Validation of critical thinking test to senior high school students

T Triastuti*, E W Laksono
Chemistry Education, Universitas Negeri Yogyakarta, Sleman, Indonesia
*titintriastuti.2018@student.uny.ac.id

Abstract. The purpose of the research is to examine the development of instruments for measuring the critical thinking abilities of high school students. This research is a quantitative descriptive study. Aspects of critical thinking skills used are to identify problems, reconstruct arguments, evaluate arguments, determine solutions and conclude. Data are analyzed based on criteria such as unidimensionality, local independence assumption test, item fit, and level of difficulty, reliability, and information function. The analysis results show that instruments can be accepted for use and are useful for educators to measure the critical thinking skills of high school students.

1. Introduction
As the main goals of future-oriented, contemporary science and chemistry education, students' critical thinking is very important because allows students to deal effectively with problems also to contribute directly to the community [1]. It is necessary to develop students’ ability to think critically in every aspect of life. Critical thinking is reflective and reasonable thinking that is asked to decide what to believe or do [2]. Critical Thinking is a prose of thinking that has the purpose of proving a problem, interpreting word questions and solving problems [3]. In addition, critical thinking is a metacognitive process through the purpose of assessing to increase opportunities to produce logical conclusions for an argument or solution to a problem [4]. That way, educators should be able to develop learning that can improve students' critical thinking skills. With the ability to think critically can help students overcome the problems obtained in learning. In addition, practicing critical thinking skills can help people become more adaptable, flexible, and better able to cope with the rapid development of information that continues to develop [4].

Critical thinking is not a method that must be learned, but rather a process, orientation of thought that includes the realm of cognition and affective [5]. Measurement of thinking ability in the cognitive realm can be done by tests. If tests are understood to encourage curriculum and teaching, then an easy way to renew teaching and improve the quality of education is to develop better tests [6]. Tests under state supervision must focus on critical thinking, therefore every test conducted by the government contains critical thinking tests [6]. So, assessment and measuring instruments need to be developed to measure students' critical thinking skills, especially in subjects that are usually tested such as chemistry.

Development of measuring tools and assessments based on thinking skills, obtained from the results of student analysis, synthesis and assessment steps in accordance with bloom's taxonomy by revising
the measuring and assessment tools used to help students acquire critical thinking skills attitudes [7]. The instrument of critical thinking ability test in this study was developed based on indicators of critical thinking abilities synthesized from several experts, there are five indicators namely identifying problems, reconstructing arguments, evaluating arguments, determining solutions and concluding [3][8][9]. The development of this critical thinking ability test instrument is used to assess and measure students' critical thinking skills on chemical equilibrium material. The development of this instrument is expected to help educators in assessing and measuring only which critical thinking abilities possessed by students.

The test instrument that will be used first performs a validity and reliability test. The validity comparisons of several experts where construction validity and reliability are mandatory for research instruments [10]. Validity is defined as the extent to which the concept is measured accurately in quantitative research and reliability is related to the consistency of a measure [11]. Assumptions of modern theory as a part of construct validity are mandatory for research instruments, but it is important to further analyze test instruments analysis with statistical methods such as item response theory models such as item fit, index difficulty and reliability [12]. So, this study aims to test the validity and reliability of the critical thinking skills test instrument.

2. Methods
This research uses descriptive quantitative research methods. The sample consisted of 178 high school students of class XI located in Yogyakarta. This instrument measures the ability to think critically on chemical equilibrium materials. The critical thinking ability test consists of eight essay questions.

The test instrument consists of five indicators of critical thinking namely identifying problems, reconstructing arguments, evaluating arguments, determining solutions and concluding [3][8][9]. This research was conducted by compiling the research instrument first then validated by two experts and through the revision stage for the instrument then tested on respondents. The instrument was analyzed with the Rasch model. Assumptions of modern theory are tested which consist of unidimensionality and local independence [13]. If the assumption test is met then further analysis can be done to determine the suitability of the item, the level of difficulty, reliability and information function.

3. Result and Discussion

3.1 The Unidimensionality Assumption Test
Unidimensionality test is part of the construct validity. Construct validity involves assessing the extent to which a measurement can measure its target variable correctly [14]. Unidimensionality tests are conducted to prove that each item in the test kit only measures one ability [15]. The data obtained were analyzed using Exploratory Factor Analysis with the IBM SPPS Statistics 23. The factor analysis was used to examine relationships within a group of observed variables [16]. The initial stage of the analysis of the adequacy of the sample is seen from the value of the KMO (Kaiser Mayer Olkin) and Bartlett's test to determine the data obtained in accordance with factor analysis. If the KMO value is greater than 0.6 and the Bartlett's test value is smaller than 0.05 then the assumption is accepted [16]. The KMO test results and Bartlett's test are presented in table 1.

| Measure                                | Value |
|----------------------------------------|-------|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy | .638  |
| Bartlett’s Test of Sphericity          | .000  |

The results showed the KMO value was (0.638 > 0.6) proving that the sample used was adequate and the Bartlett test value was (0.000 < 0.05) proving that the correlation matrix observed was different from
a single matrix, meaning and linear combination, so the data obtained based on factor analysis. To determine items that measure the same dimensions can be done with an extraction process that produces a new factor. The results of factor analysis obtained three eigenvalues that have a score greater than 1, which indicates there are three factors formed. These three factors can explain 62.140% of the total variance. The dominant factor is 31.541% of the total variance. Assumptions can be met if the dominant factor with a percentage greater than 20%, then the assumption of Unidimensionality can be met [17]. For more details, it can be seen in the results of the scree plot presented in Figure 1.

![Figure 1. Screen plot unidimensionality](image)

The results of the scree plot show that the second factor and so on begin to slope, even though the eigenvalues of the second factor and subsequently are too strict. Basically, the Unidimensionality assumption is very difficult to fulfill [18]. So, there is only one dominant factor that proves that the instrument has met the assumption of the unidimensionality test, which means the instrument of critical thinking ability only measures one dimension.

### 3.2 The Local Independence Assumption Test

The next assumption is a local independent test carried out with the help of Winstep and Microsoft Excel programs.

| Columns | K1  | K2         | K3         | K4         | K5         | K6         | K7         | K8         | K9         | K10        |
|---------|-----|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| K1      | 0.048658 | 0.008603  | 0.007529   | 0.008848   | 0.011505   | 0.008327   | 0.010009   | 0.014468   | 0.056351   | 0.131547   |
| K2      | 0.0023391 | 0.002144  | 0.002214   | 0.002216   | 0.002839   | 0.002214   | 0.002629   | 0.003172   | 0.014693   | 0.0269547  |
| K3      | 0.0019512 | 0.0019352 | 0.0020496  | 0.0025792  | 0.002216   | 0.0021208  | 0.002629   | 0.003582   | 0.0135191  | 0.0269547  |
| K4      | 0.002476   | 0.0002991 | 0.000291   | 0.000291   | 0.002291   | 0.002214   | 0.002629   | 0.003481   | 0.0135191  | 0.0269547  |
| K5      | 0.00383    | 0.002214   | 0.0002291  | 0.0002291  | 0.002839   | 0.002214   | 0.002629   | 0.003481   | 0.0135191  | 0.0269547  |
| K6      | 0.002441   | 0.000291   | 0.000291   | 0.000291   | 0.002291   | 0.002214   | 0.002629   | 0.003481   | 0.0135191  | 0.0269547  |
| K7      | 0.0038903  | 0.0002291  | 0.0002291  | 0.0002291  | 0.002839   | 0.002214   | 0.002629   | 0.003481   | 0.0135191  | 0.0269547  |
| K8      | 0.00581    | 0.0002291  | 0.0002291  | 0.0002291  | 0.002839   | 0.002214   | 0.002629   | 0.003481   | 0.0135191  | 0.0269547  |
| K9      | 0.1038299  | 0.0002291  | 0.0002291  | 0.0002291  | 0.002839   | 0.002214   | 0.002629   | 0.003481   | 0.0135191  | 0.0269547  |
| K10     | 0.6432063  | 0.0002291  | 0.0002291  | 0.0002291  | 0.002839   | 0.002214   | 0.002629   | 0.003481   | 0.0135191  | 0.0269547  |

Local independent testing aims to prove those participants’ responses in answering an item do not affect participants’ responses to other items. The analysis was carried out by compiling a matrix divided
by high to low levels, then grouped and processed by covariance analysis. If the diagonal value in the matrix approaches 0.00 then local independence is fulfilled [13]. In table 2, we can observe the value of the covariant matrix variance on the diagonal line close to 0.00. Indicates that there is no relationship between the participant’s responses to the item. In other words, each instrument item does not correlate with each other.

### 3.3 Item Fit

Item fit aims to find out whether the instrument functions as a measurement tool or not. Item fit is analyzed using the Winstep program. Item fit analysis is said to be valid if it meets the criteria value as follows: (a) MNSQ Outfit value received $0.5 < \text{MNSQ} < 1.5$; (b) the received ZSTD Outfit value $-2.0 < \text{ZSTD} < +2.0$; (c) Point Measure Correlation value: $0.4 < \text{Pt Measure Corr} < 0.85$ [19]. If it does not meet the criteria then the item must be discarded or repaired.

**Table 3. Item fit of the measure instrument**

| Item | Outfit MNSQ | Outfit ZSTD | Pt-Measure Corr. | Conclusion |
|------|-------------|-------------|------------------|------------|
| 8*  | 1.77        | 3.2         | 0.46             | Not fit    |
| 6   | 1.20        | 1.4         | 0.47             | Fit        |
| 7   | 1.08        | 0.7         | 0.72             | Fit        |
| 2   | 0.96        | -0.3        | 0.58             | Fit        |
| 3   | 0.96        | -0.3        | 0.54             | Fit        |
| 4   | 0.83        | -1.4        | 0.59             | Fit        |
| 1   | 0.84        | -0.6        | 0.34             | Fit        |
| 5   | 0.78        | -2.1        | 0.70             | Fit        |

*not fit

Based on Table 3, there is one item that is not appropriate, item 8, because it only meets one criterion, namely the Pt-Measure Corre value. While the items considered to be fit, at least meet two criteria. Therefore item 8 must be removed or cannot be used. The critical thinking aspect that contains in item 8 is determining the solution. Even so there are still other items that contain these aspects. So the instrument still contains five aspects of critical thinking. Other items qualify for item fit. It can be concluded that 7 items fit the criteria and are suitable for use to measure critical thinking skills.

### 3.4 Level of Difficulty

Difficulty level analysis is performed to find out which category of items is the difficulty level. Good instrument items have criteria for the level of difficulty of questions that are in the logit range from -2 to +2. The difficulty level of an item is measured according to the following categories: (a) value $b > 1$, difficult; (B) $0.5 \leq b < 1$, moderate; (c) $-0.5 \leq b < 0.5$, easy; (d) $b \leq -1$, very easy [20]. Table 4, shows the difficulty level of items in the range of 1.84 to -0.97.

**Table 4. Item difficulty**

| Item number | Item Measure | Category         |
|-------------|--------------|------------------|
| 1           | 1.84         | Very difficult   |
| 3           | 0.07         | Medium           |
| 8           | 0.06         | Medium           |
| 6           | 0.05         | Medium           |
| 4           | -0.24        | Medium           |
| 2           | -0.25        | Medium           |
| 5           | -0.57        | Easy             |
| 7           | -0.97        | Easy             |
The instrument of critical thinking skills test developed has various levels of difficulty. There is one item that has very difficult criteria, namely item 1, item 3, 8, 6, 4 and 2 with medium criteria and other items namely 5 and 6 easy criteria. Based on this, the instrument items have a good level of difficulty.

3.5 Reliability
Reliability analysis is valued with the Rasch program. Reliability measures ranging from 0 to 1. Measurement of the reliability of all instruments can be determined by Cronbach’s alpha value, acceptable if Cronbach’s alpha is greater than 0.70 [21]. Cronbach’s alpha value on this critical thinking ability instrument is \( \alpha = 0.70 \). That way the instrument can be relied on, which shows the suitability between the respondent and the item. With the Rasch model, instrument reliability can be determined based on respondents and items. The reliability value of the respondent was 0.63, the reliability value of the reliability item was 0.98, very reliable. Display these items as needed [22].

3.6 Information Function
The measurement function shows the item’s overall test and reliability. The information function tells how well each level of capability is estimated, the information function does not depend on the distribution of respondents on the ability scale [23]. In figure 2, it can be seen that the level of ability is moderate, the information function obtained by measurement is high. Shows that instruments about critical thinking abilities produce optimal information when given to respondents with moderate abilities.

![Figure 2. Information test for critical thinking skill](image)

4. Conclusion
This study presents psychometric characteristics for critical thinking skills instruments for secondary school students. The assumption test of modern theory with the unidimensionality assumption test results obtained, the dominant factor is 31.541% of the total variance so that the assumption of Unidimensionality is accepted. The Local Independence Assumption test assumes that all items are independent and do not depend on the answer to the previous item. With the fulfillment of the assumption test, further analysis can be done.

Analysis of the response theory model on item fit there is 1 item that is not fit that measures the ability to think critically on aspects of determining solutions. While 7 other aspects meet the criteria. For the level of difficulty obtained various levels of difficulty between very difficult, moderate and easy, which shows the instrument has a good level of difficulty. In the reliability test the value of Cronbach’s alpha on the instrument of this critical thinking ability is \( \alpha = 0.70 \). That way the instrument has good reliability. Finally, the Information Function test shows that instruments produce optimal information when given to respondents about the ability to spring. Based on the results of the analysis shows that the instrument can be accepted for use and is useful for educators to measure the critical thinking skills of high school students.
5. Reference

[1] Shakirova D M 2007 Technology for the Shaping of College Students’ and Upper-Grade Students’ Critical Thinking *Russian Education Society* **49** 9 pp.42-52.

[2] Ennis R H 1996 Critical thinking dispositions: Their nature and assessability *Informal logic* **18** 2.

[3] Facione P A 2011 Critical thinking: What it is and why it counts *Insight assessment* **2007** 1 pp.1-23.

[4] Dwyer C P, Hogan M J and Stewart I 2014 An integrated critical thinking framework for the 21st century *Thinking skills and Creativity* **12** pp.43-52.

[5] Simpson E and Courtney 2002 Critical thinking in nursing education: Literature review Elaine *Int. J. Nurs. Prat.* **8** p. 98–98.

[6] Yeh S S 2001 Tests worth teaching to: Constructing state-mandated tests that emphasize critical thinking *Educational Researcher* **30** 9 pp. 12-17.

[7] Ali A Z A R. 2010 The effect of critical thinking dispositions on students achievement in selection and placement exam for university in Turkey *Journal of Turkish science education* **7** 1 pp.61-73.

[8] Bowell T and Kemp G 2014 *Critical thinking: A concise guide*. Routledge.

[9] Watson G 2010 Technical manual and user guide: Watson–Glaser™ II Critical Thinking Appraisal.

[10] Taherdoost H 2016 Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research *How to Test the Validation of a Questionnaire/Survey in a Research*

[11] Heale R and Twycross A 2015 Validity and reliability in quantitative studies *Evidence-based nursing* **18** 3 pp.66-67.

[12] Sadhu S Adhiya E and Laksono E W 2019 Exploring and Comparing Content Validity and Assumptions of Modern Theory of an Integrated Assessment: Critical Thinking-Chemical Literacy Studies *Jurnal Pendidikan IPA Indonesia* **8** 4 pp.570-581.

[13] Hambleton R K, Swaminathan H and Rogers H J 1991 *Fundamentals of item response theory*. Sage.

[14] O'Leary-Kelly S W and Vokurka R J 1998 The empirical assessment of construct validity. *Journal of operations management* **16** 4 pp.387-405.

[15] Reckase M D 1979 Unifactor latent trait models applied to multifactor tests: Results and implications *Journal of educational statistics* **4** 3 pp.207-230.

[16] Beavers A S, Lounsbury J W, Richards J K and Huck S W 2013 Practical considerations for using exploratory factor analysis in educational research *Practical Assessment, Research, and Evaluation* **18** 1 p.6.

[17] Brown R L, Obasi C N and Barrett B 2016 Rasch analysis of the WURSS-21 dimensional validation and assessment of invariance *Journal of lung, pulmonary & respiratory research* **3** 2

[18] Lestyarini, B., 2016 Partial Credit Model (PCM) For Item Response Theory: Improving Teacher’s Competence In Language Learning Assessment *Proceeding*, p.200.

[19] Boone W J, Staver J R and Yale M S 2013 *Rasch analysis in the human sciences*. Springer Science & Business Media.

[20] Adedoyin O O and Mokobi T 2013 Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items *International Journal of Asian Social Science* **3** 4 pp.992-1011.

[21] Bland J M and Altman D G 1997 Statistics notes: Cronbach’s alpha *Bmj*, **314** 7080 p.572.

[22] Maat S M and Rosli M K 2016 The Rasch model analysis for statistical anxiety rating scale (STARS) *Creative Education* **7** 18 p.2820.

[23] Baker F B 2001 *The basics of item response theory*. Eric.