JPII 8 (4) (2019) 570-581

Jurnal Pendidikan IPA Indonesia

http://journal.unnes.ac.id/index.php/jpii

EXPLORING AND COMPARING CONTENT VALIDITY AND ASSUMPTIONS OF MODERN THEORY OF AN INTEGRATED ASSESSMENT: CRITICAL THINKING-CHEMICAL LITERACY STUDIES

S. Sadhu*[1], E. Ad'hiya[2], E. W. Laksono[3]

[1]Summit Institute of Development, Indonesia
[2]Department of Teacher Training and Education, Universitas Sriwijaya, Indonesia
[3]Department of Chemistry Education, Faculty of Mathematics and Science, Universitas Negeri Yogyakarta, Indonesia

DOI: 10.15294/jpii.v8i4.20967

Accepted: August 18th, 2019. Approved: December 27th, 2019. Published: December 31st, 2019

ABSTRACT

The development of chemical literacy as well as critical thinking, are a prominent objective of science education and become essential skills in the 21st century. The stress has not been given to the measurement of both skills together in chemistry content in high school. Another problem shows in developing an instrument in which teachers often do not know whether the content and construct that they developed in the tool can measure the skill that supposed to be measured. The first step the teacher should do is studying the content validity when an instrument has been constructed and also psychometric testing is needed. Hence, this is addressed for exploring the content validity evidence, even assumptions for the modern theory in integrated assessment for measuring students' chemical literacy and critical thinking. Also, comparing a result of content validity and assumptions for modern theory to know which is mandatory for research instrument. The initial integrated assessment consists of 37 items. Content validity was first piloted under a review of six experts and also 133 participants involved to determine the assumptions of modern theory in integrated assessment. From this study, it can be confirmed that the development of the integrated assessment on the content validity can be used to measure 13 integrated skills of chemical literacy and also critical thinking in chemical equilibrium, and an assumption for modern theory met with the standard of the valid instrument. In general, to know the quality of the items test using more sophisticated statistical analysis, the integrated assessment is appropriate for measuring the big-scale of the paper-pencil test. The outcome of this research will help teacher/lecturer in the analysis quality of an assessment that will be used in student's evaluation

Keywords: assumption of modern theory, chemical equilibrium, chemical literacy, content validity, critical thinking

INTRODUCTION

Nowadays, the primary activator as an effort to reform teaching and science education in the entire levels is by developing student's capacity of high order thinking skills (HOTS) in the era of complex science as well as technology-based society. Developing students' ability to think critically in every aspect of life is the primary goal of future-oriented science, contemporary, and chemical education. In pair with it, Shakirova (2007) stated that critical thinking is crucial because it allows students to effectively handle the problems also to make their contribution directly to society. Critical thinking belongs to essential and promi-

***Correspondence Address**
E-mail: satya0185pasca2016@student.uny.ac.id

nent skills since it is compulsory for every single person in workplace leadership, decision making, professional success, and clinical judgment. Therefore, critical thinking is the essential skill that needs to teach and train toward students. Yet, a lot of teachers deal with the problem of how to engage students in critical thinking activities. Also, students rarely think critically to find a solution for a real-world and challenging problem (Bartlett, 2002). It indicates that students rarely apply the skill of critical thinking in the classroom or toward the test. An appropriate instrument needs for the students to improve critical thinking. However, there are only some pieces of evidence that showed tests could be used to assess and train students' skills in thinking critically nowadays. Almost the entire struggle was because of the lack of instrument for measuring high school student's ability to think critically when learning science in the specific content, for instance, chemistry in high school. It aligns with Fensham & Bellocchi (2013) that the learning outcomes assessment in the science majority still focuses only on low-level thinking skills (LOTs).

Critical thinking skill needs to be implemented in real life. Regarding it, the knowledge and the ability to solve real-world problems is called scientific literacy (Bond, 1989). Experts argue one of the significant objectives in current science education reforms is by training students to enable them understood science literacy (DeBoer, 2000; Norris & Phillips, 2003; Shwartz et al., 2006; Show-yu, 2009). Critical thinking is closely related to the phenomenon in everyday life (Fives et al., 2014; Turiman et al., 2012; Taylor, 2012). The chemistry subject aims to build student's science literacy, for instance. Therefore, it is assumed that both critical thinking and science literacy are closely related, particularly in the chemistry subject in high school (Cigdemoglu & Geban, 2015). It has now been suggested that science literacy, for instance, chemical literacy is essential for a student. Both chemical literacy and chemistry learning have a direct relationship and application in everyday life and enable the student to become a more informed citizen (Show-yu, 2009). A fundamental problem with much of the literature regarding chemical literacy is only some evidence that showed tests could be used to assess and train students' skills in thinking critically.

Recently researchers have tended to focus on improving the skill of students to think critically and improve students' chemical literacy separately. Indeed, the study highlighted that both critical thinking and chemical literacy are related (Groser & Nel, 2013). Consequently, researchers believe that improve both skills together in a single instrument will make a significant result for the student. Chemistry subjects in senior high school enroll and can be found in daily activities/issues. Thus it is necessary to improve the students' critical thinking and chemical literacy. Relevant chemistry content in high school is needed to combine well both skills. One of the chemistry contents in high school that requires conceptual involved daily life and algorithmic understanding that will be matched to blend the critical thinking and chemical literacy is chemical equilibrium. Chemical equilibrium consisting of five sub-contents, which are the student is compulsory to comprehend the concepts of dynamic equilibrium , equilibrium constants, equilibrium shifts, and equilibrium in the industry and also understand mathematical calculations for determining the value of equilibrium constants. Also, A lot of researchers found that in high school, chemical equilibrium belongs to one of the crucial topics but is also tricky in chemistry content to teach (Bergguist & Heikkinen, 1990; Camacho & Good, 1989). Comprehending chemical equilibrium concepts means an understanding of further concepts; for instance, the topic of oxidation-reduction, reactions, acid-base, and also solubility (Bergguist & Heikkinen, 1990). Therefore, in this study, it is necessary to develop an integrated assessment instrument for measuring critical thinking and chemical literacy in chemical equilibrium.

Both students' critical thinking and also students' chemical literacy should improve; however, there is still a lack of development of integrated assessment to assess student's critical thinking and chemical literacy until today. The success of the assessment and training students' skills in thinking critically and comprehend chemical literacy can be seen in the assessment when conduct on test-takers. The developed instrument that uses in the learning process must be able to measure the desired skill because an instrument can be called as valid when it can measure what should be measured (Mujis, 2011). Precisely, the ultimate tool can accurately measure any prescribed variable, or it measures what should be measured. There are four types of validity; criterion, face, content, and construct validity (Jackson, 2003; Mujis, 2011). One of the validities that are used to examine the developed instrument is content validity. In contrast, content validity is looking at the content of items, whether it measures the concept being measured in the study.

Consequently, content validity is an essential examination to know whether the skills

of critical thinking and chemical literacy are contained and given what is being measured in the integrated assessment instrument. The most striking problem to emerge that content validity examination is still not profoundly understood by teachers in developing a tool. It can be stated that content validity is one source of evidence that allows the teacher to make claims about what a test measure. It is the degree to which the content of a test is representative of the domain it is intended to cover, like critical thinking skill and chemical literacy. Several authors reported the process of content validity measuring in their articles. Most of them are in developing an instrument, while others did not. Indeed, measuring and reporting the content validity of the tools is crucial. This kind of validity is also enabling readers and researchers to ensure construct validity and give confidence about their developed instruments. Therefore, it is essential to know whether the content validity in the instrument meets the standards that already planned in measuring students' skills in thinking critically and chemical literacy comprehension. Item Response Theory (IRT) is a modern theory based on the relationship between individuals' performance on a test item and the test takers' level of performance on an overall measure of the ability that the item was designed to measure skill. IRT needs three assumptions, namely a *unidimensional* trait denoted, local independence of item, and parameter invariance (Qasem, 2013). Regarding the assumptions of modern theory in this study, researchers believe that it is a kind of construct validity for research instruments. It is because the assumptions of modern theory addressed to determine if the item is independent if the item's characteristics are independent towards the student's characteristics (item parameter invariance assumption), and if the characteristics of the students in the test are independent towards the distribution of the item's characteristics (ability parameter invariance assumption). The main limitation of recent research is only focused on the content validity that involves experts only rather than test-takers or students. There is still uncertainty concerning exploring and comparing the content validity and assumption of modern theory as long as they validate the same aspects. Another point worth noting is there is a slightly match between content validity and assumption of modern theory. Content validity consists of substance, language, construct, and appearance aspect, which is almost the same as the assumption of contemporary theory that includes element and constructs. It has been suggested by Jin & Jeon (2018) in their paper that local dependence among people can be useful for test construction. In this study, the assumption of modern theory believed to construct validity.

In conclusion, this study aims to explore the content validity evidence as well as assumptions of modern theory in integrated assessment for measuring students' critical thinking and chemical literacy. The researcher also compares the result of content validity and assumptions of contemporary philosophy to know which is mandatory for research instruments.

METHODS

Participants

For exploring the content validity, this study uses the definition from the 'expert' of the word: they are professional and field experts as a panel of experts (Rubio et al., 2003). Professional experts help determine whether the integrated assessment is well constructed used for examining students' skills in thinking critically, comprehending chemical literacy, and also understanding the actual concept of the chemical equilibrium. Field experts are the teacher conducting the learning process in the classroom. It helps determine whether the integrated assessment is appropriate with the learning indicators and materials to be taught. Also, the field experts help to obtain advice that binds to the visibility and effectiveness of the integrated assessment. The experts in this study were two chemistry lecturers from the chemistry department as professional experts who experienced more than ten years and four field experts. The sampling technique used was purposive sampling, which is judgment sampling. This study used a purposive sampling technique: judgment sampling. In selecting a panel of professional and field experts were based on the following criteria: expertise in chemistry and education, academic qualification, and experiences.

In order to explore the assumptions of the modern theory, a total of 133 students of 12th grade took part comprising three high schools in the Special Region of Yogyakarta, Indonesia. The participants in each school selected using purposive sampling techniques. The purposive was used to select the participants by grade/level of the academic qualification to verify whether the integrated instrument be able to measure students in low, medium, and high levels of educational qualification.

Procedure

Two procedures were involved in this study to get the data following are content validity to expert and test to the students.

The content validity of the items was proved by presenting them to the chemistry experts. In line with the statement of Gregory (2007), she states that content validity examination can do by asking an expert in the field to be examined to provide an assessment of the items that have been made. Based on Lynn (1986), researchers have calculated two kinds of content validity. They were the content validity from individual items and the content validity from the overall scale. In this study, the content validity of personal items was performed. In this study, the process could be broken down into four steps.

First, choosing two professional experts based on consideration, that is an expert in the subject matter in the chemistry field and in evaluation, especially for integrated assessment.

Second, choosing four field experts based on the expertise and experiences in teaching chemistry in high school. Both professionals and field experts were asked to review each item, which is related to the overall integrated instrument objective. Particularly, the content, experts were asked to review every single item against the following criteria: substance, construct, language, and appearance. The content validity questionnaire is presented to the expert as the assessment sheet to the initial product of integrated assessment.

Third, calculating and analyzing the result of the experts

Fourth, doing some necessary revisions based on the comments and opinions from the experts.

The content validity was analyzed using Aiken's V Content Validity Index. Modern theory assumptions were collected from 133 student's responses that involved and analyzed using the modern theory calculation.

Research Instrument

The Integrated Assessment

The integrated assessment was designed to prove the aspect of quantitative. In this study, it was developed using the 4-D model by Thiagarajan et al. (1974). The integrated assessment was developed within three stages because it focused on showing the steps that should be considered in the process of obtaining content validity evidence in the assessment instrument and exploring the assumptions of the modern theory. The stages involved define, design, and develop.

Accomplished the construction of the instrument, the instrument of integrated assessment consisted of 37 open-ended multiple-choice questions, with five alternatives each one lettered A-E out of which only one was the correct as of the accepted answer. The more options within will increase reliability (Thorndike & Thorndike-Christ, 2010) and reduce success in guessing the answers (Dehnad et al., 2014).

The sample item of the integrated assessment instrument construct is described below to illustrate the item format. The integrated assessment was given to the students to determine the assumption of modern theory.

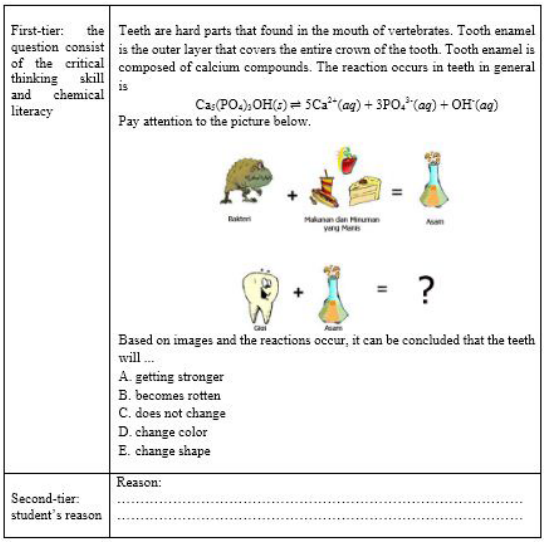

Figure 1. The Item Construct and Format of Integrated Assessment

The Content Validity Questionnaire

The content validity questionnaire is presented, two professional experts and four field experts. The content validity questionnaire contains four areas. The content validity questionnaire is the assessment sheet for the integrated assessment to determine content validity. The four areas are (a) the substance of the items to the essential competencies and its indicators to be achieved; (b) construction of the information which presented in the items included clarity of the words/phrases/diagrams; (c) Correctness of Indonesia grammar rules on each items; and (d) the appearance of the item. Professional and field experts were asked to review every single item by giving an initial score. The first score from each item will be categorized again using the scoring guide. The content validity question-

naire consists of two categories of the decision; the following are valid and not valid. If the item is authentic, this indicates that the item meets the standard of the content validity. The content validity questionnaire contains 13 statements to assess the initial product of integrated assessment. The briefed item validation sheet component is presented in Table 1.

Table 1. The Areas of Content Validity Questionnaire

Areas of the Content Validity	Statements
Substance	The item in the integrated assessment instrument is in accordance with the basic competency and indicator to be achieved in the learning process
	The item in the integrated assessment instrument is in accordance with the integrated skill to be achieved (critical thinking and chemical literacy)
	The key reason for the answer to the presented item in the scoring guide is in line with the substance/material in chemical equilibrium
Construct	The item is formulated clearly and does not give direction to the correct answer
	The alternative choice is logic in terms of material and the length is relatively same
	The question does not depend on the answer to the previous question
Language	Each statement in the integrated assessment is written using good and correct Indonesian language rules
	Each statement in the integrated assessment instrument uses a communicative language according to the age of the learners development in high school
	The using of language does not lead to multiple-interpretations, does not use figurative words, and easy to understand
Appearance	Each item provides enough space to write down the answers
	The type and size of letters on the item are used proportionally, clearly legible, and do not disturb the appearance
	The drawings, graphs, tables, diagrams, discourses, and the like contained in the item have a function as explanatory

Data Analysis

To explore the content validity, after review by the experts, data were tabulated in Microsoft Excel 2010 and analyzed from reflective reading and descriptive statistics. The content validation was accomplished with the formula of Aiken's V Content Validity Index to verify the level of agreement among the experts as judges according to the presence or not of the items of the integrated instrument. To evaluate the instrument by Aiken's V, the form used in this study was the average of the items calculated separately. Thus, it was added all Aiken's V figured independently and divided by the number of the item on the instrument about the criteria of the tool. The guidance of the scoring for the item was written to determine the score. The guidance of the scoring is described in Table 2.

Table 2. Scoring Guide

Evidence	Score
If three statements are met	3
If two statements are met	2
If only a statement is met	1
No statement is met	0

The V coefficient for the item was computed using the Aiken's V formula.

$$V = \sum S/[n(c-1)]$$

To stipulate the acceptable rate of agreement among the judges, considering the value based on the statistical significance of V. The discrete, right-tail probabilities associated with selected amounts of V for 2 to 7 categories (c) and 2 to 25 items (m) or raters (n) are given in

the table of Aiken's V (1985). The two V values (for each c and m or n) selected for inclusion in the table are those having right-tail probabilities close to but no higher than the 0.01 and 0.05 value, respectively. It short, to know the statistical significance of V, it can be determined by correlating the number of rating category (c) with the number of raters (n). In this study, the number of assessors (n) was six raters with three rating categories. Therefore, in the 0.05 level, the value of V for Aiken's V per item is 0.83. It was established the recommended amount of 0.83 as the minimum to serve as a decision criterion on the appropriateness and acceptance of each item. It indicates if the value of the items in the integrated assessment is higher than the level of agreement of content validity index (0.83) which can be concluded that the item is appropriate with the standard of the content. The Item CRI score is less than 0.83, which means the item was either not relevant to the thematic domain, nor it required verbiage revision to delete ambiguity and also ensure an accurate response. After analyzing the data, the instruments have been reformulated in accordance with the guidelines and suggestions of the judges. Criteria assessed in the integrated assessment include the substance, construction, language, and appearance of integrated assessment instruments.

The assumptions of the modern theory are from the collected student's response. The data then were scored by scoring guidelines by Bayrak (2013). Scored responses were analyzed by using Statistical Package for the Social Sciences (SPSS) Version 17.0 program, WINSTEPS, and Microsoft Excel. SPSS was used to get the total variance explained to perform the unidimensionality assumption. WINSTEP was used to get the score of measure, model S.E, infit, outfit, PT-measure, and the exact match for each item to perform the independence local and parameter invariance assumption. Microsoft Excel was used to facilitate to get the covariance matrix for independence local and correlation between scattering for parameter invariance.

RESULT AND DISCUSSION

The Content Validity

The demographic profile for professional experts is (N=2). The area of expertise covers evaluation in integrated assessment (1.50%) and subject matter content in chemistry in chemical equilibrium (1.50%). All of them are from the university as a lecturer. For the field experts (N=4) distribution shows Male (1.25.0%) dominated female experts (3.75.0%). It has the same area of field expertise in learning and teaching chemistry in high school. Content validity examination aims to produce a valid instrument in terms of content so it can be used to measure critical thinking and chemical literacy. The value of Aiken's V in every item is clarified in Figure 2.

As shown in Figure 2, research has found that the score of Aiken's V for each item in each area was more than 0.83. Based on recommendations from analysis data, the minimum level of agreement between six experts at $\geq$ 0.83 at 0.05 level of significance was set. It means that five out of the six experts for content validity must agree for the items to be part of the final instrument. A favorable rating by six of the experts yielded the score CRI of greater than 83%, or 0.83 denoted a high-level agreement is a high value. It meant that if the significant majority of the experts' opinions agree, items were considered relevant to concepts being investigated, which is the integrated skill of critical thinking and chemical literacy. It underlines that all items in the integrated assessment instrument in the chemical equilibrium proved valid in the content validity. The greatest result to emerge from the data is that all the items in the integrated assessment instrument can measure students' critical thinking and chemical literacy skills together in chemical equilibrium.

As shown in Figure 2 (a) in the area of substance, what we are unable to account for is that the fact that the score in three items almost near with the minimum score (0.83). It attributed to the wrong in writing the equation. The majority of experts felt that there was a wrong in writing the equation in the items. The experts stated that

"as shown in item number 3, you writing the wrong equation partially, this is the wrong of written the phase…

$$Ca_5(PO_4)_3OH_{(s)} \rightleftharpoons 5Ca^{2+}_{(aq)} + 3PO_4^{3-}{}_{(aq)} + OH^-_{(aq)}$$

It is important that we identify the phase of the reactants and products in a chemical reaction. To show these phases in the reaction, you have to use the same size of phase in the chemical equilibrium. Using round brackets and write the phase within it. In addition, the phase should be in italic. The right way to writing the equation as following"

$$Ca_5(PO_4)_3OH(s) \rightleftharpoons 5Ca^{2+}(aq) + 3PO_4^{3-}(aq) + OH^-(aq)$$

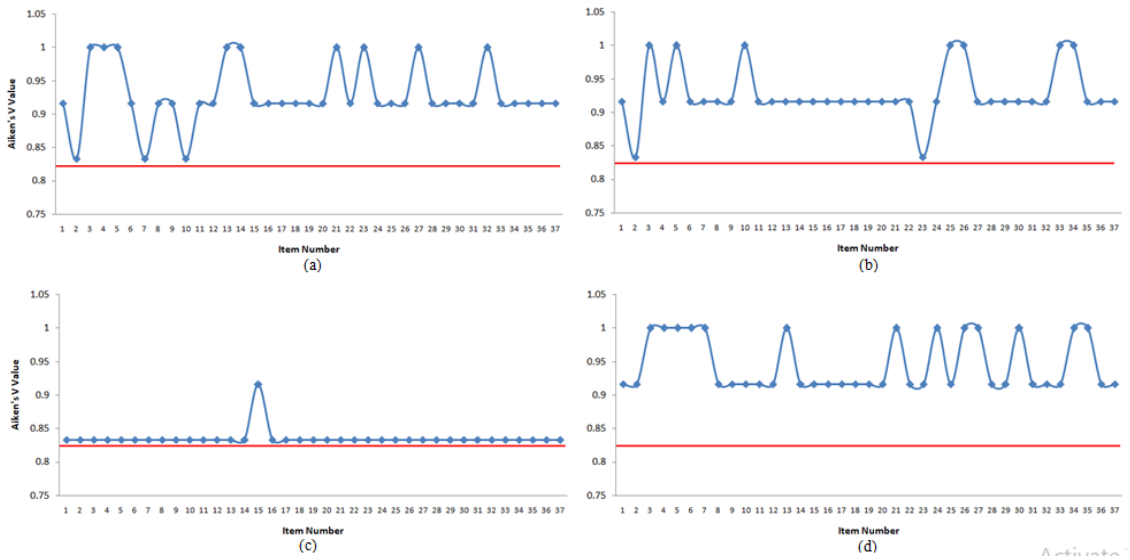

Figure 2. The Score of Content Validity in Area of (A) Substance, (B) Construct, (C) Language, and (D) Appearance

The experts argue the concurs well with Karpudewan et al. (2015) in his study that shows the equation writing in chemical equilibrium. As shown in Figure 2(c) in the area of language, it showed that the score of content validity in each item almost near with the minimum score (0.83) to serve as a decision criterion on the appropriateness and acceptance of each item. However, they still need improvements for some numbers based on expert comments and suggestions (Medeiros et al., 2015). Our result was inadequate. However, this is not particularly surprising because the language in the integrated assessment was slightly hard to understand for the student in high school because it consisted of chemical literacy. As a result, the experts gave a low score and recommendations to modify and improve the language of the item. It supports in the literature that the use of right and correct grammar under the level of test-takers is essential because the test takers will be easy to understand the contents and command of the questions that enable them to answer according to the right answers that have been set. It supports previous findings by Xie (2018) who argued that language is essential when developing an instrument includes length and vocabulary of the sentences that should appropriate with the characteristic of test-takers.

Both quantitative and qualitative were obtained in the content validity examination. The qualitative data included comments and suggestions. Experts' advice was expected to refine the integrated assessment instrument. The comments of the experts are briefly shown in Table 3 to help and improve how to develop an ethical construct and content in the designing of an integrated assessment instrument. The comments refer to the suggestion of the experts to enrich the quality of the item.

Table 3. The Suggestions in the Developing Integrated Assessment

Aspect	Suggestions
Substance	In writing down the formulation of learning indicators should begin with the operational verb
	Improving the quality of the answer to fit in well with the complete concept. Concerning with it, the answer in the assessment rubric should complete with the variance of possibility of student's response
	Reviewing the equation of the reaction that written on the stem. Making sure that the equations of the chemical reaction is written in the right way because the errors in the writing the symbols, phase, etc gives a different meaning in chemistry
Language	Need to do a correction in sentence structure
	Using words that match with the level of test takers so they can understand the meaning on the questions clearly
	Improve the sentence structure to be more consistent
Appearance	Enlarge the answer space especially for questions which need the calculation as solving

The involvement of expert lecturers in the validation of the item aims to obtain advice relating to the area of expertise. The teacher's participation in the validation of the item's goal is to get information that binds to the visibility and effectiveness of product implementation. Other suggestions that are expected are the completeness of the item about the learning indicators, the determination of time allocation, the number of items that are ideal to be tested to the students, and instructions for use.

Wynd et al. (2003) argued that the content validity of an instrument often discloses through a qualitative expert review, but quantitative analysis of review agreements is also recommended. Two quantitative approaches are used by them to develop the scale of Osteoporosis Risk (ORAT) i.e., the Content Validity Index (CVI) and the multilateral kappa coefficient. In par with it, Rubio et al., (2003) stated in his article provide a simulation of how to test the content validity. Although content validity is subjective by using several techniques, for instance, integrated agreement, Content Validity Index (CVI), and factorial validity index can improve the objectivity of test results. Content validity is still essential to be used as an initial step in developing an instrument. While there is no expert agreement on the techniques employed, recent research generally uses a combination of qualitative methods (interview, FGD, expert opinion) to obtain adequate content validity. In a nutshell, the characteristic makes the content validity very robust in that it eliminates ambivalence and allows straightforward interpretation, which helps in constructing more reliable and valid data concerning content validity in the research instrument.

The Assumptions of Modern Theory

The Unidimensionality Assumption Test

The unidimensionality assumption is the first assumption. Reckase (1979) states the unidimensionality assumption is an underlying assumption that a test measures only one variable or one dimension. It assumes that a set of a test measure only one latent trait (Kyung, 2013). In this study, the unidimensionality assumption reveals the instrument only measures a dimension, i.e., the aspects of student's integrated skill of critical thinking and chemical literacy in chemical equilibrium materials. The method used to assess unidimensional in this study was exploratory factor analysis. It was to determine how many factors existed among all items (Adedoyin & Mokobi, 2013), while an analysis of exploratory factor is more appropriate when the scale is relatively unexplored in terms of factor (Toland, 2014). Additionally, It is to extract the new factor structure and to examine the construct validity (Hyunho-Kim et al., 2016). This factor would represent the construct underlining the integrated skill of critical thinking and chemical literacy measured by the exami

The result of the unidimensionality assumption revealed that an integrated assessment instrument in chemical equilibrium contains nine eigenvalues that have a score higher than 1,000. It shows that there are nine factors formed. These nine factors can cause about 64.55% of the total variance with a dominant factor. One of the dominant factors out of 9 eigenvalues revealed that the integrated assessment instrument met with the unidimensionality assumption, which is the instrument only measure one dimension. The dominant factor was 20,238% out of 64.55% (a total of 9 elements). According to Comrey and Lee (Yilmaz et al., 2011) in factor analysis that the factor loading value is considered to be in an excellent category if it has a value of 0.71 (which explains 50% variance), the class is quite useful if the value 0.63 (which explains 40% variance), good category if the value is 0.55 (which explains 30% variance), average or moderate category if the value is 0.45 (which explains 20% variance) and bad category if the value is 0, 32 (which explains 10% variance. Therefore, 20,238% obtained in this study are classified into the average grade.

Hambleton & Swaminathan (1985) stated that unidimensionality assumptions are challenging to fulfill, ideally. Eleje et al. (2018) also state that since individuals' cognitive and personal characteristics influence test performance and cannot often be controlled, but it is not always possible to meet this assumption. Therefore, unidimensionality assumptions can be considered met the requirement if the instrument has a dominant component in measuring students' abilities (Adedoyin & Adedoyin, 2013; Guller et al., 2014; Hambleton et al., 1991; Wiberg, 2004; Wu et al., 2013). In par with it, Rijn et al. (2016) also state that one dominant factor is evidence that constructs the validity of the test is ensured. This finding correlates satisfactorily with the result in the content validity in the aspect of substance. The experts argued that all the item in the integrated assessment instrument is in accordance with the integrated skill to be achieved (critical thinking and chemical literacy). The most significant result to emerge from the data is that the result

in the content validity of the experts and the unidimensionality assumption by the students was significant.

The Local Independence Assumption Test

The local independence assumption test addressed to see whether the student's ability is independent of each item. The local independence assumption indicates that if the ability to affect the performance of the test has been deemed unchanged or constant, so the student's response towards each item is not statistically related. In other words, it can be stated that each item does not correlate with each other.

The result of the local independence assumption analysis is obtained through the variance-covariance matrix of the person measure. The local independence assumption is satisfied if the score below the diagonal line on the variance-covariance matrix is under zero. The result of the analysis of the presumption of local independence in the terms of the covariant matrix of person measure was shown in Table 4.

Table 4. The Local Independent Assumption in Covariant Matric of Person Measure

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10
Column 1	0.509734									
Column 2	-0.09803	0.023323								
Column 3	0.144978	-0.02081	0.152688							
Column 4	0.308564	-0.03076	0.450183	1.488029						
Column 5	0.225562	-0.03911	0.178619	0.519164	0.354395					
Column 6	0.352677	-0.04688	0.381605	1.22292	0.643332	1.333061				
Column 7	-0.00133	-0.00187	-0.03476	-0.11704	-0.06695	-0.14292	0.025416			
Column 8	0.079224	-0.01521	0.022536	0.045553	0.035113	0.05327	8.96E-05	0.012793		
Column 9	-0.14147	0.023462	-0.05449	-0.14431	-0.07887	-0.14798	0.006257	-0.02269	0.048648	
Column 10	-0.13206	0.023352	-0.04086	-0.08891	-0.06004	-0.09708	0.000668	-0.0215	0.040119	0.037442

Parameter Invariance Assumption Test

The parameters invariance assumption consists of two parts. They are item parameter invariance and ability parameter invariance. Item parameter invariance is an assumption to determine whether the item's characteristics are independent of the student's personalities. The ability parameter invariance is an assumption to decide whether the students' features in the test are independent towards the distribution of the item's attributes. It merely states that in item parameter invariance, the characteristics of an item will not only change because they are tested on a different group of participants. In the ability parameter invariance, the ability of student will not merely change because they are working on the various index difficulty levels (Duskri et al., 2014). Another point worth noting is Rasch modeling has advantages, which are "sample free" and "independent item." Sample free is defined as the characteristics of the item parameters that are not influenced by the sample. The independent item is defined as the item characteristics that are not dependent on the participants within the test.

The item parameter invariance assumption is evidenced by estimating item parameters in different groups of participants. In this study, the calculation of data of parameter assumption was grouped into two groups, namely the odd number of participants and even the number of participants. Based on the estimation of both groups, the scatter diagram and the correlation coefficient was determined and presented. If the linear line equation obtained has a high correlation, then it can be stated that the item parameter invariance assumption has met. After analyzing the item parameter invariance assumption, later the ability parameter invariance assumption was evidenced by estimating the ability parameters in different items. In this study, the calculation of data to obtain the ability parameter invariance assumption

were grouped into two groups, namely the odd item number and the even item number. Based on the estimation of both groups, the scatter diagram and the correlation coefficient was determined and presented. If both groups of items are highly correlated, then; as a result, the ability parameters have met the ability parameter invariance assumption. The results of the item and ability parameter invariance assumptions were shown in Figure 3.

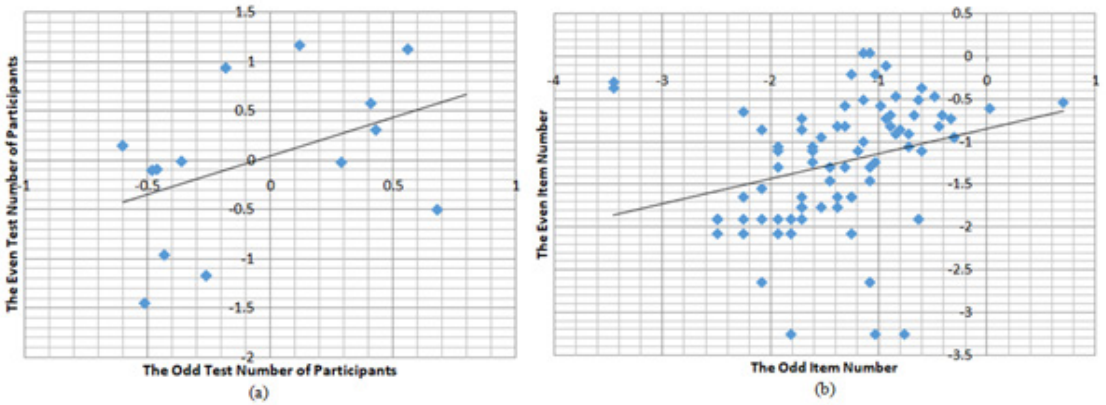

Figure 3. The Result of (a) Item Parameter Invariance Assumption and (b) Ability Parameter Invariance Assumption

As shown in Figure 3a, it can be seen that there was a positive significant among the groups. It can be seen in the number of approaching points on the linear line. As a result, the item parameter invariance assumption has met. It reveals that even though the integrated assessment instrument will be tested on the different students, later on, the item's characteristics will not change. As shown in Figure 3(b), it can be seen that there was a positive significant among the groups. It can be seen in the number of approaching points on the linear line. As a result, the ability parameter invariance assumption has met. It reveals that the participant's ability does not change while working on the different index difficulty levels.

In general, after exploring the content validity and assumption of modern theory as construct validity, our study provides considerable insight into the comparison of two validities that are undertaken to the research instrument. To sum up, the content validity is highly recommended for research instruments because the professional and field experts can give a recommendation for improving the quality of the substance, construct, language, and appearance of the research instrument before the instrument will be tested to the students. An increasing number of studies have found that content validity is the crucial factor for identifying measuring concepts, even though it is not an adequate indication that the instrument measures what should be measured (Yaghmale, 2003). The assumption of modern theory as construct validity is mandatory for research instruments because it represents the item construct of the student's response. Assumptions of modern theory reflect the quality of the tool because students involved to determine it. If the elements of the test do not match the primary construct, then the test instead measures something else, so it is ultimately creating potential bias. As a result, the student's response reveals the real construction of a research instrument, and in fact, the tool will be administrated to the student. In par with it, Taherdoost (2016) outlines a summary of the comparison of validity that is undertaken from several experts in which content validity is highly recommended, and construct validity is mandatory for research instruments.

In short, content validity could contribute to support the construct validity of an instrument (Yaghmale, 2003). Using a single approach is not enough. Yet, it is necessary to test many kinds of strategies before those were applied in the study. By reporting and exploring the instrument's content validity, the reader can easily understand the process of measuring it. The interpretation of the results will be more precise if content validity has been measured.

CONCLUSION

In conclusion, as much as 37 open-ended multiple-choice item questions required refinement, thus showing that the items were built with a good operationalization and conceptualization in the content validity. In short, the whole items of integrated assessment can measure thirteen integrated skills of student's critical thinking and chemical literacy skills on chemical equilibrium.

The local independence assumption reveals that the complete item is independent and does not depend on the answer to the previous item. The Item parameter invariance assumption shows that even though the integrated assessment instrument will be tested on different student, later on, the item's characteristics will not change. The ability parameter invariance assumption reveals that the participant's ability does not change while working on the different index difficulty levels.

This study has highlighted that content validity is highly recommended for research instruments. The assumptions of the modern theory are mandatory for research instruments. It is necessary to carry out further analysis of the tool by statistical methods such as item response theory model. The researchers suggest that 37 items that were refined would undergo a pilot study by using the IRT model. Through the IRT model, the items were selected after some due consideration, such as item fit, index difficulty, and reliability.

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. (2013). 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.

Bartlett, J. E (2002). Analysis of Motivational Orientation and Learning Strategies of High School Business Students. *Business and Education Forum*, *56*(4), 18-23.

Bayrak, B.K. (2013). Using Two-Tier Test to Identify Primary Students' Conceptual Understanding and Alternative Conceptions in Acid Base. *Mevlana International Journal of Education, 3*(2), 19-26.

Bergquist, W., & Heikkinen, H. (1990). Student Ideas Regarding Chemical Equilibrium: What Written Test Answers Do Not Reveal. *Journal of chemical Education*, *67*(12), 1000.

Bond, D. (1989). In Pursuit of Chemical Literacy: A Place for Chemical Reactions. *Journal of Chemical Education*, *66*(2),157-160.

Camacho, M., & Good, R. (1989). Problem Solving and Chemical Equilibrium: Successful Versus Unsuccesful Pcrformance. *Journal of Research in Science Teaching*, *26*, 251-272.

Cigdemoglu, C., & 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.

DeBoer, G. E. (2000). Scientific literacy: Another Look at Its Historical and Contemporary Meanings and Its Relationship to Science Education Reform. *Journal of Research in Science Teaching*, *37*(6), 582-601.

Dehnad, A., Nasser, H., & Hosseini, A.F. (2014). A Comparison between Three-and Four-Option Multiple Choice Questions. *Procedia-Social and Behavioral Sciences, 98*,398-403.

Demars, C. (2010). *Item Response Theory.* New York: Oxford University Press.

Duskri, M., Kumaidi., & Suryanto. (2014). Pengembangan Tes Diagnostik Kesulitan Belajar Matematika di SD. *Jurnal Penelitian dan Evaluasi Pendidikan, 18*(1), 44-56.

Eleje, L. I., Onah, F. E., & Abanobi, C. C. (2018). Comparative Study of Classical Test Theory and Item Response Theory Using Diagnostic Quantitative Economics Skill Test Item Analysis Results. *European Journal of Educational and Social Sciences*, *3*(1), 71 – 89.

Fensham, P.J., & Bellocchi, A. (2013). Higher Order Thinking in Chemistry Curriculum and Its Assessment**.** *Thinking Skills and Creativity, 10,* 250–264.

Fives,H., Huebner, W., Birnbaum, A.S., & Nicolich, M. (2014). Developing a Measure of Scientific Literacy for Middle School Students. *Science Education*, *98(4),* 549–580.

Gregory, R.J.(2007). *Psychological testing: history, principles, and applications (5th edition)*. New York: Pearson Education Group, Inc.

Grosser, M.M. & Nel, M. (2013). The Relationship Between the Critical Thinking Skills and the Academic Language Profiency of Prospective Teachers. *South African Journal of Eduvation*, *33*(22), 246-262.

Guller, N., Uyank, G.K., & Teker, G.T. (2014). Comparison of Classical Test Theory and Item Response Theory in Terms of Item Parameter. *European Journal of Research on Education*, *2*(1), 1-6.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory*. New York: Kluwer Inc.

Hambleton, R., K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamental of Item Response Theory*. Los Angeles: Sage Publication, Inc.

HyunhoKim., Boncho-Ku., Yeol-Kim, J., Young-Jae, P., & Young-Bae,P. (2016). Confirmatory and Exploratory Factor Analysis for Validating the Phlegm Pattern Questionnaire for Healthy Subjects. *Evidence-Based Complementary and Alternative Medicine*, *2016*(2016), 1-8.

Jackson, S. L. (2003). *Research Methods and Statistics, a Critical Thinking Approach*. USA: Thomson Wadsworth.

Jin, I. H., & Jeon, M. (2018). A Doubly Latent Space Joint Model for Local Item and Person Dependence in the Analysis of Item Response Data. *Psychometrika, 83*(333) 1-41.

Karpudewan, M., Treagust, D. F., Mocerino, M., Won,

N., & Chandrasegaran, A. L. (2015). Investigating high school student's understanding of chemical equilibrium concepts. *International Journal of Environmental and Science Education, 10*(6), 845-863.

Kyung, T. H. (2013). Windows Software that Generates IRT Parameters and Item Responses: Research and Evaluation Program Methods (REMP). Retrieved 10 March, 2017 from https://www.umass.edu/remp/software/simcata/wingen/homeF.html

Lynn, M. R. (1986). Determination and Quantification of Content Validity. *Nursing Research*, *35*, 382–385.

Medeiros, R.K.S., Junior, M.A.F., Torres, G.V, Vitor, A.F., Santos, V.E.P., & Barichello, E. (2015). Content Validity of an Instrument about Knowledge on Nasogastricintubation. *Bioscience Journal, 31*(6), 1862-1870.

Mujis, D. (2011). *Doing Quantitative Research in Education with SPSS.* London: SAGE Publications, Ltd

Norris, S. P., & Phillips, L. M. (2003). How Literacy in Its Fundamental Sense is Central to Scientific Literacy. *Science Education*, *87*, 224–240.

Qasem, M. A. N. (2013). A Comparative Study of Classical Theory (CT) and Item Response Theory (IRT) in Relation to Various Approaches of Evaluating the Validity and Reliability of Research Tools. *Journal of Research & Method in Education, 3*(5), 77-81.

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal Of Educational Statistics*, *4*(3), 207-230.

Rijn, R. W. V., Sinharay, S., Haberman, S.J., & Johnson, M.S. (2016). Assessment of Fit of Item Response Theory Models Used in Large-Scale Educational Survey Assessments. *Large-Scale Assessments in Education*, *4*(10), 1-23.

Rubio, D. M., Berg-Weger, M., Tebb, S.S., Lee, E.S., & Rauch, S. (2003). Objectifying Content Validity: Conducting a Content Validity Study In Social Work Research. *Social Work Research: Oxford Journals*, *27*(2), 94-105.

Shakirova, D. M (2007). Technology for the Shaping of College Students' and Upper-Grade Students' Critical Thinking. *Russian Education Society*, *49*(9), 42-52.

Show-Yu, L. (2009). Chemical Literacy and Learning Sources of Non-Science Major Undergraduates on Understandings of Environmental Issues. *Chemical Education Journal*, *3*(1), 1-6.

Shwartz, Y., Ben-Zvi, R., & Hofstein, A. (2006). Chemical Literacy: What Does This Mean to Scientists and School Teachers?. *Journal of Chemical Education, 83*(10), 1557-1561.

Taherdoost, H. (2016). Validity and Reliability of The Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. *International Journal of Academic Research in Management*, *5*(3), 28-36.

Taylor, J. (2012, August 14). Philosophical Teaching Will Get Students Thinking for Themselves Again. The Guardian. Retrieved from https://tinyurl.com/ybsn4de

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

Thorndike, R. M., & Thorndike-Christ, T. (2010). Measurement and Evaluation in Psychology and Education (8th Ed). Upper Saddle River, NJ: Pearson/ Merrill Prentice Hall.

Toland, M. D. (2014). Practical Guide to Conducting an Item Response Theory Analysis. *Journal of Early Adolescence*, *34*(1), 120 –151.

Turiman, P., Omar, J., Daud, A.M., & Osman, K. (2012). Fostering the 21st Century Skills through Scientific Literacy and Science Process Skills. *Procedia - Social and Behavioral Sciences,59*, 110 – 116.

Wiberg, M. (2004). *Classical Test Theory VS Item Response Theory.* Umea: Umea University Press.

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.

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two Quantitative Approaches for Estimating Content Validity. *Western Journal of Nursing Research, 25*(5), 508-518.

Xie, R. (2018). Investigating the Content Validity of the Junior Middle School Entrance English Testing. *Journal of Education and Development, 2*(1), 45-54.

Yaghmale, F. (2003). Content Validity and Its Estimation. *Journal of Medical Education, 3*(1), 25-27.

Yilmaz, K., Altinkurt, Y., & Cokluk, O. (2011). Developing the Educational Belief Scale: The Validity and Reliability Study. *Educational Sciences: Theory and Practice*, *11*(1), 343-350.