IJER Indonesian Journal of Educational Review
p-ISSN 2338-2018 | e-ISSN 2335-8407
Vol. 6, No.1 , July 2019, p 44-55

Available online at
http://pps.unj.ac.id/journal/ijer

A RASCH MODEL MEASUREMENT ANALYSIS ON SCIENCE LITERACY TEST OF INDONESIAN STUDENTS: SMART WAY TO IMPROVE THE LEARNING ASSESSMENT

Rosita Uli Sihombing[1], Dali S. Naga[1], Wardani Rahayu[1]
State University of Jakarta
rositasihombing@gmail.com

Abstract

Report of a test assessment is usually carried out separately between the level of difficulty of the test items and the student's ability. In addition, the analysis of test results is sometimes not clearly presented in a complete and in-depth explanation. This study aimed to explain the assessment process on the instrument of Science Literacy Test for Indonesian Students (SLTIS) conducted by 94 students from the 9th grade of junior high school. Those were 41 students from private schools and 53 from public schools. There were 36 test items with four possible answers for each item. In details, the findings were reported as follows: the item reliability was very good; two items were indicated to have different item functioning (DIF), and the data were fit to the model; public school students were slightly better than private school students, and there was no difference in achievement between male and female students. SLTIS instrument was suitable for diagnostic test and had a high information value on the students with moderate ability. The current study presented a smart way on how to apply the objective measurement to improve the education assessment.

Keywords: *science literacy, indonesian schools, Science Literacy Test for Indonesian Students (SLTIS), Rasch Measurement Model (RMM)*

The 21st century offers unlimited world life explosion on globalization, internationalization, and information and communication technology (ICT). For this reason, students are required to excel in academic performance and master 21st century skills to be able to face such challenges (Bybee, McCrae, & Laurie, 2009). Science and technology education has the capacity for competency of critical thinking, complex communication skills, and structured problem-solving skills known as Higher Order Thinking Skills (HOTS), which are required to achieve better work in the future (Binkley et al., 2012; Turiman, Omar, Daud, & Osman, 2012). Science literacy is another important ability in the 21st century needed to creatively utilize knowledge (Gormally, Brickman, & Lut, 2012).
Recently, many countries have applied uniform international standards in assessing their education (Kamens & McNeely, 2010). Trend in Mathematics and Science Studies (TIMSS) and the Program for International Student Assessment (PISA) have contributed the efforts to improve students' higher order thinking skills. PISA 2015 reported that the scientific literacy ability of Indonesian students, which are further mentioned by Science Literacy Ability (SLA) was at the point of 403 which was below the OECD standard points (493). Even Thailand and Vietnam surpassed Indonesia (point of SLA is 421 and 525 respectively). The scale of international scientific literacy skills was divided into 6 levels and the ability of Indonesian students was at level 1 (low). This indicated that ± 41.3 % of Indonesian students only have limited scientific knowledge, which was applied to some familiar situations (Tjalla, 2010).

The important thing about improving the quality of Indonesian teaching is, besides providing essential and strategic learning material, on enhancing the assessment result of classroom's daily learning (Tjalla, 2010; Tohir, 2018). Assessment should be carried out properly and correctly, so that it can measure students' ability to solve problems thoroughly, apart from assessing the ability to the

extent of concept. It is also essential that assessment on various competencies be developed owing to how important it is to test the difficulty of test instruments and determine students' ability through assessment of education. The assessment process should be able to produce the right measurement and analysis as a means to determine the quality of results and efforts in improving the learning process (Chan, Ismail, & Sumintono, 2014).

This study evaluated and analyzed the instrument items of Science Literacy Test for Indonesian Students (SLTIS) that have been developed. The quality/difficulty of the test items and the student's ability were analyzed using Rasch Model Measurement (RMM) approach. Numerous studies in education sector have applied RMM approach and provided positive information in the advancement of education. Some of the examples are assessment of statistical reasoning skills in junior high school students (Chan et al., 2014), development of science education assessment instruments (X. Liu, 2009; Sondergeld & Johnson, 2014; XiufengLiu, 2012), assessment of language tests (McNamara & Knoch, 2012), assessments on educational ability test (Connelly, Warren, Kim, & Di Domenico, 2016; Engelhard, 2009), and others. In particular, this research outlines a systematic analysis of the Evaluation of SLTIS instrument items. Then it is followed by a literature review, research methodology, research data analysis and interpretation. At the end of article, the conclusion of the SLTIS instrument analysis will be reviewed using RMM approach.

Literature Review

TIMSS is an international comparative study to assess the achievement in mathematics and science for students from the 4th and 8th grade (aged 10 and 14 years respectively). It aims at gathering information about the educational context related to students' achievement. TIMSS assesses students' knowledge and abilities in mathematics and science and their ability to apply knowledge to solve problems. The test items are designed to measure what they know and do through mathematics and science content on process and cognitive abilities such as knowledge, applications, and reasoning (Provasnik & Malley, 2016). TIMSS 2015 reported that Indonesia obtained a score of 397 for the science category (the lowest category), below the standard score of 500. Indonesia ranked 43rd out of 47 participating countries (Nizam, 2016). This information is important because by linking the results of national education with TIMSS, PISA, and other potential international assessments, we can compare and evaluate the achievement in national levels with the one in international assessments and Indonesian students' performance with students from other countries (Cresswell, Schwatner, & (Cresswell, Schwantner, & Waters, 2015; Mcconney, Oliver, Woods-Mcconney, Schibeci, & Maor, 2014; Neidorf, Binkley, Gattis, & Nohara, 2006).

PISA assessment was carried out by 15-year-old students focusing on the mastery of scientific competencies, understanding concepts, and the ability to apply these concepts and competencies in a variety of life situations. PISA assesses students more deeply in the knowledge and skills they need in adult life. This is what distinguishes knowledge assessment on PISA from TIMSS. High participation in PISA and the perceived progress is a clear sign of the importance of scientific literacy as a result of education and a progressive alternative to school-based science assessment (Bybee et al., 2009; R. V. Olsen & Grønmo, 2004). TIMSS and PISA are better known as a type of test that measures high-level thinking skills (HOTs).

The test items have several implications that a) it is important for everyone when facing complicated decisions/situations, b) the complex situations in HOTs evaluation need to be presented. The research shows that failure to master HOTs aspect can be a source of major difficulties in learning. HOTs in learning mathematics and science are so important, especially in developing students' ability to analyze, evaluate, and create useful new things, that it leads students to be more critical and creative in solving problems. Some efforts that can be taken to improve students' HOTs in mathematics and science are by (1) involving students in non-routine problem solving activities; (2)

facilitating students to develop the ability to analyze and evaluate (critical thinking) new things and the ability to create a new discovery (creative thinking); and (3) encouraging students to build their own knowledge in making learning be more meaningful for students (Apino & Retnawati, 2017). HOTs item in the assessment context are measuring students' abilities in: 1) transferring one concept to another; 2) processing and applying information; 3) looking for links from different information; 4) using information to solve problems; and 5) critically reviewing ideas and information (Widana, 2017).

METHOD

In this part, it was explained about three things namely, the development of instrument - Participants and procedures - and Rasch Measurement Model (RMM) Analysis.

1. Development of Instrument

The instrument tested in this research was Indonesian Science Literacy Test (ISLT) developed based on the Curriculum of Indonesian Secondary Education using higher order thinking learning approach (HOTs) and PISA framework. The subject matter tested was integrated science in grade 9 of junior high school. It is necessary to test the ISLT instrument in order to get a good and proper instrument, especially for testing students' scientific literacy skills.

2. Participants and procedures

Total participants in this study were 94 students from the 9^{th} grade, with 41 students from private junior high school and 53 from the public one. Students were coordinated by their teacher to participate in completing 36 test items of ISLT instrument which had 4 possible answers for each item. To identify the participants' demographics, the researchers set number 01 to 94 as codes for participant numbers and L code for males and P code for females. The item label applied code S1 to S36. In the instrument testing process, participants conducted tests in separate schools. Each student got a hardcopy of ISLT, an answer sheet, blank paper, pencil test, and the allocation time of 80 minutes for taking the test. The completed test instruments were then collected. Afterwards, the data were inputted into Excel worksheet, screened, and validated. The total number of the data was 3,384 (94 participants x 36 items). There were 38 missing data (1.12%) because 29 participants did not complete the test or were not able to answer the test. Table 1 presents the participants' demographic data.

Table 1. Participants' demographic data (n = 94)

School Type	Demographic (Gender)	Number of Students	Percentage (%)
Public	Male	29	30.85%
	Female	24	25.53%
Private	Male	14	14.89%
	Female	27	28.72%

Table 1 presents that women participants (54.3%) were more in numbers than men (45.7%). Public junior high school students were (56.4%) more in numbers than private junior high school students (43.6%). Both schools are located in the rural area (Pangkalpinang city, Bangka-Belitung Islands Province).

3. Rasch Model Measurement (RMM) Analysis

This study analyzed data using Rasch Model Measurement (RMM) approach. This model was initiated in 1960 by Georg Rasch who developed an analytical model of item response theory (IRT) which was originally called 1PL (one logistics parameter). The raw data processing (dichotomous data) was made into a formulation model indicating the correlation between students' ability and the level of items' difficulty (Linacre, 2004; L. Olsen, 2003). Through RMM analysis dichotomy data, the information about reliability, level of items' difficulty, person' ability, DIF item distractors, and others,

were obtained (Bond & Fox, 2007). Furthermore, this mathematical model was popularized by Ben Wright. The basic principle of Rasch modeling is to convert raw/ordinal data into interval data for statistical analysis purposes. The raw data is seen as a probability (odd probability), which is a comparison between true and false answers. The logarithm function is used to produce measurements with the same interval in the form of log odds units, indicating the student's ability and item's difficulty which has the same scale and to draw conclusion on the level of students' achievement which depends on level of item's difficulty (Olsen, 2003).

The RMM concept is to make a measurement scale with the same interval since raw scores do not have intrinsic properties. They are not used directly to provide interpretations on students' abilities. RMM jointly uses data based on the person and the item. The scores become the basis for estimating a true score that shows the level of person's ability and the level of item's difficulty (Chan et al., 2014). For dichotomy data, RMM combines an algorithm that states the results of probabilistic expectations of item 'i' and person 'n' which are mathematically formulated as follow:

$$Pni\ (Xni=1, \beta n\ di)\ ^{\wedge} = (e\ (\beta n\ -di))\ /\ (1+e\ (\beta n\ -di)) \qquad (1)$$

Where:

Pni (Xni = 1, bn, di) is the probability of the person 'n' in item 'i' to produce a correct answer (x = 1); with the person's ability, βn, and the level of item's difficulty δi (Bond & Fox, 2007).

RMM is very appropriate to be used in educational research especially in testing the development of instrument (Smith & Barnes, 2007; Sondergeld & Johnson, 2014). Compared to classical test theory, RMM analysis can show its advantages such as the ability to predict the missing data for more accurate analysis results, comprehensive testing to respondent (person and item), the ability to be done in quantitative and qualitative research, and the ability to calibrate three things in once, i.e. measurement scale, respondent, and item. Measurement of instrument in research is important to be calibrated for producing valid data and bringing about the best instrument (Chan et al., 2014; Myford, 2010; Van Zile-Tamsen, 2017).

RESULT

In this part, it was explained about five things namely, Summary of Statistics of ISLT Test Item - Psychometric Attributes of Item and Person – Item Bias - Differences in achievement between schools and gender – and about Information in test function.

1. Summary of Statistics of ISLT Test Item

RMM was used to analyze 36 ISLT test items tested on 94 participants. Summary of statistics is shown in Table 2.

Table 2. Summary of statistics of ISLT test item by RMM analysis

Information	ISLT Test Item	Person
N	36	94
Measures		
Mean	0.00	0.26
SD	0.95	0.72
SE	0.25	0.40
Outfit MNSQ		
Mean	1.05	1.05
SD	0.30	0.40
Intfit MNSQ		
Mean	1.00	0.99
SD	0.10	0.13

Outtfit ZSTD		
Mean	0.00	0.10
SD	1.20	1.10
Intfit ZSTD		
Mean	1.00	- 0.10
SD	0.10	1.10
Separation	3.79	1.82
Reliability	0.93	0.77
KR-20 Person Raw Score	0.79	

Table 2 shows the measure of person with 0.26 logit indicating that the mean value of all students was greater than the value of item measure which was 0.00 logit. It means the tendency for students' ability was slightly higher than the item's difficulty. The Cronbach alpha value aimed to measure reliability, which is the interaction between the person and the item as a whole = 0.79 showing that the reliability was quite good. The value of the person's reliability was 0.77 and the item's reliability was 0.93 pointing out that the consistency of students' answers was quite good, and the quality of instrument items in reliability aspects was very good. The mean of infit and outfit MNSQ for person was 0.99 and 1.05 respectively, while that of MNSQ infit and outfit data for items was 1.00 and 1.05. This is a sign of data being pursuant to the model as it was almost adjusted to the ideal value of 1.00. The mean value of ZSTD infit and outfit for person was 0.00 and -0.10, while that of ZSTD infit and outfit for items was 1.00 and 0.00 which indicated that the item and person's ZSTD were good because they could equalize the ideal value which was 0.00.

The grouping of person and item can be known from the separation value. The greater the value of separation, the better the quality of person and item's instrument, because it can clearly identify the groups of person and items. The item separation of 3.79 or 4 indicated that there were four groups of items classified into very difficult, difficult, moderate, and easy. The person separation of 1.82 or 2 showed that there were two groups of person classified in high ability and low ability (Saito, 2008; Van Zile-Tamsen, 2017).

2. Psychometric Attributes of Item and Person

One of RMM competences in analyzing data was able to obtain the information on psychometric attributes of item and person such as item's difficulty, person's ability, item's and person's fit, and the item's bias.

Table 3. Level of item difficulty and item fit

Information	Item Number	Measure	Outfit MNSQ	Outfit ZSTD
Highest Level of Item's Difficulty	S26	3.03	2.50	2.4
Lowest Level of Item's Difficulty	S21	-2.14	0.75	-0.6
Item unfit	S26	3.03	2.50	2.4
	S31	2.04	1.60	2.0
	S34	-0.60	0.69	-2.5
	S20	0.87	1.38	2.8

Based on Table 3, it is obtained that the highest level of item's difficulty was at S26 (3.03 logit), while the lowest one was at S1 (-2.14 logit). The four item unfit were S26, S31, S34 and S20 because the MNSQ and ZSTD items were not in the standard value area (Boone & Scantlebury, 2006).

Table 4. Level of student's ability and person fit

Information	Person Number	Measure	Outfit MNSQ	Outfit ZSTD
Highest Level of Person's Ability	37 LN	1.88	0.94	0.0
Lowest Level of Person's Ability	79PS	-1.87	1.04	0.3
Person unfit	53PN	0.10	1.78	3.3
	13LN	-0.30	1.66	2.6
	16LN	-0.30	1.85	3.2
	74LS	-0.70	1.58	1.9
	63LS	-0.84	1.63	1,9

Table 4 presents that the highest level of person's ability was 37 LN (1.88 logit), and the lowest was 79PS (-1.87 logit). Five students unfit were 53P, 13LN, 16LN, 74LS, and 63LS because the person's MNSQ and ZSTD values were not in the standard value area. Figure 1 explains the distribution of items difficulty and person abilities.

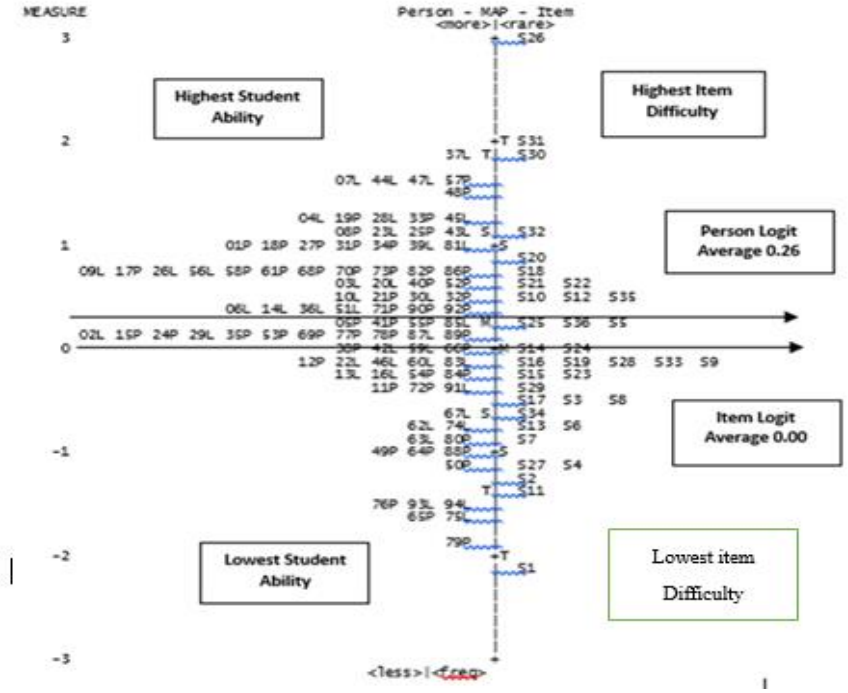


Figure 1. Wright map

3. Item Bias (DIF)

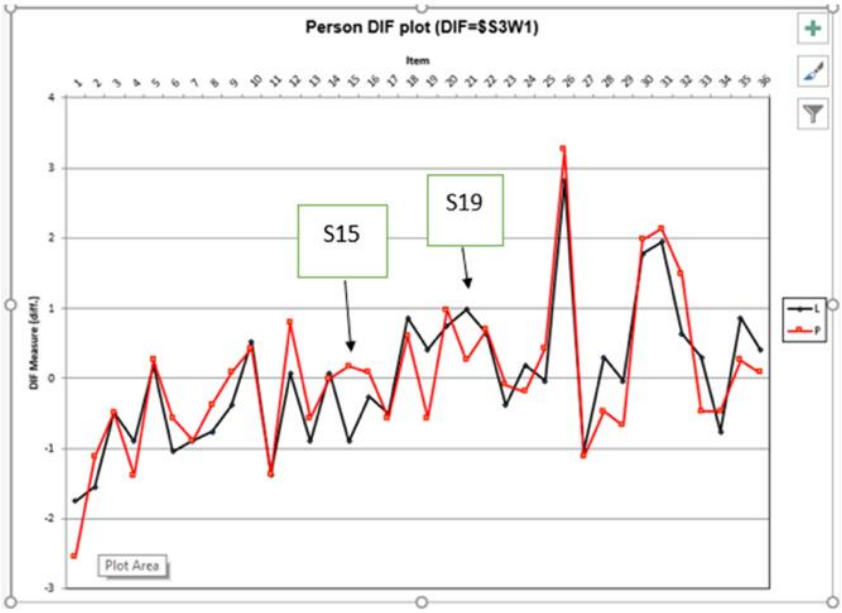
Item bias is also called Differential Item Functioning (DIF) which is a condition of two groups of respondents having the same ability, but the item bias can produce different scores. Some groups of respondents were benefited and some were not (Co-Author, 2013). Table 5 shows the item bias on the ISLT instrument. Figure 2 illustrates the diagram of DIF condition.

Table 5**.** DIF of Test items

Item Number	Probability	MNSQ	ZSTD
S15	0.0282	5.0039	2.0001
S19	0.0359	4.6310	1.8860

In Figure 2, the curve closed at the upper limit was item S26, which showed that the item's difficulty was high, while the curve closed at the lower limit was item S1 indicating that the test item was easy. Item S15 and S19 were DIF items; it appears that these items were easy to do by males

(black lines) compared to females (red lines) as it is presented in the figure that the black line is below the red line.



4. Differences in achievement between schools and gender

The data obtained were tested using independent sample of t test. Data were processed using the SPSS application. Table 6 and Table 7 show a summary of SPSS outputs for school types and gender.

Table 6. Summary of output of school type by t-test

Information	School Type	
	Public	Private
Mean	0.57	-0.15
SD	0.66	0.85
F	3.142	
p-value	0.80	
Sig (2 tailed)	0.00	

Table 6 shows that the results of the mean value of SPSS output for public school's performance was 0.57 and the standard deviation was 0.66. The private school obtained - 0.15 and 0.85. Descriptively, public school was better than private school. The homogenity test was 0.80 > 0.05, pointing out that both groups were mutually homogeneous. Sig in t-test for equality obtained a mean of 0.00 / 2 < 0.05 or sig < alpha. Accordingly, there were differences between public and private schools.

Table 7. Summary of output of gender type by t-test

Information	Gender	
	Male Student	Female Student
Mean	0.29	0.23
SD	0.87	0.81
F	0.389	
p-value	0.53	
Sig (2 tailed)	0.74	

Table 7 shows that the average ability of male students was 0.29 and the standard deviation was 0.87 while that of female students was 0.23 and 0.81 respectively. This indicated that descriptively male students were better than female students. The homogenity test was 0.53 > 0.05, pointing out that those two groups were homogeneous. Sig in the t-test for equality obtained a mean of 0.74 / 2 > 0.05 or sig > alpha. Accordingly, there was no difference in achievement between male and female students.

5. Information in test function

Each measurement always produces information about the measurement results. Measurement information depends on the relationship between the test and the individual measured. Figure 3 shows a graph of the test information function.

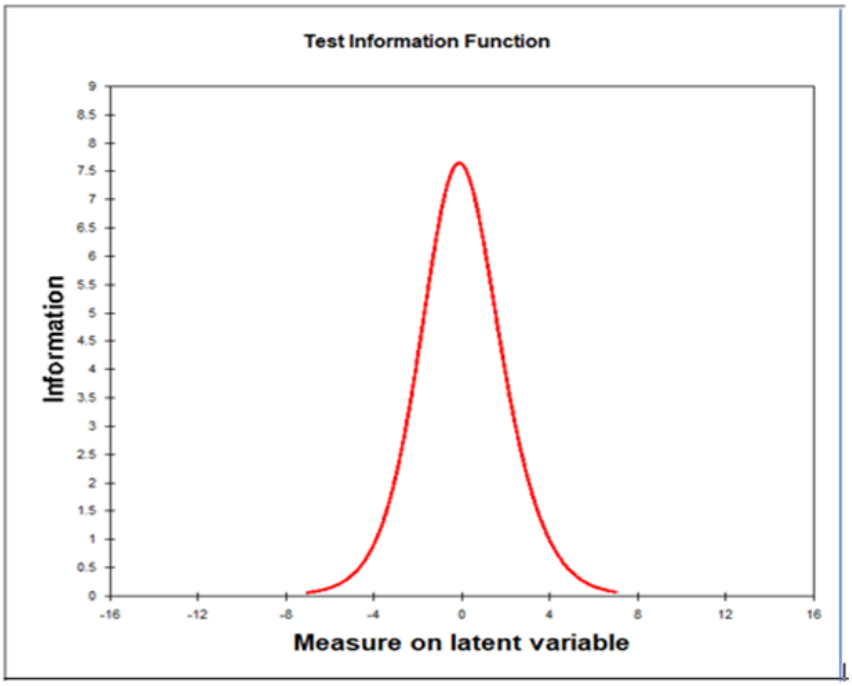


Figure 3. Test information function

The X axis shows the level of student's ability, the Y axis explains the magnitude of information function. At low level of ability, the measurement information obtained was low. At high level of ability, the measurement information obtained was also low. At medium level of ability, the measurement information obtained was high. The height of information function achieved was also quite high, which was 7.5 on the Y axis. This indicated that the ISLT instrument had a high information value if tested on students with moderate ability and it was feasible as a diagnostic test (Perera, Sumintono, & Jiang, 2018).

DISCUSSION

Education assessment is a process that is inseparable from the education itself owing to the fact that it places the learner in what he/she knows or doesn't know and what they are able to do or not. The standard test needs to be well-designed, and the aspects of validity and reliability are the so essential that those must be fulfilled. The classical test theory (CTT) only emphasizes on the score of an exam which is commonly called as personal ability. That is why the RMM approach is good to be used to provide accurate information about quality of participants and test item (Carvalho, Primi, &

Meyer, 2012; He, Liu, Zeng, & Jia, 2016; C. W. Liu & Wang, 2017).

Data analysis resulted from 36 dichotomy items that conducted by 94 participants, obtained a lot of information about psychometric attributes of the item and person. Based on Table 3 there were four items unfit because the item's MNSQ and ZSTD were not in standard value. The items unfit asked about local wisdom to protect the marine areas and biodiversity (S20), comparison of movement between ancient animals and modern animals based on the homologous principle (S26), important factors influencing metabolic processes of green bean seeds (S31), and main cause of the Arctic ice melting (S34). Item S26 and S31 were categorized as difficult item, because no student was able to do the items. Item S20 was also categorized as difficult item because it was done only by 23.40% of students. Item S34 was done by 84% of them so it was categorized as easy item but still hardly understood by students due to lack of instruction and clarity of the question.

There were five students unfit because the person's MNSQ and ZSTD values were not in the standard value. To verify the items and person fit, the same criteria were used, in which $0.5 < MNSQ < 1.5$; $-2.0 < ZSTD < +2.0$ (Boone & Scantlebury, 2006). The five students unfit were 53PN, 13LN, 16LN, 74LS, and 63LS. Four male students had ability below the level of item's difficulty (minus logit), and one female student was on the average logit (0.25 logit). Those four students were from public school and two students were from private school.

Two items DIF were S15 and S19, whose MNSQ and ZSTD values did not match the standard values. Item S15 asked about the solution for water pollution. Item S19 asked about the right efforts in marine/beach protection. Both items were easier to be done by males than by females, because the test items majorly concentrated on environment (outdoor activity) which is usually more understood by males. S15 and S19 test item were categorized as DIF items, because the probability values of the items were as follows: 0.0282 and 0.0359. The item's probability value < 0.05 which was relevant to (Bond & Fox, 2007) opinion that an item is bias if the item's probability value < 5%.

The information function aims to show what measurement functions are performed. In this study the information function graph obtained shows that the ISLT instrument was intended to provide information about the ability of students in moderate level of abilities. The information function shows the reliability of measurement made because Rasch modeling emphasizes on the separation of coefficient. The higher the peak of the information function achieved, the higher the reliability value of measurements taken (Sumintono & Widhiarso, 2015). Figure 3 concludes that the ISLT instrument was suitable for diagnostic tests and had a high information value if tested on students with moderate ability.

Result from t-test showed that there were differences between public and private schools. The SPSS output indicated that public school had better performance than private schools. Both schools were accredited "A" and located in the same area which is in the rural area. However, public schools were better in preparing the materials than private schools due to a number of guidance for the teaching staff. The SPSS output pointed out that there was no difference in achievement between male and female students. This condition was in view of the fact that male and female students shared the same motivation, time management, confidence, self-testing strategies, and positive competitive attitude (Dabbagh & Khajehpour, 2011).

CONCLUSION

Based on the analysis using the RMM approach, the results revealed in depth and detail explanation about the important processes and analyses related to the assessment of learning. The information obtained out of the research is the item's difficulties and person's abilities, items and person fit, fit models, DIF distractor items, comparison of performance between private and public schools, and comparison of performance between male and female students. This study found of how detailed the RMM approach can be measured when it is implemented on objective measures to

improve educational assessment through a systematic and good assessment process. Furthermore, the results provided more information about the list of levels of student learning with the highest and lowest ability, and unfit person based on the measure score, score of MNSQ outfit and ZSTD outfit. From the level of test items' difficulty given by the teacher to students, some were categorized the most difficult and easiest items and further found the unfit items based on the measured score, score of MNSQ outfit and ZSTD outfit. The Differential Item Functioning (DIF) was also found based on probability scores, MNSQ, and ZSTD scores. The achievement of overall teaching and learning activities between school types and gender can also be obtained based on statistical information. Information test function explains that the instrument used to test students has a high information value when tested on students with moderate abilities.

REFERENCES

Apino, E., & Retnawati, H. (2017). Developing Instructional Design to Improve Mathematical Higher Order Thinking Skills of Students. *Journal of Physics*, 1–8. https://doi.org/10.1088/1742-6596/755/1/011001

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). *Defining Twenty-First Century Skills*. *Assessment and Teaching of 21st Century Skills*. https://doi.org/10.1007/978-94-007-2324-5_2

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch Model: Fundamental Measurement in the Human Sciences. *Journal of Educational Measurement*, *2*, 360. https://doi.org/10.1111/j.1745-3984.2003.tb01103.x

Boone, W. J., & Scantlebury, K. (2006). The Role of Rasch Analysis When Conducting Science Education Research Utilizing Multiple-Choice Tests. *Science Education*, *90*(2), 253–269. https://doi.org/10.1002/sce.20106

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An Assessment of Scientific Literacy. *Journal of Research in Science Teaching*, *46*(8), 865–883. https://doi.org/10.1002/tea.20333

Carvalho, L. de F., Primi, R., & Meyer, G. J. (2012). Application of the Rasch model in measuring personality disorders. *Trends in Psychiatry and Psychotherapy*, *34*(2), 101–109. https://doi.org/10.1590/S2237-60892012000200009

Chan, S. W., Ismail, Z., & Sumintono, B. (2014). A Rasch Model Analysis on Secondary Students' Statistical Reasoning Ability in Descriptive Statistics. *Procedia - Social and Behavioral Sciences*, *129*, 133–139. https://doi.org/10.1016/j.sbspro.2014.03.658

Co-Author. (2013). *Teori Sekor Pada Pengukuran Mental (Scores Theory on Mental Measurements)*. Jakarta: Nagarani Citrayasa.

Connelly, B. S., Warren, R. A., Kim, H., & Di Domenico, S. I. (2016). Development and Validation of Research Scales for the Leadership Multi-rater Assessment of Personality (LMAP). *International Journal of Selection and Assessment*, *24*(4), 362–367. https://doi.org/10.1111/ijsa.12154

Cresswell, J., Schwantner, U., & Waters, C. (2015). *A Review of International Large-Scale Assessment in Education: Assessing Component Skill and Collecting Contextual Data, PISA*. paris. Retrieved from http://dx.doi.org/10.1787/9789264248373-en

Dabbagh, S., & Khajehpour, M. (2011). Gender Differences in Factors Affecting Academic Performance of High School Students. *Procedia - Social and Behavioral Sciences*, *15*, 1040–1045. https://doi.org/10.1016/j.sbspro.2011.03.236

Engelhard, G. (2009). Applied Measurement in Education The Measurement of Writing Ability With a Many-Faceted Rasch Model. *Applied Measurement in Education*, *5*(3), 171–191. https://doi.org/10.1207/s15324818ame0503

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

He, P., Liu, X., Zeng, C., & Jia, M. (2016). Using Rasch Measurement to Validate an Instrument for Measuring the Quality of Classroom Teaching in Secondary Chemistry Lessons. *Chemistri Education Research and Practice*, 381–393. https://doi.org/10.1039/C6RP00004E

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the Growth of International Educational Testing and National Assessment. *Comparative Education Review*, *54*(1), 5–25. https://doi.org/10.1086/648471

Linacre, J. M. (2004). Rasch Model Estimation: Further Topics. *Journal of Applied Measurement*, *5*(1), 95–110.

Liu, C. W., & Wang, W. C. (2017). Parameter Estimation in Rasch Models for Examinee-Selected Items. *Journal of Educational Measurement*, *54*(4), 518–549. https://doi.org/10.1111/jedm.12159

Liu, X. (2009). Beyond Science Literacy: Science and The Public. *International Journal of Environmental and Science Education*, *4*(3), 301–311.

Mcconney, A., Oliver, M. C., Woods-Mcconney, A., Schibeci, R., & Maor, D. (2014). Inquiry, Engagement, and Literacy in Science: A Retrospective, Cross-National Analysis Using PISA 2006. *Science Education*, *98*(6), 963–980. https://doi.org/10.1002/sce.21135

McNamara, T., & Knoch, U. (2012). The Rasch Wars: The Emergence of Rasch Measurement in Language Testing. *Language Testing*, *29*(4), 555–576. https://doi.org/10.1177/0265532211430367

Myford, C. M. (2010). Applied Measurement in Education Investigating Design Features of Descriptive Graphic Rating Scales Investigating Design Features of Descriptive Graphic Rating Scales. *Measurement*, *7347*(April 2011), 37–41. https://doi.org/10.1207/S15324818AME1502

Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments. Technical Report. NCES 2006-0. *National Center for Education Statistics*, 1–176. Retrieved from http://eric.ed.gov/?q=assessment&pr=on&ft=on&ff1=subMathematics+Achievement&pg=3&id=ED491692

Nizam. (2016). *Summary of Learning Assessment Result from National Examination, PISA, TIMSS, and INAP. Seminar Puspendik 2016*.

Olsen, L. (2003). *Essays on Georg Rasch and His Contributions to Statistics*. Copenhagen. Retrieved from http://www.rasch.org/olsen.htm

Olsen, R. V., & Grønmo, L. S. (2004). TIMSS Versus PISA : the Case of Pure and Applied, (October 2013), 1–16.

Perera, C. J., Sumintono, B., & Jiang, N. (2018). The Psychometric Validation of The Principal Practices Questionnaire Based on Item Response Theory. *International Online Journal of Educational Leadership*, *2*(1), 21–38. https://doi.org/10.22452/iojel.vol2no1.3

Provasnik, S., & Malley, L. (2016). *Highlights From TIMSS and TIMSS Advanced 2015*. Washington DC. Retrieved from papers3://publication/uuid/52810E71-38BA-436F-84B2-0708645B317F

Saito, H. (2008). EFL Classroom Peer Assessment: Training Effects on Rating and Commenting, *25*(4), 553–581. https://doi.org/10.1177/0265532208094276

Smith, B., & Barnes, G. V. (2007). Development and Validation of a Orchestra Performance Rating Scale. *Journal of Research in Music Education*, *55*(3), 268–280. https://doi.org/10.2307/3345801

Sondergeld, T. A., & Johnson, C. C. (2014). Using Rasch Measurement for the Development and Use of Affective Assessments in Science Education Research. *Science Education*, *98*(4), 581–613. https://doi.org/10.1002/sce.21118

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch Pada Assessment Pendidikan (Rasch Modeling Application on Educational Assessment)* (1st ed.). Cimahi: Trim Komunikata.

Tjalla, A. (2010). *A Portrait of Indonesian Education Quality in Terms of Results of International Studies. National Seminar FKIP-UTFKIP-UT*, (3), 1–22. Retrieved from http://pustaka.ut.ac.id/pdfartikel/TIG601.pdf

Tohir, M. (2018). The Results of 2015 Indonesia PISA Has Increased. *Reseach Gate*, (January), 2015–2017. Retrieved from https://www.researchgate.net/publication/322420745

Turiman, P., Omar, J., Daud, A. M., & Osman, K. (2012). Fostering the 21st Century Skills through

Scientific Literacy and Science Process Skills. *Procedia - Social and Behavioral Sciences*, *59*, 110–116. https://doi.org/10.1016/j.sbspro.2012.09.253

Van Zile-Tamsen, C. (2017). Using Rasch Analysis to Inform Rating Scale Development. *Research in Higher Education*, *58*(8), 922–933. https://doi.org/10.1007/s11162-017-9448-0

Widana, I. W. (2017). *Module for Preparing Higher Order Thinking Skill (HOTs) Test.* Jakarta: Directorate General of Primary and Secondary Education at the Ministry of Education and Culture.

XiufengLiu. (2012). *Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach*. *Science Education* (Vol. 96, No. 1,). https://doi.org/10.1002/sce.20477