5th ICRIEMS Proceedings
Published by Faculty Of Mathematics And Natural Sciences
Yogyakarta State University, ISBN 978-602-74529-3-0

A Training Model for Pre-Service Science Teacher to Develop The Competency of Test Instrument Arrangement Based on International Mapping and Benchmarking

Dadan Rosana[1,a)], Eko Widodo[2,b)], Didik Setyawarno[3,c)], and Wita Setianingsih [4,d)]

[1,2,3,4] Science Education Study Program, Faculty of Mathematics and Science
Yogyakarta State University
Jl Kolombo No 1, Karangmalang, Depok, Sleman, Yogyakarta, Indonesia

[a)]Corresponding author: danrosana@uny.ac.id
[b)]eko_widodo@uny.ac.id
[c)]didiksetyawarno@uny.ac.id
[d)]wita@uny.ac.id

Abstract. The standards of learning process for pre-service science teachers in Indonesia need to be improved, especially related to the international benchmarking surveys. One of surveys is the Program for International Student Assessment (PISA) which measures the aspect of what the students know and what they can do (applications) with their knowledge. This study aimed at producing a training model for pre-service science teachers in developing an assessment with the standard of international benchmarking survey in this case PISA in order to be able to compete in global market. The subject of this research was the students of pre-service teacher in the Study Program of Science Education, Faculty of Mathematics and Science, UNY. The research method was using Research and Development (Thiagarajan, S., et al, 1974) with Four-D Models, which was modified through Barg & Gall's R & D model (1983). The research stages included Define; Design; Develop and Disseminate. The results of this study are (1) the valid pre-service training model to develop the competency of test instrument arrangement using the standard of international mapping and benchmarking with the V'Aikens coefficient of 0.91-0.97, where the inter-rater reliability can be categorized as "very good", (2) the effectiveness results of the pre-service training model implementation in developing the competency of test instrument arrangement with the standard of international mapping and benchmarking, and (3) the practicability level of the model, which can be categorized as "very good" according to the lecturer and the student.

Keyword: *pre-service training model, pre-service science teachers, test with international mapping standard*

INTRODUCTION

The era of disruption 4.0 occurs when the movement of industrial world or work competition is no longer linear. The change is very fast and fundamental which substituting the old patterns to create a new order. It overturns the systems that already exist since ten or even hundreds of years and replaced with a new system. The system is driven by innovative and creative young generations with their digital literacy.

This era is like a double-edged sword which has positive and negative impact. For example, the change in the international world directly influence most of countries. It give effect to the social, politic, even mental and nation. It is a real challenge for education to prepare the innovative and creative young people. Therefore, it is very important to develop science literacy and high-order thinking skills, especially related to international benchmarking surveys such as Program for International Student Assessment (PISA).

PISA is a test system organized by the Organization for Economic Cooperation and Development (OECD) [1}, to evaluate the education system of 72 countries around the world. Every three years, a 15-year-old student is randomly selected, to take the tests of three basic competencies, namely reading, math and science. The test measures what students know and what they can do (applications) with their knowledge. The theme of the survey is changed every 3 years and in 2015 the focus is on the competence of sciences.

The Indonesian government begin to give serious attention to international surveys or mapping as it relates to the nation's competitiveness in the global era. Head of Research and Development Board, Ministey of Education and Culture, Totok Suprayitno, says that the improvement of Indonesia's achievements in 2015 is well-enough though the results are still below the OECD mean. Based on the mean score, there is an increase in the score of PISA Indonesia in the three tested competences, especially in science competence, from 382 points in 2012 become 403 in 2015, the mathematics competence increase from 375 into 386. Meanwhile, reading competence has not shown significant increase, i.e 396 to be 397. It elevates Indonesia's position in 6th place compared to the second-ranked of the lowest in 2012 [2].

Moreover, based on the median score, the achievement of Indonesian student on reading is getting higher, from 337 in 2012 to be 350 in 2015. The mathematical score raise 17 points from 318 into 335. The highest improvement is in the science area which increase from 327 into 359. This higher median comparing to the mean can become a good indicator to improve the access and the quality distrubution inclusively [2].

Further, Head of Educational Assessment Center of Research and Development Board, Ministry of Education and Culture says that there is a consistent increase on the sampling coverage of Indonesian students, i.e. 46 percent in 2003 to 53 percent in 2006. Furthermore, the score rise from 63.4 percent in 2012, and became 68.2 percent in 2015. "Increasing the coverage of this sampling is an evidence that the program of 9-year compulsory education and the expansion towards a compulsory learning of 12 years as well as the inclusion of Indonesian student participation in education is fruitful," as he said in Jakarta on Tuesday [3].

The most important thing of these international benchmarking surveys, such as PISA, is the information that can be followed-up based on the diagnoses from the survey. The achievements must be gradually improved through the enhancement of the education quality in Indonesia. If the increase rate in 2012-2015 can be maintained, then, by 2030 our achievement will be equal to the average achievement of OECD countries. Therefore, it is important to include PISA assessment in the learning process, especially for pre-service teacher program becasue the quality standard of sciences teachers in Indonesia need to be improved, particularly related to international benchmarking surveys.

Based on the above problem analysis, the purpose of this research is to improve the competence of professional pre-service teacher of sciences field in the assessment development of international standard benchmarking for global competitiveness. The strategic target is the students of pre-service teachers in the Institute of Teacher Education (LPTK).

RESEARCH METHOD

The research was using Research and Development method, as the research flow illustrated in Figure 1. The phase of "define" or "research and information collectio"n [4] was the initial research and data collection through literature study, needs analysis and field study. The design or planning phase was the product design including the aim of the product use, the product user and the description of the product components. The stage of develop or develop preliminary form of product was an early product development. The disseminate phase had four developmental steps, namely preliminary field testing which were initial field trials, main product revision or test results revision, main field testing or field trials and operational product revisions [4] or refinement of field test results.

Portfolio documentation techniques were used to collect data related to the Research implementation, such as test guidelines, test materials, answer keys, and student responses as the research sample [5]. The response in this study was obtained after the students worked on a set of international benchmarking PISA survey instruments containing various test item, i.e. multiple choice, matching, essay, and other types. This instrument was made by a collaborative team of researchers and students.

The PISA international benchmarking survey instrument that had been arranged was, then, validated to make sure the instrument can actually describe the aspect being measured [6]. The items were made based on the distributed guideline which was proportional based on the description of the listed material in the curriculum, so that the content validity or theoretical validity is eligible. The coefficient of the content validity in this research was processed based on the given score from expert judgment. After that, the judgment results were computed using the Aiken formula [7]:

$$V = \frac{\sum s}{[n(c-1)]}$$

s = r – lo
lo = the lowest validity score
c = the highest validity score
r = the score from the expert

Four ratings categories were used, namely "irrelevant", "less relevant", "relevant", and "highly relevant"" and the Aiken index should be in the score of 0.87 (α = 0.05) or 0.93 (α = 0.01) [7]. However, according to [8], the validity coefficient around 0.7 is still acceptable and considered satisfactory. Based on the analysis with aiken formula, it was obtained 0.935 for index average of the content validity [11]. The content validity for the instruments of the international benchmarking survey of PISA increased from 0.867 into 1. Thus, it can be concluded that the items of the test instrument are valid.

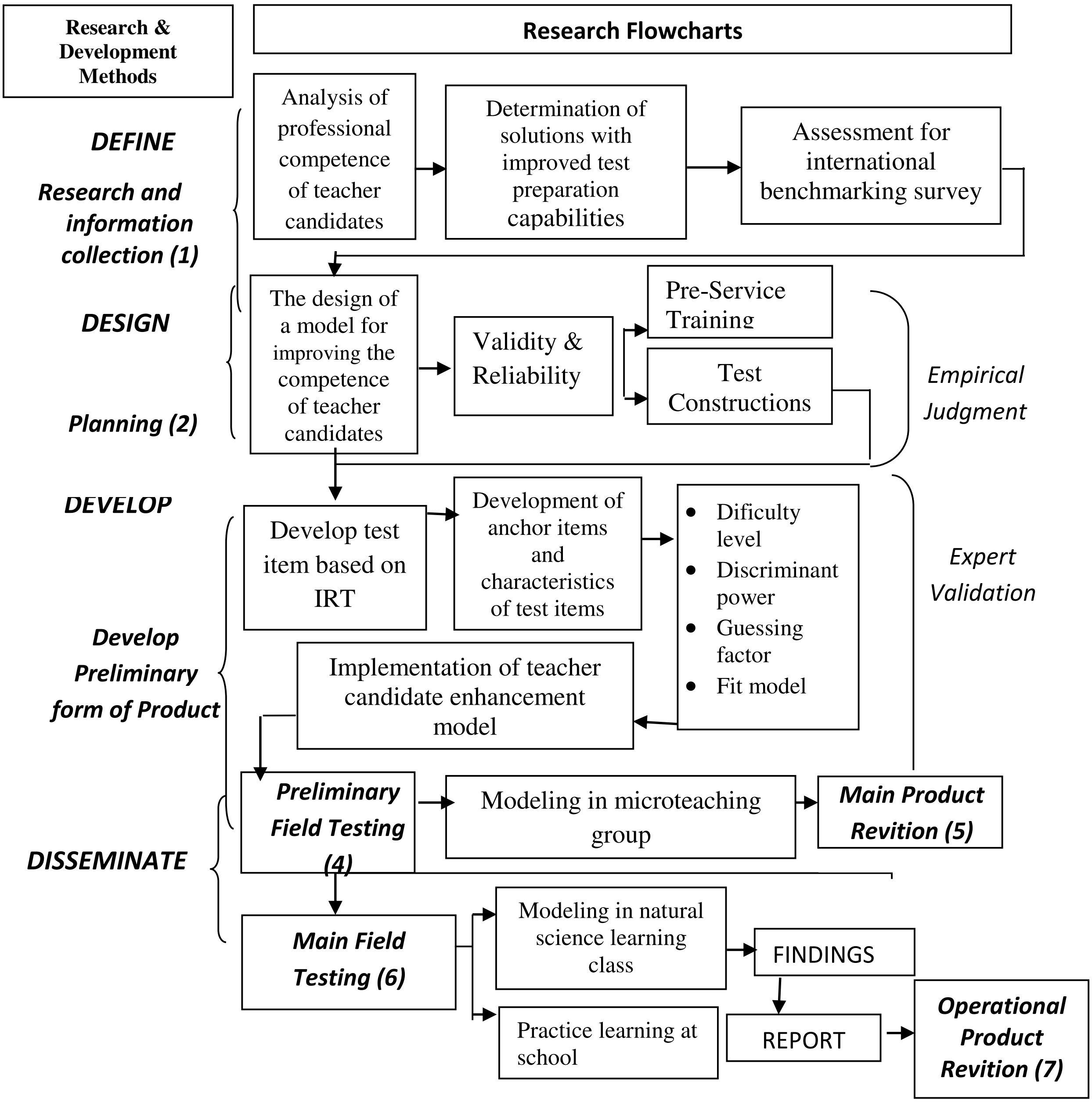
Research & Development Methods
Research Flowcharts
DEFINE
Research and information collection (1)
DESIGN
Planning (2)
DEVELOP
Develop Preliminary form of Product
DISSEMINATE
Analysis of professional competence of teacher candidates
Determination of solutions with improved test preparation capabilities
Assessment for international benchmarking survey
The design of a model for improving the competence of teacher candidates
Validity & Reliability
Pre-Service Training
Test Constructions
Empirical Judgment
Develop test item based on IRT
Development of anchor items and characteristics of test items
Dificulty level
Discriminant power
Guessing factor
Fit model
Expert Validation
Implementation of teacher candidate enhancement model
Preliminary Field Testing (4)
Modeling in microteaching group
Main Product Revition (5)
Main Field Testing (6)
Modeling in natural science learning class
Practice learning at school
FINDINGS
REPORT
Operational Product Revition (7)

FIGURE 2. Research Desgn

In the initial field trials and limited trials of PISA international benchmarking survey instrument for integrated science learning, the research subjects involved 6th semester students of Sciences Education Study Program in the year of 2018 who joined microteaching course and the samples were 2 classes , i.e. A and I class. Each class was devided into experimental groups and control group for 5 different subjects. In the microteaching class, the two classes were divided into groups with 10 members of each group that were taught by two lecturers. The experimental design during preliminary field testing was as follows.

TABLE 1. Preliminary field testing design

Class	Pre-test	5 Initial Main Topics	Post-test (1)	5 Following Main Topics	Post-test (2)
A	T_1	X	T_2	X	T_3
I	T_1	X	T_2	X	T_3

RESULT AND DISCUSSION

The basic development of pre-service trainning model is the ability to make sciences literacy assessment in PISA which contains knowledge within the curriculum and cross-curriculum. Moreover, the measured scientific literacy aspects are as follows: using knowledge and identifying problems to understand facts, making decisions about nature and changes that occur in the environment. The questions of PISA really requires reasoning and problem-solving abilities. A student is considered to be able to solve problems if he/she can apply their acquired knowledge previously to the new unknown situations. In the PISA test intens, there are eight characteristics of cognitive ability, such as (1) thinking and reasoning, (2) argumentation, (3) communication, (4) modeling, (5) problem posing and solving, (6) representation, using symbolic, (7) formal and technical language and operations, (8)) use of aids and tools

Those eight cognitive characteristics are really matched to the learning objectives of sciences based on the curriculum. It means that PISA problem not only demand the concept application but also how the concept can be applied in various situations, as well as the students' ability of reasoning and arguing in solving a problem.

The PISA Framework of sciences is based on three dimensions: (1) the content (2) the process, from phenomenon observation, connecting the phenomenon with sciences, till solving the problem being observed; and (3) the situations and contexts, as shown in the picture below:

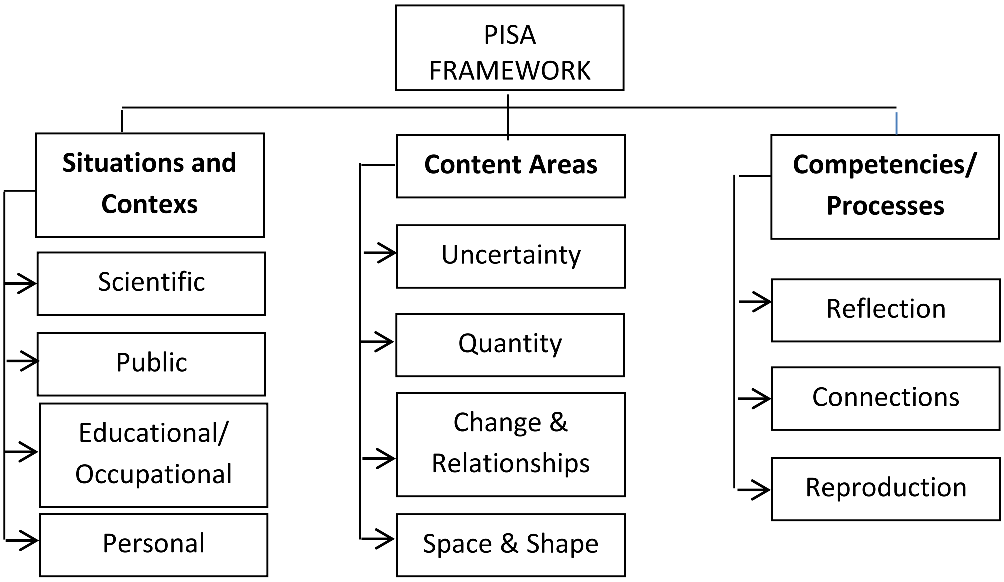

FIGURE 2. PISA IPA *Framework*

The research results on “define” or “research and information collection” stage [4] begin by analyzing the sciences PISA Framework in the high order thinking skills domain. [9] and [10] defines high order thingking as the use of complex, nonalgorithmic thinking to solve a Task, in which there is not a predictable, well-rehearsed approach or pathway explicitly suggested by the task instruction, or a worked out example. According to Stein, high-level thinking uses complex, non-algorithmic thinking to accomplish a task, some unpredictable, using different approaches for the existing and different tasks from previous examples.

Furthermore, the "design or planning" phase [4] covers a product design that can be produced including the purpose of product use in the form of pre-service training model, product users and description of product components. In this stage, an international standar-benchmarking assessment is prepared based on the specified format in the instrument specification stage, as set out in Table 1.

TABLE 1. Items included in pilot instruments by response types

	Number of Item	%
Essay - Constructed Response	10	20
Essay - Reasonning	10	20
Simple Multiple Choice	10	20
Matching	10	20
True-False	10	20
Total	50	100

The assessment rubric is using the range of 1 to 4 for each indicator. The next step describes each indicator with 4 statements. The statement is adjusted to the aspect and indicator that has been set. The result of this stage is the initial design of the international standard benchmarking assessment. The items are created to fulfill the content validation to make sure that the statement can really measure the indicator of the manipulative skills.

The "develop" or "develop preliminary form of products" [4] is an initial product development. In this stage, the instrument validation of pre-service training model is in the form of assessment draft on international standard benchmarking survey by a lecturer of material expert, a lecturer of assessment expert, and four sciences teachers of junior high school. The content validation stage is performed to determine the feasibility of the instrument related to the appropriatness of the statement with the indicator, the language use in case of communicative aspect. The validators use the sheets in evaluating the developed tool. The sheet itself constitutes the aspects of substance, language and construction.

The obtained scores from all validators for each statement are analyzed using Aiken's formula to calculate the content validity coefficient (V) for each statement. Validator ratings are converted into 4 categories: valid categories without revisions, valid with a little revision, valid with multiple revisions and invalid [12]. The validator input is used to revise the assessment instrument. The V'aiken coefficient is formulated into the mean score to be confirmed with the numbers limit based on table of Aiken's V for the number of categories ranges of 4 with 6 assessors, i.e. 0.78. The result of the validation of the international standard benchmarking assessment with V de Aiken coefficient is 0.91-0.97, while International standard benchmarking assessment obtains the assessment results above the minimum criteria that means the developed assessment is declared valid to be used in the study. However, a valid assessment instrument cannot be considered feasible if it is not reliable so its reliability is clarified through product trials, in preliminary field testing.

The result of item analysis shows that the level of item difficulty in the international standard benchmarking survey begin from 0.189 to 0.889 with the mean score of 0.623. The average level of difficulty categorized as good because according to [12], [13] and [14] for multiple choice with five alternative answers, the optimal difficulty level is 0.59 . Furthermore, referring to [15] criteria, the categorization of the difficulty degree in each items is as follows.

TABLE 2. The categorization of the item difficulty level

Catagory	Persentage (%)	Item number
Easy	30 (15 items)	3,4,8,9,14,20, 25,27,32,33,38,39,44, 49,45
Moderate	42 (21 items)	5, 7,10,13,17,18,19, 22,23,24, 28,29,30,34, 35,37,40, 42,43, 47,48
Difficult	28 (14 items)	1,2, 6,11,12, 15, 16,21,26,31,36,41,46, 50

The discriminating power of international benchmarking standard-assessment ranges from 0.148 to 0.592 with the mean score of of 0.376. The analysis result for discriminating power of an item use classical approach (biserial point correlation) indicating that there are 5 items (10%) that can not fully distinguish the ability of pre-service students. This is because those ten items have discriminating power index below the referred criteria i.e. 0.3 ([16], [17]).

The "disseminate" phase in this stage is only 2 stages of the four-step development. They are the preliminary field testing [4] as initial field trial as well as the main product revision or test results revision. Meanwhile, the stages that have not been done yet are the main field testing or field trials and the operational product revision [4] or product improvement based on field test results.

The obtained data in this study is the material mastery in the form of the assessment results on the final and initial assessment of international benchmarking survey standards based on the results of pre-test and post-tes. The pretest is given to the student before the benchmarking benchmarking standard assessment development to determine the students' initial mastery of the student material while the post-test is held after the instrument development study in order to know the student mastery of materials after haing treatment. Below is the results description for each data.

The initial data of the students' science literacy ability and high order thinking skills can be known through pre-test. It consists of 50 items which is given to the experimental group and control group. In summary, the preliminary student ability can be seen in Table 3.

TABLE 3. Parameter Data of Students' Pre-test for international benchmarking standard

Variable	Score			
	Highest	Lowest	Mean	Std. deviation
Control Class	48	24	38,5	4,32
Experiment Class	51	22	39,2	4,26

The final assessment data of international student benchmarking survey is obtained from post-test. It is employed to the control and experimental group. The test item is similar to the pre-test but the item order is set random. To sum, the data is presented in Table 4.

TABLE 4. Parameter Data of Students' Post-test for international benchmarking standard

Variable	Score			
	Highest	Highest	Highest	Highest
Control Class	82	52	64	5,72
Experiment Class	90	66	76	6,31

Hypothesis testing is done by using manova test and, based on the above analysis, the data has been known to be normal distribution, homogeneous and independent. Hypothesis testing is done on science literacy data and procedural ability. Based on the calculation resultan, it can be seen that the F test is significant at α 5% so it is not equal to 0. It means the ability of science literacy and high order thinking skills getting influence from the model availability for autonomous learning in case of the development of international standard benchmarking assessment. Having known that the multivariate test is signifikan, Then, the univariate test F is implemented.

TABEL 5. Levene,s test of equality of error variance

Test Type	F	df1	df2	Sig.
science literacy	3.421	1	56	.059
High Order Thinking	3.236	1	56	.062

Based on the table, it can be seen that the significance score between the ability of science literacy and high order thinking skills is not similar, i.e. the significance of science literacy of 0.059 and high order thinking was 0.062. In addition, the manova test can be seen in the following table.

TABLE 5. Test of between-subjecs effect

Source	Dependent Variable	Type III Sum of Squares	Df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	Science literacy	.089[a]	1	.089	5.162	.025	.081
	High Order Thinking	6346.321[b]	1	6346.321	164.532	.000	.762
Intercept	Science literacy	15.364	1	15.364	832.864	.000	.898
	High Order Thinking	399616.814	1	399616.814	1.131E4	.000	.986
Perlakuan	Science literacy	.094	1	.094	5.221	.028	.084
	High Order Thinking	6311.038	1	6311.038	162.348	.001	.762
Error	Science literacy	1.024	56	.019			
	High Order Thinking	2098.322	56	39.072			
Total	Science literacy	16.421	57				
	High Order Thinking	425073.023	57				
Corrected Total	Science literacy	1.216	56				
	High Order Thinking	8468.499	56				

Manova analysis is conducted to know whether the independent variable influences the dependent variable. It can be revealed from the corrected models and the treatments. Based on the table, both present the same F-test information. The result of F univariate test shows that it has significance level which less than 0.05, it indicates the model use influences the ability of scientific literacy and procedural ability. Partial Eta Square (PES) scores of science literacy and high order thinking are 0.081 and 0.762, respectively. This means that model usage affect science literacy by 8.1% and by 76.2% for high order thinking [18].

From the analysis result, it can be seen that learning with the model is able to influence the ability of science literacy and high order thinking skills (8.1%) and high order thinking (76.2%). The results of the analysis explain that the students involvement in learning by applying the learning model as an indicator of the learning effectiveness. The students do not only receive the materials from lecturers, but students also try to gain knowladge and to develop themselves. Therefore, learning outcomes is not just about score but it can truly increase the students' science literacy and high order thinking skills.

The students' ability need to be trained by working with the test item of international standard benchmarking assessment, so that the application of learning model can be optimal. It requires not only hard-skills but also soft-skills for hard work and smart work in groups. This is in line with [19], that student soft-skills can be improved by context-based learning, for example the application of procedural knowledge in sciences learning.

That ability can be seen clearly when the students are able to finish the test individually or in groups. It can be detected through their ability to explain the exercise completion [20]. In accordance with the principle of learning model, it urges students to be active during the learning process because it requires students to make their own questions and answers based on the given questions by the lecturer through stimulus in the form of pictures, stories, diagrams, etc.

The students are also trying to gain knowladge and develop themself and by applying the learning model, it can encourage students to do their best. Through application of learning with the model, the students are not only enthusiastic in doing assessment based on international benchmarking surveys, but also train them to learn in groups. The international standard benchmarking assessment make students to convey ideas, ideas, opinions. Purwoko, etc. [21] show that the frequency of student involvement in learning in line with the improvement of teachers' competency. In addition, the students also learn to appreciate ideas, and opinions from others.

CONCLUSION

Based on the description of research findings and discussion above, it can be concluded that the developed pre-service trainning model has been able to improve the professional competence of pre-service science teacher in the development of international standard benchmarking survey (PISA) assessment. The indicators include (1) the valid pre- service trainning model to develop the competence of test instrument arrangement based on the standard international benchmarking with V'Aikens coefficient of 0,91-0,97, and inter-rater reliability acieve the category of excellent, (2) the effectiveness of model application performed in good category, and (3) the practicality level can be classified as very good according to lecturer and student, and (4) the model of pre-service teachers competence improvement in developing assessment with international benchmarking standard by employing the pre-service trainning model has not shown significant improvement, therefore it needs further treatment.

ACKNOWLEDGMENT

Our gratitude goes to Universitas Negeri Yogyakarta, especially Faculty of Mathematics and Sciences which has provided opportunities as well as financial and managerial support for this research implementation.

REFERENCES

1. OECD , PISA 2015 Results (Volume III): Students’ Well-Being, OECD Publishing (2017), Paris, http://dx.doi.org/10.1787/ 9789264273856-en.pp: 12-15
2. OECD, PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy, OECD Publishing (2013), http://dx.doi.org/10.1787/9789264190511-en, pp: 123-128
3. Biro Komunikasi dan Layanan Masyarakat Kementerian Pendidikan dan Kebudayaan. https://www.kemdikbud.go.id/main/blog/2016/12/peringkat-dan-capaian-pisa-indonesia-mengalami-peningkatan (6 Desember 2016)
4. Borg, W. R. and Gall, M. D. Educational Research An Introduction. New York: Longman. (1983).
5. *F. Kurnia, D. Rosana,* Supahar. Developing evaluation instrument based on CIPP Models on the Implementation of Portfolio Assessment. *AIP Conference Proceedings* 1868, 080003; doi:10.1063/1.4995187. aip.scitation.org/doi/abs/10.1063/1.4995187 (2017).
6. *L.R.* Aiken. Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, 40, 955-959. (1980)
7. *L.R.* Aiken. Three Coefficients for Analyzing the Reliability, and Validity of Ratings. *Educational and Psychological Measurement*, 45, 131-142. (1985).
8. Sireci, G .Stephen & Geisinger, F.Kurt . Using subject-matter experts to assess content representation: An MDS analysis. *Applied Psychological Measurement*, 19, 241-255. doi:10.1177/014662169501900303. (1995).
9. M.K. Stein & S. Lane. Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. Educational Research and Evaluation, 2, 50-80. (1996)
10. T. Thompson. Mathematics teachers’ interpretation of higher-order thinking in bloom’s taxonomy. International Electronic Journal of Mathematics Education. Volume 3, Number 2, July (2008).
11. M. R. Lynn. Determination and quantification of content validity. *Nursing Research, 35*(6), 382-385.http://dx.doi.org/10.1097/00006199-198611000-00017. (1986).
12. L. M.Crocker. & J. Algina. *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston. pp: 313 (1986)
13. Ridho, Ali. Karakteristik Psikometrik Tes Berdasarkan Pendekatan Teori Tes Klasik dan Teori Respon Aitem. *Jurnal Insan Media.* II (2): 1-28. (2007).
14. B. Wright &J. Linacre. "Combining and Splitting Categories". Rasch Masurement Transactions, 6, 233-235. (1992).
15. J.M. *Allen* & W. M. *Yen*. “Introduction to Measurement Theory” ... Universitas Muhammadiyah. Surakarta. Singarimbun, Masri dan Masri Effendi, Sofian, (eds.). (*1989).*
16. C.R.Reynold, R. B.Livingstone, & V. Wilson. *Measurement and Assesment in Education*: Second Edition. London: Pearson Education. (2010)
17. B. Kartowagiran. Kinerja Guru Profesional Pasca Sertifikasi. Jurnal Pendidikan. (2011). (online) (http//www.uny.co.id, 1 April 2017).
18. Holland, P. W., & Dorans, N. J.. *Linking and equating*. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 187-220).Westport, CT: Praeger.(2006)

19. D. Rosana, Jumadi. Pujianto. Pengembangan soft skills mahasiswa program kelas internasional melalu pembelajaran berbasis konteks untuk meningkatkan kualitas proses dan hasil belajar mekanika. *Jurnal Pendidikan IPA Indonesia*, 3(1), 12–21. https://doi.org/10.15294/jpii.v3i1.2896. (2014).
20. S. Kim, A. A.von Davier, & S. Haberman. ll-sample equating using a synthetic linking function. *Journal of Educational Measurement,* 45, 325{342}. (2008)
21. Purwoko,A.A, Y. Andayani, M. Muntar, I. N. Diartha. *Efforts* in Improving Teachers Competencies Through Collaboration between Teacher Forum on Subject Matter (MGMP) and Pre-Service Teacher Training Institution (LPTK). *Jurnal Pendidikan IPA Indonesia*, *6*(1), 11-15. https://doi.org/10.15294/jpii.v6i1.8858. (2017)