Jurnal Pendidikan MIPA
Volume 16. Nomor 4, Tahun 2026 | ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

Development of Reading and Numeracy Literacy Assessment Instrument in Electrolyte and Non-Electrolyte Solutions for Critical Thinking Skills

Amelia Ineza Karima[1)], Triannisa Rahmawati[1),*], Wiwi Siswaningsih[1)], Nahadi[1)]

[1)]Universitas Pendidikan Indonesia

*Coresponding Author: triannisa.ra@upi.edu

ABSTRACT

The development of assessment instruments that can measure reading literacy, numeracy literacy, and critical thinking skills is still a necessity in chemistry learning. This study aims to develop reading literacy and numeracy assessment instruments to measure critical thinking skills in the material of electrolyte and non-electrolyte solutions. This study used a test design and development method based on the *Standards for Educational and Psychological Testing*. The instrument was validated by five experts and tested on 21 tenth-grade high school students. Data analysis included content validity (CVR), empirical validity (Pearson Product Moment), reliability (KR-20 and Cronbach's Alpha), difficulty level, and distinguishing power. The results showed excellent content validity, with CVR values of 0.992, 1.000, and 0.968, respectively. Of the 35 items, 23 (65.71%) were empirically valid. The reliability coefficients were 0.70 for multiple choice, 0.93 for essay, 0.72 for complex multiple choice – true/false, and 0.15 for matching. The difficulty level was dominated by the moderate category (57%). distinguishing power shows that 55% of the test items are in the fair to good category. Overall, the instrument met most of the quality criteria and showed potential for assessing reading comprehension, numeracy, and critical thinking skills, although further improvements were suggested.

Keywords: Assessment; Reading Literacy; Numeracy Literacy; Critical Thinking Skills; Electrolyte and Non-Electrolyte Solutions

INTRODUCTION

Education in the 21st century increasingly focuses on providing essential skills such as critical thinking, creativity, communication, and collaboration, known as the "4C" (Sulistyaningrum et al., 2024). To determine the achievement of these skills, an assessment system is needed that not only measures mastery of the material but also evaluates students' ability to understand, interpret, and apply knowledge in various life contexts (Putri et al., 2021). Therefore, developing assessments that are appropriate to the demands of the 21st century is a crucial aspect of improving the quality of learning.

In this context, reading literacy and numeracy assessments are the answer to improving the quality of learning because they enable students to critically evaluate information, interpret data, and engage in logical problem-solving (Rahmawati & Kamaludin, 2023). In line with this, chemistry learning requires competencies in reading literacy and numeracy because both support conceptual understanding, data interpretation, and accurate problem-solving in chemistry (Shidiq et al., 2022). Furthermore, the Indonesian policy context positions minimum competency assessments as a replacement for national exams, emphasizing literacy and numeracy as core competencies to prepare students for 21st-century challenges, including scientific literacy and problem-solving in real-world contexts (Inganah et al., 2023; Jahro et al., 2024). However, achievement in reading literacy and numeracy competencies remains suboptimal. Based on the 2025 Indonesian Education Report, the reading literacy rate for high school students reached 72.89%, while numeracy literacy reached 70.31%, indicating that approximately 27–30% of students have not yet achieved the expected competencies (Kementerian Pendidikan Dasar dan Menengah, 2025). This situation demonstrates the need to develop assessments capable of measuring and encouraging improvements in students' reading and numeracy skills.

The application of reading literacy and numeracy assessments in chemistry learning requires assessments that not only measure memory but also the ability to understand scientific texts, interpret data, and apply

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

mathematical reasoning to everyday problems. Therefore, in developing chemistry questions, it is necessary to focus on reading literacy and numeracy to align with current assessment requirements in chemistry learning (Shidiq et al., 2022). Reading literacy and numeracy assessments are also oriented towards the application of knowledge in everyday life contexts, beyond mere memorization, thus supporting the development of students who are able to reason critically in everyday life (Couture et al., 2017; Evans et al., 2012). These critical thinking skills are formed when students can understand and interpret information from various sources, then use quantitative reasoning to analyze data and draw conclusions based on scientific evidence (Hapudin & Mujazi, 2024).

In the context of chemistry, one suitable material for measuring students' critical thinking skills is electrolyte and non-electrolyte solutions. This material not only requires students to understand theoretical concepts but also requires them to connect everyday phenomena with scientific evidence by analyzing observations, interpreting experimental data, and explaining the relationship between electrical conductivity, types of substances, and ionization processes. In this process, students need to be able to read scientific information, interpret tables or graphs of experimental results, compare various data, and draw conclusions based on existing evidence (Eljinsa & Zamhari, 2024; Silvia et al., 2021). These characteristics indicate a close relationship between electrolyte and non-electrolyte solutions and reading literacy, numeracy, and critical thinking skills. Therefore, an assessment instrument is needed that can measure these three abilities in an integrated manner (Nasruddin & Jahring, 2024; Prasetyani et al., 2024).

In addition to the need for instruments that can measure reading literacy and numeracy skills, as well as critical thinking skills, developments in educational technology are also driving the transformation of assessment implementation towards digital in chemistry learning because it supports the creation of more meaningful and student-centered learning, thus having the potential to increase student engagement during the learning process (Iqlima et al., 2026). The use of digital platforms allows assessments to be carried out online, in real time, and equipped with an automatic data analysis system that can identify patterns and trends in student learning outcomes (Ilahi et al., 2025; Sudirman et al., 2022). Previous studies in chemistry education that emphasize literacy-based and technology-based approaches report improvements in students' analytical and reflective thinking when assessments are contextualized with real-world phenomena and supported by digital tools. However, an integrated instrument that combines reading literacy, numeracy, and chemistry content in a digital format is still needed (Ridhoni et al., 2022; Saputri et al., 2022).

Among the various digital platforms available, learning management systems (LMS), such as Moodle, provide assessment features that can facilitate teachers in creating, managing, and evaluating student tests and assignments. Implementing this technology not only makes the assessment process more efficient but also supports the implementation of authentic assessments that meet the needs of 21st-century learning (Cuhanazriansyah et al., 2025). However, many educators have not integrated digital technologies, such as LMSs, into the assessment process, which reduces student engagement and the provision of timely feedback (Rahmawati & Kamaludin, 2023; Widianto et al., 2024).

Several studies have developed various assessment instruments in chemistry learning that are oriented towards reading literacy, numeracy literacy, critical thinking skills, and the use of LMS. However, the focus and scope of development are still diverse so they are not able to accommodate the needs of the latest assessment that integrates reading and numeracy literacy, critical thinking skills, and the use of digital platforms in a single instrument. Siswaningsih et al. (2023) conducted research on the implementation of reading and numeracy literacy assessment instruments for the topic of electrolyte and non-electrolyte solutions in the context of ethnoscience, but only focused on reading and numeracy literacy skills and did not utilize digital platforms such as LMS. In addition, Amelia & Saputri (2026) developed a minimum competency assessment instrument for the topic of materials, elements, compounds, and mixtures using the E-Ujian CBT platform that only focused on science literacy but did not focus on numeracy literacy and critical thinking skills. Ware et al. (2024) developed an assessment instrument that integrates chemical literacy and HOTS in acid-base material, but the focus is still on the development of chemical literacy instruments in the form of conventional assessments and has not integrated reading and numeracy literacy and the use of LMS in the assessment process. Based on these studies, research on the development of LMS-based assessment instruments that integrate reading and numeracy literacy skills with critical thinking skills in electrolyte and non-electrolyte solution material is still limited. Therefore,

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

the novelty of this study lies in the development of LMS-based reading and numeracy assessment instruments to measure students' critical thinking skills in electrolyte and non-electrolyte solution material.

Based on the description, this research was conducted with the title "Development of Reading and Numeracy Literacy Assessment Instrument in Electrolyte and Non-Electrolyte Solutions for Critical Thinking Skills".

METHOD

This study employed the test design and development method based on the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014). This method was chosen because it provides systematic guidelines in the development and validation of educational assessment instruments, from the preparation of test specifications to the evaluation of instrument quality. This framework consists of four stages, namely: (1) the stage of developing and evaluating test specifications; (2) the stage of developing, testing, and evaluating test items; (3) the stage of compiling and evaluating new test forms; and (4) the stage of developing procedures and materials for administration and assessment. However, this research is limited to only the first two stages because it is focused on the development and testing process. Participants in this study involved five experts in the field of chemistry education, namely three lecturers from the Chemistry Education Study Program, FPMIPA UPI and two high school chemistry teachers as validators, as well as 21 high school students in grade X at a school in Bandung. This study only involved 21 high school students in grade X as respondents for the limited trial because this study aimed to conduct an initial trial of the developed instrument before testing on a wider scale. This research was conducted for two days in May 2026. The research procedure in this study can be seen in Figure 1 below.

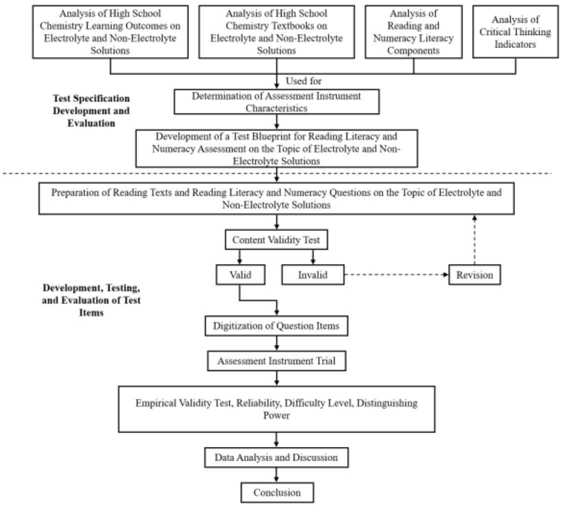

Figure 1. Research Procedure

The development and evaluation stage of the test specifications was carried out through an analysis of learning outcomes in the electrolyte and non-electrolyte solution material, an analysis of high school chemistry textbooks on the electrolyte and non-electrolyte solution material, an analysis of reading literacy and numeracy components, and an analysis of critical thinking indicators in the electrolyte and non-electrolyte solution material. Next, the characteristics of the assessment instrument to be developed were determined. The determination of these characteristics included the form of questions, test format, cognitive level, aspects measured, reading text approach, and expected instrument quality criteria. The results of this characteristic determination were then used as the basis for developing the test blueprint. The test blueprint was developed as a systematic guide for constructing the test items. The main components contained in the test blueprint are the title of the information text, learning outcomes, sub-material, reading literacy or numeracy components, critical thinking indicators, test item indicators, test format, and question number. The test blueprint was then used as the basis for developing the test items for the reading literacy and numeracy assessment instrument.

The next stage is the development, testing, and evaluation of test items. At this stage, reading literacy and numeracy instruments are prepared in various question formats based on the test specification stage that has been carried out previously. The items are designed to measure reading literacy and numeracy, as well as critical

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

thinking skills based on predetermined indicators. The developed instrument is then validated by an expert validator to assess the alignment between the text and the test items, the alignment between the item indicators and the test items, and the alignment between the critical thinking indicators and the test items. The validity used is content validity. The data analysis technique for content validity is carried out using the Content Validity Ratio (CVR) method according to(Lawshe, 1975) using the following equation.

$$CVR = \frac{ne - \frac{N}{2}}{\frac{N}{2}} \tag{1}$$

Where ne is the number of validators who declared the item valid and N is the total number of validators. In this study, five validators were involved, so the minimum CVR threshold for declaring a test item valid according to (Lawshe, 1975) is 0.99 at a significance level of 0.05.

The reading literacy and numeracy assessment instruments that had undergone validation and were declared suitable by the validators were then converted into the Moodle LMS. Subsequently, a pilot test and data analysis were conducted. This instrument pilot test and data analysis were conducted to obtain empirical data related to the instrument's use in the field and to assess the quality of the developed test items. The pilot test data were then comprehensively analyzed to determine the quality of the test items, including empirical validity, reliability, difficulty level, and distinguishing power. The item quality analysis in this study used the Classical Test Theory (CTT) approach because the research objective was to evaluate the initial quality of the instrument through an analysis of empirical validity, reliability, difficulty level, distinguishing power. The CTT approach was chosen because it is commonly used in educational evaluation and is suitable for assessing item characteristics in the early stages of assessment instrument development (Sarea & Ruslan, 2019). The data analysis techniques used to determine the quality of the test items were as follows:

Empirical Validity

The calculation of empirical validity was carried out using the Pearson's Product Moment Correlation Coefficient technique with the help of Microsoft Excel to obtain the calculated r value with the equation referring to (Riduwan, 2013) as follows.

$$r_{xy} = \frac{n\sum XY - (\sum XY)(\sum Y)}{\sqrt{\left((\sum X^2) - (\sum X)^2\right)(n\sum Y^2 - (\sum X)^2)}} \tag{2}$$

Where r_{xy} is the correlation coefficient between variable X and variable Y, n is the number of samples, $\sum X$ is the sum of item scores, and $\sum Y$ is the sum of total item scores. Empirical validity in this study used a significance level of 0.05 with a total of 21 students as respondents, so the critical r value used was 0.433. The test item is declared valid if the calculated r value is greater than the critical r value. Interpretation of the validity correlation coefficient value refers to the categories according to (Arikunto, 2018).

Reliability

In this study, the reliability calculation of the test items used the Kuder-Richardson 20 (KR-20) and Cronbach's Alpha techniques with the help of Microsoft Excel according to the form of the test and the type of scoring. According to Ramadhan et al. (2024), the KR-20 technique is used to measure the reliability of test items that have a maximum score of one or are dichotomous, such as in the form of multiple-choice questions. The Cronbach's Alpha technique is used to measure the reliability of test items with a maximum score of more than one or are polytomous, such as in the form of essay questions, matching, and complex multiple-choice questions. For the Kuder-Richardson 20 (KR-20) technique, the equation used refers to (Sugiyono, 2023) as follows.

$$r_i = \frac{k}{(k-1)}\left\{\frac{S_t^2 - \sum p_i q_i}{S_t^2}\right\} \tag{3}$$

Where r_i is the instrument reliability coefficient, k is the number of test items, p_i is the proportion of correct answers, q_i is 1 – p_i, dan S_t^2 is the test variance. For the Cronbach's Alpha technique, the equation used refers to (Slamet & Wahyuningsih, 2022) as follows.

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

$$r_x = \frac{n}{n-1}\left(1 - \frac{\sum \sigma_t^{\,2}}{\sigma_t^{\,2}}\right) \tag{4}$$

Where r_x is item reliability, n is the number of items, $\sum \sigma_t^{\,2}$ is the score variance for each item, and $\sigma_t^{\,2}$ is the total score variance. The interpretation of the KR-20 reliability coefficient refers to the criteria according to Djaali & Muljono (2008), while the interpretation of the Cronbach's Alpha coefficient refers to the criteria according to Gliem & Gliem (2003).

Difficulty Level

In this study, the difficulty index was calculated using Microsoft Excel. To calculate the difficulty index of a question item with a maximum score of one, the equation used refers to (Fatimah & Alfath, 2019) as follows.

$$P = \frac{B}{JS} \tag{5}$$

Where P is the difficulty index, B is the number of students who answered the question correctly, and JS is the total number of students who took the test. To calculate the difficulty level of a question with a maximum score of more than one, the equation used refers to (Arifin, 2012) as follows.

$$P = \frac{\bar{x}}{X_{max}} \tag{6}$$

Where P is the difficulty index, $\bar{x}$ is the average score, and X_{max} is the maximum score for the item. The classification of difficulty levels refers to the categories according to (Arikunto, 2018).

Distinguishing Power

In this study, the discrimination index was calculated using Microsoft Excel. Before calculating the discrimination index, the students were divided into two groups: an upper group and a lower group. Based on a total of 21 students, the 10 students with the highest total scores were grouped into the upper group, and the 10 students with the lowest total scores were grouped into the lower group. One other student was in the middle group and was not included in the discrimination index calculation. This division was carried out to identify differences in answers between students with high and low abilities on each question. This way, the discrimination index value for each question could be analyzed more accurately. To calculate the discrimination index for a question item with a maximum score of one, the equation used refers to Arikunto (2018) as follows.

$$D = \frac{B_A}{J_A} - \frac{B_B}{J_B} \tag{7}$$

Where D is the discrimination index, J_A is the number of participants in the upper group, J_B is the number of participants in the lower group, B_A is the number of participants in the upper group who answered correctly, and B_B is the number of participants in the lower group who answered correctly. To calculate the discrimination index for an item with a maximum score of more than one, the equation used refers to Arifin (2012) as follows.

$$D = \frac{average\ score\ of\ the\ upper\ group - average\ score\ of\ the\ lower\ group}{maximum\ score\ of\ the\ question} \tag{8}$$

Where D is the discrimination index. The classification of the discrimination index range refers to the categories according to (Arikunto, 2018).

In this study, 40 questions were tested. However, five multiple-answer multiple-choice (MAMC) questions were not included in the item quality analysis. This was due to technical constraints in the Moodle LMS during the trial, such as the system being unable to process answers when students selected more than two options. As a result, the answers obtained from these five questions did not fully represent students' responses and potentially affected the accuracy of the item quality analysis. Therefore, these five questions were not included in the item quality analysis.

RESULTS AND DISCUSSION

The development and evaluation stage of test specifications is carried out through an analysis of learning outcomes in the electrolyte and non-electrolyte solution material by referring to the Decree of the Head of the

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

Education Standards, Curriculum, and Assessment Agency of the Ministry of Primary and Secondary Education Number 046/H/KR/2025 (BSKAP, 2025). This analysis aims to adjust the material to the applicable learning outcomes, determine the scope of relevant material, and formulate the depth and breadth of the material. Based on the results of the analysis, the sub-materials of electrolyte and non-electrolyte solutions used are: (1) electrolyte solutions; (2) strong electrolyte solutions; (3) weak electrolyte solutions; (4) non-electrolyte solutions; (5) ionization; and (6) dissociation. The electrolyte and non-electrolyte solution material is analyzed thoroughly by referring to high school Chemistry textbooks and General Chemistry books so that the assessment instruments developed remain aligned with the curriculum, have conceptual suitability, and avoid misconceptions or inappropriate presentation of the material.

The development of test items is based on the analysis of the reading and numeracy literacy framework. The content used in this reading and numeracy literacy assessment is informational text containing a scientific context. In the reading literacy aspect, the competencies measured are: (1) access and retrieve; (2) interpret and integrate; and (3) evaluate and reflect. In the numeracy literacy aspect, the domain used is data & uncertainty with the subdomains of data & representation and uncertainty & opportunity. The cognitive levels used in the numeracy literacy aspect are: (1) knowing; (2) applying; and (3) reasoning. This reading and numeracy literacy aspect is also integrated with critical thinking skills so that the developed instrument can measure students' critical thinking skills. The indicators of critical thinking skills used are based on Facione (2011), namely interpretation, analysis, inference, evaluation, explanation, and self-regulation. A description of each critical thinking indicator according to Facione (2011) is presented in Table 1 below.

Table 1. Description of Critical Thinking Indicators

Indicator	Indikator Description
Interpretation	The ability to understand and express the meaning or significance of various experiences, situations, data, events, assessments, rules, procedures, and so on
Analysis	The ability to identify intended and actual inferential relationships among statements, concepts, descriptions, or other forms of representation to express beliefs, judgments, experiences, reasons, information, or opinions
Inference	The ability to identify the elements needed to draw reasonable conclusions, to form conjectures and hypotheses to consider relevant information
Evaluation	The ability to assess the credibility of statements or other representations that are descriptions or descriptions of a person's perceptions, experiences, situations, judgments, beliefs, or opinions to assess the logical strength of actual inferential relationships
Explanation	The ability to state and justify reasoning based on evidential, conceptual, methodological, criteria, and contextual considerations and to present such reasoning in the form of convincing arguments
Self-Regulation	The ability to monitor one's cognitive activity by applying analytical and evaluation skills to one's own inferential judgments with the aim of questioning, confirming, validating, or correcting reasoning or results obtained

Thus, the developed instrument not only serves to measure students' reading literacy and numeracy skills, but also their critical thinking skills.

Next, the characteristics of the assessment instrument to be developed are determined. This determination is made with the aim of establishing the initial general design of the instrument. These characteristics include the question format, test format, cognitive level, aspects to be measured, reading text approach, and expected instrument quality criteria. The results of this characteristic determination are then used as the basis for developing the test blueprint.

The test blueprint was developed as a systematic guideline for constructing test items. The test blueprint was developed to ensure that the test items align with the reading literacy, numeracy, and critical thinking skill indicators. The main components of the test blueprint are the information text title, learning outcomes, sub-topics, reading literacy and numeracy components, critical thinking indicators, test item indicators, test format, and test number.

 ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

The next stage in this research was the preparation of reading and numeracy literacy instruments based on the test specification stage that had been carried out previously. The process of preparing this instrument began with the creation of reading texts in the form of informational texts relevant to the material of electrolyte and non-electrolyte solutions, totaling five reading texts. Next, the preparation of questions was carried out, the questions consisted of five types of questions, namely multiple choice, complex multiple choice, matching, short answer, and essay. The total number of questions prepared was 50 items, consisting of 25 reading literacy questions and 25 numeracy literacy questions. When viewed from the critical thinking indicators, there were 12 interpretation questions, 15 analysis questions, 7 inference questions, 5 evaluation questions, 7 explanation questions, and 4 self-regulation questions. All questions were spread across the five reading texts that had been prepared previously.

The developed instrument was then validated by five expert validators. Validation was conducted to assess the alignment between the text and the test items, the alignment between the item indicators and the test items, and the alignment between the critical thinking indicators and the test items. The validity used was content validity. Content validity indicates the extent to which the test items in an instrument can describe the competencies to be measured (Utami et al., 2024). At this stage, a content validation sheet was prepared as a reference for the validators in assessing the instrument. This validation sheet contained several assessment components, namely the reading text in the form of information text, test items, answer keys, and scoring guidelines. Data analysis techniques for content validity were carried out using the Content Validity Ratio (CVR) method according to Lawshe (1975). The summary of the results of this validity analysis is presented in Table 2 below.

Table 2. Summary of Content Validity Analysis Results

Content Validity Category	Number of Texts/Quest Items	CVR Value	Remarks
Alignment between text and test items	49	0.992	49 valid questions with a note of 1 valid question with revision and 1 invalid question
Alignment between item indicators and test items	50	1*	50 valid questions with a note that 21 questions are valid with revisions
Alignment between critical thinking indicators and test items	47	0.968	47 valid questions with a note that 24 questions are valid with revision and 3 questions are invalid

The minimum CVR limit value for declaring a test item valid in this study is 0.99. Table 2 shows that in the category of alignment between text and test items, 49 test items were declared valid, with 1 test item declared valid with revision and 1 test item declared invalid, so that the overall CVR value in this category is 0.992. In the category of alignment between item indicators and test items, 50 test items were declared valid, but there were 21 test items declared valid with revision, so that the overall CVR value in this category is 1. The * symbol on this CVR value indicates that although all are valid, there are still some questions that need to be revised. Input from these experts is important so that bias in the preparation of test items can be reduced and to minimize significant errors, such as irrelevant statements that can affect the measurement results. In the category of alignment between critical thinking indicators and test items, 47 test items were declared valid, but there were 24 test items declared valid with revision and 3 test items declared invalid, so that the overall CVR value in this category is 0.968. The overall CVR value for the category of alignment between critical thinking indicators and test items indicates that the CVR value is still below the minimum CVR limit for a test item to be declared valid. This is because there are 3 test items that are declared invalid, which affects the overall CVR value. Therefore, the test items that are declared invalid will not be included in the trial stage. This also applies to the category of alignment between text and test items, namely 1 test item that is declared invalid will not be included in the trial stage.

The reading literacy and numeracy assessment instruments that have been validated and declared suitable by the validators are then converted into the Moodle LMS. The use of the Moodle LMS not only facilitates online assessments but also supports the measurement of reading and numeracy literacy skills, as well as critical thinking skills, by presenting various stimuli in a single platform. Stimuli in the form of reading texts, tables, graphs, images, and videos can be presented in an integrated manner with test items, allowing students to analyze

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

information, interpret data, and draw conclusions based on evidence before answering questions. Compared to conventional paper-based assessments, the Moodle LMS also enables more systematic assessment management through time management, automatic response recording, and real-time storage of assessment results. Using this LMS allows students to access and work on questions online, making the assessment process more efficient, organized, and easier to monitor. The Moodle LMS also supports automatic answer collection and result checking. In addition to containing the assessment instruments, the Moodle LMS also includes learning materials, allowing the assessment process to be integrated with learning activities. The assessment instruments displayed in the Moodle LMS are shown in Figure 2 below.

Figure 2. Assessment Instrument Display on Moodle LMS

Then, before the trial was conducted, students first participated in learning activities through the prepared LMS. Next, the trial and data analysis were conducted. Forty questions were tested because four items were declared invalid in the content validity test and several questions measured the same reading literacy, numeracy literacy, and critical thinking indicators. This instrument trial and data analysis were conducted to obtain empirical data related to the use of the instrument in the field and to assess the quality of the developed items. The data from the trial were then analyzed comprehensively to determine the quality of the items, including empirical validity, reliability, difficulty level, and discriminating power. The analysis of item quality will be presented below.

Empirical Validity

The results of the empirical validity analysis of the instrument are presented in Table 3 below.

Table 3. Empirical Validity Analysis Results

Item Number	Critical r Value	Calculated r Value	Validity Status	Interpretation
1	0.433	0.251649	Invalid	Low
2	0.433	0.187441	Invalid	Very Low
3	0.433	-0.21763	Invalid	Very Low
4	0.433	0.494243	Valid	Moderate
5	0.433	-0.21857	Invalid	Very Low
6	0.433	0.449101	Valid	Moderate
7	0.433	0.497347	Valid	Moderate
8	0.433	0.464587	Valid	Moderate
9	0.433	0.553147	Valid	Moderate
10	0.433	0.491678	Valid	Moderate
11	0.433	0.076131	Invalid	Very Low
12	0.433	0.513018	Valid	Moderate
13	0.433	0.337479	Invalid	Low
14	0.433	0.197963	Invalid	Very Low
16	0.433	0.559176	Valid	Moderate
17	0.433	0.269871	Invalid	Low
18	0.433	0.352691	Invalid	Low
19	0.433	0.628621	Valid	High
20	0.433	0.637333	Valid	High
21	0.433	0.838041	Valid	Very High

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

Item Number	Critical r Value	Calculated r Value	Validity Status	Interpretation
22	0.433	0.385564	Invalid	Low
23	0.433	0.626911	Valid	High
24	0.433	0.723867	Valid	High
25	0.433	0.555433	Valid	Moderate
26	0.433	0.702024	Valid	High
27	0.433	0.503344	Valid	Moderate
28	0.433	0.606906	Valid	High
29	0.433	0.664769	Valid	High
30	0.433	0.730731	Valid	High
31	0.433	0.81244	Valid	Very High
33	0.433	0.28674	Invalid	Low
34	0.433	0.073221	Invalid	Very Low
35	0.433	0.773585	Valid	High
38	0.433	0.543155	Valid	Moderate
40	0.433	0.719141	Valid	High

Table 3 shows that 23 items were declared valid, with 11 items categorized as moderate, 10 as high, and 2 as very high. This indicates that most of the items adequately measured the intended construct. However, 12 items were declared invalid, with 6 items categorized as low and 6 as very low. This indicates that most of the items met the empirical validity criteria. Items deemed empirically invalid require improvement in further research.

The analysis results show that the 12 items that did not meet empirical validity were characterized by relatively complex cognitive demands, requiring students to integrate information from the reading stimulus with the concepts of electrolytes and non-electrolytes to evaluate several statements simultaneously. In one item, students were asked to determine whether four statements concerning the relationship between electrical conductivity, dissolved ions, and water quality were true or false. This type of question requires the ability to comprehend the reading, connect several concepts, and evaluate each statement, potentially increasing response errors if there is information that is not understood or ambiguous wording. Therefore, the revision process can be carried out by simplifying the wording of the statements and clarifying the relationship between the stimulus and the question. These findings are in line with (Nahadi & Firman, 2019) regarding various factors that cause low empirical validity, namely factors inherent in the test text itself, such as unclear sentence structure, overly complex sentence structure, and difficult-to-understand reading material. Furthermore, factors influencing empirical validity can also be observed in student responses, such as the large number of students who did not answer or who answered randomly. This condition is due to a lack of understanding of the material, low accuracy in reading questions, low motivation in doing the test, the physical condition of students when doing the test, and errors in understanding the meaning of the questions.

Reliability

The results of the instrument reliability analysis are presented in Table 4 below.

Table 4. Reliability Analysis Results

Item Format	Analysis Technique	Reliability Coefficient	Reliability Status	Interpretation
Multiple Choice (Maximum Score = 1)	Kuder-Richardson 20 (KR-20)	0.75	Reliable	Moderate
Essay (Maximum Score = 2)	Cronbach's Alpha	0.93	Reliable	Very High
Complex Multiple Choice – True/False (Maximum Score = 4)	Cronbach's Alpha	0.72	Reliable	Acceptable
Matching (Maximum Score = 5)	Cronbach's Alpha	0.15	Not Reliable	Unacceptable

Table 4 shows that three question types were found to be reliable: multiple choice (maximum score = 1), with a reliability coefficient of 0.75, categorized as moderate; essay (maximum score = 2), with a reliability coefficient of 0.93, categorized as very high; and complex multiple choice – true/false (maximum score = 4), with

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

a reliability coefficient of 0.72, categorized as acceptable. However, one question type was deemed unreliable: matching (maximum score = 5), with a reliability coefficient of 0.15, categorized as unacceptable. These analysis results indicate that most of the items in the developed instrument had good internal consistency. However, the matching item format still requires improvement.

The low reliability coefficient for the matching questions is thought to be influenced by the characteristics of the items developed. Both matching items require students to connect several pairs of related concepts within a single context. In one item, students are asked to match the concept of a weak electrolyte with its characteristics, while in the other, students are asked to match the relationship between the working process of a battery with its scientific reasoning. These characteristics lead to a single error in understanding the main concept, which can influence several answer pairs simultaneously, resulting in low response consistency between students. Furthermore, the matching questions in this study only consisted of two items, resulting in relatively limited score variation. This condition is thought to contribute to the low reliability coefficient obtained. Therefore, the matching questions need to be refined by increasing the number of items and arranging more varied concept pairs to improve the instrument's internal consistency. These findings align with Muharromah et al. (2025) regarding various factors that contribute to low reliability coefficients, namely too few items, unclear instructions, and the physical condition of the students during the test. An instrument with too few items can potentially provide unstable measurements and underrepresent true abilities. Unclear test instructions can lead to students' inability to understand the meaning of the questions. Students who are not feeling well during the test can potentially lead to inconsistent answers, which can reduce the stability of the measurement results.

Most of the questions in the developed instrument can be considered reliable because the reliability coefficient is above the minimum limit of 0.70 for an instrument to be considered reliable (Fraenkel et al., 2012).

Difficulty Level

The results of the analysis of the level of difficulty are presented in the form of a table of distribution of the level of difficulty of the test items in Table 5 below.

Table 5. Distribution of Level of Difficulty of Question Items

Interpretation	Number of Items	Percentage (%)
Easy	8	23
Moderate	20	57
Difficult	7	20
Total Number of Item	35	

Table 5 shows that there are 8 easy items, representing 23%, 20 moderate items, representing 57%, and 7 difficult items, representing 20%. The percentage ratios of these distributions indicate that the difficulty levels of the items are close to the recommended levels. Arifin (2012) explains that a good difficulty level is one that is evenly distributed, consisting of 25% easy items, 50% moderate items, and 25% difficult items.

Overall, the developed instrument demonstrates a fairly balanced difficulty level. This indicates that the instrument is of good quality because it is dominated by moderate items, allowing students' cognitive abilities to be differentiated proportionally (Arikunto, 2018).

Distinguishing Power

The results of the discriminatory power analysis are presented in the form of a table of the distribution of the discriminatory power of the test items in Table 6 below.

Table 6. Distribution of Distinguishing Power of Question Items

Interpretation	Number of Items	Percentage (%)
Good	10	29
Fair	9	26
Poor	13	37
Very Poor	3	9
Total Number of Items	35	

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

Table 6 shows that 10 items were categorized as good (29%), 9 items were categorized as fair (26%), 13 items were categorized as poor (37%), and 3 items were categorized as very poor (9%). This indicates that more than half of the items in the developed instrument, totaling 55%, had fair to good discriminatory power and met the criteria for discriminatory power, making them reliable for distinguishing between high-ability and low-ability students. However, another 45% of the items had poor to very poor discriminatory power, making them unable to effectively differentiate student abilities. Therefore, item improvements are needed in future research.

The average discrimination index value for all items was 0.25, categorized as fair, indicating that, in general, the items in the developed instrument had fairly good discriminatory power. However, 45% of the items still needed improvement. Analysis of several items with low discriminatory power indicates that question characteristics also influence the item's ability to differentiate between high- and low-ability students. Some items use a complex multiple choice – true/false format, requiring students to evaluate several statements based on a single core concept. In this context, misunderstandings of the underlying concept lead to multiple statements being answered incorrectly simultaneously, resulting in low score variation among students. Furthermore, several essay questions require open-ended responses, resulting in more diverse student responses and an inability to fully differentiate between student ability levels. There are also items that measure the basic concept of electrolytes through relatively simple statements that can be answered correctly by most students. These characteristics are thought to contribute to the low discriminatory power of some items. Therefore, revisions were made through refining the item wording, improving the quality of distractors, and adjusting the level of question complexity so that each item is more capable of differentiating students based on their ability levels. These findings align with Warju et al. (2020) regarding various factors contributing to low discrimination index scores, namely questions that are too easy or too difficult and distractors that are not functioning properly. If a question is too easy, it will be answered correctly by both high-ability and low-ability students. Conversely, if a question is too difficult, it will be answered correctly by both high-ability and low-ability students. A distractor that is not functioning properly means that the wrong answer choices on a question are too easy to guess or are not attractive to students. As a result, the question cannot differentiate between high-ability and low-ability students.

CONCLUSION

This research successfully developed an LMS-based assessment instrument that integrates reading literacy, numeracy, and critical thinking skills in the topic of electrolyte and non-electrolyte solutions. Overall, the developed instrument demonstrated good quality based on content validity, empirical validity, reliability, difficulty level, and discriminatory power. However, several items still require refinement to meet all instrument quality criteria. Therefore, this instrument has the potential to be used as an alternative 21st-century assessment tool that supports the measurement of reading literacy, numeracy, and critical thinking skills in chemistry learning.

This research has limitations because instrument development only covered the first two stages of the Standards for Educational and Psychological Testing, namely (1) developing and evaluating test; and (2) specifications and developing, testing, and evaluating test items. The stages of developing a new test format and developing administration and scoring procedures have not yet been implemented. Therefore, further research is recommended to continue these development stages, involve a larger number of respondents, and refine items that do not meet the instrument quality criteria to obtain a more comprehensive instrument ready for wider implementation.

Bibliography

Amelia, R. N., & Saputri, N. Y. (2026). Developing a Minimum Competency Assessment in Science Literacy: A Rasch-Model Approach on Elements, Compounds, and Mixtures. *Unnes Science Education Journal*, *15*(1), 1–12. https://doi.org/10.15294/usej.v15i1.34242

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Arifin, Z. (2012). *Evaluasi pembelajaran*. Direktorat Jenderal Pendidikan Islam, Kementerian Agama RI.

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

Arikunto, S. (2018). *Dasar-Dasar Evaluasi Pendidikan* (3rd ed.). Bumi Aksara.

BSKAP. (2025). *Keputusan Kepala Badan Standar, Kurikulum, dan Asesmen Pendidikan Nomor 046/H/KR/2025 tentang capaian pembelajaran pada pendidikan anak usia dini, jenjang pendidikan dasar, dan jenjang pendidikan menengah.*

Couture, É. M., Chouinard, M.-C., Fortin, M., & Hudon, C. (2017). The relationship between health literacy and quality of life among frequent users of health care services: a cross-sectional study. *Health and Quality of Life Outcomes*, *15*(1), 137. https://doi.org/10.1186/s12955-017-0716-7

Cuhanazriansyah, M. R., Perta, H., & Wildan, A. (2025). Platform Penilaian Pembelajaran Berbasis Teknologi: Tinjauan Bentuk dan Tantangan. *Journal of Educational Research and Community Service*, *1*, 190–196.

Djaali, & Muljono. (2008). *Pengukuran dalam Bidang Pendidikan*. Grasindo.

Eljinsa, S. M., & Zamhari, M. (2024). A Powtoon Animation Video Development Utilizing a Scientific Approach for Teaching Electrolyte and Non-electrolyte Solutions. *Jurnal Penelitian Pendidikan IPA*, *10*(11), 8985–8995. https://doi.org/10.29303/jppipa.v10i11.9545

Evans, K. R., Lewis, M. J., & Hudson, S. V. (2012). The Role of Health Literacy on African American and Hispanic/Latino Perspectives on Cancer Clinical Trials. *Journal of Cancer Education*, *27*(2), 299–305. https://doi.org/10.1007/s13187-011-0300-5

Facione, P. A. (2011). *Critical thinking: What itCritical Thinking: What It Is and Why It Counts is and why it counts*.

Fatimah, L. U., & Alfath, K. (2019). ANALISIS KESUKARAN SOAL, DAYA PEMBEDA DAN FUNGSI DISTRAKTOR. *Al-Manar*, *8*(2), 37–64. https://doi.org/10.36668/jal.v8i2.115

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*.

Hapudin, M. S., & Mujazi, M. (2024). In House Training (IHT) Increasing Teacher Competencies in Building Literacy and Numeration Learning Strategies. *Aktual: Jurnal Pengabdian Kepada Masyarakat*, *2*(1), 55–59. https://doi.org/10.58723/aktual.v2i1.171

Ilahi, D. S. K., Zaini, F. M., Muhammad, B., Humaidi, H., Zakiyullah, A., & Sofa, A. R. (2025). Penilaian Pembelajaran PAI Berbasis Google Forms, Quizizz, dan Grade Scope: *Indonesian Research Journal on Education*, *5*(2). https://doi.org/10.31004/irje.v5i2.2308

Inganah, S., Utomo, D. P., Baiduri, B., & Djunaidi, A. (2023). Accompaniment of Contextual Learning Oriented Minimum Competency Assessment of Mathematics through Lesson Study. *Berdikari: Jurnal Inovasi Dan Penerapan Ipteks*, *11*(2), 234–246. https://doi.org/10.18196/berdikari.v11i2.17692

Iqlima, I., Haryanto, H., Miharti, I., Asrial, A., & Ekaputra, F. (2026). Pengembangan PowerPoint Interaktif Berbantuan Canva pada Materi Ikatan Kimia untuk Menumbuhkan Motivasi Belajar Siswa. *JURNAL PENDIDIKAN MIPA*, *16*(4), 1351–1361. https://doi.org/10.37630/jpm.v16i4.4833

Jahro, I. S., Sutiani, A., Suryani, S., & Nugraha, A. W. (2024). Improving Students' HOTs-Literacy Abilities Through Making Summaries at The End of Learning. *Jurnal Inovasi Pembelajaran Kimia*, *6*(1), 153. https://doi.org/10.24114/jipk.v6i1.57484

Kementerian Pendidikan Dasar dan Menengah. (2025). *Rapor Pendidikan Indonesia 2025: Capaian dan Agenda Perbaikan*. Pusat Standar dan Kebijakan Pendidikan, Badan Standar, Kurikulum, dan Asesmen Pendidikan, Kementerian Pendidikan Dasar dan Menengah.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4).

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

Muharromah, F., Al Fani, F., & Ridho, A. R. (2025). RELIABILITAS ALAT UKUR: JENIS-JENIS, CARA PENGUKURAN, DAN FAKTOR-FAKTOR YANG MEMPENGARUHINYA DALAM PEMBELAJARAN. *JIPDAS (Jurnal Ilmiah Pendidikan Dasar*, *4*(1), 63–73.

Nahadi, & Firman, H. (2019). *Asesmen Pembelajaran Kimia*. UPI Press.

Nasruddin, N., & Jahring, J. (2024). Profile of Junior High School Students' Creative Thinking Ability and Numerical Literacy in Solving Mathematical Problems. *Prisma Sains : Jurnal Pengkajian Ilmu Dan Pembelajaran Matematika Dan IPA IKIP Mataram*, *12*(1), 22. https://doi.org/10.33394/j-ps.v12i1.8577

Prasetyani, K., Ati, L. L., Rondli, W. S., & Kanzunnudin, M. (2024). Quality Analysis of Grade 6 Literacy and Numeracy End-of-Semester Exam Questions Based on Classical Test Theory. *Jurnal Guru Sekolah Dasar*, *1*(1), 1–9. https://doi.org/10.70277/jgsd.v1i1.0001

Putri, M, Krisnaningsih, E, Suprapro, N, Deta, U, Dwikoranto, & D. (2021). Project -based learning (PjBL)-STEM: Bibliometric analysis and research trends (2016 –2020). *Berkala Ilmiah Pendidikan Fisika, 9(3), 368.* https://doi.org/10.20527/bipf.v9i3.11184

Rahmawati, A., & Kamaludin, A. (2023). Development of Motion Graphic Video for Electrolyte and Non-Electrolyte Solutions to Increase Senior High School Students' Interest in Learning. *Jurnal Penelitian Pendidikan IPA*, *9*(4), 1753–1764. https://doi.org/10.29303/jppipa.v9i4.1777

Ramadhan, M. F., Siroj, R. A., & Afgani, M. W. (2024). Validitas and Reliabilitas. *Journal on Education*, *6*(2), 10967–10975. https://doi.org/10.31004/joe.v6i2.4885

Ridhoni, W., Punaji Setyosari, Dedi Kuswandi, Saida Ulfa, & Dahlia Janan. (2022). CC Thinker: Mobile-based Assessment to Train Creative and Critical Thinking in Students. *International Journal of Interactive Mobile Technologies (IJIM)*, *16*(15), 16–29. https://doi.org/10.3991/ijim.v16i15.29991

Riduwan. (2013). *Dasar-dasar Statistika*. Alfabeta.

Saputri, S. V., Dayanti, N., Sari, R. K., Sartika, R. P., & Hadi, L. (2022). DEVELOPMENT OF CRITICAL THINKING ABILITY ORIENTED TEXTBOOK ON ELECTROLYTE AND NON-ELECTROLYTE SOLUTION MATERIALS. *Jurnal Pendidikan Matematika Dan IPA*, *13*(1), 13. https://doi.org/10.26418/jpmipa.v13i1.34741

Sarea, M. S., & Ruslan, R. (2019). KARAKTERISTIK BUTIR SOAL: CLASSICAL TEST THEORY VS ITEM RESPONSE THEORY? *DIDAKTIKA : Jurnal Kependidikan*, *13*(1), 1–16. https://doi.org/10.30863/didaktika.v13i1.296

Shidiq, A. S., Yamtinah, S., Masykuri, M., Ulfa, M., & Mulyani, B. (2022). Penguatan Kemampuan Guru-Guru Kimia SMA di Surakarta dalam Menyiapkan Instrumen AKM Literasi Membaca dan Numerasi berbasis Sains Kimia. *Bubungan Tinggi: Jurnal Pengabdian Masyarakat*, *4*(4), 1402. https://doi.org/10.20527/btjpm.v4i4.6158

Silvia, S., Hamzah, B., & Mustapa, K. (2021). Thinking Ability and Student Learning Outcomes in the Material of Electrolyte and Non Electrolyte Solutions at Madrasah Aliyah Negeri. *Jurnal Riset Pendidikan MIPA*, *5*(1), 1–8. https://doi.org/10.22487/j25490192.2021.v5.i1.pp1-8

Siswaningsih, W., Susetyo, B., Ariesta, A. S., & Rahmawati, T. (2023). *Implementation of minimum competency assessment (MCA) containing ethnoscience on the topic of electrolyte and non-electrolyte solutions*. 090009. https://doi.org/10.1063/5.0113856

Slamet, R., & Wahyuningsih, S. (2022). VALIDITAS DAN RELIABILITAS TERHADAP INSTRUMEN KEPUASAN KER. *Aliansi : Jurnal Manajemen Dan Bisnis*, *17*(2). https://doi.org/10.46975/aliansi.v17i2.428

Sudirman, S., Sarjan, M., Rokhmat, J., Hamidi, H., & Fauzi, I. (2022). Penilaian Pendidikan IPA secara Realtime dan Terintegrasi dengan Artificial Intelligence: Perspektif Filsafat. *Jurnal Ilmiah Profesi Pendidikan*, *7*(4b). https://doi.org/10.29303/jipp.v7i4b.888

Sugiyono. (2023). *Metode penelitian kuantitatif, kualitatif, dan R&D* (2nd ed.). Alfabeta.

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i4.4958

Sulistyaningrum, H, Nuraida, D, Ismail, A, Andik, & M. (2024). The effectiveness of use of PowerPoint–iSpring interactive learning media on student creativity. *Journal of Humanities and Social Sciences Studies, 6 (3), 83 – 90.* https://doi.org/10.32996/jhsss.2024.6.3.9

Utami, L., Festiyed, Dian Purnama Ilahi, Arista Ratih, Elvi yenti, & Lazulva. (2024). ANALISIS INDEKS AIKEN UNTUK MENGETAHUI VALIDITAS ISI INSTRUMEN SCINETIFIC HABBITS OF MIND. *Journal of Research and Education Chemistry*, 6(1), 59. https://doi.org/10.25299/jrec.2024.vol6(1).17430

Ware, K., Mole, P. N., & Nirmalasari, M. A. Y. (2024). Validitas dan Keterbacaan Integrated Assesment Literasi Kimia dan HOTS Bermuatan Etnosains Daerah Sikka pada Materi Asam Basa. *JURNAL PENDIDIKAN MIPA*, *14*(4), 1103–1112. https://doi.org/10.37630/jpm.v14i4.2073

Warju, W., Ariyanto, S. R., Soeryanto, S., & Trisna, R. A. (2020). ANALISIS KUALITAS BUTIR SOAL TIPE HOTS PADA KOMPETENSI SISTEM REM DI SEKOLAH MENENGAH KEJURUAN. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *17*(1), 95. https://doi.org/10.23887/jptk-undiksha.v17i1.22914

Widianto, P. C., Fadhilawati, D., Sutanti, N., & Supriyono. (2024). Examining Quizalize's Impact on Students' Reading Comprehension Mastery. *Jurnal Pendidikan : Riset Dan Konseptual*, *8*(2), 338. https://doi.org/10.28926/riset_konseptual.v8i2.983