Development of Written Test Instruments to Measure Basic Skills of STEAM *(Science, Technology, Engineering, Arts, and Mathematics)* Junior High School Students on Science Concept

Nada Amira*
Prodi Pendidikan IPA, Sekolah Pascasarjana
Universitas Pendidikan Indonesia
Bandung, Indonesia
*nadaamira@upi.edu

Ana Ratna Wulan
Departemen Pendidikan Biologi
Universitas Pendidikan Indonesia
Bandung, Indonesia

Abstract—**Constructing a high quality of test items needs a substantial time and efforts in developing and examining each test item. Literature suggested this limitation can be remedied with a large collection of test items along with their measurement characteristics which termed as item bank. However, a well-developed item bank is formed by a set of high quality of test items using a rigorous development and validation procedures. In achieving this, this study attempts to describe the development of written test instruments to measure basic skills of STEAM (Science, Technology, Engineering, Arts and Mathematics) in science concept. Accurate measure of students' achievement is very crucial for curriculum and instruction planning and educational-related program evaluation. Therefore, the written test instruments to measure basic cognitive skills of STEAM is urgently needed as it usually only assesses the product or the project which being learned. Research data was analyzed using RASCH Model Analysis with Ministep program which investigate the Wright response map, difficulty level, reliability and discrimination index. The sample consisted of 30 students answering 3 science concept sets of 30 test items each; Energy in Life System (reliability value = 0.83), Pressure in Substance and Plants (reliability value = 0.88) and Volcanoes and Earthquake (reliability value = 0.89). Moreover, these reliability value of the test items are considered as a very good quality, with 75 out of 90 items were stored in the item bank and empirically fulfilled the quality as suggested by the Rasch measurement model.**

Keywords—*development, written test instruments, STEAM basic skills, Rasch Measurement Model, item reliability*

I. INTRODUCTION

Science and technology contribute to economic and social progress and assist the sustainable development of the country, and Natural Science Education has always been the foremost priority in educational planning and development [1]. Natural Science Education is essentially a product, process and application. Nowadays, a growing focus has been to prepare better students for higher education and equip them with the skills and knowledge that they need to become successful innovators in the 21st century work force. So-called 21st century skills, like problem solving ability, communication skills, critical and creative thinking are considered to be essential for successful participants in our society. The integration of creativity still confronts teachers with the pedagogical dilemma of implementing it in the classroom. One option to integrate creativity is expanding STEM subjects (Science, Technology, Engineering, Mathematics) to interdisciplinary STEAM education (STEM with Arts) [2]. In line with this movement, STEAM education emerged within the scientific disciplines. It is nowadays recognized and widely used as an over-arching discipline that connects the various content topics and incorporates the above-mentioned 21st century skills. STEAM has also been conceptualized as a transdisciplinary teaching and learning approach where you start with the "issue or problem and, through the process of problem solving, bring to bear the knowledge of those disciplines that contributes to a solution or resolution [3].

STEAM learning often connected with the project based learning which usually evaluate its product that being made by students. Meanwhile, evaluation in education is not limited to that aspect and there is only few research studied about the STEAM basic skills using the written test. The written test instruments with combined HOTS skills assessment is usually corresponding to the STEAM learning model which is designed to train students level of thinking which composed by the ability to (1) transfer a concept to the other, (2) process and apply information, (3) find a relationship between ideas, (4) use information to solve problems, and (5) explore idea and a concept critically [4]. Therefore, a re-conceptualization of STEAM in terms of developing written test instruments to measure the basic skills of STEAM in science concept is essentially needed in order to be able to reach the above-mentioned goals. The study advocates adding an integrated

STEAM assessment model to see if students can use the various STEAM basic skills and knowledge in an interdisciplinary way [5]. The written test instruments being developed is analyzed using a Rasch model measurements. Rasch model provides equal interval scales which each items define the latent trait in the same manner from time point to time point [6]. Rasch model can analyze the validity, reliability, discrimination value, unidimensionality, response dependency, equity an also detect the bias items using differential item functioning (DIF) [7].

The research objectives are to develop a written test instruments to measure the basic skills of STEAM which being processed its validity and reliability. This written test instruments then being tested to 30 students as the limited trial step of research. This research also aims to describes the quality of the blueprint test and the items developed to measure the basic skills of STEAM.

II. METHODS

This research is off-learning which does not involve the application of STEAM-based learning in the classroom, but students will be given a written test which must be answered and then analyzed the basic STEAM skills possessed by students. The research method used in this research is Research and Development (R&D) defined as a process used to develop and validate products used in education and learning. This so called limited development design is a development model where research is used to design products (written test instruments) and new procedures that are tested in the field, evaluated, validated and refined to meet certain criteria to produce valid and reliable written test instruments. The development of a written test instrument to measure basic STEAM skills is carried out according to the following flow. The written test instrument was used to assess students' knowledge based on the knowledge dimension and cognitive level on the assessment of STEAM basic skills. The dimensions of knowledge in the written test instrument that were developed include factual, conceptual, and procedural knowledge regarding the concept of science. The form of written test that is being developed was 30 items of multiple choice. Science concepts that being developed are Energy in Life System, Pressure in Substance and Plants, and also Volcanoes and Eartquake.

This research and development study was used Rasch Model Measurement approach which was initiated in 1960 by George Rasch who developed an analytical model of item response theory (IRT) which was originally called 1PL (one logistics parameter). 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 [8]. The figure below shows the research procedures used in the research.
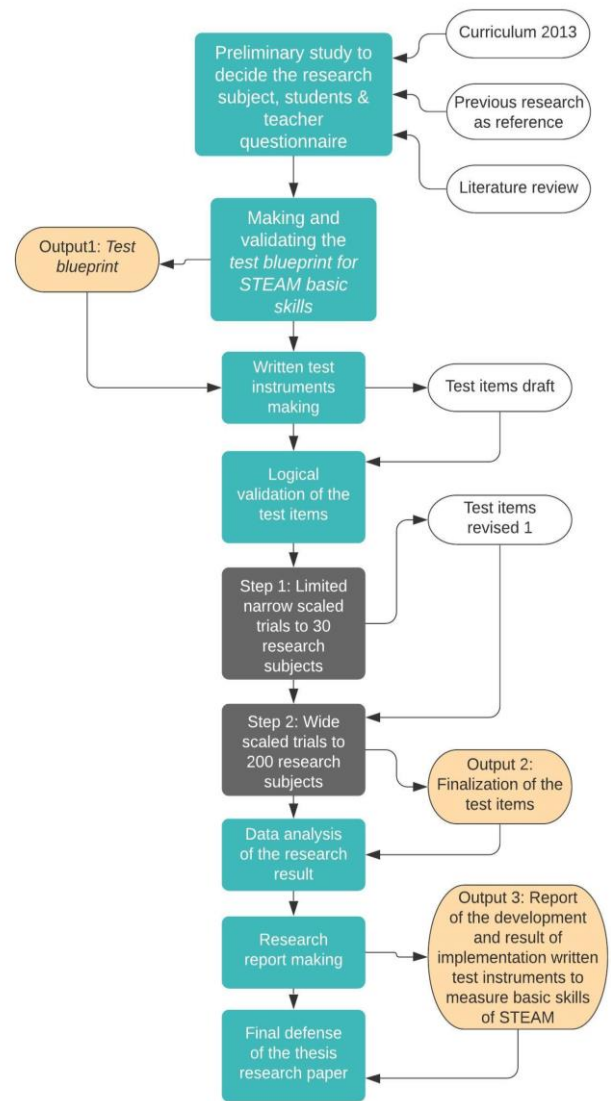

Fig. 1. Research procedure.

The research subjects involved in this study were grade IX from the school that has been selected. The sample was selected using purposive sampling method. Purposive sampling is a technique of taking samples not based on random, region or strata, but based on reasons of consideration that focus on specific research objectives [9]. The consideration of choosing this sampling technique is that a school that meets the following requirements is needed: (1) Has or has not implemented STEAM-based learning, constructivism learning or learning with a scientific approach to the science concept, (2) Students are expected to have the basic STEAM skills mentioned above, (3) The school has science teachers with qualifications and characteristics that are approximately the same as the minimum teaching period of 5 years.

III. RESULTS AND DISCUSSION

National and global interest in STEAM education is growing, and predictive reports suggest the trend will likely continue; however, there is surprisingly little research on the efficacy of instructional approaches. Furthermore, there is a lack of research in determining and measuring essential 21st century skills (e.g., creativity, innovation, communication, collaboration) important for success in STEAM activities. As STEAM education is a newly emerging field in education, there is almost no data available for an international comparison. The same holds true for STEM education. Because there is no generally accepted framework, cross-country comparisons are difficult to make. A first point of discussion is the interpretation of the 'A' in STEAM. A narrow view interprets 'A' as the visual and performing arts. It has been argued that these latter offer a means for conceptualizing, understanding and expressing science and that 'they cultivate a particular kind of ontology, a complex combination of being lost in the moment and utterly present, an experience that lies at the heart of inquiry based science education' [10]. As noted in the introduction, STEAM is considered to be an excellent vehicle to introduce the 21st century skills in education. Because art work is more about questioning and understanding concepts than finding answers to a given problem, it is in essence inquiry-based and as such is analogous with principles of critical thinking [11]. STEAM basic skills are a set of knowledge, skills and cognitive abilities that are associated with each other which then be represented by test items that is being developed. Ann Arbor Public Schools (2014) and OEDC (2016) also suggests basic STEAM skills which include:

- Skills to analyze scientific phenomena (research skills, develop plans and timelines, analyze data and draw conclusions from the results of the analysis).

- Problem solving skills (skills to understand complex systems and diagrams into small parts, process and identify causal relationships, and propose solutions for project progress).

- Mathematical skills (skills to calculate and analyze mathematical data).

- Engineering design skills and technology use (skills in selecting software and technical tools in solving problems / projects).

- Creative thinking skills (skills to solve problems, design new products and ideas and communicate findings effectively).

- Skills to organize data and scientific evidence (skills to collect and differentiate validity of information, record accurate data and complete tasks within the specified time).

Research data was analyzed using RASCH Model Analysis with Ministep program. Item analysis of a test comes after the preliminary draft of a test has been constructed, administered on a sample and scored out. Tabulation is done to determine the following important characteristics of the Wright response map, difficulty level, reliability and discrimination index [12]. There are 3 sets of test instruments with the summary statistics as follow:

A. Energy in Life System

Summary statistics provide overall information about the quality of the overall student response patterns, the quality of the instruments used, as well as the interactions between people and items. Test item reliability for the concept of Energy in Life System is 0.86 which concluded as a very good quality. Meanwhile, the average value of INFIT MNSQ and OUTFIT MNSQ in the item table are 0.98 and 1.07, respectively, while the ideal value is 1.00 (the closer to the value 1.00, the better). The value of INFIT ZSTD and OUTFIT ZSTD values are 0.02 and 0.21, in this case the ideal value is 0.0 (the closer to the value 0.00, the better the quality). The Cronbach alpha value aimed to measure the reliability, which is the interaction between the person and the item as a whole [8]. Based on the data, the reliability value is 0.86 showing that the reliability was very good. The summary statistics table is represented as follow Table 1.

TABLE I. SUMMARY STATISTICS (ENERGY IN LIFE SYSTEM)

	Total score	Count	Measure	Infit		Outfit	
				MNSQ	*ZSTD*	*MNSQ*	*ZSTD*
MEAN	20.0	30.0	.00	.51	.02	1.07	.21
SEM	2.0	.0	.40	.02	.14	.09	.18
P.SD	7.3	.0	1.44	.09	.49	.34	.66
S.SD	7.6	.0	1.49	.09	.51	.36	.68
MAX.	28.0	30.0	2.83	.74	1.27	1.94	1.29
MIN.	6.0	30.0	-1.94	.41	-.78	.59	-.64
ITEM RELIABILITY			.86	SEPARATION		2.53	

The grouping of persons and items can be seen from the value of separation mentioned in table 1. The greater the separation value, the better the quality of the instrument in terms of overall respondents and items, because it can identify respondent groups and item groups. The separation value is 2.53. The equation used is:

$$H = \frac{[(4 \; x \; SEPARATION) + 1)]}{3} \qquad (1)$$

$$H = \frac{[(4 \; x \; 2.53) + 1)]}{3}$$

$$H = \frac{[(10.12) + 1)]}{3}$$

$$H = 3.70$$

The equation (1) resulted value of 3.70 can be concluded as 4, which means that there are four groups of questions that can be interpreted as difficult, medium, easy, and very easy questions. Examples of science questions on the concept of Energy in Life System from the Written Test Instruments that is being developed (Figure 2):

Fig. 2. Examples of Test Instruments based STEAM.

B. Pressure in Substance and Plants

As for the question set of Pressure in Substance and Plants concept has item reliability 0.82 which means the quality of the items in the instrument concluded as a good reliability. Meanwhile, the INFIT MNSQ and OUTFIT MNSQ average values are 0.99 and 0.97, respectively, while the ideal value is 1.00 (the closer to 1.00, the better). The value on MNSQ Outfit and Infit means that this instrument is a good productive measure to assess the basic STEAM skills. As for the INFIT ZSTD and OUTFIT ZSTD values are 0.06 and 0.11, in this case the ideal value is 0.0 (the closer to 0.00, the better the quality). The summary statistics table is represented as follow Table 2.

TABLE II. SUMMARY STATISTICS (PRESSURE IN SUBSTANCES AND PLANTS)

	Total score	Count	Measure	Infit		Outfit	
				MNSQ	*ZSTD*	*MNSQ*	*ZSTD*
MEAN	16.5	30.0	.00	.44	.99	0.06	.11
SEM	1.5	.0	.29	.02	.04	.23	.27
P.SD	5.5	.0	1.05	.07	.14	.84	.97
S.SD	5.7	.0	1.09	.07	.14	.88	1.01
MAX.	27.0	30.0	2.11	.62	1.18	1.17	2.33
MIN.	6.0	30.0	-2.16	.40	-.79	-1.18	-.92
ITEM RELIABILITY			.81	SEPARATION		2.04	

Based on the overall Rasch model analysis mentioned in table 2 above, some questions need to be deleted while some need to be revised. At the end, 5 of 30 questions need to be deleted because of its unacceptable value of Infit and Outfit MNSQ and ZSTD as well as the Point Measure Correlation value.

C. *Volcanoes and Earthquake*

This question set has item reliability 0.89 which means that the quality of the items in this instrument is a good reliability aspect. Meanwhile the average values of INFIT MNSQ and OUTFIT MNSQ are 1.00 and 0.94 respectively, while the ideal value is 1.00 (the closer to the value 1.00, the better). While INFIT ZSTD and OUTFIT ZSTD the average values are 0.05 and 0.00, in this case the ideal value is 0.00 (the closer to the value 0.00, the better the quality) (Table 3).

TABLE III.   SUMMARY STATISTICS (VOLCANOES AND EARTHQUAKE)

	Total score	Count	Measure	Infit		Outfit	
				MNSQ	*ZSTD*	*MNSQ*	*ZSTD*
MEAN	19.1	30.0	.00	1.00	.05	.94	.00
SEM	2.0	.0	.46	.05	.26	.11	.26
P.SD	7.4	.0	1.70	.20	.99	.43	.97
S.SD	7.7	.0	1.76	.20	1.02	.44	1.00
MAX.	26.0	30.0	4.38	1.44	2.32	2.03	2.11
MIN.	3.0	30.0	-1.58	.64	-2.39	.43	-2.16
ITEM RELIABILITY			.89	SEPARATION		2.82	

According Rasch Model Analysis in overall question sets, there are some questions which can be answered correctly by all respondents. For example, on the question set of Energy in Life System particularly on the question number S2 about wind power plant which assess the STEAM basic skills of analyze scientific phenomena (applying the scientific phenomena in a real life). As well as the questions number S28 about photosynthesis practical activity which represent the skills to organize data and scientific evidence. Then, as for the question set of Pressure in Substances and Plants, question number S9 about the relationship of the height and pressure can be answered perfectly by all respondents. This question represents the skills of problem solving especially in identify and evaluate data reliability, objectivity and generalization of problem solving. Each of the subjects in STEAM focuses on problem solving in different ways [13]. For example, engineering and art are similar in their design processes: identify a problem, brainstorm possible solutions, prototype, test, and improve based on data or feedback. The question number S16 about the Pressure in Substance project about making a simple water dispenser can be answered by all respondents. This question is signifying the engineering design skills and technology use (skills in selecting software and technical tools in solving problems / projects). Engineering design has been treated as a pedagogical strategy to bridge science and mathematics concepts in use of solving ill-defined (open-ended) problems, developing creative thinking, formulating solutions and making decisions, and considering alternative solutions to meet a variety of constrains [14]. In the future, assessment for integrated learning must prioritize not only the aspects of learning objectives but also a valid format and type of assessment that can reflect the students' basic STEAM skills, while this has not been fully realized in current STEAM learning. So hopefully, this research can be at least beneficial in providing the framework of written test instruments to measure the basic skills of STEAM.

IV. CONCLUSION

STEAM has also been conceptualized as a transdisciplinary teaching and learning approach where you start with the "issue or problem and, through the process of problem solving, bring to bear the knowledge of those disciplines that contributes to a solution or resolution [3]. According to Rasch Model Analysis stated above, Energy in Life System (reliability value = 0.86), Pressure in Substance and Plants (reliability value = 0.81) and Volcanoes and Earthquake (reliability value = 0.89) considered as a very good quality of test items instruments with very good value of Infit and Outfit MNSQ and ZSTD. This means that the test items have the very good qualification of item fit, item measure, item validity, item difficulty rate, and item discrimination. In conclusion, 75 out of 90 items were stored in the item bank and empirically fulfilled the quality as suggested by the Rasch measurement model. The Rasch Measurement Analysis works as the reference to revised the test items, which will be used in the next process of research that is wide-scale trials with 200 respondents for a better and comprehensive result. Also, having the test items more concrete can make it easier to show the student how things are going, and helps adjust learning objectives and outcomes of STEAM curriculum.

REFERENCES

[1] I.M. Saleh and M.S. Khine, Fostering scientific habits of mind: Pedagogical knowledge and best practices in science education. Sense Publishers, 2009.

[2] C. Conradty, S.A. Sotiriou, and F.X. Bogner, "How Creativity in STEAM Modules Intervenes with Self-Efficacy and Motivation," Educ. Sci., vol. 10, no. 3, p. 70, 2020.

[3] D. Herro, C. Quigley, J. Andrews, and G. Delacruz, "Co-Measure: developing an assessment for student collaboration in STEAM activities," Int. J. STEM Educ., vol. 4, no. 1, p. 26, 2017.

[4] U. Rosidin, A. Suyatna, and A. Abdurrahman, "A combined hots-based assessment/stem learning model to improve secondary students' thinking skills: a development and evaluation study," J. Educ. Gift. Young Sci., vol. 7, no. 2, pp. 435–448, 2019.

[5] A. Bicer, R.M. Capraro, and M.M. Capraro, "Integrated STEM assessment model," EURASIA J. Math. Sci. Technol. Educ., vol. 13, no. 7, pp. 3959–3968, 2017.

[6] W.J. Boone and K. Scantlebury, "The role of Rasch analysis when conducting science education research utilizing multiple-choice tests," Sci. Educ., vol. 90, no. 2, pp. 253–269, 2006.

[7] A. Tennant and P.G. Conaghan, "The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?," Arthritis Care Res. (Hoboken)., vol. 57, no. 8, pp. 1358–1362, 2007.

[8] R.U. Sihombing, D.S. Naga, and W. Rahayu, "A Rasch Model Measurement Analysis On Science Literacy Test Of Indonesian Students: Smart Way To Improve The Learning Assessment," IJER-INDONESIAN J. Educ. Rev., vol. 6, no. 1, 2019.

[9] A. Suharsimi, "Dasar-dasar evaluasi pendidikan edisi 2," Jakarta PT Bumi Aksara, 2013.

[10] T. Barone and L. Bresler, "International Journal of Education and the Arts," Res. Stud. Music Educ., vol. 15, no. 1, p. 71, 2000.

[11] S. Ghanbari, "Learning across disciplines: A collective case study of two university programs that integrate the arts with STEM.," Int. J. Educ. Arts, vol. 16, no. 7, 2015.

[12] C. Boopathiraj and K. Chellamani, "Analysis of test items on difficulty level and discrimination index in the test for research in education," Int. J. Soc. Sci. Interdiscip. Res., vol. 2, no. 2, pp. 189–193, 2013.

[13] J.W. Bequette and M.B. Bequette, "A place for art and design education in the STEM conversation," Art Educ., vol. 65, no. 2, pp. 40–47, 2012.

[14] N.J. Kim, B.R. Belland, and D. Axelrod, "Scaffolding for optimal challenge in K–12 problem-based learning.," Interdiscip. J. Probl. Learn., 2018.