

## Development of Integrated Assessment Instrument to Measure Critical Thinking Skills and Self-Efficiency in Acid-Base Concept

*Nuryuliana<sup>1</sup> and Anti Kolonial Prodjosantoso<sup>1</sup>*

*<sup>1</sup>Departement of Chemistry Education, Postgraduate Program, Yogyakarta State University, Jl. Colombo No.1, Karang Malang, Yogyakarta, 55281, Indonesia*

*\*E-mail: [nuryuliana.2019@student.uny.ac.id](mailto:nuryuliana.2019@student.uny.ac.id)*

Received: 23 November 2021; Accepted: 16 December 2021; Published: 31 Desember 2021

### Abstract

The purpose of this research is to describe the stages of developing and validating a product which includes content validity and constructs validity. The method used is Research and Development (R&D) with define, design, develop, and disseminate (4D) model. The initial product was validated by material and learning evaluation expert lecturer. The suitability of the items evaluated by chemistry teacher based on indicators of competency achievement. The sample selection was determined by stratified random sampling of 170 students of 11<sup>th</sup> grade senior high school in Yogyakarta City. Content validity was analyzed using Aiken's V formula. The results of data analysis show that all components of the test instrument are valid with Aiken's V value greater than 0.80 with a significance of 0.05. Verification of construct validity was analyzed using exploratory factor analysis with the help of the SPSS version 21.0 program. The results prove that all components in the integrated assessment instrument are suitable to be used to measure the critical thinking skills and self-efficacy of students on acid-base concept.

Keywords: acid-base, integrated assessment, self-efficacy

DOI: <https://doi.org/10.15575/jtk.v6i2.14872>

### 1. Introduction

Assessment is an important stage in the learning process. Assessment is the process of gathering information to measure the achievement of participants' learning outcomes before, during, and after the learning process (Astuti et al., 2017). Assessment not only assesses student learning outcomes but also assesses student learning processes (Gerritsen-van Leeuwenkamp et al., 2019). Assessment of student learning outcomes consists of several aspects, namely aspects of knowledge, attitudes, and skills. These three aspects are very important in chemistry education because the nature of chemistry is both a process and a product. Therefore, the assessment must cover all these aspects in the learning process.

The knowledge dimension consists of several, namely factual knowledge and procedural knowledge (Asy'ari et al., 2018). Factual knowledge is cognitive knowledge that includes facts, principles, concepts, theories, and laws. While procedural knowledge is knowledge-oriented to process skills. Process skills are very important skills in understanding and applying chemical concepts. So, the assessment in the field of chemistry learning must cover these two aspects, namely the knowledge and process skills aspects (Reynders et al., 2019).

Chemistry is one part of natural science that requires students to have high-level thinking skills for each material being studied (Sumarni et al., 2018). Chemical materials have characteristics that are very closely related to

the phenomena of everyday life. One of the chemicals that are very closely related to everyday life is acid-base material. Acid-base material is contextual material that is easy to relate to and apply in everyday life. Some phenomena of acid-base material that can be found in everyday life, for example, determining the nature of acids and bases using natural materials that exist in the environment of students and acid and alkaline solutions needed by the human body could develop students' critical thinking skills (Rasmawan, 2020).

Critical thinking is one of the thinking skills in higher-order thinking skills and is a 21<sup>st</sup> century learning goal (Dwyer et al., 2014; Kivunja, 2014; Redhana, 2019). Critical thinking skills are very important for individuals to master in order to succeed in facing challenges, life problems, and careers in the rapidly changing era of science and technology (Sari et al., 2020). In addition, critical thinking skills are skills that can develop students' abilities in interpreting, applying, and concluding a problem based on scientific evidence (Živković, 2016). Students who have higher critical thinking skills tend to criticize information and provide accurate explanations. According to Danczak (2020) 30% of Indonesian graduates state critical thinking as one of the top five skills that they want to develop further. However, in reality, learning only prepares students to continue their education or intended career, even though other things are also important, namely the ability and skills to solve problems faced in everyday life, especially in the world of work (Tan et al., 2006). Therefore, developing critical thinking skills is becoming increasingly important because students need to adapt to these changes in an active and skilled way, for example in conceptualizing, applying, analyzing, evaluating information, and drawing conclusions.

Critical thinking is considered a skill and focuses on the thinking process (Utami et al., 2017). The thinking process starts from low-level thinking to higher-order thinking. Critical thinking skills help students make decisions appropriately and efficiently (Danczak, 2017).

Critical thinking is self-awareness of one's thinking in assessing a problem based on orientation, accuracy, processes, theories, methods, and background problems so that they can make the right decisions (Shavelson et al., 2019). Critical thinking is an essential way of thinking in analyzing, investigating, and evaluating problems based on scientific evidence (Danczak et al., 2020). Critical thinking includes the process of concrete thinking and abstract thinking. Critical thinking can search for, understand, and evaluate relevant statements logically and rationally during the process of problem-solving and decision-making (Shaw et al., 2020; Thomas, 2011). Mahanal (2019) showed that it is not enough for students to have critical thinking skills, but students need to apply their critical thinking skills efficiently.

Critical thinking consists of three aspects, namely identifying problems, constructing arguments, and evaluating arguments (Sarigoz, 2012). The classification of critical thinking according to Watson-Glaser consists of defining a problem, determining solutions with strong arguments, drawing valid conclusions based on solutions, and evaluating conclusions (Sternod et al., 2015). Students' critical thinking skills are seen from the students' arguments against the problem and conclusions about these arguments. Therefore, critical thinking indicators used in the research and development of integrated assessment instruments include (1) identifying problems, (2) reconstructing arguments, (3) determining arguments, (4) analyzing arguments, (5) drawing conclusions, and (6) represents the argument.

Improving students' critical thinking skills is an important thing to do. Therefore, reviewing the supporting aspects is one way that can be done. The supporting aspect of critical thinking skills in question is the personality factor. Hoffman and Schraw (2009) state that one of the most important personality factor personality factors is self-efficacy. Self-efficacy derived from social cognitive theory is an individual's self-confidence in his ability to carry out certain tasks (Bandura, 1990).

Students who are less confident in their abilities tend to avoid the assigned tasks (Flaherty, 2020). Avoidance behavior shows that student have low efficacy, while students with high confidence tend to be confident in completing the given task (Xu et al., 2016; Zimmerman, 2000). Self-efficacy is an assessment of students' abilities that are influenced by positive or negative experiences that have been experienced before (Dalgety et al., 2006). Positive experiences that have been experienced by students can increase self-efficacy, while negative experiences that have been experienced can reduce students' self-efficacy or confidence in the tasks given. (Dalgety et al., 2006). It is therefore important to develop an assessment instrument to measure student efficacy.

A good assessment is an assessment that can reflect all the skills that will be assessed both in terms of cognitive, affective, and psychomotor (Saputri et al., 2018; Sumarni et al., 2018). However currently, the assessment is only centered on one aspect, namely the cognitive aspect (Zoller, 2001), while the affective and psychomotor aspects are still lacking (Hensen et al., 2019). In addition, these three aspects are still assessed separately. Assessment of student learning outcomes on cognitive aspects and affective aspects can be assessed simultaneously (integrated). An integrated assessment of student learning outcomes is used to assess the abilities and skills of students (Sumarni et al., 2016). Ferrell et al (2016) states that the affective aspect plays an important role in the learning process where students with good affective abilities tend to have good cognitive abilities than students with low affective abilities who only focus on memorizing (Frey et al., 2017). This shows that there is a relationship between cognitive and affective aspects. In addition, cognitive aspects and affective aspects are still assessed separately (Galloway & Bretz, 2015; Sadhu & Laksono, 2018). Therefore, based on the problems above, it is important to research the development of assessment instrument that are used to measure critical thinking skills objectively, and which include affective aspects, one of which is self-efficacy.

## 2. Research Method

### 2.1. Participants

The subjects used in this study were 170 student of 11<sup>th</sup> grade from three schools selected by stratified random sampling during 2<sup>nd</sup> semester of the 2021/2022 academic year in the Special Region of Yogyakarta. Subjects in this study have different levels of ability.

### 2.2. The Development Framework

This research model is a procedural research and development model, namely descriptive research that describes the research steps that must be followed to produce the final product. This type of research is a type of Research and Development (R&D). The development model used is the 4D development model (Four-D) developed by Thiagarajan (1947) which consists of four stages, namely define, design, develop, and disseminate. In this study, the integrated assessment was developed in three stages.

#### 2.2.1. Define

The define stage is the first stage in the development of an integrated assessment instrument. The define stage aims to conduct a needs analysis, review various literature studies, and formulate a research framework. At the define stage, the emphasis is on reviewing various literature studies such as international journal articles, as well as relevant books on critical thinking skills, self-efficacy, and the development of integrated assessment instruments, and acid-base materials

#### 2.2.2. Design

The second stage of developing an integrated assessment instrument is the design stage. This stage aims to design a content framework which is divided into three stages. The first stage is to analyze the teaching and learning objectives referring to the chemistry syllabus for 2013 revised 2017 curriculum and textbooks in 11<sup>th</sup> grade. Next, the researcher determines competency standards, basic competencies, and learning indicators. The second stage is to describe the indicators of critical thinking skills, indicators of self-

efficacy, design items, arrange questions, and scoring guidelines. The third step is an integrated assessment instrument written in Indonesian.

### 2.2.3. Develop

The third stage is the development stage. At this stage, the test instruments are reviewed, assessed, and validated by experts. Types of item validity consist of content validity and construct validity. Content validity aims to test the validity of a test instrument based on its content. At the content validity stage, the test instrument was determined and given to a panel of experts, namely chemists, evaluation experts, and chemistry teachers. Experts are selected based on their education, knowledge, and experience while working in compiling good and quality test instruments. The initial stage of content validity is that the researcher explains the main purpose of the integrated assessment to the expert panel, then the expert panel is asked to review the suitability of each item with the overall learning objectives. In particular, the expert panel was asked to review each item, based on the following criteria: (a) basic competencies and indicators to be achieved, (b) clarity of words/phrases/sentences in each item, (c) rules of using correct grammar, (d) and display of questions. In line with the advice obtained from the experts then used to refine the items.

Based on the results of the initial product validation, the integrated assessment instrument consists of 18 questions. These 18 items have met the valid criteria so that they can be used for field trials. However, several things must be considered with the number of essay questions as many as 18 items. Based on the advice of chemistry experts and teachers the number of questions as many as 18 items is too many to be tested with a time allocation of two hours of lessons (90 minutes). In addition, currently, the learning process is still carried out online (remotely) which causes a reduction in the time allocation to two lesson hours (60 minutes). This is in line with the opinion (Salirawati, 2011) which states that the time needed by test takers to take the exam ranges from six to ten minutes based on the level of difficulty of the questions or two until

three times the time needed by the teacher in solving the same questions. In addition, Utomo and Ruijter (1994) stated that the effective time for test-takers to complete the exam ranged from 1.5 to 2.5 hours. If the time used is more than the estimated time allocation, it will cause the reliability of the items to decrease. Based on the results of content validation on the initial product, the items used at the trial stage were nine items.

After all the suggestions from the experts were revised, then a field trial was carried out to 170 student of 11<sup>th</sup> grade who had studied acid-base material. The purpose of the field trial was to test the construct validity and item parameter analysis (item model fit, item difficulty, and reliability analysis). Before carrying out field trials, students are given information about the purpose of the test. General information related to procedures for working on description questions, and asking students to take the exam seriously and carefully. The data from the field trials were then analyzed. After all the items are declared valid, then they are reassembled so that they can be used at the next stage.

## 2.3. Integrated Assessment Instrument

The product developed in this study is an integrated assessment instrument to measure students' critical thinking skills and self-efficacy on acid-base material.

### 2.3.1. Item Construct

The items are arranged based on the contents of 2013 revised 2017 Curriculum. To obtain the results of critical thinking skills and self-efficacy expected of students, the stems of each item are arranged in such a way that it can provoke critical thinking and self-efficacy of students simultaneously.

In this study, researchers combined several critical thinking frameworks from several experts, namely, Ennis (1985), Facione (1990), (Saxton et al., 2012). Determination of the framework is based on the suitability and needs in the preparation of an integrated assessment instrument related to critical thinking skills. Aspects of critical thinking consist of: (1) identifying problems, (2)

reconstructing arguments, (3) evaluating arguments, (4) determining solutions, and (5) concluding. Meanwhile, the self-efficacy framework is combined from several experts namely, (Bandura, 1990), (Zimmerman, 2000), (Uzuntiryaki-Kondakçi & Çapa-Aydın, 2013). The dimensions of self-efficacy consist of

confidence, persistence, and resilience. In this study, the critical thinking aspect is integrated with the self-efficacy aspect into one aspect. In line with that, there are eight integrated aspects (critical thinking and self-efficacy) that can be measured using an integrated assessment.

**Table 1. Integrated Aspects of Critical Thinking Aspects and Self-Effective Aspects**

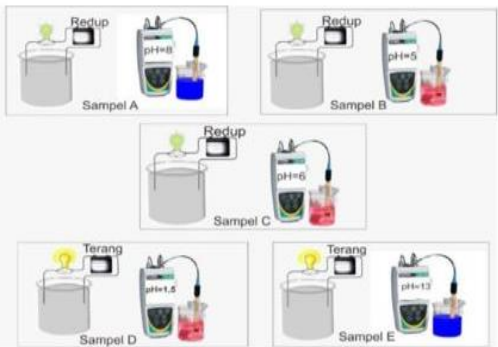
No	Aspects of Critical Thinking	Aspects of Self Efficacy	Integrated Aspect
1.	Firm in deciding something by utilizing chemical concepts	Describing a conclusion	Draw conclusions based on phenomena that occur in everyday life by utilizing chemical concepts
2.	Firm in deciding something by utilizing chemical concepts	Argument Evaluation	Evaluating arguments based on phenomena discovered using chemicals
3.	Confident in doing something	Identify the problem	Identify phenomena that occur in everyday life and then connect them with their understanding
4.	Firm in deciding something by utilizing chemical concepts	Describing a conclusion	Draw conclusions based on phenomena that occur in everyday life by utilizing chemical concepts
5.	Confident in doing something	Identify the problem	Identify phenomena that occur in everyday life and then connect them with their understanding
6.	Firm in deciding something by utilizing chemical concepts	Argument Evaluation	Evaluating arguments based on phenomena discovered using chemicals
7.	Doing something based on the benefits of applying chemical concepts a	Determine the solution	Determine the benefits of applying chemical solutions and application examples
8.	Doing something based on the benefits of applying chemical concepts	Reconstructing the argument	Reconstruction of arguments from the benefits of applying chemical solutions and application examples

### 2.3.2. Item Format

The integrated assessment instrument consists of nine items of description. The format for filling in the description questions consists of descriptive questions that must be answered by students and students are asked to write answers properly and correctly. Meanwhile, the format for collecting student

answers is asked to load the folder provided then the answers are photographed and uploaded on the answer link that has been shared. An example of an integrated assessment instrument item is explained as follows to describe the format of the items above.

A student conducted an experiment related to classifying and distinguishing acidic and basic solutions with the same concentration based on the strength of their acidity so that the following observations were obtained.



a. Calculate the concentration of each sample through calculations!

b. Present the observational data in the form of a table containing the sample code, pH, light bulb flame and concentration!

Figure 1. Example of question 1 integrated assessment instrument for acid-base materials

A group of students was asked to conduct an experiment to identify the acidic and basic properties of a solution. The teacher prepares a colorless acid and base solution without explaining the two solutions. How do students determine the nature of the acid and base of the two, explain!

Figure 2. Example of question 2 integrated assessment instrument for acid-base materials

### 2.3.3. Scoring Procedure

This scoring guide was created to make it easier for teachers to assess test results. The assessment guide is described as follows.

Table 2. Scoring Procedure Example Question 1

Criteria	Score
Write what you know in the question	1
Calculating the concentration of each sample	2
Make a table based on the results of observations with a description of the sample code, pH, light bulb flame, and concentration	3
<b>Total Score</b>	<b>6</b>

Table 3. Scoring Procedure Example Question 2

Criteria	Score
Determine which indicators to use	1
Write down the experimental results with descriptions	1
Explaining the results of the experiment	2
<b>Total Score</b>	<b>4</b>

### 2.4. Data Analysis

Item Response Theory (IRT) is mostly used to analyze assessment instruments in education. Item response theory is used in designing and analyzing a test instrument. The reason for using item response theory is that the test taker's ability level is expected to remain invariant with any group of items to be measured. Based on the research objectives above, content validity was analyzed using the Aiken's V formula (1985), and construct validity was analyzed using factor analysis, and item quality was analyzed using the Rasch model. Factor analysis and the Rasch model are part of item response theory analysis.

Construct validity using an exploratory approach aims to see how many factors are needed to explain the relationship between variables to be measured. The initial stages carried out on construct validity are as follows: (a) looking at the value of the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO-MSA) test and Bartlett's Test of Sphericity, (b) looking at the eigenvalues of the variance-covariance correlation matrix. This aims to see the adequacy of the sample size used so that it can be continued in the next analysis using unidimensional assumptions (basic assumptions). The theory of unidimensional assumptions or basic assumptions is unidimensionality, which means that the test items only measure one ability. This means that the integrated assessment instrument only measures critical thinking skills and self-efficacy. In addition, the quality of the items is assessed by looking at the fit of the item model (Item fit), the level of item size, and reliability. The data from the field trials were analyzed with the help of the SPSS version 21.0 and Quest.

### 3. Result and Discussion

#### 3.1. Content Validity

The coefficient of content validity is determined based on the results of the assessment of  $n$  experts with the number of categories given (Aiken, 1985). In this research, there are five validators to provide an assessment of the test instrument with five rating categories. Aiken's  $V$  formula to calculate content validity is described as follows:

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

Information:  $s = r - l_o$ ,  $l_o$  = the lowest value of the validity assessment,  $c$  = the highest number of validity,  $r$  = the number given by the rater, and  $n$  = the number of assessments. This means that all indicators in the integrated assessment instrument can be used to measure critical thinking skills and students' self-efficacy on acid-base material.

#### 3.2. Construct Validity

Construct validity was analyzed using factor analysis with the help of SPSS version 21.0. The unidimensional assumption is proven using factor analysis. The output results of factor analysis include the Kaiser-Mayer-Olkin Measure of Sampling Adequacy (KMO-MSA) test, Bartlett's test, variance-covariance matrix, and scree plot. The initial step is to fulfill the unidimensional assumption, namely by looking at the adequacy of the sample size used. The adequacy of the sample size is proven by the KMO-MSA test and Bartlett's test of sphericity is used to see if there is a correlation between variables.

**Table 3. Results of KMO-MSA and Bartlett Tests**

Kaiser-Meyer-Olkin Measure of Sampling Adequacy	0.858
Approx. Chi-Square	442.471
Bartlett's Test of Sphericity df	36
Sig.	0.000

To prove the unite assumption, the KMO-MSA test must be more than 0.5 and the significance of Barlett's test must be less than 0.05 (Field, 2009). Based on the data from the analysis of the KMO-MSA test and Barlett's test in Table 3, it shows that the KMO-MSA value is 0.858 with Bartlett's test significance of 0.000, this proves that the KMO-MSA test and Bartlett's test have been fulfilled. This means that the sample size used is sufficient. Therefore, factor analysis can be continued with the interpretation of the eigenvalues of the variance-covariance correlation matrix.

**Table 4. Results of Eigen Value and Correlation Matrix of Variance-Covariance Matrix**

Component	Nilai Eigen Value		
	Total	% of Varince	Cumulative %
1	3.114	34.597	34.597
2	1.149	12.762	47.360
3	1.029	11.396	58.755
4	0.926	10.284	69.039
5	0.792	8.805	77.845
6	0.703	7.806	85.650
7	0.537	5.964	91.615
8	0.412	4.584	96.198
9	0.342	3.802	100.000

Exploratory factor analysis aims to analyze the relationship between variables by using a correlation test, to obtain a new variable named factor. There are several ways to interpret the fulfillment of the unidimensional assumption, one of which is by looking at the contribution of the first eigenvalue of the test variance (Kartowagiran et al., 2019). Based on the output results in Table 4, shows that the student's response to the test instrument contains three Eigenvalues that are more than one (Eigenvalue < 1). These three factors can explain 58.755% of the total variance. This means that the integrated assessment instrument can explain the ability of students by 58.755% (valid).

Based on the Kaiser Criteria, that there are three factors formed, but there are dominant factors (Beavers et al., 2013). The dominant factor that is expected is knowledge of

chemistry because the instrument developed is based on the grid contained in the basic competencies in the field of chemistry, especially on acid-base materials. Knowledge of chemistry consists of mathematical ability and language ability (Pyburn et al., 2013). While the other two factors that were measured included personality factors and test implementation factors such as anxiety, motivation, and tendency to guess answers (Retnawati, 2014).

The results of factor analysis can be explained with a scree plot to visualize the Eigenvalue as shown in Figure 1, the eigenvalues start slightly at the third eigenvalue. So, it shows that there are dominant factors that are measured by the integrated assessment instrument and these two factors contribute to the instrument.

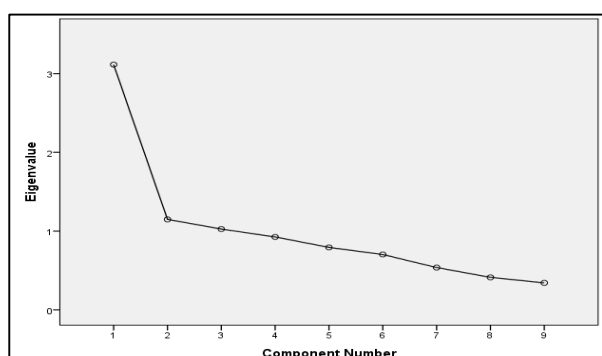


Figure 3. Scree Plot

The unidimensional assumption is very difficult to fulfill ideally (Retnawati, 2014). However, the unidimensional assumption can be said to be fulfilled if the test instrument contains one dominant component (Hambleton et al., 1985). When referring to the first factor, the unidimensional assumption is said to be fulfilled if "the first factor should account for at least 20 percent of the test variance". This shows that the first factor has a cumulative percentage of 20% (Hambleton et al., 1985). As shown in the output results of Table 4, the percentage of the first factor is 43.041%, so it can be stated that the

unidimensional assumption has been fulfilled. If the unidimensional assumption has been met, then the construct validity has also been met. In short, nine items can be used for further analysis.

### 3.3. Item Parameter Analysis

#### 3.3.1. Analysis of Instrument Model Fit

Item fit is a method to describe whether an item can work optimally so that it can be used as a good measurement tool. The level of difficulty of the items with the test takers is said to be suitable if the established model has an INFIT MNSQ price in the range of 0.5 to 1.5 (Linacre, 2002). Standardized item parameters mean that the items have good psychometric characteristics so that if measurements are made using these items, the measurement results obtained will be more accurate (Kriswantoro dkk, 2016). Based on the results of the item fit analysis in Table 5, the MNSQ outfit value shows that the total items are included in the fit category, so it can be used for measurement.

Table 5. Results of Model Fit Analysis (*Item Fit*)

Item Number	INFIT MNSQ	Information
Item 1	0.77	Fit
Item 2	1.02	Fit
Item 3	0.94	Fit
Item 4	0.95	Fit
Item 5	1.01	Fit
Item 6	0.94	Fit
Item 7	0.97	Fit
Item 8	1.10	Fit
Item 9	1.01	Fit

#### 3.3.2. Item Difficulty

The level of item difficulty is an indicator to describe whether the item is classified as easy or difficult. An item can be said to be "good" if it meets two requirements for a good level of difficulty, which are -2 logit bit +2 logit (Hambleton & Swaminathan, 1985). The results of the item difficulty level analysis are presented in full in Table 17 as follows.



**Table 6. Results of the Analysis of Item Difficulty**

Item Number	Item Difficulty	Information
Item 1	0.37	Good
Item 2	0.31	Good
Item 3	0.36	Good
Item 4	0.41	Good
Item 5	0.35	Good
Item 6	0.30	Good
Item 7	0.24	Good
Item 8	0.34	Good
Item 9	0.25	Good

The level of difficulty that has been formulated by the teacher does not mean that it is following the level of difficulty of the empirical results (Kriswantoro et al., 2016). This is because when making items, the teacher classifies items based on the level of difficulty (easy, medium, and high) based on their intuition. The quality of a test item can be determined from the level of difficulty possessed by each item. An item is said to be good if the item is not too difficult and not too easy, or in other words, the item's difficulty level is moderate or sufficient (Kartowagiran et al., 2019). Based on the results of the analysis of the difficulty level of the items in Table 5, it shows that the nine items are included in the good category.

### 3.4. The Reliability of Integrated Assessment Instruments

Cronbach's Alpha internal consistency coefficient was found to be 0.82. This means that there is about 82% certainty about the consistency of the items in obtaining more or less the same results over and over again. This shows that the test instrument is very reliable. This finding is supported by Ceniza and Cereno (2012) which states that if the reliability coefficient is in the range of 0.81 to 1.0, this indicates high reliability, 0.61 to 0.80 indicates moderate reliability, 0.41 to 0.60 indicates reasonable reliability, 0.10 to 0.40 indicates low reliability, and less than 0.10 indicates no reliability. Therefore, the reliability coefficient on the integrated assessment instrument is high

## 4. Conclusion

Based on the findings in this study, the integrated assessment instrument that has been prepared is based on indicators of acid-base learning, indicators of critical thinking skills, and indicators of self-efficacy. Overall, the integrated assessment instrument that has been developed is declared valid based on content validity and construct validity, so that it can be used to measure critical thinking skills and self-efficacy on acid-base materials.

Based on the results of the study, discussion, and conclusions that have been found above, the suggestions made are it is necessary to pay attention to the requirements in compiling tests in general so that the validity of the test kits that have been developed have a quality test quality. One of them is the language used when compiling the test must use language that is difficult for students to understand so as not to cause double meaning (ambiguous), making it easier for students to answer the tests given and the answers from test-takers as expected.

## References

- Astuti, S. R. D., Suyanta, Lfx, E. W., & Rohaeti, E. (2017). An integrated assessment instrument: Developing and validating instrument for facilitating critical thinking abilities and science process skills on electrolyte and nonelectrolyte solution matter. *AIP Conference Proceedings, 1847*, 1-6. Retrieved from <https://aip.scitation.org/doi/abs/10.1063/1.4983909>
- Asy'ari, M., Ikhsan, M., & Muhali, M. (2018). Validitas instrumen karakterisasi kemampuan metakognisi mahasiswa calon guru fisika. *Prisma Sains: Jurnal Pengkajian Ilmu Dan Pembelajaran Matematika Dan IPA IKIP Mataram, 6*(1), 19-26. <https://doi.org/10.33394/j-ps.v6i1.955>
- Bandura, A. (1990). Perceived self-efficacy in the exercise of personal agency. *Journal of Applied Sport Psychology, 2*(2), 128-

163.<https://doi.org/10.1080/10413209008406426>

- Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research and Evaluation, 18*(6), 1–13. <https://doi.org/10.7275/qv2q-rk76>
- Chi, M. T. H., De Leeuw, N., Chiu, M. H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439–477. [https://doi.org/10.1016/0364-0213\(94\)90016-7](https://doi.org/10.1016/0364-0213(94)90016-7)
- Danczak, S. M., Thompson, C. D., Overton, T. L. (2017). 'What does the term critical thinking mean to you?' A qualitative analysis of chemistry undergraduate, teaching staff and employers' views of critical thinking. *Chemistry Education Research and Practice, 18*, 420-434. <https://doi.org/10.1039/C6RP00249H>
- Danczak, S. M., Thompson, C. D., & Overton, T. L. (2020). Development and validation of an instrument to measure undergraduate chemistry students' critical thinking skills. *Chemistry Education Research and Practice, 21*(1), 62-78. <https://doi.org/10.1039/C8RP00130H>
- Dalgety, J., & Coll, R. K. (2006). Exploring first-year science students' chemistry self-efficacy. *Chemistry Education Research and Practice, 1*(1), 2-17. Retrieved from <https://link.springer.com/article/10.1007/s10763-005-1080-3>
- Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity, 12*, 43–52. <https://doi.org/10.1016/j.tsc.2013.12.004>
- Ennis, R. H., Millman, J., & Tomko, T. N. (1985). *Cornell critical thinking tests (3rd ed.)*. Pacific Grove, CA: Midwest Publications.
- Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. Millbrae, CA: California Academic Press.
- Ferrell, B., Phillips, M. M., & Barbera, J. (2016). Connecting achievement motivation to performance in general chemistry. *Chemistry Education Research and Practice, 17*(4), 1054-1066. <https://doi.org/10.1039/C6RP00148C>
- Field, A. (2009). *Discovering statistics using SPSS (3<sup>rd</sup> ed)*. London: Sage Publication Ltd.
- Flaherty, A. A. (2020). A review of affective chemistry education research and its implications for future research. *Chemistry Education Research and Practice, 21*(3), 698–713. Retrieved from <https://doi.org/10.1039/C9RP00200F>
- Frey, R. F., Cahill, M. J., & McDaniel, M. A. (2017). Students' concept-building approaches: A novel predictor of success in chemistry courses. *Journal of Chemical Education, 94*(9), 1185-1194. <https://doi.org/10.1021/acs.jchemed.7b00059>
- Galloway, K. R., & Bretz, S. L. (2015). Development of an assessment tool to measure students' meaningful learning in the undergraduate chemistry laboratory. *Journal of Chemical Education, 92*(7), 1149–1158. <https://doi.org/10.1021/ed500881y>
- Gerritsen-van Leeuwenkamp, K. J., Joosten-ten Brinke, D., & Kester, L. (2019). Students' perceptions of assessment quality related to their learning approaches and learning outcomes. *Studies in Educational Evaluation, 63*(1), 72–82. <https://doi.org/10.1016/j.stueduc.2019.07.005>
- Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory*. Boston, MA: Kluwer Inc.

- Hensen, C., & Barbera, J. (2019). Assessing affective differences between a virtual general chemistry experiment and a similar hands-on experiment. *Journal of Chemical Education*, *96*(10), 2097–2108. <https://doi.org/10.1021/acs.jchemed.9b00561>
- Hoffman, B., & Schraw, G. (2009). The influence of self-efficacy and working memory capacity on problem-solving efficiency. *Learning and Individual Differences*, *19*(1), 91–100. <https://doi.org/10.1021/acs.jchemed.9b00561>
- Kartowagiran, B., Mardapi, D., Purnama, D. N., & Kriswantoro, K. (2019). Parallel tests viewed from the arrangement of item numbers and alternative answers. *Research and Evaluation in Education*, *5*(2), 169–182. <http://dx.doi.org/10.21831/reid.v5i2.23721>
- Kivunja, C. (2014). Innovative pedagogies in higher education to become effective teachers of 21st century skills: Unpacking the learning and innovations skills domain of the new learning paradigm. *International Journal of Higher Education*, *3*(4), 37–48. Retrieved from <https://eric.ed.gov/?id=EJ1067585>
- Kriswantoro, K. Amelia, R. N. & Irwanto (2016). Peningkatan kompetensi calon pendidik kimia melalui item response theory: strategi menghadapi masyarakat ekonomi asean. *Prosiding Seminar Nasional Kimia dan Pendidikan Kimia VIII Program Studi Pendidikan FKIP UNS*, 64–73. Retrieved from <https://osf.io/preprints/inarxiv/nqf79/download>
- Lewis. R. Aiken. (1985). Three Coefficients For Analyzing The Reliability And Validity Of Ratings. *Educational and Psychological Measurement*, *45*, 131–141. <https://doi.org/10.1177%2F0013164485451012>
- Mahanal, S., Zubaidah, S., Sumiati, I. D., Sari, T. M., & Ismirawati, N. (2019). RICOSRE: A learning model to develop critical thinking skills for students with different academic abilities. *International Journal of Instruction*, *12*(2), 417–434. Retrieved from <https://eric.ed.gov/?id=EJ1211048>
- Pyburn, D. T., Pazicni, S., Benassi, V. A., & Tappin, E. E. (2013). Assessing the relation between language comprehension and performance in general chemistry. *Chemistry Education Research and Practice*, *14*(4), 524–541. <https://doi.org/10.1039/C3RP00014A>
- Rasmawan, R. (2020). Development of SETS-based teaching materials in acid-base accompanied by critical thinking exercises and moral forming. *EduChemia (Jurnal Kimia Dan Pendidikan)*, *5*(2), 134. <http://dx.doi.org/10.30870/educhemia.v5i2.7934>
- Redhana, I. W. (2019). Mengembangkan keterampilan abad ke-21 dalam pembelajaran kimia. *Jurnal Inovasi Pendidikan Kimia*, *13*(1), 2239–2253. <https://doi.org/10.15294/jipk.v13i1.17824>
- Retnawati, H. (2014). *Teori respon butir dan penerapannya (untuk peneliti, praktisi pengukuran, dan pengujian mahasiswa pascasarjana)*. Yogyakarta: Nuha Medika.
- Reynders, G., Suh, E., Cole, R. S., & Sansom, R. L. (2019). Developing student process skills in a general chemistry laboratory. *Journal of Chemical Education*, *96*(10), 2109–2119. <https://doi.org/10.1021/acs.jchemed.9b00441>
- 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. Retrieved from <https://eric.ed.gov/?id=EJ1183420>

- Salirawati, D. (2011). Pengembangan instrumen pendeteksi miskonsepsi kesetimbangan kimia pada peserta didik SMA. *Jurnal Penelitian dan Evaluasi Pendidikan*, 15(2), 232-249. <http://dx.doi.org/10.21831/pep.v15i2.1095>
- Saputri, N., Adlim, A., & Inda Rahmayani, R. F. (2018). Pengembangan Instrumen Penilaian Psikomotorik Untuk Praktikum Kimia Dasar. *JTK (Jurnal Tadris Kimiya)*, 3(2), 114-124. Retrieved from <https://core.ac.uk/download/pdf/234031096.pdf>
- Sari, L. P. N., Fajariningtyas, D. A., & Hidayat, J. N. (2020). Validitas instrumen penilaian kemampuan berpikir kritis melalui model problem based learning menuju pembelajaran IPA abad ke-21. *LENSA (Lentera Sains): Jurnal Pendidikan IPA*, 10(2), 125-136. <https://doi.org/10.24929/lensa.v10i2.121>
- Sarigoz, O. (2012). Assessment of the high school students' critical thinking skills. *Procedia - Social and Behavioral Sciences*, 46, 5315-5319. <https://doi.org/10.1016/j.sbspro.2012.06.430>
- Saxton, E., Belanger, S., & Becker, W. (2012). The critical thinking analytic rubric (CTAR): Investigating intra-rater and inter-rater reliability of a scoring mechanism for critical thinking performance assessments. *Assessing Writing*, 17(4), 251-270. <https://doi.org/10.1016/j.asw.2012.07.002>
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., & Marino, J. P. (2019). Assessment of university students' critical thinking: next generation performance assessment. *International Journal of Testing*, 19(4), 337-362. <https://doi.org/10.1080/15305058.2018.1543309>
- Shaw, A., Liu, O. L., Gu, L., Kardonova, E., Chirikov, I., Li, G., Hu, S., Yu, N., Ma, L., Guo, F., Su, Q., Shi, J., Shi, H., & Loyalka, P. (2020). Thinking critically about critical thinking: validating the russian heighten® critical thinking assessment. *Studies in Higher Education*, 45(9), 1933-1948. <https://doi.org/10.1080/03075079.2019.1672640>
- Sumarni, W., Supardi, K. I., & Widiarti, N. (2018). Development of assessment instrument to measure critical thinking skills. *IOP Conference Series: Materials Science and Engineering*, 349(1), 1-11. Retrieved from <https://iopscience.iop.org/article/10.1088/1757-899X/349/1/012066/meta>
- Sumarni, Woro, Sudarmin, S., Wiyanto, W., & Supartono, S. (2016). Preliminary analysis of assessment instrument design to reveal science generic skill and chemistry literacy. *International Journal of Evaluation and Research in Education (IJERE)*, 5(4), 331-340. Retrieved from <https://eric.ed.gov/?id=EJ1132223>
- Sternod, L., & French, B. (2015). *The Watson-Glaser™ II critical thinking appraisal*. Journal of Psychoeducational Assessment Pullman, USA: Washington State University.
- Tan, K. S., Goh, N. K., & Chia, L. S. (2006). Bridging the cognitive-affective gap: teaching chemistry while advancing affective objectives: the singapore curricular experience. *Journal of Chemical Education*, 83(1), 59-63. <https://doi.org/10.1021/ed083p59>
- Utami, B., Saputro, S., Ashadi, A., Masykuri, M., & Widoretno, S. (2017). Critical thinking skills profile of high school students in learning chemistry. *International Journal of Science and Applied Science: Conference Series*, 1(2), 124-130. <http://dx.doi.org/10.20961/ijscacs.v1i2.5134>

Utomo, T., & Ruijter, K. (1994). *Peningkatan dan pengembangan pendidikan*. Jakarta: Gramedia Pustaka Utama.

Uzuntiryaki-Kondakçi, E., & Çapa-Aydin, Y. (2013). Predicting critical thinking skills of university students through metacognitive self-regulation skills and chemistry self-efficacy. *Kuram ve Uygulamada Egitim Bilimleri*, 13(1), 666–670. Retrieved from <https://eric.ed.gov/?id=EJ1016667>

Xu, X., & Raker, J. R. (2016). Self-efficacy and academic performance in first-semester organic chemistry: testing a model of reciprocal causation. *Chemistry Education Research and Practice*, 1(1), 1-

9. <https://doi.org/10.1039/C6RP00119J>

Zimmerman, B. J. (2000). Self-Efficacy: An Essential Motive to Learn. *Contemporary Educational Psychology*, 25(1), 82–91.

Živković, S. (2016). A model of critical thinking as an important attribute for success in the 21<sup>st</sup> century. *Procedia – Social and Behavioral Sciences*, 232(4), 102–108. <https://doi.org/10.1039/C6RP00119J>

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. <https://doi.org/10.1039/B1RP90004H>