thematic case of Natural Sciences. One *testlet* consisting of 3 test items. The test items pay attention to the achievements of scientific literacy developed by PISA 2015 which consists of: explaining phenomena scientifically, interpreting data and evidence scientifically, evaluating and designing scientific investigations. Indicators of each achievement of scientific literacy according to PISA 2015 standards are presented in Table 2.

Table 2.
Level Indicators of Achievement in Science Literacy According to PISA 2015 Standards

Achievements in Science Literacy	Indicator used
Explain the phenomenon scientifically	Remember and apply appropriate scientific knowledge Identify, use and be able to produce explanatory models
Interpret scientific data and evidence	Change data from one representation to another Analyze and interpret data and draw correct conclusions Identify assumptions, evidence, and reasons in texts related to science
Evaluate and design scientific investigations	Make generalizations from explanations Identifying the questions explored in the given scientific study

Item validation uses the Partial Credit Model (Rasch for polytomous) modelling with four categories (0,1,2 and 3). In addition to the aspect of scientific literacy achievement that is considered, in this test also pay attention to aspects of the content consisting of Physics, Chemistry, Biology, and Mathematics. The four fields form an integrated knowledge that explains the thematic phenomena of science. Scoring 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 respectively as in Table 3 below.

Table 3.
Testlet Scoring Model

Score	Criteria
0	Unable to answer all items
1	Successfully answered one item in
2	Successfully answered two items in
3	Successfully answered all items (3 items)

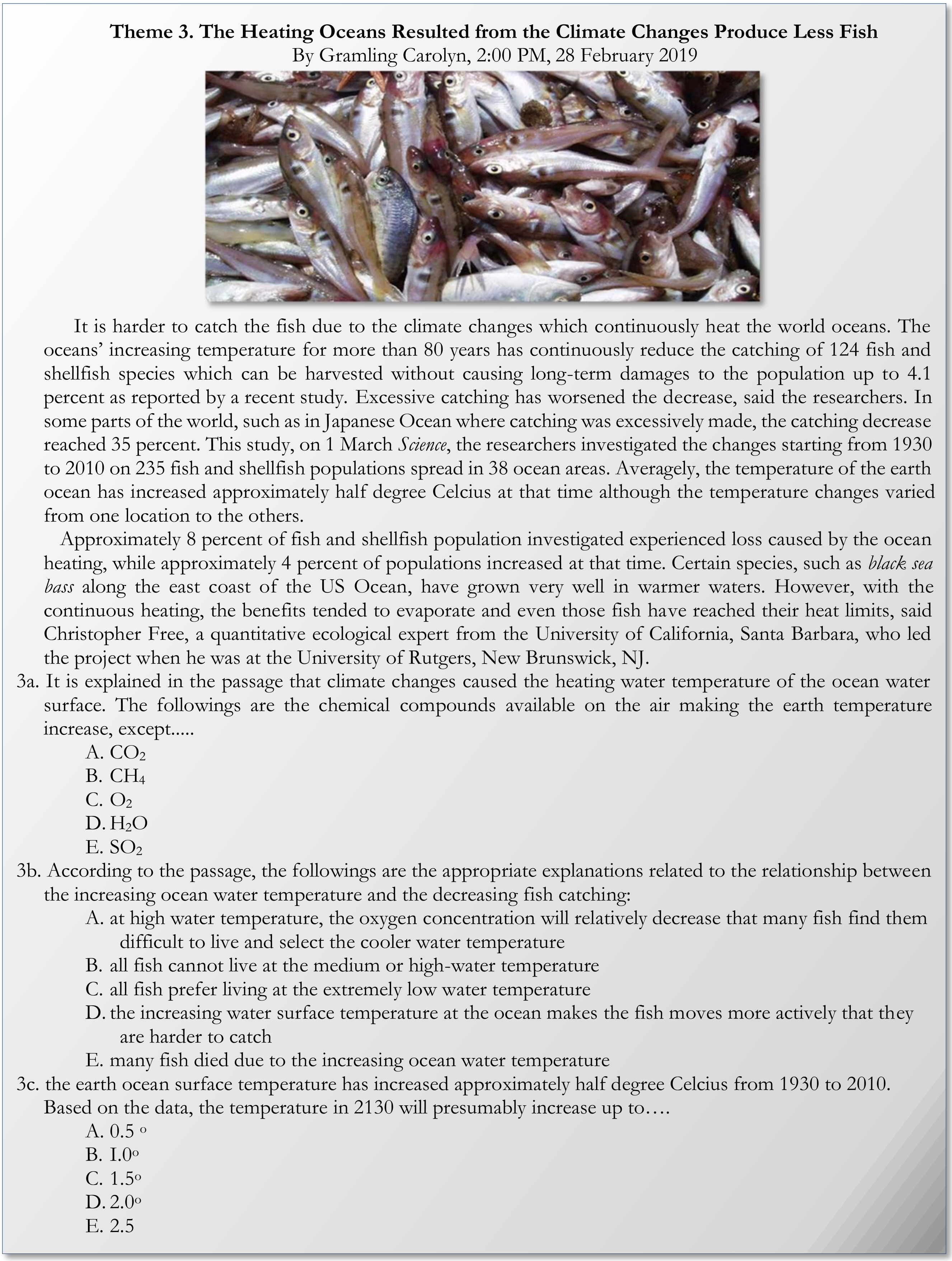
In the aspect of content, this scientific literacy instrument emphasizes the ability of students to answer problems after reading scientific news or cases of integrated science given (Runnels, 2012). For matter material obtained from scientific news such as www.sciencenews.org, www.sciencenewsforstudents.org, www.readworks.org, and a collection of integrated science questions on college entrance examinations in Indonesia. Table 4 below is a list of science news used as a matter of questions in the measurement of scientific literacy with integrated science.

Table 4.
List of Scientific News in the Measurement of ISLS

Item No	The Theme or Title of Scientific News
1	50 Years Ago, People Think MSG Causes “Chinese Restaurant Syndrome”
2	Eating Lots of Fiber Can Improve Some Cancer Treatments
3	Oceans that heat up due to climate change produce fewer fish
4	Watching TV is associated with a decrease in verbal memory in the elderly
5	Sleeping on weekends cannot make up for lost sleep
6	Understanding Tsunamis
7	How to turn a greenhouse into a powerhouse
8	Process
9	Researchers Start Understanding False Memory Formations Better
10	Carbon Dioxide in Mammals
11	Ammonia Synthesis
12	Aluminum metal
13	Fragrant Root Oil
14	Bioenergy

Development Phase

At the development stage, science-based scientific literacy skills instruments are made with reference to the 2015 PISA standards. This is because PISA defines scientific literacy as the skills to engage with issues related to science, and with scientific ideas, as a reflective form. Thus, the instrument developed is considered appropriate to be based on PISA standards. The examples of instruments developed can be seen in Figure 2.



Theme 3. The Heating Oceans Resulted from the Climate Changes Produce Less Fish

By Gramling Carolyn, 2:00 PM, 28 February 2019

It is harder to catch the fish due to the climate changes which continuously heat the world oceans. The oceans' increasing temperature for more than 80 years has continuously reduce the catching of 124 fish and shellfish species which can be harvested without causing long-term damages to the population up to 4.1 percent as reported by a recent study. Excessive catching has worsened the decrease, said the researchers. In some parts of the world, such as in Japanese Ocean where catching was excessively made, the catching decrease reached 35 percent. This study, on 1 March *Science*, the researchers investigated the changes starting from 1930 to 2010 on 235 fish and shellfish populations spread in 38 ocean areas. Averagely, the temperature of the earth ocean has increased approximately half degree Celcius at that time although the temperature changes varied from one location to the others.

Approximately 8 percent of fish and shellfish population investigated experienced loss caused by the ocean heating, while approximately 4 percent of populations increased at that time. Certain species, such as *black sea bass* along the east coast of the US Ocean, have grown very well in warmer waters. However, with the continuous heating, the benefits tended to evaporate and even those fish have reached their heat limits, said Christopher Free, a quantitative ecological expert from the University of California, Santa Barbara, who led the project when he was at the University of Rutgers, New Brunswick, NJ.

3a. It is explained in the passage that climate changes caused the heating water temperature of the ocean water surface. The followings are the chemical compounds available on the air making the earth temperature increase, except.....

A. CO_2
B. CH_4
C. O_2
D. H_2O
E. SO_2

3b. According to the passage, the followings are the appropriate explanations related to the relationship between the increasing ocean water temperature and the decreasing fish catching:

A. at high water temperature, the oxygen concentration will relatively decrease that many fish find them difficult to live and select the cooler water temperature
B. all fish cannot live at the medium or high-water temperature
C. all fish prefer living at the extremely low water temperature
D. the increasing water surface temperature at the ocean makes the fish moves more actively that they are harder to catch
E. many fish died due to the increasing ocean water temperature

3c. the earth ocean surface temperature has increased approximately half degree Celcius from 1930 to 2010. Based on the data, the temperature in 2130 will presumably increase up to….

A. $0.5\ ^{o}$
B. $I.0^{o}$
C. 1.5^{o}
D. 2.0^{o}
E. 2.5

Figure 2.
The Example of ISLS' item

Implementation Phase

The implementation stage is phase of the instrument is given to the participants. The instrument was given to 310 grade XII students of the sciences program from two senior high schools.

Evaluation Phase

Validity of Content Aspects

A measuring instrument is considered to have content validity if the measure has the contents already able to measure the overall contents of what is to be measured. Thus, decisions based on content validity determine whether students have mastered or are proficient, or fail to answer items that measure scientific literacy in agreement with the measurement objectives outlined in the grid and the instruments. The content validity of this scientific literacy instrument can be applied because the domain to be measured can be known clearly and comprehensively so that it can (Sireci & Faulkner, 2014).

There are two types of content validity, namely, face validity and logical validity. Face validity is achieved when an examination of the test items concludes that the test measures the relevant aspects. The basis for the conclusion is more based on common sense. Face validity is the lowest type of validity. It is easy to see from the grid and instruments that the instruments made have been compiled to meet 3 aspects, namely: thematic case-based tests, in this case scientific news, test items are arranged based on scientific literacy achievements according to PISA 2015 standards, the test requires integrated Physics, Chemistry, Biology and Mathematics competencies.

Logical validity is also called Sampling validity. This validity requires careful limitations of the area (domain) of behavior measured and a logical design that can include parts of the behavioral area. The extent to which this type of validity has been fulfilled can be seen from the scope of items contained in the test. Content validity can be done by: making test questions or test specifications, asking for expert opinions/experts.

Experts involved were the two people each of whom was experts in the field of Natural Sciences, seen from the Functional Position, the Structural Position and the quality of scientific publications. The results of the instrument review can be seen in Table 5.

Table 5.

Results of the Contents Review of the ISLS for Senior High School Students of the Sciences Program

No	Indicator	Assessor 1		Assessor 2	
		Yes	Not	Yes	Not
1	News / Narratives contain scientific truths	✓		✓	
2	News / Narration based on data	✓		✓	
3	Items in one *testlet* (theme) are sorted by:	✓		✓	
	a. The ability of students to explain phenomena scientifically (first item)				
	b. Interpret scientific data and evidence (second item)				
	c. Evaluate and design scientific investigations (third item)				
4	The correct answer key	✓		✓	
5	Involves the ability of integrated science to successfully answer test items	✓		✓	

From the results of the two assessors, it can be stated that the Science Literacy Measurement Instrument for Senior High School Students of Sciences Program which has been made feasible in terms of content or in accordance with the measurement objectives.

Validity of Psychometric Aspects

Validation of psychometric aspects aims to make sure the test items meet the psychometric rules in the preparation of items. Psychometric aspects that need attention are the material, construction, language and narrative aspects of the testlet. For the process of evaluating the validity of psychometric aspects, researchers used two speakers each from psychometrics experts and school teachers who were in charge of preparing test items. The complete psychometric validation results can be seen in Table 6.

From the results of the two assessors, it can be stated that the Science Literacy Measurement Instrument for Senior High School Students of Sciences Program which has been made feasible in terms of content or in accordance with the measurement objectives.

Table 6.

Results of Evaluation of the Validity of Psychometric Aspects of the ISLS for Senior High School Students of Sciences Program

Indicator	Assessor 1	Assessor 2
Material		
1. Questions must be in accordance with the indicators.	Meet	Meet
2. The choice of answers must be homogeneous and logical in terms of material	Meet	Meet
3. Each question must have one correct or most correct answer.	Meet	Meet
Construction		
4. The subject matter must be formulated clearly and firmly.	Meet	Meet
5. The formulation of the subject matter and choice of answers must be the statement that is needed only.	Meet	Meet
6. The point is do not give directions to the correct answer.	Meet	Meet
7. The subject matter should not contain double negative statements.	Meet	Meet
8. The length of the choice of answers must be relatively the same.	Meet	Meet
9. Answer choices do not contain the statement, “All of the above answer choices are wrong”, or “All of the above answer choices are correct”.	Meet	Meet
10. Answer choices in the form of numbers or times must be arranged in the order of the size of the number, or chronologically.	Meet	Meet
11. Pictures, graphs, tables, diagrams, and the like contained in the problem must be clear and functional.	Meet	Meet
12. Item do not depend on the answer to the previous question.	Meet	Meet
Language		
13. Each question must use language in accordance with Indonesian language rules.	Meet	Meet
14. Don’t use local language, if the question will be used for other regions or nationally.	Meet	Meet
15. Each question must use communicative language.	Meet	Meet
16. Answer choices do not repeat words or phrases that are not a unity of understanding.	Meet	Meet
Narration of Testlet		
17. In accordance with the field of science studies that are multidisciplinary	Meet	Meet
18. Easy to understand for senior high school students in the Sciences program (Class XI)	Meet	Meet
19. Clear description and can be concluded	Meet	Meet

Construct Validity

In accordance with the explanation of Table 1 about the construct validity criteria in the Content aspect, the following will explain some of the data from the analysis using Rasch modeling for polytomous data (PCM). Table 7 contains the results of the analysis of item compatibility with the model (Item Fit). Item fit basically explains whether an item functions to take measurements normally or not. Quantitatively the test items that are declared fit or can function well are if the MSQ Outfit value is between 0.5 and 1.5 while the outfit t value is between -2 to 2.0 and the probability of acceptance of Ho (model compatibility) is greater than 0.05 ($p > 0.05$). Outfit is outlier-sensitive fit, which is a measure of the sensitivity of response patterns to items with a certain level of difficulty from the respondents (students) or vice versa. Outfit t is the t-test for the data suitability hypothesis with the model.

The MSQ Outfit value is calculated from the chi-square value divided by the degree of freedom (df). From Table 7 it appears that all items are generally acceptable as good items. All item number have an p-value > 0.05. The magnitude of the level of difficulty in each category (threshold) can be seen in Table 8.

Table 7.

Results of Item Fit Analysis for ISLS for Science Program of Senior High School Students

No Item	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t
1.	276.622	309	0.907	0.892	0.907	-1.358	-1.187
2.	266.734	309	0.961	0.860	0.863	-2.006	-1.998
3.	295.586	309	0.699	0.954	0.945	-0.686	-0.835
4.	344.585	309	0.080	1.112	1.079	1.478	1.060
5.	273.097	309	0.930	0.881	0.895	-1.802	-1.641
6.	266.355	309	0.962	0.859	0.871	-1.764	-1.606
7.	273.356	309	0.929	0.882	0.876	-1.759	-1.852
8.	293.061	309	0.734	0.945	0.952	-0.791	-0.696
9.	279.890	309	0.882	0.903	0.904	-1.445	-1.439
10.	308.961	309	0.490	0.997	0.993	-0.024	-0.072
11.	325.239	309	0.252	1.049	1.022	0.549	0.261
12.	301.197	309	0.614	0.972	0.987	-0.370	-0.175
13.	349.672	309	0.055	1.128	1.116	1.850	1.717
14.	274.672	309	0.921	0.886	0.890	-1.759	-1.738

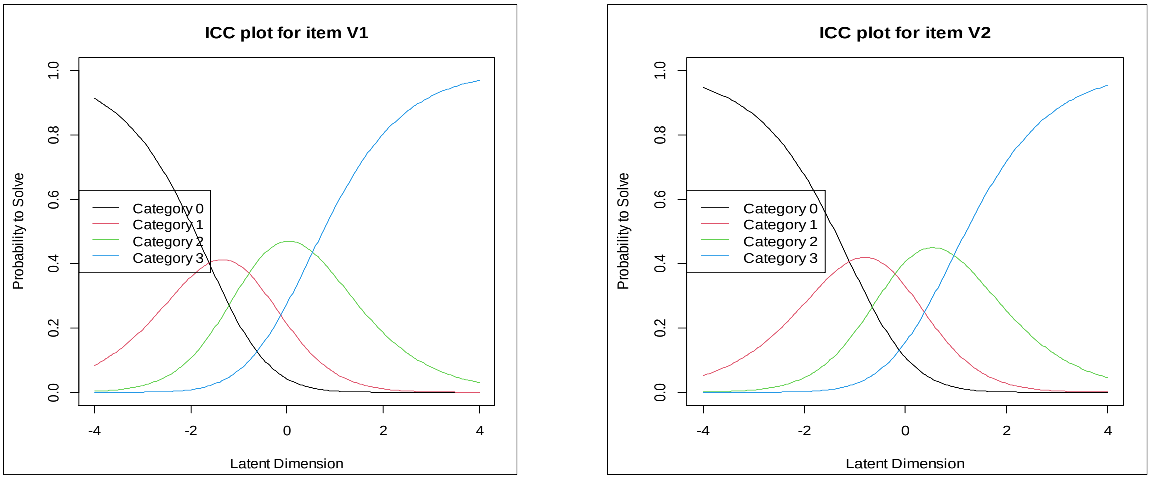
This outfit value illustrates the deviant response of test participants from the ideal model. With an outfit value that exceeds the fairness limit, it can be stated that the item has a significant deviation from the Rasch model. Deviations in this case, are some test takers who have abilities lower than the difficulty level of the item succeed in answering the item correctly or some test participants who have abilities above the difficulty level but fail to answer the item correctly. The mismatch of responses to the model can be caused by many factors such as carelessness, misconception or success in guessing (Sumintono & Widhiarso, 2015). Thus, the Rasch Model can be used to identify misconceptions.

Many studies show the Rasch Model can be used to identify the occurrence of misconceptions on tests that are large scale. This is especially true for tests of mastery in physics, chemistry, and science (e.g. Wind, & Gale, 2015; Romine et al. 2015; Sheu et al. 2013; Morris et al. 2012; Herrmann-Abell, & DeBoer, 2011; Edwards, & Alcock, 2010; Planinic et al. 2010).

PCM does not require steps to complete the test items sequentially and does not have to have the same difficulty. The PCM developed in this instrument has four categories so that the PCM analysis produces three thresholds for each item. From Table 8 it can be seen that the lowest difficulty level for item number twelve for threshold 3 is -3,104 while the highest difficulty level for item number four for Threshold 3 is 3.904. The difficulty level of 3.904 means that participants are expected to be able to work on the items correctly if they have a minimum ability of 3.904. The item difficulty level is a location parameter that shows the position of the grain characteristic curve in relation to the ability scale. The item difficulty level parameter is illustrated by a point on the capability scale where the opportunity to answer correctly is 0.5. The greater the value of the difficulty level parameter, the greater the ability needed by respondents to get the opportunity to answer the item correctly as much as 0.5. For more details, Figure 3 explain the characteristic curves of item number 1 and number 2.

Table 8.
Grain Difficulty Rating for ISLS for Science Program of Senior High School Students

Item	Threshold	Value	Item	Threshold	Value	Item	Threshold	Value
1	C1	-1.615	7	C1	0.877	13	C1	-0.016
	C2	-2.404		C2	0.278		C2	-0.652
	C3	-1.868		C3	-1.276		C3	-2.376
2	C1	-1.102	8	C1	0.811	14	C1	0.628
	C2	-1.307		C2	0.395		C2	0.281
	C3	-0.346		C3	-1.107		C3	-0.708
3	C1	-0.642	9	C1	1.525			
	C2	-0.054		C2	1.406			
	C3	1.111		C3	0.313			
4	C1	-0.242	10	C1	0.961			
	C2	1.145		C2	0.302			
	C3	3.904		C3	-1.485			
5	C1	-0.056	11	C1	1.292			
	C2	-0.132		C2	-0.664			
	C3	1.230		C3	-4.158			
6	C1	-1.592	12	C1	-0.758			
	C2	-2.680		C2	-1.364			
	C3	-1.950		C3	-3.104			



a b

Figure 3. *Characteristic Curve: (a) Number 1, and; (b) Number 2*

From Figure 3 (a) and (b) it can be seen that for category 0, the higher the respondent's ability the lower the chance, whereas for category 3 the higher the respondent's ability the opportunity to answer is correct. As for categories 1 and 2, this is not the case, but the opportunity to answer correctly increases in line with the increase in ability and will reach its peak in a certain ability, while the opportunity will go down again in line with the ability of respondents.

From Table 5 it can be seen that the level of difficulty of grain moves from -3,104 to 3.904. An effective test has an item difficulty level of -2.00 to 2.00 (e. g. Wu & Adam, 2007; Hambleton, et al. 1991; Wright, & Stone, 1979). However, tests built to measure competence as well as instruments for measuring scientific literacy for high school students in the Sciences Program should be able to measure the ability of all test takers 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 as developed by the item response theory/normal distribution, the level of difficulty of items for competency measurement can start from -3.00 to 3.00, because at that interval it can measure around 99.98% of test-takers. Thus, from the results of the analysis of all items of the scientific literacy measurement instrument test for students who have been arranged, they are at intervals of -3.00 to 3.00 so that it is effective as a competency test. This is made clear by Figure 4 which describes the item map and Figure 5 which describes the person-map where all levels of item difficulty are at predetermined intervals. Figure 6 connects test takers' abilities and item difficulty levels.

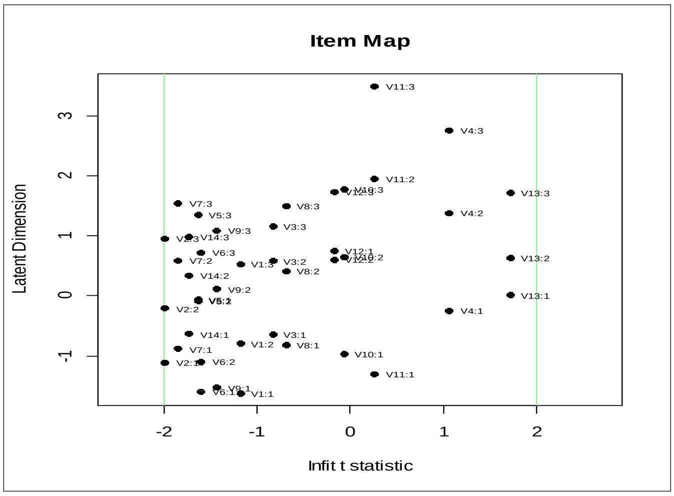


Figure 4.
Item Map of ISLS for Senior High School Students of Sciences Program

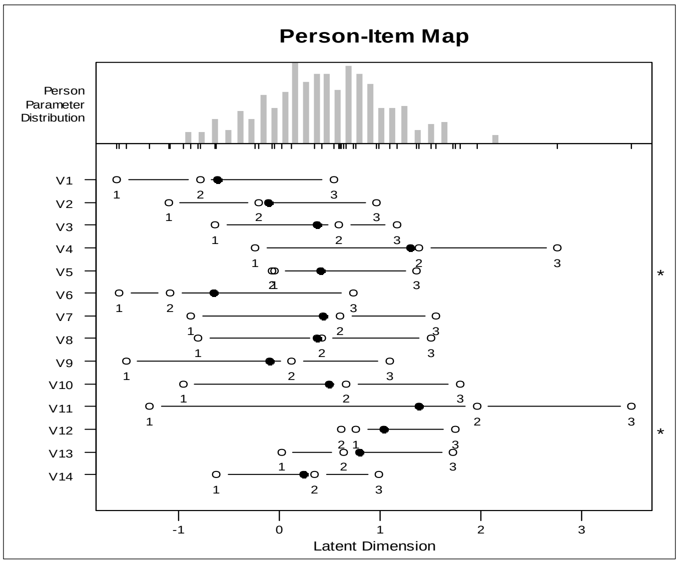


Figure 5.
Person-Item Map Item of ISLS for Senior High School Students of Sciences Program

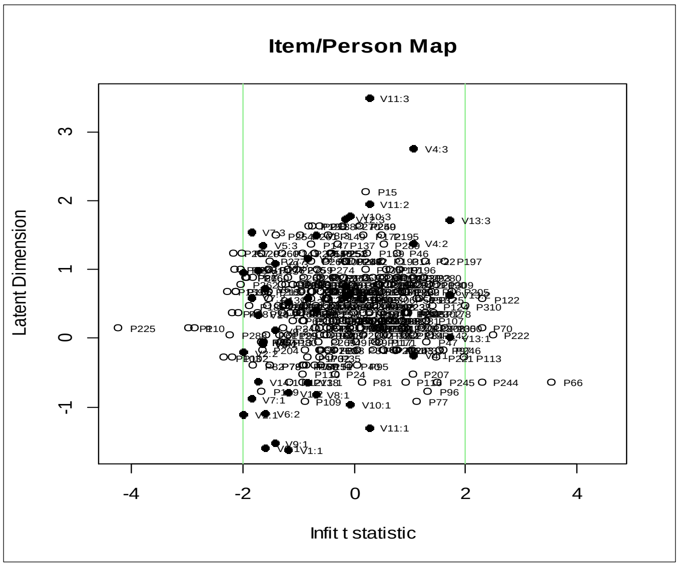


Figure 6.
Item/Person Map Items of ISLS for Senior High School Students of Sciences Program

Evidence that the items of scientific literacy measurement instruments for senior high school students in the Sciences program were used for the ability of test-takers between -3.00 and 3.00 explained by the item information function and tests (Figure 7). The figure illustrates that the information function will be maximal at the interval of student ability between 0 to 1.0 and effective between -3.0 to 3.00.

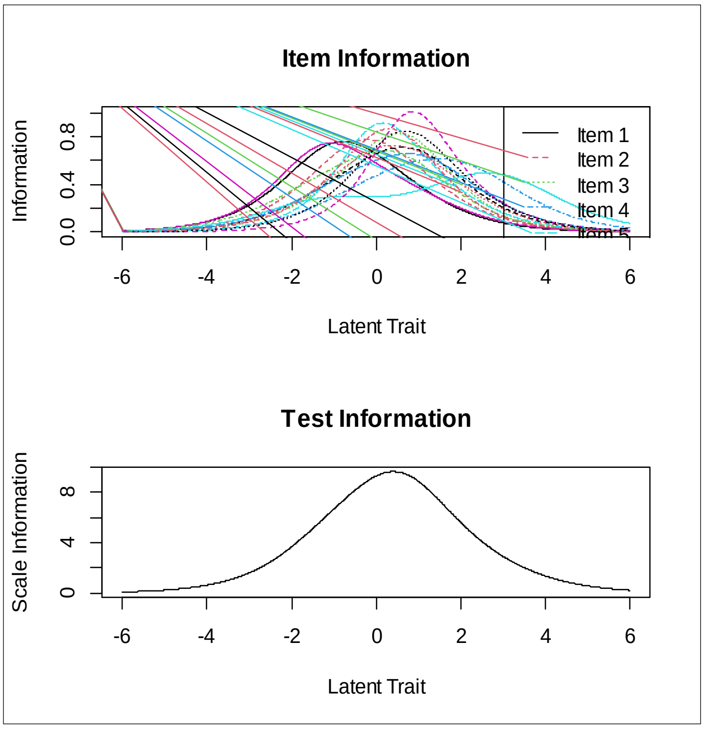


Figure 7.
Information Function of Item of ISLS for Senior High School Students of Sciences Program

The Validity of Constructive Substantive Aspects

To see the quality of the construct validity from the substantive aspect, a test of the ability of the test takers to the model is used. This test is basically to test the consistency of the response or different response patterns of participants to the test items based on the level of difficulty. Different response patterns are the mismatch of responses given based on their ability compared to the ideal model. A test participant who has an 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 cause an aberrant response. How many students experience this aberrant response is a measure of the validity of the substantive type construct.

This distorted response can be caused by inaccuracy, cheating or even misconceptions. A person's response test experiences irregularities or is not called a person fit. Criteria for acceptance of test-takers' responses are considered to be experiencing irregularities or not the same as item fit criteria. Quantitatively the response of the test taker who was declared fit or not experiencing deviation was if the MSQ Outfit value was between 0.5 and 1.5 while the outfit t value was between -2 to 2.0 and the probability of acceptance of Ho (model compatibility) was greater than 0.05 (p> 0.05). Of the 310 test-takers, there were fifteen test-takers who experienced a response that deviated from the model. It can be seen that the five test-takers did not meet as many as two (p-value and MSQ outfit) of the three-person fit criteria. The list of test-takers is described in Table 9.

From these explanations, it can be concluded that there were 95.16% responses of test-takers that were reasonable according to the model or did not experience deviations while there were 4.5% of responses experienced deviations. The large percentage of test-takers who have a reasonable response according to this model can be the basis that the test fulfils enough substantive validity. Even when using the 0.01 level of confidence, then all test participants' responses according to the model.

Student responses that deviate from the Rasch model show indications of students doing careless or lucky guess or even cheating (Sumintono & Widhiarso, 2015). Several studies have shown that person fit can be used as preliminary data for cheating, careless or lucky guess in students' test work (e. g. Shu et al. 2013; Wagner-Menghin et al. 2103; Meyer, & Zhu, 2013; Magis et al. 2012; Hohensinn, & Kubinger, 2011; Elhan et al. 2010; Lamprianou, 2010; Liu, & Yu, 2011).

Table 9.
Test Participants Who Have an Aberrant Response in ISLS

Participant	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t
P33	23.882	13	0.032	1.706	1.348	1.80	1.03
P47	22.440	13	0.049	1.603	1.454	1.60	1.30
P66	47.152	13	0.000	3.368	2.868	4.07	3.54
P70	26.728	13	0.014	1.909	1.897	2.25	2.29
P97	28.873	13	0.007	2.062	1.571	2.47	1.54
P98	23.095	13	0.041	1.650	1.145	1.75	0.54
P197	25.439	13	0.020	1.817	1.611	2.00	1.62
P205	23.695	13	0.034	1.692	1.629	1.83	1.76
P221	22.673	13	0.046	1.619	1.550	1.60	1.47
P222	28.252	13	0.008	2.018	2.013	2.44	2.49
P244	26.634	13	0.014	1.902	2.033	2.02	2.29
P245	25.489	13	0.020	1.821	1.594	1.87	1.49
P246	23.189	13	0.039	1.656	1.599	1.69	1.60
P300	23.126	13	0.040	1.652	1.583	1.73	1.62
P310	23.876	13	0.032	1.705	1.732	1.87	1.99

The Validity of Constructive Structural Aspects

There are two test indicators that have construct validity of structural aspects, namely the test is unidimensional and has stability in estimating the parameters of the items and test participants. Tests built in the one-dimensional paradigm must really have one dimension so that the measurement results obtained can have meaning. The principle of unidimensional testing is first stated by the null hypothesis which states that the second eigenvalue is not greater than the first eigenvalue with an alternative hypothesis that the second eigenvalue is greater than the first eigenvalue. The results of the unidimensional test analysis with the R program using the *ltm* package can be seen in Table 10 while the results of the analysis of the curve can be seen in Figure 8.

Table 10.
Unidimensional Test Results of Grains Instrument for Measurement of ISLS in the Science Program

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:	3.3528
Average of second eigenvalues in Monte Carlo samples:	2.2128
Monte Carlo samples:	100, p-value: 0.01

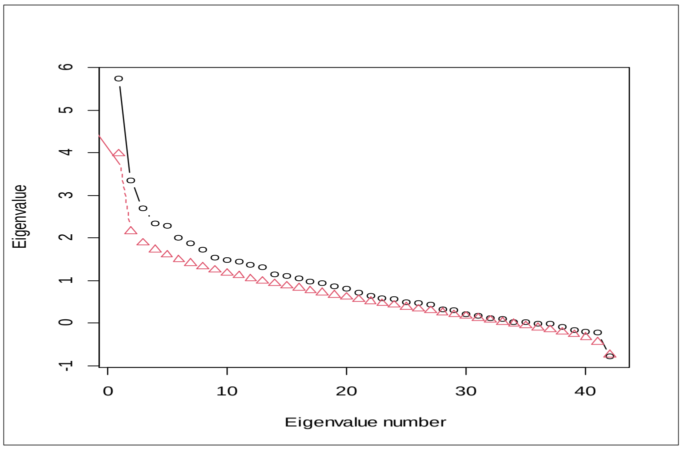


Figure 8.
Graph Analysis of Dimensionality of ISLS for Senior High School Students Sciences Program

From Table 10 it can be seen that the probability of the resulting unidimensional test is 0.01, a value greater than or equal to 0.01 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 science literacy tests for high school students in the Science program can be stated to be unidimensional.

The Validity of Constructions 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 domains) so that they can be seen whether they have 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 it could be standardized science learning achievement tests. It can be interpreted that the validity test of an external extract is basically an evaluation of an instrument that has been developed. In this regard, researchers will do the second year.

One approach to finding out the validity of the external aspect constraints in this first-year research is to use Person Separation reliability or Person Separation information. Person separation is used to classify people based on information obtained from tests. The low separation of people (less than 2) from the 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 still needed to measure it. The results of the Person separation analysis using the *eRm* package can be seen in Table 11.

Table 11.

Person Separation Reliability Test on the Items of Measurement Instruments for Scientific Literacy for Senior High School Students Sciences Program

Person Separation Reliability Test	
Separation Reliability	0.6062
Observed Variance	0.2811 (Squared Standard Deviation)
Mean Square Measurement Error	0.1107 (Model Error Variance)

From Table 11 it can be seen that the value of Person Separation reliability is 0.6062. Thus, the person separation score for the test is 1,142. From the value of the person, separation can be seen that the classification of test-takers obtained more than one or close to 2. This means that the instrument that has been created can distinguish test participants in two categories, namely literate and non-literate. The consequence is that the test results only distinguish test participants into two groups, namely test-takers who already have a minimum level of scientific literacy and who do not yet have a minimum level of scientific literacy. This information can be followed up in determining the graduation limit for science literacy tests for high school students of the Sciences Program.

The Validity of Constructive Aspects of Consequences

The consequential aspect in construct validity is the implication of the score interpretation as a source of action. Evidence regarding aspects of consequential validity also discusses the actual and potential consequences of testing and using scores, especially in terms of sources of invalidity such as bias, fairness, and distributive justice. In this regard, the measurement of scientific literacy for high school students in the Mathematics and Natural Sciences Program needs to be detected for a test bias.

In Rasch modelling with *eRm* packages, the detection of item bias can be approached by determining the items that are experiencing differential item functioning (DIF) using the Waldt Test. DIF relates to the estimation of different grain parameters in different subpopulations, in this case, the test takers are distinguished based on their sex. If an item is considered more difficult or easier by male test-takers than women or vice versa, then the item contains a DIF. DIF or also referred to as external item bias is not justification for item bias because, in order to know whether there is bias or not, an in-depth qualitative study must be conducted regarding the causes of DIF. However, the appearance of DIF can be a clue to the possibility of bias. Wald test on item level can be seen in Table 12.

When using a significance level of 0.05, there are no DIF detected. If at the significance level of 0.05, the probability of rejecting the correct Ho is 0.05. Ho here states that student responses to tests do not experience DIF. Related to this in the determination of DIF, the researchers chose a significance level of 0.05 so that there were no items that were considered to be detected by DIF.

Table 12.
Wald Test on Item Level of ISLS for Senior High School Students Sciences Program

Item Level	z-statistic	p-value	presence of DIF (5 %)
beta V1.c1	-0.285	0.775	not detected
beta V1.c2	0.089	0.929	not detected
beta V1.c3	-0.652	0.515	not detected
beta V2.c1	-0.663	0.508	not detected
beta V2.c2	-0.605	0.545	not detected
beta V2.c3	-2.301	0.021	not detected
beta V3.c1	-0.933	0.351	not detected
beta V3.c2	-0.557	0.578	not detected
beta V3.c3	-1.642	0.101	not detected
beta V4.c1	-0.556	0.578	not detected
beta V4.c2	-1.659	0.097	not detected
beta V4.c3	0.242	0.809	not detected
beta V5.c1	-1.418	0.156	not detected
beta V5.c2	-2.029	0.042	not detected
beta V5.c3	-2.907	0.004	not detected
beta V6.c1	1.611	0.107	not detected
beta V6.c2	1.945	0.052	not detected
beta V6.c3	1.483	0.138	not detected
beta V7.c1	0.511	0.610	not detected
beta V7.c2	1.645	0.100	not detected
beta V7.c3	-0.963	0.336	not detected
beta V8.c1	1.666	0.096	not detected
beta V8.c2	2.023	0.043	not detected
beta V8.c3	1.768	0.077	not detected
beta V9.c1	-0.322	0.747	not detected
beta V9.c2	-0.123	0.902	not detected
beta V9.c3	0.713	0.476	not detected
beta V10.c1	0.951	0.342	not detected
beta V10.c2	0.903	0.366	not detected
beta V10.c3	-0.140	0.889	not detected
beta V12.c1	1.113	0.266	not detected
beta V12.c2	-2.281	0.023	not detected
beta V12.c3	-0.773	0.440	not detected
beta V13.c1	1.081	0.279	not detected
beta V13.c2	0.173	0.863	not detected
beta V13.c3	1.568	0.117	not detected
beta V14.c1	-0.751	0.453	not detected
beta V14.c2	0.657	0.511	not detected
beta V14.c3	-1.528	0.127	not detected

Conclusion and Recommendations

Based on the opinion of the two assessors on the content and psychometric aspects, it can be concluded that the Integrating Science Literacy Scale (ISLS) for the Senior High School Student Science Program that has been made is feasible according to the measurement objectives. Based on the construct validity criteria on the content aspect: it appears that all test items are in accordance with the Rasch modeling used and have a difficulty level of items ranging from -3.104 to 3.904. Effective tests generally have an item difficulty level of -3.00 to 3.00 (eg Wu & Adam, 2007; Hambleton et al. 1991; Wright, & Stone, 1979). However, a test that was built to measure competence such as ISLS should be able to measure the ability of all test takers so that the distribution of difficulty levels is wider than the selection test. The paradigm of selection tests or tests that use norm references, while ability tests use reference criteria.

To see the quality of the construct validity from the substantive aspects, the test taker's response to the model is used. Of the 310 test takers, there were fifteen test takers who experienced responses that deviated from the model.

From this explanation it can be concluded that there are 95.16% of test takers 'responses that are reasonable according to the model or do not experience deviations, while 4.5% of test takers' responses have deviations.

Test indicators that have construct validity in structural aspects are unidimensional (Ravand & Firoozi, 2016). From Table 10 it can be seen that the probability of the resulting unidimensional test is 0.01, a value greater than or equal to 0.01 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 science literacy tests for high school students in the Science program can be stated to be unidimensional.

The validity of the external aspect construct is used to determine the extent to which the test results are supported by other measurements. From the person separation index, it can be seen that the classification of test participants is more than one or close to 2. This means that the instrument that has been made can distinguish test participants into two categories, namely literacy and non-literacy. Consequently, the test results only distinguish test participants into two groups, namely test takers who already have a minimum level of scientific literacy and those who do not have a minimum level of scientific literacy. This information can be followed up in determining the passing limit of the science literacy test for high school students of the Science Program.

Construct validity in the consequential aspect is the implication of the interpretation of the score as a source of action. This is related to the issue of dishonesty and test bias (Winne, 2020). In Rasch modelling with *eRm* packages, the detection of item bias can be approached by determining the items that are experiencing differential item functioning (DIF) using the Waldt Test. DIF relates to the estimation of different grain parameters in different subpopulations, in this case, the test takers are distinguished based on their sex. Related to this in the determination of DIF, the researchers chose a significance level of 0.05 so that there were no items that were considered to be detected by DIF.

From the results of the study note all items that are suitable to be used as measurement instruments for scientific literacy. While Items others with validity analysis which includes content, psychometrics, and constructs (content, substantive, structural, external, consequence) meet the requirements as good items. Scientific literacy is an important part of the education world and everyday life. With the development of Integrating Scientific Literacy Scale (ISLS) instrument, which was appropriate to the 2015 PISA standard, we hope that this instrument can be useful because it has gone through the research process.

Based on the results of this study, the researcher suggested that ISLS could be one of the tests used in the final exams for high school students in science programs. The ISLS score can be considered as one of the student graduation criteria. This has been welcomed so that starting in 2020 an ISLS test will be carried out on all senior high school students of the science program at SMAN 2 and SMAN 3, Tegal City. To strengthen the implementation of the test, a decree of the principal of SMAN 2 number 423.5 / 09/2020 and a decree of the principal of SMAN 3 number 420/020/2020 has been issued.

To find out more about the effectiveness of a test, it is necessary to evaluate, especially in relation to the validity of the resulting score. The ISLS score needs to be tested for its validity using a more valid score such as an intelligence test, as well as other potential tests that measure the same construct as ISLS. Further research is needed regarding the validity test of the scores produced by ISLS. Likewise in determining graduation, it is necessary to carry out further studies related to determining the passing grade of the ISLS.

Limitations of Study

The weakness of this study is that the validity of the criteria for the test instruments has not been done yet. Test the validity of the criteria is needed in order to ensure that the test results are in line with other standardized tests that have similar constructs. The validity test of this criterion can be done by comparing the results of this student's scientific literacy tests with the results of other tests such as intelligence tests, aptitude tests or national examination results.

As a large scale test, this test was only attended by 310 test participants from two schools in the City of Tegal. To become a standardized test that can be used at the city, provincial and even national levels, ISLS needs to be applied in a larger population. This is because the backgrounds of Indonesian students vary widely from the socioeconomic level, ethnicity, religion, gender and culture of society. However, there was an upside with Rasch modeling. Several studies have shown that the parameter estimates in Rasch modeling will be stable with a sample size of more than 100 test takers (Babcock, B., & Albano, 2012; Chen et al. 2014).

Acknowledgements

We would like to thank Kementerian Riset, Teknologi dan Pendidikan Tinggi Republic of Indonesia for providing funding for this research. Likewise, we would like to thank all parties involved, especially the Principals of SMA 2 and SMA 3 of Tegal City who has supported and granted research permits.

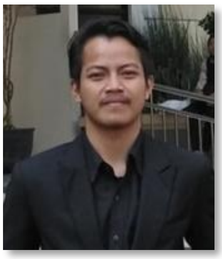
Biodata of Authors

Dr. Purwo Susongko, MPd. is a lecturer and researcher at the Universitas Pancasakti Tegal, Jl. Halmahera KM 1 Tegal , Central Java, Indonesia. He is associate professor in science educational assessment. His research focus is on Science educational Assessment especially in scientific literacy and Rasch Modeling application. **Affiliation:** University of Pancasakti Tegal, Indonesia **E-mail:** purwosusongko@upstegal.ac.id **Orcid ID**: 0000-0001-9126-1027 **Scopus ID**: 57200101798 **WoS Researcher ID:** ABB-3947-2020 **Phone:** (+62) 81802850666

Mobinta Kusuma is lecturer and researcher at the University of Pancasakti Tegal, Jl. Halmahera KM 1 Tegal, Central Java, Indonesia. She is Assist. Professor in science education department. **Affiliation**: University of Pancasakti Tegal,Indonesia **E-mail:** mobintakusuma@upstegal.ac.id **Orcid ID:** 0000 0002 5924 5075 **WoS Researcher ID:** AAA-5793-2020 **Phone:** (+62) 85640251605

Yuni Arfiani is lecturer and researcher at the University of Pancasakti Tegal, Jl. Halmahera KM 1 Tegal, Central Java, Indonesia. She is Assist. Prof. in science education department. **Affiliation:** University of Pancasakti Tegal, Indonesia **E-mail:** yuniarfiani@upstegal.ac.id **Orcid ID:** 0000-0002-9557-2102 **Phone:** (+62) 8562584220 **Scopus ID:** - **WoS Researcher ID:** -

Dr Achmad Samsudin, MPd. is lecturer and researcher at the University of Pendidikan Indonesia, Jl. Setiabudi No 229 Bandung, West Java, Indonesia. He is Assist. Prof. in Physics Education Department. His research focus is on physics education especially in the misconceptions, conceptual change and understanding. **Affiliation:** University of Pendidikan Indonesia **E-mail:** achmadsamsudin@upi.edu **Orcid ID:** 0000-0003-3564-6031 **Scopus ID:** 57191537500 **WoS Researcher ID:** E-5170-2015 **Phone:** (+62) 85225709383



Adam Hadiana Aminudin, MPd. is researcher at the University of Pendidikan Indonesia, Jl. Setiabudi No 229 Bandung, West Java, Indonesia. His research focusses on physics education. **Affiliation:** University of Pendidikan Indonesia **E-mail:** adamhadiana@upi.edu@upi.edu **Orcid ID:** 0000-0001-7409-9195 **Scopus ID:** 57216794098 **WoS Researcher ID:** S-7982-2018 **Phone:** (+62) 81313086163

References

Abdul Rahim, F., & Chun, L. S. (2017). Proposing an affective literacy framework for young learners of English in Malaysian rural areas: Its key dimensions and challenges. *Malaysian Journal of Learning and Instruction*, *14*(2), 115–144. https://doi.org/10.32890/mjli2017.14.2.5

Adeleke, A. A., & Joshua, E. O. (2015). Development and Validation of Scientific Literacy Achievement Test to Assess Senior Secondary School Students ' Literacy Acquisition in Physics. *Journal of Education and Practice*, *6*(7), 28–43.

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement*, *36*(7), 565-580.

Baghaei, P., & Amrahi, N. (2011). Validation of a Multiple Choice English Vocabulary Test with the Rasch Model. *Journal of Language Teaching and Research*, *2*(5), 1052-1060. https://doi.org/10.4304/jltr.2.5.1052-1060

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), 421–434. https://doi.org/10.1088/0143-0807/34/2/421

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Branch, R. M. (2009). *Instructional design: The ADDIE approach* (Vol. 722). Springer Science & Business Media.

Benjamin, T.E., Marks, B., Demetrikopoulos, M.K., Rose, J., Pollard, E., Thomas, A., & Muldrow, L.L. (2017). Development and Validation of Scientific Literacy Scale for College Preparedness in STEM with Freshmen from Diverse Institutions. *International Journal of Science and Mathematics Education, 15*, 607–623. https://doi.org/10.1007/s10763-015-9710-x

Bybee, R. W. (2012). Scientific literacy in environmental and health education. In *Science / Environment / Health: Towards a Renewed Pedagogy for Science Education* (Vol. 9789048139491, pp. 49–67). Springer Netherlands. https://doi.org/10.1007/978-90-481-3949-1_4

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of life research*, *23*(2), 485-493.

Dietz, R. D., Pearson, R. H., Semak, M. R., & Willis, C. W. (2012). Gender bias in the force concept inventory? *AIP Conference Proceedings*, *1413*, 171–174. https://doi.org/10.1063/1.3680022

Edwards, A., & Alcock, A. (2010). Using rasch analysis to identify uncharacteristic responses to undergraduate assessments. *Teaching Mathematics and Its Applications*, *29*(4), 165–175. https://doi.org/10.1093/teamat/hrq008

Gormally, C., Brickman, P., & Lut, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE Life Sciences Education*, *11*(4), 364–377. https://doi.org/10.1187/cbe.12-03-0026

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hanson, S. (2016). *The assessment of scientific reasoning skills of high school science students: A standardized assessment instrument*. http://ir.library.illinoisstate.edu/cgi/viewcontent.cgi?article=1505&context=etd

Hanushek, E. A., & Woessmann, L. (2016). Knowledge capital, growth, and the East Asian miracle. *Science*, *351*(6271), 344–345. https://doi.org/10.1126/science.aad7796

Heng, L. L., Surif, J., Seng, C. H., & Ibrahim, N. H. (2015). Mastery of scientific argumentation on the concept of neutralization in chemistry: A Malaysian perspective. *Malaysian Journal of Learning and Instruction*, *12*(1), 85–101. https://doi.org/10.32890/mjli2015.12.5

Herrmann-Abell, C. F., & DeBoer, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice*, *12*(2), 184–192. https://doi.org/10.1039/c1rp90023d

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), 380–393. http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/07_Hohensinn.pdf

Holbrook, J., & Rannikmae, M. (2009). The Meaning of Scientific Literacy. In *ERIC*. http://www.ijese.com/

Jufri, A. W., Hakim, A., & Ramdani, A. (2019). Instrument Development in Measuring the Scientific Literacy Integrated Character Level of Junior High School Students. *Journal of Physics: Conference Series*, *1233*(1). https://doi.org/10.1088/1742-6596/1233/1/012100

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. https://doi.org/10.1007/s10763-011-9314-z

Lamprianou, I. (2010). The practical application of Optimal Appropriateness Measurement on empirical data using rasch models. *Journal of Applied Measurement*, *11*(4), 409–423.

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

Madsen, A., McKagan, S. B., & Sayre, E. C. (2013). 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), 020121-020136. https://doi.org/10.1103/PhysRevSTPER.9.020121

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of snijders's l z* index of person fit with emphasis on response model selection and ability estimation. In *Journal of Educational and Behavioral Statistics*, 37(1), 57–81. https://doi.org/10.3102/1076998610396894

Mari, L., Carbone, P., & Petri, D. (2012). Measurement fundamentals: A pragmatic view. *IEEE Transactions on Instrumentation and Measurement*, *61*(8), 2107–2115. https://doi.org/10.1109/TIM.2012.2193693

Md-Ali, R., Karim, H. B. B. A., & Yusof, F. M. (2016). Experienced primary school teachers' thoughts on effective teachers of literacy and numeracy. *Malaysian Journal of Learning and Instruction*, *13*(1), 43–62. https://doi.org/10.32890/mjli2016.13.1.3

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256. https://doi.org/10.1177/026553229601300302

Meyer, J. ., & 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. http://www.rpajournal.com/dev/wp-content/uploads/2013/05/SF3.pdf&sa=X&scisig=AAGBfm2n7WN_mLyfzwk1qYsjHvIek10RhA&oi=scholarr&ei=JU2iUtXYMJGrhQfYz4HQCA&ved=0CDAQgAMoATAA

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. https://doi.org/10.1119/1.4731618

Neumann, I., Neumann, K., & Nehm, R. (2010). Evaluating instrument quality in science education: Rasch-based analyses of a Nature of Science Test. *Taylor & Francis*. https://doi.org/10.1080/09500693.2010.511297ï

Nordin, H., & Ariffin, T. F. T. (2016). Validation of a technological pedagogical content knowledge instrument in a Malaysian secondary school context. *Malaysian Journal of Learning and Instruction*, *13*(1), 1–24. https://doi.org/10.32890/mjli2016.13.1.1

OECD. (2015). *OECD iLibrary | PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. https://www.oecd-ilibrary.org/education/pisa-2015-assessment-and-analytical-framework_9789264255425-en

O'Neill, T. R., Gregg, J. L., & Peabody, M. R. (2020). Effect of sample size on common item equating using the dichotomous rasch model. *Applied Measurement in Education*, *33*(1), 10-23.

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). https://doi.org/10.1103/PhysRevSTPER.6.010103

Pratiwi, M., Siahaan, P., Samsudin, A., Aminudin, A. H., & Rachmadtullah, R. (2020). *Introduction , Connection , Application , Reflection , Extension-Multimedia Based Integrated Instruction (ICARE-U): A Model to Improve Creative Thinking Skills*. *24*(08).

Ravand, H., & Firoozi, T. (2016). Examining construct validity of the master's UEE using the Rasch model and the six aspects of the Messick's framework. *International Journal of Language Testing*, *6*(1), 1-18.

Romine, W. L., Schaffer, D. L., & Barrow, L. (2015). International Journal of Science Education 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. *Taylor & Francis*, *37*(16), 2740–2768. https://doi.org/10.1080/09500693.2015.1105398

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. https://doi.org/10.1002/tea.21303

Runnels, J. (2012). Using the Rasch model to validate a multiple choice English achievement test. *International Journal of Language Studies*, *6*(4), 141-155.

Rusilowati, A., Kurniawati, L., Nugroho, S. E., & Widiyatmoko, A. (2016). Developing an instrument of scientific literacy asessment on the cycle theme. *International Journal of Environmental and Science Education*, *11*(12), 5718–5727.

Rusilowati, A., Nugroho, S. E., Susilowati, E. S. M., Mustika, T., Harfiyani, N., & Prabowo, H. T. (2018). The development of scientific literacy assessment to measure student's scientific literacy skills in energy theme. *Journal of Physics: Conference Series*, *983*(1). https://doi.org/10.1088/1742-6596/983/1/012046

Saddhono, K., & Rohmadi, M. (2014). A sociolinguistics study on the use of the Javanese language in the learning process in primary schools in Surakarta, Central Java, Indonesia. *International Education Studies*, *7*(6), 25–30. https://doi.org/10.5539/ies.v7n6p25

Samsudin, A. (2020). Rasch Analysis: Measuring Students Attitudes toward Physics using the CLASS. *Test Engineering & Management*, *83*(June), 15461–15467.

Sheu, T. W., Tsai, C. P., Tzeng, J. W., Chen, T. L., & Nagai, M. (2013). An algorithm of the misconception order. *Applied Mechanics and Materials*, *284-287*, 3010–3014. https://doi.org/10.4028/www.scientific.net/AMM.284-287.3010

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. https://doi.org/10.1007/s11336-012-9311-3

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100-107.

Smith, A. B., Fallowfield, L. J., Stark, D. P., Velikova, G., & Jenkins, V. (2010). A Rasch and confirmatory factor analysis of the General Health Questionnaire (GHQ) - 12. *Health and Quality of Life Outcomes*, *8*. https://doi.org/10.1186/1477-7525-8-45

Soobard, R., & Rannikmäe, M. (2011). Assessing student's level of scientific literacy using interdisciplinary scenarios. In *Science Education International*, *22*(2), 133-144. https://eric.ed.gov/?id=EJ941672

Stenbeck, M., Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1992). Fundamentals of Item Response Theory. *Contemporary Sociology*, *21*(2), 289. https://doi.org/10.2307/2075521

Suhandi, A., & Samsudin, A. (2019). Effectiveness of the use of developed teacher's book in guiding the implementation of physics teaching that provides science literacy and instill spiritual attitudes. *Journal of Physics: Conference Series*, *1280*(5). https://doi.org/10.1088/1742-6596/1280/5/052054

Sumintono, B. (2018). *Rasch Model Measurements as Tools in Assesment for Learning*. https://doi.org/10.2991/icei-17.2018.11

Suryana, T. G. S., Setyadin, A. H., Samsudin, A., & Kaniawati, I. (2020). Assessing Multidimensional Energy Literacy of High School Students: An Analysis of Rasch Model. *Journal of Physics: Conference Series*, *1467*(1), 1-10. https://doi.org/10.1088/1742-6596/1467/1/012034

Susongko, P. (2016). Validation of science achievement test with the Rasch model. *Jurnal Pendidikan IPA Indonesia*, *5*(2), 268–277. https://doi.org/10.15294/jpii.v5i2.7690

Susongko, P., Widiatmo, H., Kusuma, M., & Afiani, Y. (2019). Development of integrated science-based science literacy skills instruments using the Rasch model. *Unnes Science Education Journal*, *8*(3), 268-277.

Tamassia, L., & Frans, R. (2014). Does integrated science education improve scientific literacy? In *Journal of the European Teacher Education Network*, *9*(1),131–141. http://62.28.241.73/index.php/jeten/article/view/44

Wagner-Menghin, M., Preusche, I., & Schmidts, M. (2013). The Effects of Reusing Written Test Items: A Study Using the Rasch Model. *ISRN Education*, *2013*, 1–7. https://doi.org/10.1155/2013/585420

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. https://doi.org/10.1103/PhysRevPhysEducRes.12.020111

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. https://doi.org/10.1002/sce.21172

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior*, *112*, 106457.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design: Rasch Measurement*. Chicago, Illinois: MESA Press

Wright, B., & Mok, M. M. C. (2004). An Overview of the Family of Rasch Measurement Models. *Introduction to Rasch Measurement*, 1–24. http://www.statistica.unimib.it/utenti/lovaglio/overview rasch.pdf

Wu, M., & Adams, R. (2007). *Applyıng the Rasch model to psycho-socıal measurement a Practıcal Approach* https://pdfs.semanticscholar.org/a263/e16ffe74da6ec87d7855ed878d6ab90acfed.pdf

Yenni, R., Hernani, & Widodo, A. (2017). The implementation of integrated science teaching materials based socio-scientific issues to improve students scientific literacy for environmental pollution theme. *AIP Conference Proceedings*, 1848, 060002 https://doi.org/10.1063/1.4983970