Journal of Engineering Science and Technology
Vol. 13, No. 6 (2018) 1476 - 1488

DEVELOPMENT AND VALIDATION OF REASONING-BASED MULTIPLE CHOICE TEST FOR MEASURING THE MASTERY OF CHEMISTRY

NAHADI*, HARRY FIRMAN, MIFTAHUL ULUM

Departemen Pendidikan Kimia, Universitas Pendidikan Indonesia,
Jl. Dr. Setiabudi no 229, Bandung 40154, Indonesia
*Corresponding Author: nahadi@upi.edu

Abstract

This study was carried out to develop a valid and reliable reasoning-based multiple-choice test on thermochemistry topic. Development and validation method was applied. This study involved 161 science students as the participants. The results showed that the test developing content validity index (CVI) has a value of 0.949 and reliability with Cronbach Alpha value of 0.955, difficulty index in the medium level and discrimination index in the high level. The results of this research clearly show that 28 items of reasoning-based multiple-choice test developed in this research meet the criteria as a valid test to measure students' reasoning that can be used as school summative and final tests. The data obtained through the interview provide reinforcement that the developed test instrument is feasible to use in both summative chemistry tests and school final examinations.

Keywords: Chemistry, Development, Reasoning-based multiple-choice test, Validation

1. Introduction

Logical reasoning skills are essential for student mastery in many concepts where more complex problem solving strategies are required to succeed in science [1, 2]. The logical reasoning has significant impact on their students' improvement in skills associated with socio-scientific reasoning and scientific creativity [1]. Improvements in the structure and complexity of students' arguments, the degree of rational informal reasoning, and students' conceptual understanding of science can occur [2-5].

Assessment of learning outcomes is used to monitor student learning development and outcomes, as well as to diagnose student needs for some ongoing

learning improvement [6, 7]. Assessment is a process of collecting, analyzing, and making conclusion about the obtained information in order to make a decision. One of the assessments which is annually administered nationally by the government of Indonesia is national exit examination. The purpose of national exit examination is to measure the student learning achievements and outcomes, by requiring students to pass the exams. For the international scale, Indonesia has joined in TIMSS (*Trends in International Mathematics and Science Study*). The framework consists of three cognitive domains, which are used as the criteria of measurement, including knowing, applying, and reasoning [8]. Assessment is also carried out in the area of practical sciences, which are related to the school activities in order to do a research study on a particular problem or issue. All the domains stated previously include some cognitive levels, starting from the low to the moderate level of domain (C1, C2, and C3) and continued from the moderate to the upper level of cognition (C4, C5, C6). Used in the practice of multiple choice test development, which requires reasoning skills; the student understanding can be measured in depth.

All this time, the practice of assessments in schools, including the national exit examinations, is commonly in the form of multiple choice test and paper and pencil test. This method is considered as the form of multiple choice test which is mostly applied in the process of assessment [9]. Multiple choice test is considered as a form of assessment method which has high objectivity level, helps students to answer the tests, is more effective and efficient in the process of assessment, promotes some ease in rating or scoring the student test results based on a certain rubrics, and provides more complete and detailed results which can be calculated by using statistical calculation [10, 11]. Because of the benefits, multiple choice test is considered as the form of assessment which is mostly applied when assessing a large number of participants or test takers [7]. However, multiple choice test also has some limitations, which are: there are possibilities that students choose the correct answers only by guessing, not by thinking; it is rather difficult to design and develop the test; and the student understanding cannot be measured in depth if the multiple-choice test is not designed well; as the result, the test is not authentic [9, 12]. However, the limitations of multiple choice test can be solved by doing some modification on the structure of distractors based on the data on misconceptions obtained from students [12].

One of the topics on chemistry subject, which is taught in senior high schools, is thermochemistry. Thermochemistry has some characteristics, including understanding the concepts, especially on the subject of system and environment, and types of enthalpy changes [13, 14]. Besides that, the main subject of thermochemistry is one of the chemistry subjects which includes the process of calculation and requires good understanding on concepts [15]. A number of researchers found that some factors can make students obtain low achievements on the subject [16]. Thus, the strategies to improve students' comprehension are required [17, 18]. One of the factors for the unsuccessful teaching and learning is student low ability in connecting the concepts. As the result, the thermochemistry concepts become more abstract and cause some misconceptions.

The paradigm shifting on assessments, from low order thinking skill assessment to high order thinking skill assessment occurs based on the Indonesian Government Program in the Ministry of Education and Culture Affair. It is stated that there is a need for the development of test instruments, which can measure high order thinking skills. Many studies have been conducted related to the development of

essay high order thinking skill test. However, there has not been development of reasoning-based multiple choice test [19, 20].

Based on some literature studies done by the researchers, it is found that the study on the development of reasoning based-multiple choice test on chemistry subject has not been conducted before. Therefore, in this paper, we intend to do a research study on the development and validation of reasoning-based multiple choice test to measure student understanding on thermochemistry material.

2. Experimental Method

This research applies the *Development and Validation method*. The development and validation method consists of four phases: (1) Defining the purpose of the test, including the theoretical foundation which underpins the development of the test and introduction; (2) The development and evaluation of test specification; (3) Try out and validation; (4) Test evaluation on the applied procedures [21]. The researcher obtained some ideas about the development of reasoning–based multiple-choice test, which refers to the characteristics of the test development based on *TIMSS Framework 2015*. The test was developed based on reasoning which includes three types of reasoning namely deductive, inductive, and abductive reasoning [22]. The three types of reasoning are covered and represented in the test items and require students to involve themselves in the process of analyzing data and information, drawing a conclusion, broadening understanding in a new situation, developing hypothesis, and planning a scientific investigation. All of these are covered in each test item, in which the development stages were included in the test guidance.

The development method is used to produce reasoning-based multiple choice test on thermochemistry subject. After that, the developed reasoning-based multiple choice test is tested to get the level of validity and reliability [5, 6]. The research procedure can be seen in Fig. 1.

This study involved 161 senior high school students (11^{th} graders of science program) and three senior high school chemistry teachers (teaching the 11 graders of science program) in one senior high school in Lembang, Bandung, Indonesia in the academic year 2016/2017. Generally, there were two phases in conducting the research of test development and validation. The development phase was started from conducting a pilot study by reviewing some related articles and previous studies. The researchers did some studies on a number of references and related journals, with both national and international scales to obtain some ideas about the development of reasoning –based multiple choice tests, which refers to the characteristics of the test development based on *TIMSS Framework 2015*. The next step was designing and creating the blueprint of the reasoning-based multiple choice test on thermochemistry, and then developing the reasoning-based multiple choice test items.

The second phase was validation. The process of content validation was carried out based on some professional considerations by a group of experts to define the content validation of each test item, either from the material aspect and test construction, or from the readability of the language used. The next step was analyzing the results of the validation which is then calculated by using CVR (*Content Validity Ratio*). Lastly, the trial of the test was conducted to get the data

whether or not the reasoning-based multiple choice test developed by the researcher is reliable and applicable.

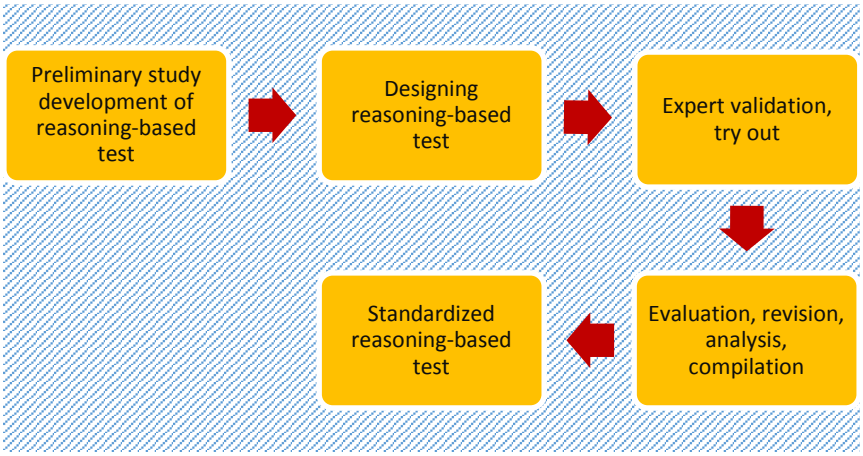

Fig. 1. Flowchart of experimental method.

In this study, the researchers used three kinds of research instrument, which are used to collect data, including content validity sheet which is filled by experts, questionnaires which are used to get the data on students' response about the reasoning-based multiple choice test, and interview guidance used to interview the teachers and three of the students.

In addition, to understand the teacher's response to the developed and solved problems, five indicators were used. (1) Teacher's response to the deepest content in the questions; (2) Teacher's response to the application of assessment; (3) teacher's response to the implementation of the assessment; (4) teacher's response to the availability of implementation of testing time; and (5) suggestions for the improvement of questions in the test. Among the indicators, we concerned only on the indicator 1 and 3, whereas indicator 2, 4, and 5 were neglected.

3. Results and Discussion

3.1. Content validity

As soon as the development process of the reasoning-based multiple choice test was finished, some experts tested check the content validity of each test item. Seven experts were involved, including experts of education evaluation, 2 experts of chemistry subjects, and 3 chemistry teachers. The results of the validation test were analyzed by using *CVR* (*Content Validity Ratio*). The obtained *CVR* values are then compared with the critical CVR values ($CVR_{critical}$) [4]. The critical CVR value ($CVR_{critical}$) from 7 responded is 0.622. Based on the critical CVR value (0.622.), it can be concluded that the test items are valid because the CVR value is greater than 0,622. The results of the CVR value analysis are presented in the following Fig. 2.

From the results of the CVR analysis in Fig. 2, it can be seen that all the test items have CVR values greater than the critical CVR value ($CVR_{critical}$); so, it can be concluded that all the test items are valid. Besides testing the relevance between the test items and the indicators, the experts also provide some ideas and advice as consideration for revising the test in order to make the test better.

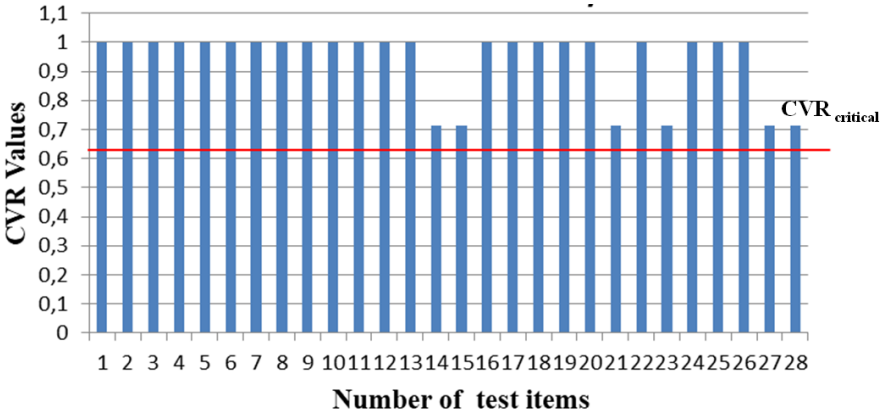

Fig. 2. Level of content validity of each test item based on judgment.

As show in Fig. 1, it can be seen 6 validators who are also legal experts agree with the suitability between the indicators with the item and there is 1 expert validator disagrees with test items number 14, 15, 21, 23 and 28. Thus, in accordance with the critical value that is equal to 0.622, the CVR is still somewhat higher than the critical value; of course all the items apply. Based on the CVR results, CVI was calculated to determine the validity of the contents of the problem as a whole. CVI is the average CVR for each item. The value of CVI divided is 0.949; thus the overall question developed meets the criteria of content validity [4].

In addition to validating the conformity with the indicators, expert validators also provide ideas and suggestions as input for refinement of criminal law-based matters. The input items are the bases for the researchers to make corrections/ revisions to validate items based on the CVR analysis results. Test legibility is done on six students to see aspects of language used in the text matter. From the results of this legibility test, it can be concluded whether the use of language is good or not, because this will affect students to the problems asked in the matter. The sentence-based remedies developed in this phase are based on suggestions and inputs from expert validators and from the legality test results. The following are suggested items for selection and improvement based on measured indicators. In indicator 1.1, for example, the reasoning skill contained in the questions that measure the achievement of this indicator is the analysis. The features in question indicator that is developed based on indicator 1.1 is a student instruction to analyze and use the information provided to be able to answer questions. A problem developed to measure the achievement of indicator 1.1 is a matter of numbers 1 and 2. From the CVR analysis results, it is shown that this problem has met the valid criteria, the expert validator will provide suggestions for improvement of the indicator to be more clear and directed [4]. The problem indicators are "Identify problems and use relevant information from the analysis results on charts or diagrams to explain exothermic and endothermic reactions" and be changed to "Identify relevant problems and information from charts or diagrams to explain exothermic reaction conditions and endotherms" In addition, expert validators also provide suggestions for improvements to the editorial question in the matter. Figure 3 is a matter of prior repair. Figure 4 is a matter of after improvement.

The main sentence in Fig. 3 still has problems, the expert validator advises against using question marks but questions that complete the entice. Then, for question number 2 in Fig. 4, there is no correction.

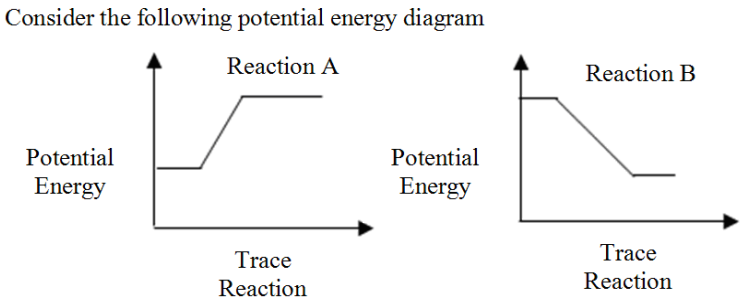

Fig. 3. Example of test items developed prior to revision.

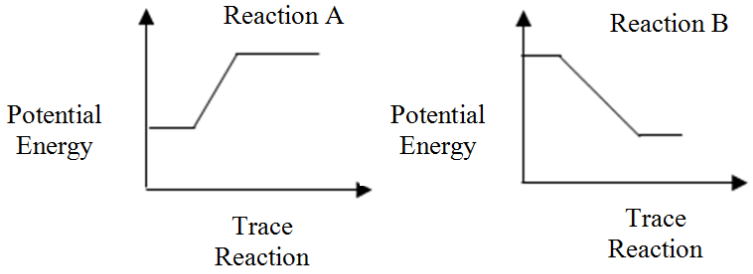

Fig. 4. Example of test items developed after revision.

3.2. Validity of the reasoning-based multiple choice test items

To get the data of the level of validity of each test item, we conducted a trial test. The analysis of the validity was done based on the results of the trial test given to 161 11th grade science students, in the senior high school. The data were then analyzed using SPSS 23 (See Fig. 5):

Based on Fig. 5, it can be seen that 20 test items contain high validity and 8 items have moderate validity. The criteria of the validity are applied based on the validity

category of test item proposed [23]. 28.57% of the test items have moderate validity, while 71.43% of the test items have high validity [11] (See Fig. 6).

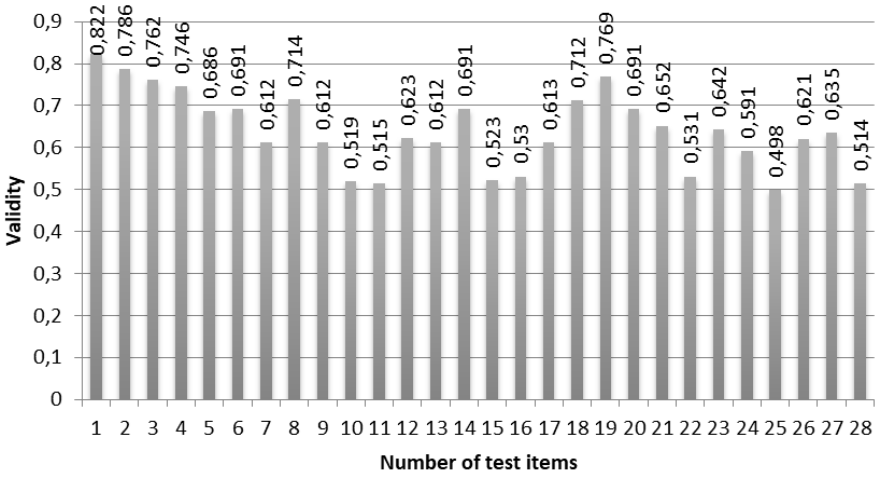

Fig. 5. Level of empirical validity of each test item based on field test.

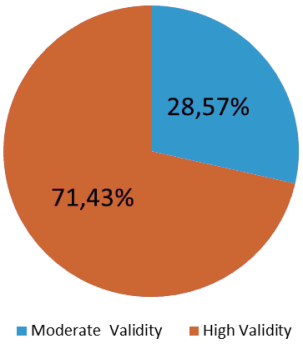

Fig. 6. Distribution of level of validity of the developed test items.

3.3. Reliability of the reasoning-based multiple choice test

Reliability of the test items was obtained through an analysis using SPSS 23. The analysis was carried out based on the data obtained through the trial test in the school. The level of the reliability, either high or low, is defined by the obtained *Alpha Cronbach's* value. Based on the result of analysis on 39 test items, which are administered to 161 students as the respondents, it is obtained that the *Alpha Cronbach's* value is 0.902.

When the Alpha Cronbach's value gets close to 1, it can be said that the test item has a good reliability. Based on the classification of reliability designed by George and Mallery [24], which categorizes the alpha Cronbach's value > 0.90 into the category of high reliability [23, 24], the data show that the reasoning-based multiple choice test, developed by the researcher, has a good or high reliability.

3.4. Level of difficulty

Level of difficulty of a test item is related to the chances in which students' answers on the questions, whether their answers are correct or incorrect. In this research, the analysis on the level of difficulty of the test items was done by using SPSS 23. When the level of difficulty value of the test item is obtained, the value is then matched with the classification of level of difficulty of test item based on the classification proposed [9]. Figure 7 shows the level of difficulty of the test items based on the analysis using SPSS.

There are 4 test items, which belong to difficult category (the level of difficulty value is between 0.11 and 0.30), and 23 items belong to moderate level of difficulty category (the level of difficulty value is between 0.31 and 0.70) [24]. And, the rest of one item belongs to easy category (the level of difficulty value is between 0.71 and 1).

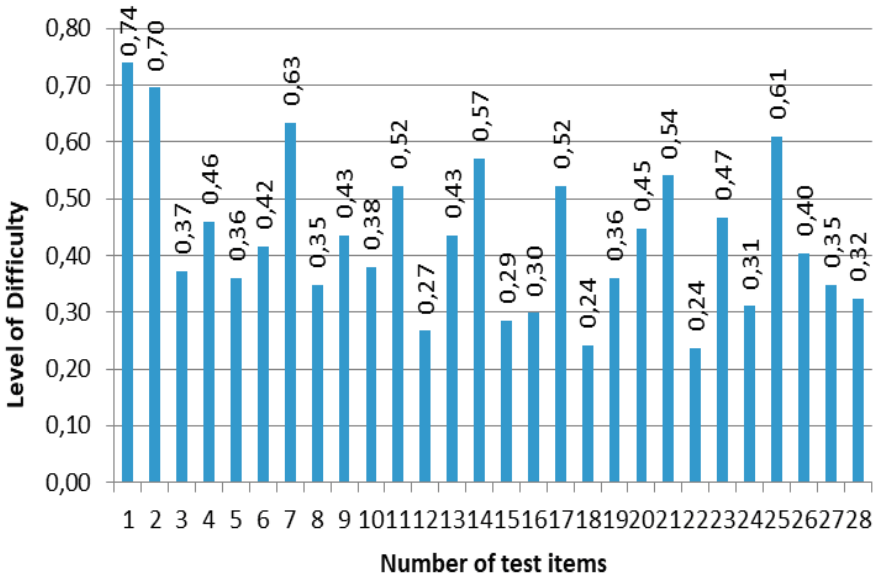

Fig. 7. Level of difficulty of each of the developed test items.

3.5. Discrimination index

Analysis of the discrimination index was done to know how the test items can distinguish between the students who master the materials and who do not. The analysis of the discrimination index was carried out using simple analysis in Excel 2013. Figure 8 describes the analysis result on the discrimination index of each test item, which is developed in the present study.

The result of the analysis is then categorized based on the classification of discrimination index proposed [9]. There is 1 test item which belongs to the category of test items with very good discrimination index or the test items with the discrimination index ranges from 0.71 – 1, while 27 test items belong to the category which has good discrimination index or the value of discrimination index ranges from 0.41 – 0.70 [11].

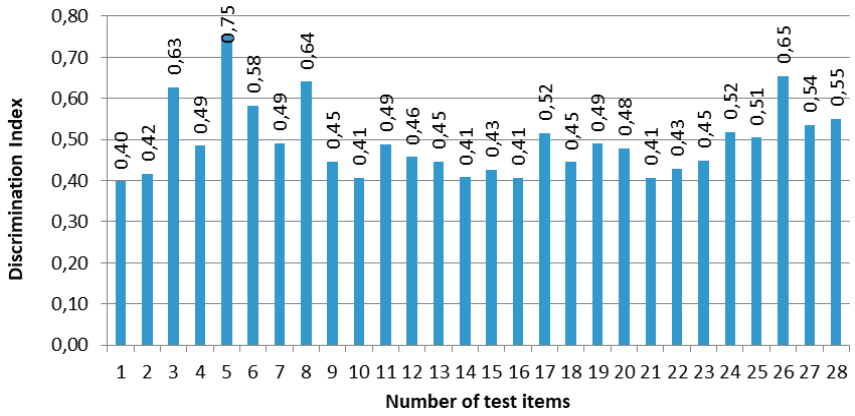

Fig. 8. Level of discrimination index of the developed test items.

3.6. Interview with teachers and students

To collect the data about the teachers' and students' responses to the reasoning-based multiple choice test developed by the researchers, interviews were carried out. This interview is important to get precise qualitative analysis for supporting the present study [3]. Three chemistry teachers and three 11th grade science students have been interviewed. Some of the results are explained based on indicator 1 and 3, whereas indicator 2, 4, and 5 were neglected. Detailed information about indicator 1 and 3 are as follows:

- Indicator 1 (Understanding the characteristics of reasoning-based multiple choice test which is developed, reviewed from the aspects of knowledge which is measured and the readability of the test items).

This question requires the teachers and students to give responses to the test items reviewed from the level of difficulty or the knowledge which is being measured and the readability of the test items. In details, the teachers 'responses about the characteristics of reasoning-based multiple choice test are described as follows:

Teacher 1: *"Viewed from the knowledge aspect which is being measured, reasoning skills are needed to answer these test items. From the readability aspect of the test items, the sentences used in each test item is adequately good, there is a key word for each problem being asked, so students can more easily understand the meaning of the questions.*

Teacher 2: *"The test items require students to think with reasoning skills because the problems being asked lead the students to use their understanding and logic. In term of the readability of the test items, overall the sentences and words used in the test items are good; however, there are some terms which are unfamiliar for students, such as control variable, independent variable, and dependent variable. Therefore, it is necessary to give some explanation about the unfamiliar terms before asking the students to do the test.*

Teacher 3: *"To do this test, higher knowledge is required because the test items can be answered by using higher order thinking skill. In term of the readability of the sentences used, it can be categorized as good and clear.*

The questions asked to the students are: "Did you find difficult to answer the questions in the test? If yes, what factors that make it difficult for you to answer the questions? The students' responses about the characteristics of reasoning-based multiple choice test is described as follows:

Student 1: *"Yes, I feel rather difficult to answer the questions because I am not used to answering questions which require reasoning skills.*

Student 2: *"Yes, such these questions are rarely exemplified in the learning process and exercises in school.*

Student 3: *"Yes, because I must have more understanding on the materials in order to be able to answer the questions; while in the learning process, such these reasoning questions are never exemplified.*

- Indicator 3 (the application of reasoning-based multiple choice test in summative test and national examination).

The question for this indicator requires the teachers and students to give their responses to whether or not the reasoning-based multiple choice test can be used in summative tests or national examinations. The teachers' responses to the application of the reasoning-based multiple choice test in summative tests and national examinations are explained as follows:

Teacher 1: *"By considering the facilities available in the school, It can be implemented for summative test, so there will be an a conformation". However, I somewhat disagree if the reasoning-based multiple choice test is included in the national examination. It is because not all schools have adequate facilities, especially laboratory, tools and chemical materials. So, it is necessary to improve the standard before applying this kind of test in the national examination.*

Teacher 2: *"Very good. So, the measurement of how far the students understand the lessons during the learning process can be observed. However, I somewhat disagree if this test is applied for the national examination; because students must learn through the same process which are asked in the test. While in some schools in rural areas, for examples, it is not possible to do that. As the result, there will be a mismatch between the learning materials if such test is applied for the national examination".*

Teacher 3: *"Agree, but it is necessary to make some compliance with the problems involved in the test items. And it can be applied in the national examination, but some adaptation and conformation need to be done in the teaching and learning process; so the problems discussed in the test items can be understood by students".*

The same question is asked to the students. The question requires the students to give their responses about whether or not such this test can be applied in the summative tests and national examinations. The three students responded positively. The students' responses are described as follows:

Student 1: *"Agree, because if it is used in the summative test, students can understand the material better and it will be something new and interesting if it is used in the national examination".*

Student 2: *"Good because it can be a learning experience for me and it can motivate me to learn more.*

Student 3: *"The questions are good; they can measure my understanding. And then, I agree if it is used for national examination. I agree. So, I must learn more diligently".*

Based on the results of the interviews with the teachers and the students, there are various types of information about the importance of developing reasoning-based tests and their roles in each moment of assessment. As shown in the results of the interview above, indicator 1 is related to the level of knowledge measured and test item readability. Teachers and students argued that the developed test items required their reasoning in answering those items. This becomes a stimulus for teachers in packaging the learning requires students to think [8, 25]. Students should be accustomed to reasoning in learning in order that they will not have difficulty and they will be able to answer such questions [1]. Similarly, in indicator 3, teachers and students argued that learning has made students accustomed to reasoning, such tests are very good to be used in school summative and final tests. This can be possible if learning activities are simultaneous with their assessments.

4. Conclusion

Based on the findings and discussion, this research came into the following conclusions: 1) All the test items in the reasoning-based multiple choice test developed by the researcher are valid; 2) Based on the criteria of reliability, the reliability of the test can be accepted; 3) The test developed fulfills the criteria of appropriateness, reviewed from the availability of time to do the test. Moreover, there are positive responses from the research respondents if the developed test is applied in chemistry summative test. The process of developing this test can become a model for researchers and practitioners in developing various reasoning-based tests. This test can be used as a model and inspiration in designing tests and learning based on reasoning as a model of knowledge required in various fields in life. From the interview, students and teachers agreed that the reasoning-based test items can be used as the reference in giving some examples in the learning process. As the result, students will be familiar and trained in doing tests which require high order thinking skills.

Acknowledgements

Rector of UPI, Dean of FPMIPA UPI, and Head Department of Chemistry Education who have provided fund and facility support for the implementation of this research are properly acknowledged.

References

1. Venville, G.J.; and Dawson, V.M. (2010). The impact of a classroom intervention on grade 10 students' argumentation skills, informal reasoning, and conceptual understanding of science. *Journal of Research in Science Teaching*, 47(8), 952-977.

2. Bird, L. (2010). Logical reasoning ability and student performance in general chemistry. *Journal of Chemical Education*, 87(5), 541-546.
3. Haristiani, N.; Aryanti, T.; Nandiyanto, A.B.D.; and Sofiani, D. (2017). Myths, islamic view, and science concepts: The constructed education and knowledge of solar eclipse in Indonesia. *Journal of Turkish Science Education*, 14(4), 35-47.
4. Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563-575.
5. Nandiyanto, A.B.D.; Asyahidda, F.N.; Danuwijaya, A.A.; Abdullah, A.G.; Amelia, N.; Hudha, M.N.; and Aziz, M. (in press). Teaching "Nanotechnology" for elementary students with deaf and hard of hearing. *Journal of Engineering, Science and Technology (JESTEC), Special Issue on AASEC'*2017.
6. Litbang Kemdikbud. (2015). Survey internasional TIMSS. Retrieved December 10, 2015, from http://litbang.kemdikbud.go.id/index.php/survei-internasional-timss/tentang-timss.
7. Liu, O.L.; Lee, H.S.; and Linn, M.C. (2011). Measuring knowledge integration: Validation of four-year assessments. *Journal of Research in Science Teaching*, 48(9), 1079-1107.
8. Jönsson, A.; Rosenlund, D.; and Alvén, F. (2017). complement or contamination: a study of the validity of multiple-choice items when assessing reasoning skills in physics. *Frontiers in Education*, 2, 1-11.
9. Creswell, J.W.; and Creswell, J.D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Los Angeles: Sage publications Inc.
10. Cetin-Dindar, A.; and Geban, O. (2011). Development of a three-tier test to assess high school students' understanding of acids and bases. *Procedia-Social and Behavioral Sciences*, 15, 600-604.
11. Towns, M.H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education*, 91(9), 1426-1431.
12. Herrmann-Abell, C.F.; and 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.
13. Cigdemoglu, C.; and Geban, O. (2015). Improving students' chemical literacy levels on thermochemical and thermodynamics concepts through a context-based approach. *Chemistry Education Research and Practice*, 16(2), 302-317.
14. Nandiyanto, A.B.D.; Munawaroh, H.S.H.; Kurniawan, T.; and Mudzakir, A. (2016). Influences of temperature on the conversion of ammonium tungstate pentahydrate to tungsten oxide particles with controllable sizes, crystallinities, and physical properties. *Indonesian Journal of Chemistry*, 16(2), 124-129.
15. Haláková, Z.; and Prokša, M. (2007). Two kinds of conceptual problems in chemistry teaching. *Journal of Chemical Education*, 84(1), 172.
16. Wigfield, A.; and Guthrie, J.T. (1997). Relations of children's motivation for reading to the amount and breadth or their reading. *Journal of Educational Psychology*, 89(3), 420.

17. Haristiani, N.; and Aryadi, S. (2017). Development of android application in enhancing learning in Japanese kanji. *Pertanika Journal of Science & Technology*, 25, 157-164.
18. Haristiani, N.; and Firmansyah, D.B. (2017). Android application for enhancing Japanese JLPT N5 kanji ability. *Journal of Engineering, Science and Technology* (*JESTEC*), *Special Issue on AASEC*'2016, 12, 106-114.
19. Cloonan, C.A.; and Hutchinson, J.S. (2011). A chemistry concept reasoning test. *Chemistry Education Research and Practice*, 12(2), 205-209.
20. Mullen, K.; and Schultz, M. (2012). Short answer versus multiple choice examination questions for first year chemistry. *International Journal of Innovation in Science and Mathematics Education (formerly CAL-laborate International)*, 20(3), 1-18.
21. Adams, W.K.; and Wieman, C.E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, 33(9), 1289-1312.
22. Fischer, H.R. (2001). Abductive reasoning as a way of worldmaking. *Foundations of Science*, 6(4), 361-383.
23. Guilford, J.P. (1942). *Fundamental statistics in psychology and education*. New York: McGraw-Hill Book Company.
24. Gliem, J.A.; and Gliem, R.R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales, 85-88. Retrived on February 2018, from https://scholarworks.iupui.edu/bitstream/handle/1805/344/Gliem%20& %20Gliem.pdf?s.
25. Laius, A.; and Rannikmäe, M. (2011). Impact on student change in scientific creativity and socio-scientific reasoning skills from teacher collaboration and gains from professional in-service. *Journal of Baltic Science Education*, 10(2), 127-137.