*International Journal of Instruction*
*e-ISSN: 1308-1470 ● www.e-iji.net*

*July 2018 ● Vol.11, No.3*
*p-ISSN: 1694-609X*
*pp. 557-572*

Received: 12/01/2018
Revision: 02/05/2018
Accepted: 07/05/2018

# Development and Validation of an Integrated Assessment for Measuring Critical Thinking and Chemical Literacy in Chemical Equilibrium

**Satya Sadhu**
Department of Chemistry Education, Post Graduate Programme, Yogyakarta State University, Indonesia, *satya0185pasca2016@student.uny.ac.id*

**Endang W Laksono**
Assoc. Prof., Department of Chemistry Education, Faculty of Mathematics and Science, Yogyakarta State University, Indonesia, *endang_widjajanti@uny.ac.id*

Although the development of critical thinking and chemical literacy is a major goal of science education, the adequate emphasis has not been given to the measurement of both skills. This study reports the development and validation of an integrated assessment instrument to measure students' critical thinking skill and chemical literacy together in chemical equilibrium. The development of the framework in this study was implemented based on 4-D model. The stage of development begins with define, design, develop, and disseminate. The initial integrated assessment consisted of 37 open-ended two-tier multiple-choice question items. The preliminary version of the integrated assessment was initially piloted on review with experts and the paper-pencil test was administered to a group of students. The data were analyzed using Item Response Theory (IRT). Analysis factor was involved to find out the construct validity and strengthen the finding of content validity with the experts. In additionally, PCM 1-PL model was conducted to examine the quality of parameter estimates of the test items. The findings of the content validity, construct validity and quality of the items, overall suggest that the integrated assessment has relatively high validity and reliability and can be used for measuring the 13 integrated skills can be measured in chemical equilibrium.

Keywords: content validity, construct validity, integrated assessment, critical thinking, chemical literacy, chemical equilibrium, item response theory (IRT)

## INTRODUCTION

Advances in the field of science and technology have greatly improved the quality of various fields such as communication, health, agriculture, educational environment, and lifestyle. However, rapid changes in science and technology also come at a cost for future generations will face even more challenging decisions than the current one

**Citation:** Sadhu, S., & Laksono, E. W (2018). Development and Validation of an Integrated Assessment for Measuring Critical Thinking and Chemical Literacy in Chemical Equilibrium. *International Journal of Instruction*, 11(3), 557-572. https://doi.org/10.12973/iji.2018.11338a

(Alghafri & Bin Ismail, 2014; Akgun & Duruk, 2016; Mapaella & Siew, 2015). Changes, challenges, and the fast flow of knowledge caused a shift in the focus of education institutions from developing theoretical knowledge to developing thinking ability or skill, e.g. critical thinking is widely claimed as a primary goal of science education. Consequently, practicing critical thinking skills is growing more important as students need to adjust to such change by actively and skillfully identifying problem, reconstructing arguments, evaluating arguments, determining solution, and drawing conclusion which will result in individuals capable to think critically (Alghafri & Bin Ismail, 2014; Akgun & Duruk, 2016).

A good education is the one that not only prepares its students to continue their study to a higher level, but also to solve challenges and problems its students face in everyday life or workplace. These challenges and problems are the ones that need critical thinking skill to solve. Research has found that Indonesia placed in 4$^{th}$ from the bottom at TIMSS (Trends in International Mathematics and Science Study) (Institute of Education Science, 2016). It showed that Indonesian students only mastered the theories of basic knowledge in the field of biology, chemistry, physics, and geography at low cognitive level. To develop in students critical thinking skills, science teachers in Indonesia have a crucial responsibility for the next generation and need to encourage students both to improve their critical thinking skills. Therefore, an instrument is needed to assess and train critical thinking skill to students in the field of science, especially chemistry. Pacific Policy Research Center (2010) stated that there are 24 standards that focus on the assessment of student skills, one of which is critical thinking skill assessment.

Critical thinking is closely related to everyday life. Thus, there is a correlation between critical thinking and chemistry literacy. Literacy is a knowledge and skill for the life of people, where chemistry literacy is closely related to an individual's ability to use chemistry knowledge in identifying various questions and drawing conclusions based on evidence which can help that individual to make a decision regarding the scientific world and the relationship between human and nature (Organization for Economic Cooperation and Development, 2006). Understanding chemistry is very important matter because our physical environment is decisively influenced by chemistry and filled with chemicals (Gilbert & Treagust, 2009). However, understanding chemistry and the ability to apply it in daily life which is also known as chemistry literacy is a more important matter (Tsaparlis, 2000). Based on Organization for Economic Cooperation and Development (2015) in the Program for International Student Assessment from 70 countries participating, Indonesia placed 62$^{nd}$. Indonesian science literacy was placed in level 1a. It proves that students are required to train and develop their science literacy.

However, there is little evidence that instruments are being used to assess student' critical thinking and chemical literacy in chemistry. Much of the difficulty lies in the lack of integrated assessment to assess student's critical thinking and chemical literacy in high school science classrooms.

Regulation of Ministry of Education and Culture Indonesia number 66 in 2013 regarding the standard for education assessment stated that learning assessment should include three learning aspects which are cognitive, affective, and psychomotor.

However, currently learning assessment still focuses on cognitive aspect only (Ministry of Education and Culture Indonesia, 2013). The process of learning assessment that is currently used only focus on cognitive aspect without including affective and psychomotor aspect (Zoller, 2001). Even though those three aspects are assessed, they are still performed separately. The assessment of the cognitive and affective aspect of learning outcome should be done simultaneously. Austin et.al (2015) state the students with good affective aspect in learning will have good cognitive skill compared to students without affective aspect and only depend on rote. It shows that there is a correlation between affective and cognitive aspects. Currently, the assessment of cognitive and affective aspects was done separately (Saribaz & Bayram, 2009; Tosun & Taskesenligil, 2013; Keil, Haney & Zoffel, 2009). Thus, an integrated instrument needs to be developed that aim to assess critical thinking and chemical literacy in a single test only.

Critical thinking and chemistry literacy instrument in a specific domain of chemistry is still in the form of one-tier, which is essay only or multiple choice only. This is not effective enough to train and measure students' critical thinking skill. The drawbacks of the use of the one-tier instrument, in the form of multiple choice or essay, become an important matter to innovate in developing other forms of tests. An assessment instrument needs to be developed to improve those drawbacks in the form of multiple choices including response and alternative to students' conception. The students are required to support their answer choice by giving a reason. The use of reason in answering multiple choice test items is an effective way to assess meaningful learning. All things considered, in this study developed an integrated assessment instrument in the form of open-ended two-tier multiple choice questions.

Open-ended two-tier multiple choice questions have two main benefits over conventional one-tier questions. The first is a decrease in the measurement error. In a general multiple choice question with 5 possible choices, there is a 20% chance of correctly guessing the answer by the student. These random, correct guesses must be accounted for in the measurement error. A two-tier multiple-choice question is considered correct only if both tiers are answered correctly by the student. As a result, a student responding to a question with 5 choices in the first tier and the correct reason in the second has less than 20% chance of randomly correct guessing. The second benefit to the two-tier multiple choice questions format is that it allows for the probing of two aspects of the same phenomenon. In the first tier, students are asked to predict the outcome of a chemical change, and the second tier asks for an explanation. This allows the probing of the phenomenological domain with the first tier and the conceptual domain with the second (Tüysüz, 2009).

The purposes of this study were to report on the development and validation of an integrated assessment for measuring critical thinking and chemical literacy in $11^{th}$ grades in chemical equilibrium. The study would provide the quality of the items using item response theory analysis. Additionally, the study would provide information regarding its construct validity by means of Kaiser–Meyer–Olkin (KMO) test and Bartlett Sphericity test to identify whether the data were appropriate for factor analysis.

## METHOD

### Participants

The participants in this study were 158 of 12$^{th}$ grade students purposively selected from three high schools during the first semester of 2017 in the special region of Yogyakarta, Indonesia. The 158 students have different ability level from three different high schools. The three high schools were selected based on the grade in the school by seeing the result of the recent score national exam. In this study, the researcher used one school for each of high, middle and low-grade school.

### The Development Framework

The test and its set-up were designed to reveal quantitative aspect in integrated assessment. This research model is a procedural development research, which is a descriptive research, showing the steps which have to be followed to produce the final product. The development of the integrated assessment used the 4-D model which was developed by Thiagarajan, Semmel and Semmel (1974). In this study, the integrated assessment was developed within three stages, as follows.

*Define*

The first stage in developing the integrated assessment was "define". There are some parts on the first stage of developing the integrated assessment. First, analyzed the characteristics of students in accordance with the design of the development of integrated assessment instrument which relating to learning material and indicators would be used to achieve the learning objectives. Second, front-end analysis was conducted to develop an integrated assessment based on a literature of assessment used by teachers to measure student's critical thinking skills and chemical literacy. Third, specifying instructional objectives was conducted to formulate learning objectives to be achieved and determine the purpose of developing the integrated assessment. Fourth, concept analysis was conducted to identify the subject matter that would be used, adapted to the learning process that will be done by the teacher in the learning process. Additionally, task analysis to identify indicators of critical thinking skills and chemical literacy to be achieved.

*Design*

The second stage in developing the integrated assessment was design the construct of the items on the integrated assessment. They were divided into three parts. The first parts involved analyzing the main teaching and learning objectives of Indonesia chemistry curriculum content and textbooks in grades 11. The content was characterized according to the scientific knowledge and skills. The researcher determined the standards of competence, basic competence, and indicators to be developed on integrated assessment. The second part involved describing the indicators of critical thinking skills and chemical literacy, designing the initial questions, script questions, and rubric assessments. The third part involved writing up the items. The items were written in Indonesia language.

*Develop*

The third stage is develop the integrated assessment instrument included a review by expert judgments. Content validity of the items was established by presenting them to a panel of experts in the areas of chemistry content and evaluation experts and also chemistry teachers. The experts were selected based on their experiences in the field of education, their knowledge in research work as well as their familiarity with the subject. The main purpose of the integrated assessment was initially explained to them, and subsequently, they were requested to review each item in relation to the overall purpose of the instrument. Specifically, the content experts were requested to review each item based on the following criteria: (a) substance of the items to the basic competencies and indicators to be achieved, (b) construct of the information presented in the items include clarity of the words/ phrases/diagrams of each item, (c) Language of the items to the correct Indonesia grammar rules, and (d) appearance of the items. In line with the comments, all the necessary revisions were made. Their feedback was used for refinement of the items. The initial product results from the review would be used in the trial implementation.

After incorporating all the expert comments, the paper-pencil administration was conducted with a group of third-year science students (N=158) who had recently completed the chemical equilibrium course. The goal of the paper-pencil administration was to examine construct validity and parameter analysis of items (analysis of instrument model fit, item difficulty, and reliability). Prior to the paper-pencil test administration, the students were provided oral instruction regarding the purpose of the test, general instruction on how they should respond to the open-ended two-tier multiple-choice items, and a request to take the test seriously. The results of the paper-pencil test were analyzed. Revisions were made on the final version of the items based on the result on the paper-pencil test.

**Instrument**

*Item Construct*

The items were constructed based on the contents of Chemistry Curriculum 2013 in Indonesia. In order to elicit the desired critical thinking and chemical literacy from the students, the stem of each item was written in such way that provoked critical thinking and chemical literacy together.

In this study, the researchers combined some critical thinking framework from several experts: Facione (2013), Bowel and Kemp (2005), and Watson-Glaser (2008). The retrieval of the framework is based on the suitability and importance in making integrated assessment instrument related to chemical literacy. The aspects of critical thinking following (1) identifying a problem, (2) reconstructing an arguments, (3) evaluating an arguments, (4) determining a solution, and (5) drawing a conclusion. The definition of chemical literacy consists of four dimensions: understanding chemical ideas, contextual aspects, cognitive aspects, and affective aspects (Shwartz et.al., 2005). In this study, the aspect of chemical literacy was limited to the contextual aspect. The aspects of chemical literacy in the contextual aspects following (1) explain the

phenomenon by using the chemical concept, (2) solve a problem using a chemical understanding, and (3) analyze the strategies and benefits of chemical applications. The aspects of critical thinking were integrated with the aspects of the chemical literacy to be one aspect. Concerning with it, there are 13 integrated skills or aspect of critical thinking and chemical literacy that can be measured using the integrated assessment.

Table 1
The Integrated Aspects of Critical Thinking and Chemical Literacy Aspects

| No | Aspects of Chemical Literacy | Aspects of Critical Thinking | Integrated Skill |
|---|---|---|---|
| 1 | Explain the phenomenon by using the chemical concept | Draw a conclusion | Drawing conclusions based on phenomena by utilizing the chemical concept |
| 2 | Explain the phenomenon by using the chemical concept | Evaluate arguments | Evaluate arguments based on phenomena by utilizing the chemical concept |
| 3 | Solve a problem using a chemical understanding | Identifying problems | Identify problems by utilizing chemical understanding |
| 4 | Solve a problem using a chemical understanding | Evaluate arguments | Evaluate arguments by utilizing chemical understanding |
| 5 | Explain the phenomenon by using the chemical concept | Draw a conclusion | Drawing conclusions based on phenomena by utilizing the chemical concept |
| 6 | Solve a problem using a chemical understanding | Draw a conclusion | Drawing conclusions by exploiting the understanding of chemistry |
| 7 | Explain the phenomenon by using the chemical concept | Identifying problems | Identify problems based on phenomena by utilizing the chemical concept |
| 8 | Solve a problem using a chemical understanding | Identifying problem | Identify problems by utilizing chemical understanding |
| 9 | Solve a problem using a chemical understanding | Draw conclusion | Drawing conclusions by exploiting the understanding of chemistry |
| 10 | Analyze the strategies and benefits of chemical applications | Identifying problem | Identify problems from chemical application instances |
| 11 | Explain the phenomenon by using the chemical concept | Draw conclusion | Drawing conclusions based on phenomena by utilizing the chemical concept |
| 12 | Analyze the strategies and benefits of chemical applications | Determining solution | Determine the solution of the chemical application example |
| 13 | Analyze the strategies and benefits of chemical applications | Reconstructing arguments | Reconstruct arguments from chemical app instances |

*Item Format*

The initial integrated assessment instrument consisted of 37 open-ended two-tier multiple-choice question items with five possible choices. The item format of the open-ended two-tier multiple choice question, the first-tier consisting of a multiple choice question and the second-tier consists of a blank for the desired reason. The students were asked to select a correct answer among the distracters on the first tier and then give explanations for their choices on the second-tier. The sample item of integrated assessment construct is described below to illustrate the item format.

Ani checked her tooth condition in the hospital. At the time of the consultation, the dentist explained that in the human mouth there is a layer of tooth enamel containing calcium hydroxyapatite compound ($Ca_5(PO_4)_3OH$) as much as 95% and water as much as 5%. Tooth enamel is the outermost layer of the tooth that covers the entire crown of the tooth and is the hardest part of the body and is formed by cells called ameloblast. The reaction can be written

as follows

$$Ca_5\,(PO_4)\,3OH_{(s)} \rightleftharpoons 5\,Ca^{2+}{}_{(aq)} + 3PO_4^{3-}{}_{(aq)} + OH^-{}_{(aq)}$$

The dentist states if Ani should reduce the foods containing acid ($H^+$). By using the concept of dynamic equilibrium, when Ani eats many dishes containing acid, then ....

A. Email gear is getting stronger
B. Email tooth decreases
C. Email gear does not change
D. Dental email changes color
E. Dental email is deformed

Reason.....
.....
.....

Figure 1
The Open-Ended Two-Tier Multiple Choice Questions

*Scoring Procedure*

The scoring guide for the item was created to make the teacher easy to assess the test. The scoring guide is described below.

Table 2
Scoring Guide

| First Tier | Second Tier | Score |
|---|---|---|
| No answer | No answer | 0 |
| False | False | 0 |
| Correct | False | 1 |
| Correct | Correct with the uncompleted reason | 2 |
| Correct | Correct with the complete reason | 3 |

*Data Analysis*

Item Response Theory (IRT) is mostly used for modeling responses to items and scoring of educational tests (Adedoyin & Mokobi, 2013). Item response theory is modern analysis of the item. The reason of using the item response theory is, the ability level of an examinee is accurately estimated with any set of items that would be measured. In the light of the aforementioned purposes of the study, the content validity was conducted by using Aiken's V formula (1985), the construct validity was conducted by using factor analysis, and the quality of the items was conducted using the Rasch model. Factor analysis and Rash model are part of the item response theory analysis.

In the construct validity, the exploratory approach was used to see how many factors are needed to explain the relationship between a set of indicators by observing greater load factors. The initial stages of construct validity conducted in this study following (a) Check the appropriate value of the Kaiser–Meyer–Olkin (KMO) and Bartlett Sphericity test, (b) Check the appropriate value of the anti-image correlation on varimax rotation. It stages were conducted in order to identify the advisability of the data to be able carry on the further analysis using the a basic assumptions. This theory is uni-dimensionality which means items of test measure only one ability. It would indicate that the integrated assessment only measure the integrated skill of ciritcal thinking and chemical literacy. In

additionally, quality of items was assessed by conducting item analysis in terms of conformity of item (item fit), difficulty index and reliability. The collected data were computed in SPSS 16 and Winsteps. The items which didn't met with the passing score in every stage of analysis would not be stored in the next analysis.

## FINDINGS AND DISCUSSION

### *Content Validity*

In this study, 37 open-ended two-tier multiple-choice question items in the preliminary version of the instrument has been validated by a panel of expert judgments. The experts validated the suitability of instrument test to be used in the context. The experts' response revealed that the integrated assessment is a highly suitable instrument to every aspect which were substance, construct, language, and appearance. The value of the items are clarified on the figure 2.

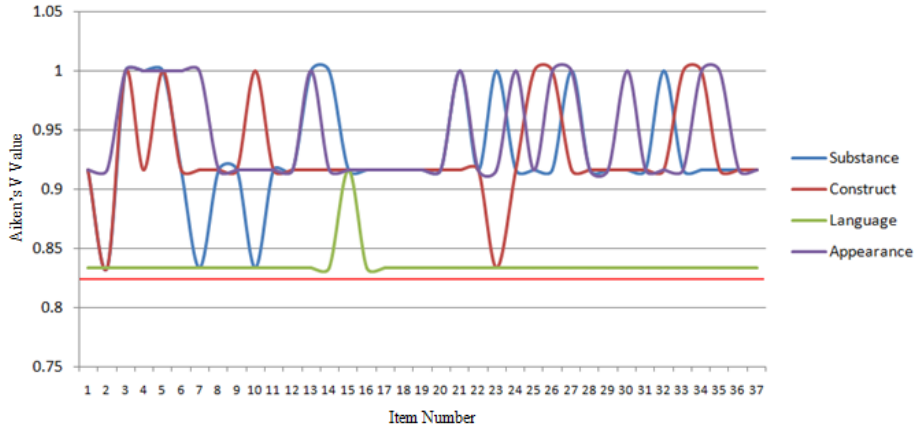


Figure 2
The Graph of Aiken's V Value

The red line is the critical value of the item acceptance in the content validity analysis. According on the Figure 1, all the score items was over the critical value. It indicates that 37 items in the preliminary version of the integrated assessment instrument were stored to be carried on the next analysis.

### *Construct Validity*

Construct validity was conducted using the factor analyis. Factor analysis aims to identify the relationship between variables by looking at the eigenvalue in the matrix of intercultural covariance-variance based on the computational result. The collected data were analyzed to find out whether they are suitable for factor analysis or not. The construct validity of the instrument was determined by using principal component analysis. In this study, for the construct validity was conducted two times. First, there were 8 items that discarded and could not be stored because did not met with the

standard value of anti-image correlation values. Therefore, the 8 items discarded and the 29 item left was calculated again . The second calculation after 8 items discarded explains gradually below.

In order to identify whether the data were appropriate for factor analysis, Kaiser–Meyer–Olkin (KMO) Test and Bartlett Sphericity test were conducted in the principal component analysis. Also, varimax rotation method was used to give a better picture of factors in this analysis. The details of these analyses are as follows:

1) The computational result of KMO value was found 0,739. Leech et al. (2005) stated that factor analysis could not be conducted under the value of 0.50 because it is the critic value. The KMO value at least above 0.50 shows that the data are appropriate for factor analysis
2) Bartlett Sphericity Test score result was 1.491E3 and significant at 0.000 level ($X^2406$ = 1.491E3). The significance value was found lower than 0.05, which means factor analysis can be conducted for the further analysis.

The anti-image correlation values were analyzed after the KMO and the Bartlett Sphericity test was accepted. The anti-image for each of the 29 items range from 0.554 to 0.849. None of the items have a value below 0.50, which shows that the load values of these items highly contribute to the factor structure of the instrument. Consequently, no items were discarded.

The unidimensionality assumption test was conducted after estimation and acceptance of the Kaiser–Meyer–Olkin (KMO) test, Bartlett sphericity test, and the anti-image value on the factor analysis.

Unidimensionality can be tested using two approaches which are exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Confirmatory factor analysis is appropriate when the scale has known factor properties, while an exploratory factor analysis is more appropriate when the scale is relatively unexplored in terms of factor (Toland, 2014). In this study, EFA approach was conducted to know how many the factor would be appear. Therefore, to know whether the exploratory factor analysis approach can be done, the value of KMO should met with the passing score. If the KMO value is lower than 0.50, then the exploratory factor analysis approach can not be done (Yilmaz, Altinkurt & Cokluk, 2011). In this study the KMO value obtained was 0.739 so that the exploratory factor analysis approach can be done. Therefore, the unidimensionality assumption can be analyzed by using the EFA. The results of the unidimensionality assumption test are presented briefly in Table 3

Tabel 3
Total Variance Explained

| Component | Initial Eigenvalues | | Component | Initial Eigenvalues | |
|---|---|---|---|---|---|
| | Total | % of Variance | | Total | % of Variance |
| 1 | 5.869 | 20.238 | 6 | 1.353 | 4.664 |
| 2 | 2.834 | 9.771 | 7 | 1.343 | 4.631 |
| 3 | 2.094 | 7.220 | 8 | 1.145 | 3.949 |
| 4 | 1.585 | 5.466 | 9 | 1.085 | 3.741 |
| 5 | 1.412 | 4.870 | …..29 | 0.188 | 0.649 |

Based on table 3 shows that there are 9 eigenvalues that have values greater than 1,000. Consequently, based on Kaiser criterion (Beavers et al., 2013), the variance of the 9 factors appeared as the response of test participants to the integrated assessment. These nine factors can account for about 64.55% of the total variance. As a result of this method, nine factors were identified in the scale. It implies that there exists a dominant component or factor referred to as the ability measured by the integrated assessment instrument (Hambleton, Swaminathan & Rogers, 1991). Thus, integrated assessment instrument developed had quantitative skill as a dominant factor.

The result of unidimensionality assumption can also be observed through scree plot. The scree plot graph in Figure 3 confirms that the integrated assessment instrument includes nine factors. The scree plot is able to clarify the visualization of the eigenvalues by the number of components maintained into the factor. The determination of the number of factors formed from the items is done by considering the initial constructs underlying the development of instruments, eigenvalues, and scree plot images. The eigenvalue is clarified on the scree plot of Figure 3.

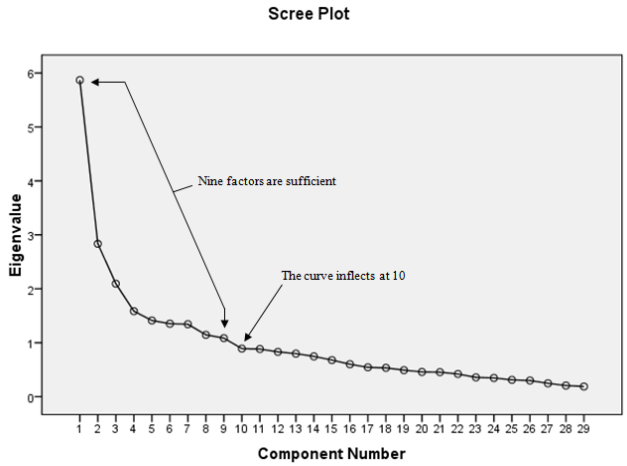


Figure 2
Scree Plot of Factor Analysis

Figure 3 shows that the relative curve starts to slop on the tenth factor, so it can be stated that at least 9 factors were formed with the first factor as the dominant factor (20.238%). If the output of factor analysis in exploratory factor analysis generated by the first factor is capable of explaining variance greater than 20%, then the unidimensionality assumption has been accepted (Reckase, 1979; Smits et al., 2011; Wu et al., 2013). According to the experts, Hambleton and Swaminathan (1985: 16) assert that unidimensionality assumptions are very difficult to fulfill ideally. Therefore, unidimensionality assumptions can be considered to be fulfilled if the instrument contains a dominant component that measures the ability of students who have been tested (Hambleton, Swaminathan & Rogers, 1991; Wu et al., 2013; Adedoyin & Adedoyin, 2013; Guler et al., 2014). In short, 29 items met with the passing score and stored to be carried in the next analysis.

**Parameter Analysis of Items**

*Analysis of Instrument Model Fit*

The analysis of statistical fit is a check on the internal validity. Within the latent trait test model, the internal validity of a test was assessed in terms of the statistical fit of each item to the model. The form of the item in the integrated assessment is open-ended two-tier multiple-choice question items with five possible choices. Consequently, the data is in the polytomus. The polytomus data in the Rasch model is appropriate to be analyzed by using the Partial Credit Model 1 Parameter Logistic (PCM 1-PL). The model in this study referes to the PCM 1-PL. The item fit describes whether the item is functioning normally in measuring or not. If the item is outliers or misfits, it indicates that there is a misconception on the students to the item. This information is very useful for teachers to improve the quality of teaching so misconceptions can be avoided. According to Korashy (1995), if the fit statistic of an item is acceptable, then the item can be said as a valid item. According to Boone et al. (2014), the criteria used for the suitability of the outliers or misfits items following: (a) the value of outfit mean square (MNSQ) is received if 0.5 <MNSQ <1, (b) the value Z-Standard Outfit (ZSTD) is accepted if -2.0 <ZSTD <+2,0, (c) the value of Correlation Points (Pt Mean Corr) is accepted if 0.4 <Pt Measure Right <0.85. The results of the item fit are presented briefly for some items in table 4.

Table 4
Analysis of Item Fit

| No | Item No | Value of Output MNSQ | Value of outfit z-standard | Value of point measure correlation | Conclusion |
|---|---|---|---|---|---|
| 1 | Item 1 | 1,18 | 1 | 0,37 | Fit |
| 2 | Item 2 | 0,71 | -2,2 | 0,56 | Fit |
| 3 | Item 5 | 1,07 | 0,5 | 0,46 | Fit |
| 4 | Item 6 | 0,68 | -2,8 | 0,56 | Fit |
| 5 | Item 7 | 0,90 | -0,4 | 0,44 | Fit |
| ….29 | Item 37 | 0,84 | -0,3 | 0,27 | Fit |

Sumintono (2015) states that ZSTD is highly appreciated by the sample size. If the sample size is very large, it can be ascertained that the ZSTD value will always be above 3. Thus, some experts recommend not be able to use ZSTD when the size of the samples are in the large period (N> 500). On par with Sumintono (2015) states that if the item only meets one criterion only, then the item is fit. Consequently, no items were discarded on this analysis. It indicates that each item has been fit with the model of PCM 1-PL, so the items in the integrated assessment instrument are feasible to be used for further analysis. In addition, all the items can be said to be the final product of an integrated assessment instrument that has been empirically valid.

*Item Difficulty*

The difficulty index of each item was conducted to establish an additional measure of the features of the integrated assessment instrument. The difficulty index was computed to get an idea of the proportion of the test takers correctly responding to the item. However, it is important to recognize that aims to measure acquisition of critical thinking and chemical literacy.

In this study, the interpretation of values for difficulty according to Adedoyin and Mokobi (2013) with addition. The items are categorized very easy if the value of b (measure) is more than -1. Items are categorized easy if the value of b (measure) is less than -1. Items are categorized medium if the value of b (measure) at intervals of -0.5< b <+0.5. Items are categorized difficult if b is less than 1. Items are categorized very difficult if b is more than 1**.** The difficulty index for each of the 29 items range from -1.29 to 1.03. An item can be categorized to be good if it has an index of difficulty more than -2.0 logit or less than +2.0 logit **(**Baker, 2001; Hambleton & Swaminathan, 1985: 36). As a results of item difficulty, 29 items are obtained viable in the measurement phase. The results of the item difficulty is presented in Figure 3.

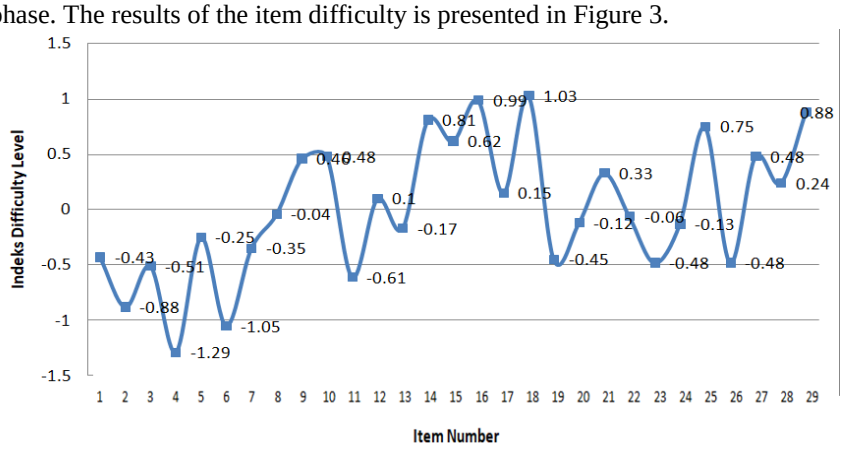


Figure 3
The Graph of Item Difficulty

The figure 4 shows the deployment of index difficulty item in the integrated assessment instrument.

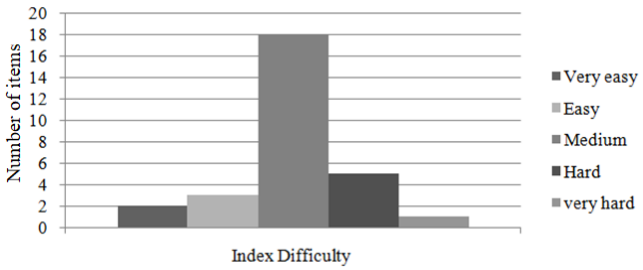


Figure 4
Deployment of Index Difficulty Items

Based on the Figure 4, the deployment of index difficulty items was evenly distributed in the integrated assessment. The well deployment of index difficulty item on an instrument is important because the instrument can measure whether student really understand the chemical equilibrium content in the low and high level.

*The Reliability of The Integrated Assessment Instrument*

The Cronbach's Alpha internal consistency coefficient was found 0.850. This implied that there was 85% certainty of the consistency of the test items in yielding approximately same result repeatedly. This implied that the test was very reliable. The finding was supported by Ceniza and Cereno (2012) who states that if the reliability coefficient within the range of 0.81 to 1.0 indicated high reliability, 0.61 to 0.80 signified a moderate reliability, 0.41 to 0.60 signified fair reliability, 0.10 to 0.40 signified slight reliability, and less than 0.10 signified no reliability. Thus, the integrated assessment instrument reliability was high.

## CONCLUSION

The findings of this study ascertained that the developed integrated assessment has relatively high validity and reliability. Overall, the integrated assessment is suitable for testing the critical thinking and chemical literacy in one test. There are 13 integrated skills of critical thinking and chemical literacy that can be measured by using the integrated assessment. Thus, the researchers believe that the integrated assessment could be used to assess critical thinking and chemical literacy of the chemical equilibrium on third graders in high school.

The valid integrated assessment can be applied in the learning process using the appropriate teaching model and method to swiftly improve student's critical thinking and chemical literacy.

## REFERENCES

Adedoyin, O.O., & Adedoyin, J.A. (2013). Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters. *Herald Journal of Education and General Studies, 2*(3), 107-114.

Adedoyin, O.O., & Mokobi, T. (2013a). Using IRT psychometric analysis in examining the qulity of junior certificate mathematics multiple choice examination test items. International Journal of Asian Social Science, 3(4), 992-1011.

Akgun, A., & Duruk, U. (2016). The investigation of preservice science teacher's critical thinking disposition in the context of personal and social factors. *Journal of Science Education International*, *27*(1), 3-15. Doi: https://files.eric.ed.gov/fulltext/EJ1100164.pdf.

Alghafri, A. S. R., & Bin Ismail, H. N. (2014). The effects of integrating creative and critical thinking on schools student' thinking. *International Journal of Social and Humanity*, *4*(6), 518-525. Doi: 10.7763/IJSSH.2014.V4.410.

Austin, A.C., Ben-Daat, H., Zhu, M., Atkinson, R., Barrows, N., R.Gould, I. (2015). Measuring student performance in general organic chemistry. *Chemistry Education Research and Practice*, *1*(16), 168-178. Doi: 10.1039/C4RP00208C

Baker, F. B. (2001). *The basic of item response theory*. New York: ERIC Clearinghourse on Assessment and Evaluation.

Beavers, A.S., Lounsbury, J.W., Richard, J.K., et al. (2013). Practical considerations for using factorial analysis in Educational Research. Practical Assessment, Research & Evaluation, 18(6), 1-13.

Boone, W.J., Staver, J.R., & Yale, M.S. (2014). Rasch Analysis in the Human Sciences. Dordrecht: Springer Netherlands.

Bowell, Tracy., & Kemp, Gary. (2005). Critical thinking: a concise guide second edition. Taylor & Francis e-Library. Retrieved 22 December, 2017 from http://vct.qums.ac.ir/portal/file/?180488/Critical-thinking-a-concise-guide.pdf

Ceniza, J.C., & Cereno, D.C. (2012). Development of mathematic diagnostic test for DORSHS. Retrieved 22 December, 2017 from http://www.doscst.edu.ph/index.ph[/academics/graduateschool/publication/category/5-volum-1-issue-12012?

Facione, P. A. (2013). *Critical Thinking: What It Is and Why It Counts*. Measured Reasons and The California Academic Press. Retrieved 22 December, 2017 from https://www.nyack.edu/files/CT_What_Why_2013.pdf

Gilbert J.K., Treagust D.F. (2009). Introduction: macro, submicro and symbolic representations and the relationship between them: key models in chemical education. In: Gilbert J.K., Treagust D. (eds) multiple representations in chemical education. Models and Modeling in Science Education, *4*. Springer, Dordrecht. Doi: https://doi.org/10.1007/978-1-4020-8872-8_1

Guller, N., Uyank, G.K & Teker, G.T. (2014). Comparison of Classical Tes Theory and Item Response Theory in Terms of Item Parameter. *European Journal of Research on Education*, *2*(1), 1-6. Doi: http://iassr2.org/rs/020101.pdf

Hambleton, R.K., Swaminathan, H. (1985). *Item Response Theory Principles and Applications*. Boston: Kluwer Nijhoff Publising.

Hambleton, R.K., Swaminathan, H., Rogers, H.J. (1991). Fundamental of item response theory. Los Angeles: Sage Publication, Inc.

Institute of Education Science. (2016). *Highlights* From TIMSS And TIMSS Advanced 2015. Retrieved 22 December, 2017 from https://nces.ed.gov/pubs2017/2017002.pdf

Keil, C., Haney, J., & Zoffel, J. (2009). Improvements in students achievement and science process skills using environmental health science problem-based learning curricula. *Electronic Journal of Science Education*, *13*(1), 1-18. Doi: http://ejse.southwestern.edu/article/view/7782/5549.

Korashy, A.F. (1995). Applying the Rash model to the Selection of items for mental ability test. *Educational and Psychological Measurement*, *55*(5), 753-763. Doi: https://doi.org/10.1177/0013164495055005006.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for intermediate statistics: Use and interpretation*. New Jersey: Lawrence Erlbaum Associates, Inc.

Mapaella, R., & Siew, N. M. (2015). The development and validation of a test of science critical thinking for fifth grades. *SpringerPlus*, *4*(741), 1-13. Doi: 10.1186/s40064-015-1535-0.

Ministry of Education and Culture Indonesia. (2013). *Regulation of ministry of education and culture number 66 in 2013 regarding the standard for education assessment.* Jakarta: Ministry of Education and Culture Indonesia.

Organization for Economic Cooperation and Development. (2006). Science Competencies for Tommorow's Word. Volume 1: Analysis. Retrieved 10 March, 2017 from http://www.oecd.org/pisa/pisaproducts/39703267.pdf

Organization for Economic Cooperation and Development. (2015). Program for International Student Assessment (PISA), Retrieved 10 March, 2017 from https://nces.ed.gov/.

Pacific Policy Research Center. (2010). *21st century skills for students and teachers*. Honolulu: Kamehameha Schools Research & Evaluation Division

Reckase, M.D. (1979). Unifactor laten trait models applied to multifactor test: result and implications. *Journal of Educational Statistics, 4*(3), 207-230. Doi: 10.2307/1164671.

Saribaz, D., & Bayram, H. (2009). Is it possible to improve science process skill and attitudes towards chemistry through the development of metacognitive skills embedded within a motivated chemistry lab?: A self-regulate learning approach. *Procedia Social and Behavioral Science, 1*(1), 61-72.

Shwartz, Yael., Ben-Zvi, Ruth., & Hofstein, Avi. (2005). The importance of involving high-school chemistry teachers in the process of defining the operational meaning of

‘chemical literacy. *International Journal of Science Education*, *27*(3), 323–344. Doi: https://doi.org/10.1080/0950069042000266191.

Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: a simulation study. *Pyschiatry Research*, *188*, 147-155. Doi: 10.1016/j.psychres.

Sumintono, B., Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan.* Indonesia: Trim Komunikata.

Thiagarajin, S., Semmel, D., & Senmel, M. (1974). Instructional Development for Training Teachers of Exceptional Children: A Sourcebook, Retrieved 10 March, 2017 from https://files.eric.ed.gov/fulltext/ED090725.pdf

Toland, Michael D.(2014). Practical guide to conducting an item response theory analysis. Journal of Early Adolescence, 34(1), 120 –151. Doi: 10.1177/0272431613511332.

Tosun, C., & Taskesenligil, Y. (2013). The effect of problem-based learning on undergraduate students’ learning about solution and their physical properties and scientific processing skills. *Chemistry Education Research and Practice, 14*, 36-50. Doi: 10.1039/C2RP20060K.

Tsaparlis, G. (2000). The states-of-matter approach (SOMA) to introductory chemistry. *Chemistry Education Research and Practice*, *1*, 161-168. Doi: 10.1039/A9RP90017A.

Tuysuz, C. (2009). Development of two-tier diagnostic instrument and assess students’ understanding in chemistry. *Academic Journal*, *4*(6), 626-631. Doi: http://www.academicjournals.org/article/article1380558833_Tuysuz.pdf.

U.S. Department of Education. (2016). *Highlights From TIMSS And TIMSS Advanced*. NCES 2017-002.

Watson, Goodwin., & Glaser, Edward M. (2008). *Critical thinking appraisal: Short form*. United States of America: Pearson Education, Inc.

Wu, Q., Zhang, Z., Song, Y., Zhang, Y., Zhang, Y., Zhang, F., LI, R., Miao, D. (2013). The Development of Mathematical Test Based on Item Response Theory. *International Journal of Advancements in Computing Technology*, *5*(10), 209-216.

Yilmaz, Kursad., Altinkurt, Yahya., & Cokluk, Omay. (2011). Developing the educational belief scale: the validity and reliability study. Educational Sciences: Theory & Practice, 11(1), 343-350.

Zoller, U. (2001). Alternative assessment as (critical) means of facilitating HOCS-promoting teaching and learning in chemistry education. *Chemistry Education Research and Practice in Europe, 2*(1), 9-17. Doi: 10.1039/B1RP90004H.