JPPIPA 10(3) (2024)

Jurnal Penelitian Pendidikan IPA

Journal of Research in Science Education

http://jppipa.unram.ac.id/index.php/jppipa/index

Digital Assessment Tools: As a Media for Assesing High School Science Learning

Riza Andriani[1], Mellyzar[2*], Isna Rezkia Lukman[2], Muttakin[2], Evawati[2], Ali Imron Pasaribu[2], Mhd. Ridwan Fadli[2]

[1] Department of Physics Education, Universitas Malikussaleh, North Aceh, Indonesia.
[2] Department of Chemistry Education, Universitas Malikussaleh, North Aceh, Indonesia.

Received: December 6, 2023
Revised: January 12, 2024
Accepted: March 25, 2024
Published: March 31, 2024

Corresponding Author:
Mellyzar
melyzar@unimal.ac.id

DOI: 10.29303/jppipa.v10i3.6416

Abstract: This research is to produce a suitable digital assessment tool for high school students' in chemical literacy with validity, reliability, levels of difficulty and discrimination power. The research metodology is Research and Development (R&D) using the Oriondo and Dallo-Antonio model. The research stages include planning the test instrument, testing, determining empirical validity, determining reliability, and interpreting test scores. In the planning (instrument creation), where the validation score estimates by experts in content, language, construction, and media are above 0.8, indicating that the test instrument is considered suitable for further testing to estimate validity, reliability, difficulty levels, and discrimination power. Out of the 25 developed questions, 24 fit the Rasch model. The reliability of Instrument Part I is 0.67, and Part II is 0.65, falling into the "sufficient" category. The questions exhibit varied difficulty levels, ranging from very easy to very difficult, with a standard deviation (SD) of 0.21 for Part I and 0.14 for Part II. The separation achieved by Instrument Part I is into 4 ability groups, while Instrument Part II separates into 2 ability groups. The distractors for multiple-choice questions fall into the "good" and "very good" categories, fulfilling their intended function effectively.

Keywords: Digital assessment; Evaluation; Quizizz; R&D

Introduction

Assessment is an important thing in learning, as a process of collecting data to determine developments and efforts to improve the quality of education. The quality of learning can be seen from the assessment results (Indahri, 2021). Learning outcomes assessment aims to measure the success of the learning process at school, so teachers need to use strategies so that students can learn effectively and efficiently and learning goals can be achieved (Rahayu, 2021). Therefore, a teacher needs a learning evaluation tool, namely an assessment instrument (Prawesti et al., 2021). After the learning process, teachers need to measure the results of how far students have developed during the learning process by giving evaluation tests. Learners is considered qualified if it has gone through a series of examination stages.

In this way, assessment has become the center of special attention in the world of education. Evaluation instruments usually use printed form, this is one of the causes of weaknesses in the implementation of tests such as using a lot of paper, lack of motivation, not immediately collecting answers when the allotted time is up and the assessment process takes a long time so it is felt to be less effective (Iqbal et al., 2018). Apart from that, evaluation can also be carried out using *Tecnology Information and Communication* (ICT) (Purnamawati et al., 2019). This is demonstrated by the increasing development of online exams, both computer-assisted and using Android in the learning process, including in the implementation of graduation exams (Pratiwi, 2017).

The use of ICT as a learning evaluation instrument not only provides convenience and saves time, but can also support programs to increase environmental

How to Cite:
Andriani, R., Mellyzar, Lukman, I. R., Muttakin, Wati, E., Pasaribu, A. I., & Fadli, M. R. (2024). Digital Assessment Tools: As a Media for Assesing High School Science Learning. *Jurnal Penelitian Pendidikan IPA*, *10*(3), 1362–1374. https://doi.org/10.29303/jppipa.v10i3.6416

awareness by minimizing the use of paper (Mudrikah, 2021). Apart from that, the use of ICT-based evaluation instruments is also seen as being able to provide varied instruments and reduce the weaknesses of printed evaluation systems (Iqbal et al., 2018). Considering the importance of assessment instruments in the evaluation process, a teacher as an instructor is required to be able to develop good assessment instruments (Alvina et al., 2022; Oktharia et al., 2017).

Based on the results of direct observations at SMA Negeri 3 Putra Bangsa, especially in class XI Science, the results were that the assessment instrument used was a printed sheet. Researchers also interviewed study teachers and students. Based on interviews with teachers in the field of study, it can be obtained that the assessment instruments use printed form, resulting in wasteful use of paper, cheating occurs during the test, and when the evaluation test is given, some students get bored during the test, this is due to the test being given. less effective. Based on interviews with students, especially those who have studied colloid system material, namely class The form of questions used is multiple choice and essays so that students feel bored and not enthusiastic when taking the test.

Furthermore, the researchers conducted observations at SMA Negeri 2 Seunuddon, specifically in class XI IPA, and obtained results indicating that they face similar issues as SMA Negeri 3 Putra Bangsa. The issue lies in the evaluation process, where assessment instruments are in printed form. This causes problems during test activities, such as students copying answers or failing to submit their responses promptly when the allotted time has elapsed. Based on the observation results, it is evident that the assessment instruments used are not interactive. Therefore, there is a need to develop interactive assessment software for class XI IPA at SMA Negeri 2 Seunuddon and SMA Negeri 3 Putra Bangsa, aiming to make the evaluation process effective and facilitate teachers in collecting students' test results directly.

One of the interactive software that teachers can develop as an assessment instrument is the quizizz application (Dzikrullah & Syafi'i, 2021). Quizizz is an application that facilitates both teachers and students to access evaluations and learning materials anywhere and anytime (Amany, 2020). Furthermore, quizizz can be used as an assessment instrument with a design that is attractive, creative, innovative, and enjoyable (Ramadhani & Ardi, 2022). The hope is that the use of assessment instruments utilizing technology can address the weaknesses of the traditional printed form system still in use today. Some advantages offered by quizizz-ased assessment instruments include greater efficiency, minimizing paper usage, the ability to quickly randomize questions to reduce cheating, adherence to predetermined time plans, immediate visibility and downloadable results in Excel format, making it easier for teachers to conduct corrections (Capuno, 2023). Therefore, it is necessary to develop a digital assessment for student learning, especially in science learning, in the form of scientific literacy. With this research, we will know the effectiveness of Quizizz as a learning assessment media tool in terms of time efficiency, objectivity, and reducing cheating during assessment implementation.

This research aims to ensure that the quizizz-based assessment instrument provides convenience for both students and teachers in conducting tests or assessments on Colloid System materials. What distinguishes this research from previous studies is the use of literacy-based questions, the research object, the subject or material developed, and the development objectives that the researchers have undertaken.

Method

According to Sugiyono (2016) research methods are a scientific way to research, design, produce and test the validity of products that have been produced. This research is a form of development research or Research and Development (R & D) using the Oriondo and Dallo-Antonio development models. The procedure in this development consists of five stages:

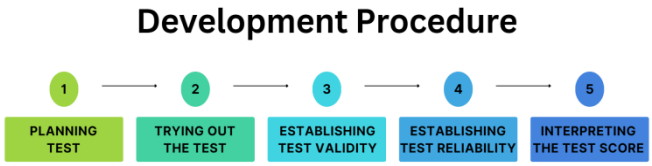

Figure 1. Development Procedure

The data collection instrument using instrument Validation Sheet (validating content, construction and language of the questions) to assess the quality of the assessment instrument prepared by the researcher before being included in the media or application *Quizizz,* Media validation sheet to assess the quality of media in terms of language use, typography, color, and design for the Quizizz-assisted digital assessment tool developed, Student response questionnaire on questions and media from the Quizizz-assisted Digital Assessment Tool developed to assess the practicality of using media as a tool for evaluating student learning outcomes, and Quiz questions for the Digital Assessment Tool assisted by Quizizz to measure the chemical literacy skills of high school students in North Aceh. Validation by 2 lecturers from Samudra Langsa University and 2 chemistry teachers from SMAN 1

Muara Batu and SMAN Modal Bangsa Arun. The test trials in class XI Science (2 Classes) at SMAN 2 Seunuddon and SMAN 3 Putra Bangsa. The data processing technique in this research uses the Winstep program with Rasch (Item Fit) model.

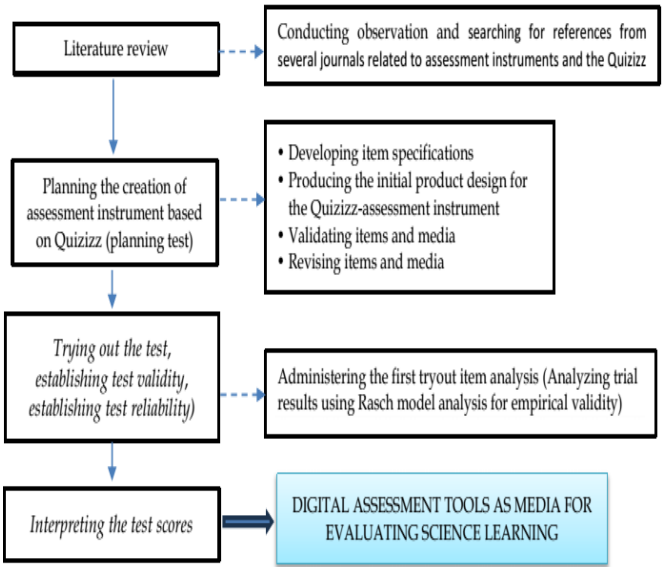

Figure 2. Research process

Result and Discussion

Test Creation Planning

Preliminary Study Determination

Based on the results of observations and interviews by researchers at 2 schools, SMA Negeri 2 Seunuddon and SMA Negeri 3 Putra Bangsa, the results were that in carrying out evaluation tests, the assessment instruments used were printed or paper, this was the main cause of the assessment process being less effective, especially in the classroom. XI IPA on Colloidal materials.

From the results of this review, it is evident that popular digital tools used in formative assessment include Kahoot (194) based on the total research articles indexed in Scopus. However, when considering website visitors, Quizizz (58.4M) emerges as the more popular digital tool. Scopus-indexed studies have extensively utilized Kahoot in formative student assessments, but there is a limited number of researchers (26) who have explored Quizizz as an alternative for formative assessment. Socrative has also gained popularity among researchers in Scopus-indexed articles (Andriani et al., 2024).

Table 1. Popular Digital Tools based on Total Visitors and Scopus Publications Category

Tools	Website	Category Rank (Education)	Total Visit	Scopus
Kahoot	https://kahoot.it	3	40.4M	194
Quizizz	https://quizizz.com/	37	58.4M	26
Edulastic	https://edulastic.com/	121	1.8 M	0
Google Form	https://docs.google.com/forms	N/A	N/A	39
Mentimeter	https://www.mentimeter.com/	353	4M	18
Plickers	https://get.plickers.com	5.789	105.2K	28
Socrative	https://socrative.com/	423	2.2 M	83
Nearpod	https://nearpod.com/	90	5.8 M	20
Formative	https://goformative.com/	592	524.4 K	3
Classflow	https://classflow.com/	3.850	46.2 K	1
Quizalize	https://www.quizalize.com/	3.824	404.1 K	1

In terms of accessibility, all these digital tools are easily accessible by providing standard/basic and premium versions. They are generally compatible with devices operating on both iOS and Android, and are also available in web versions. Therefore, these tools can be accessed on various devices such as laptops, PCs, or smartphones (iOS and Android).

Table 2. Accessibility Digital Tools

Tools	Web Version	Android Version	Ios Version	Version
Kahoot	https://kahoot.it	√	√	Basic, Pro, Premium, Premium+
Quizizz	https://quizizz.com/	√	√	Basic, Premium
Edulastic	https://edulastic.com/			Free, Premium, Enterprise
Google Form	https://docs.google.com/forms	√	√	Free
Mentimeter	https://www.mentimeter.com/	√	√	Basic, Pro
Plickers	https://get.plickers.com	√	√	Basic, Pro
Socrative	https://socrative.com/	√	√	Basic, Pro
Nearpod	https://nearpod.com/	√	√	Basic, Pro
Formative	https://goformative.com/	-	-	Bronze, Silver, Gold
Classflow	https://classflow.com/	√	-	Basic, Pro
Quizalize	https://www.quizalize.com/	√[2]	√[2]	Basic, Premium

In terms of the types of tests that can be carried out by digital tools, Quizizz is superior to others in the basic/standard version. For the paid version, all tools on average offer the same and almost the same types of tests (Andriani et al., 2024).

Table 3. The Types of Tests that Can be Conducted Using Digital Tools Are

Type Test/ Digital Tools	Kahoot	Quizizz	Edulastic	Google Form	Mentimeter	Plickers	Socrative	Nearpod	Formative	Classflow	Quizalize
Multiple Schoice/ True-False	√	√	√	√	√	√	√	√	√	√	√
Short Answer	√*		√	√	√		√		√	√	
Fill in the blank		√	√					√			
Multi Select (Checkbox)	√	√	√	√	√						√
Open Ended		√[1]	√	√	√[2]			√	√	√	
Discussion								√			
Matching			√					√	√*	√	
Sorting	√*								√*	√	
Ordering (Sequencing)			√						√*	√	√
Polling (Survey)	√*				√	√		√		√	
Image Supported	√	√	√	√	√	√*	√*				√
Video Supported	√			√							

Based on the results of this analysis, the researchers decided to develop a digital assessment tool assisted by Quizizz to carry out formative assessments on high school students.

Preparation of Question Grids and Questions

The preparation of the questions is carried out in stages: Preparation of the question grid, preparation of content, linguistic and construction validation sheets, preparation of media validation sheets, and the preparation of questionnaires for student and teacher responses. The assessment instrument developed consists of chemistry literacy questions using the Quizizz application. The resulting final product is used as an assessment instrument to streamline the evaluation test process on colloid system material which consists of 5 variations of 25 questions, including multiple choice questions, rearrangement, short answer 1, matching and short answer 2.

Item Scoring Techniques

Scoring is the first step in the process of processing test results. Scoring is a process of converting test answers into numbers. The measurement results are then converted into values through a certain processing process (Purwanto, 2009).

Table 4. Item Scoring Techniques

Test Formats	Scoring	Score for each item	Percentage
Multiple choice	Each correct answer is given a score 4 and wrong answers are given a score of 0	4	20%
Rearrange	One correct answer = score 1 Two correct answers = score 2 Three correct answers = score 3 Four correct answers = score 4 No correct answers = score 0	4	20%
Short answer 1	Each correct answer is given a score 4 and wrong answers are given a score of 0	4	20%
Matching	One correct answer = score 1 Two correct answers = score 2 Three correct answers = score 3 Four correct answers = score 4 No correct answers = score 0	4	20%
Short answer 2	Each correct answer is given a score 4 and wrong answers are given a score of 0	4	20%
	Total		100%

Initial Product Design of the Quizizz-Based Assessment Instrument User Interface

The initial product study is a media-based assessment instrument *quizizz* which contains colloidal material.

Home Page

The main page consists of questions that have been created with a total of 25 questions with 5 variations. There are three options for quizzes that are ready to be played. First, you can start the quiz directly. Second, it can be used as homework (PR). Third, it can be used in paper mode.

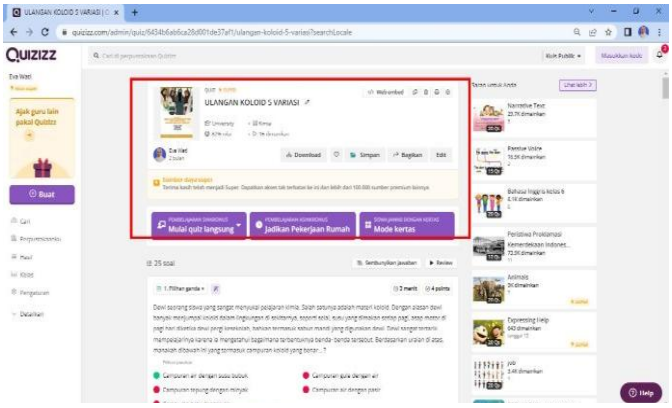

Figure 3. Image of the quiz main page

Quiz Settings Page

On the quiz settings page there are four modes of quiz work, group, individual, classic and exam models. This research uses exam mode so that the assessment process takes place seriously. Modes other than exams are modes that are used specifically for games, such as adding scores if you finish work more quickly, etc.

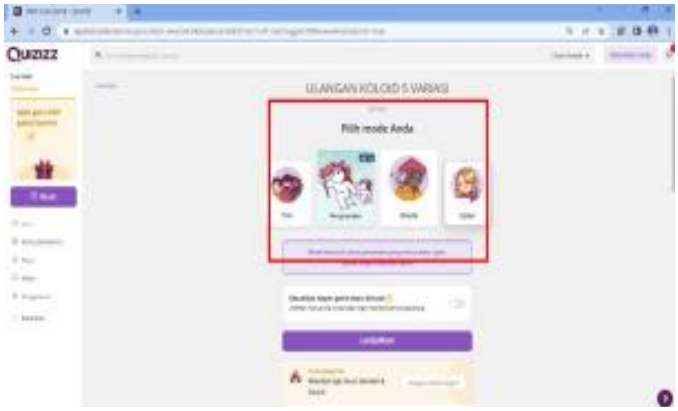

Figure 4. Image of the quiz settings page before sharing

Quiz Work Page

When students enter the quiz page it will look like the one below waiting for a count from 1-5.

Figure 5. Quiz work page

Logical Validity of the Instrument

In the analysis activity, the validity scores of the material and media in this research were analyzed using a formula *Aiken'S V* which is adjusted to the Aiken index. Aiken's index will produce questions that

are valid and worthy of use when evaluating $V \geq 0.80$ for five *raters* with a four scale questionnaire. Meanwhile, media validity, the Aiken index which produces media that is valid and suitable for use is based on the validator score if the value $V \geq 0.79$ for six *raters* with a five scale questionnaire. The validation results for each question item using the Aiken V formula, the results can be obtained that the analysis of the question items with using the Aiken formula is declared valid. The Aiken values that must be achieved are $V \geq 0.80$ for 4 *raters* with a four scale questionnaire. In this study, the highest Aiken score obtained for a question item was 1, while the lowest was 0.94. Media validation analysis results can be obtained that the analysis of each rater's assessment scores on the media using the Aiken formula is declared valid and some are invalid. The Aiken values that must be achieved are $V \geq 0.84$ for four *raters* with a five scale questionnaire. In this study, the highest Aiken value obtained for the media was 0.94, while the lowest Aiken value was 0.84. Therefore, the basic thing that becomes media revision is suggestions and input from validators as support for the media being developed.

Validity and Reliability Test

The results of the next product trial are analyzed using Modern Item Response Theory (IRT), with Rasch modeling using Winstep program version 3.73. The aim of analyzing each item is to obtain characteristics of each question item so that the items used are of good quality. Empirical data in this study consist of ordinal data with values ranging from 1-5. The questions developed consist of 2 parts, Part I and Part II. Part I consists of 3 types of questions, namely multiple-choice, Fill-in the Blank I, Fill in the Blank II, with a total of 15 question items. Part II consists of two types of questions, namely matching and rearranging, with a total of 10 question items. Modern analysis is carried out in stages: analysis of unidimensionality, the level of suitability of question items with the Rasch model: *output Table 10 item coulomb: fit order*, and reliability and separation or groups *item* and *person*.

Unidimensionality Test

Unidimensionality analysis of the assessment instruments was carried out with the help of the Winstep program onoutput tables 23 on analysis unidimensionality. In the unidimensionality test, the minimum limit that must be obtained is 20% (Brentani & Golia, 2007). The results obtained from the unidimensionality test of the Part I instrument were 24.2%, which means that 15 questions met the unidimensionality requirements. Results of assumption testing analysis. The unidimensionality of the Part II instrument is 23.7%, the 10 questions in Part II also meet the requirements for unidimensionality.

		-- Empirical --		Modeled
Total raw variance in observations	=	19.8	100.0%	100.0%
Raw variance explained by measures	=	4.8	24.2%	24.5%
Raw variance explained by persons	=	1.4	7.0%	7.1%
Raw Variance explained by items	=	3.4	17.2%	17.4%
Raw unexplained variance (total)	=	15.0	75.8% 100.0%	75.5%
Unexplned variance in 1st contrast	=	3.1	15.7% 20.7%	
Unexplned variance in 2nd contrast	=	2.4	12.2% 16.1%	
Unexplned variance in 3rd contrast	=	2.4	12.0% 15.8%	
Unexplned variance in 4th contrast	=	1.5	7.8% 10.3%	
Unexplned variance in 5th contrast	=	1.4	6.9% 9.1%	

(a)

		-- Empirical --		Modeled
Total raw variance in observations	=	13.1	100.0%	100.0%
Raw variance explained by measures	=	3.1	23.7%	24.2%
Raw variance explained by persons	=	.8	6.0%	6.1%
Raw Variance explained by items	=	2.3	17.7%	18.1%
Raw unexplained variance (total)	=	10.0	76.3% 100.0%	75.8%
Unexplned variance in 1st contrast	=	3.0	22.6% 29.6%	
Unexplned variance in 2nd contrast	=	2.7	20.7% 27.1%	
Unexplned variance in 3rd contrast	=	1.7	13.0% 17.1%	
Unexplned variance in 4th contrast	=	1.4	10.5% 13.8%	
Unexplned variance in 5th contrast	=	.4	3.1% 4.0%	

(b)

Figure 6. Instrument dimensionality test results part I and part II

Level of Conformity of Question Items to the Rasch Model (Item Fit)

The analysis of the appropriateness of a question item, or whether an item is considered good or not, is referred to as Item Fit. In item fit analysis, it is determined whether a question item falls within the normal category and measures how well the question item can be understood by students. The measurement of item fit for question items can be observed through values such as outfit mean square, outfit z-standard, and Pt-measure correlation. Question items that are considered acceptable are those that meet the measurement accuracy criteria in item fit.

The characteristics for MNSQ, ZSTD, and Pt-Mean Corr values can be stated as follows: MNSQ values are considered acceptable if they fall within the range of 0.5 to 1.5, ZSTD values are considered acceptable if they fall within the range of -2.0 to +2.0; and, Pt-Mean Corr values are considered acceptable if they fall within the range of 0.4 to 0.85 (Amelia, 2021). A question item is deemed unfit (misfit) if it does not meet at least two of the aforementioned criteria and should be replaced with another question item.

Table 5. Instrument Fit Item Analysis

Question Items	Outfit MNSQ	Outfit ZSTD	PT-Measure Correlation
S1	0.87	-0.9	0.46
S2	0.8	-0.7	0.4
S3	1.14	0.6	0.33
S4	1.03	0.2	0.22
S5	1.19	1	0.3
S6	1.16	1	0.4
S7	0.8	-1.3	0.48
S8	0.9	-0.6	0.49
S9	0.78	-1.5	0.46
S10	1.1	0.8	0.4
S11	0.87	-0.9	0.49
S12	0.82	-0.5	0.4
S13	0.91	0.4	0.42
S14	0.64	-2	0.61
S15	2.6	5	-0.01
S16	1.15	0.9	0.4
S17	0.85	-0.9	0.44
S18	0.99	0	0.47
S19	0.92	-0.5	0.42
S20	1.09	0.7	0.4
S21	0.93	0.3	0.4
S22	1.03	0.2	0.32
S23	0.59	0.8	0.34
S24	0.76	-1.6	0.55
S25	0.5	-2	0.63

Based on the description of the diagram above, it can be concluded that the question items, in general, meet the criteria for good item fit in terms of Outfit MNSQ, Z-STD, and Pt-Measure Correlation values. Thus, it can be indicated that there is no misconception among students. A total of 25 question items generally meet the criteria, except for question item number 15. The outfit index for question item number 15 is too high, exceeding the specified outfit value limit, making it classified as a misfit item. Question item 15 cannot be categorized as acceptable; this may be due to two opinions: the item has a defect with poor discriminant power, or it measures a different ability than intended. Misfit items should be monitored as they contribute less to the reliability of test scores.

The minimum criteria for Outfit MNSQ, Z-STD, and Pt-Measure Correlation values that must be met are Outfit MNSQ between 0.5 and 1.5, Z-STD between -2.0 and +2.0, and Pt-Measure Correlation between 0.4 and 0.85. Furthermore, the requirements for the suitability of question items to the Rasch model also indicate that each question item can measure students' mastery of the developed instrument, which is a colloid material-based chemistry literacy question using the Quizizz application, totaling 24 question items that will be used as the final product in assessment instrument development.

Reliability Test

Reliability can be examined in the summary statistic output table in the Winstep output to determine person reliability, item reliability, and Cronbach's Alpha. Student response data is divided into two parts based on the same scoring. Part I includes the scores for multiple-choice questions, fill in the blank 1, and fill in the blank 2, while part II includes rearranging and matching questions. The results of the analysis can be observed in the following table.

Table 6. Instrument Reliability Test

Category	Part I	Part II
Person Reliability	0.53	0.65
Reliability	0.89	0.65
Cronbach Alpha	0.62	0.60

Instrument Part I is found to have a person reliability of 0.53, categorized as moderate, item reliability of 0.89, categorized as high, and Cronbach's Alpha of 0.62, categorized as moderate. Instrument Part II has a person reliability of 0.67, categorized as moderate, item reliability of 0.65, categorized as moderate, and Cronbach's Alpha of 0.60, categorized as fair.

The characteristics of reliability values during testing categorize the adequacy and consistency of both individuals taking the test and the test instrument items themselves. Persons/students taking the test are consistent in responding, and items are consistent in measuring the intended abilities. Therefore, it can be concluded that the developed assessment instrument is reliable as it meets the minimum reliability coefficient value of 0.6 (Lukman et al., 2022; Sujarwanto & Rusilowati, 2015).

Analysis of the Level of Difficulty and Accuracy of Question Items

The results of instrument testing on students will inevitably have a level of difficulty for each question item with specific characteristics. In this study, Rasch modeling analysis was used, allowing the difficulty level of question items to be observed on the Logit scale in the Winstep application. In the Winstep application, the Logit scale can be seen in the Output: Item Measure, providing insights into difficulty levels along with error rates. The Logit scale is divided into groups representing the difficulty level of items and respondents, indicating items from easiest to most challenging and respondents from lowest to highest ability. Essentially, the difficulty level in Rasch modeling theory is similar to the classical test level, comparing the number of questions tested with correct answers. If the Logit values for the difficulty level of items and abilities increase, the question items are

considered better. The value of each difficulty level for items and abilities in students' responses improves as the Logit value increases. A question item is considered good if its error rate is smaller. A good question item is one that can be used to measure and differentiate the abilities of each student. The quality of a question item can be assessed through the standard error (SE). An item or question item is considered good or ideal if SE < (0.5-1.00). An SE value < 0.5 indicates that the item is precise in measuring ability, a value of 0.5 < SE < 1.00 means the item is reasonably precise in measuring ability, and if SE > 1.00, it means the item is not precise in measuring ability.

Table 7. Criteria for Level of Difficulty Questions Parts I and II

Part I		Part II	
Criteria	Decision	Criteria	Decision
$X > 0.21$	Very difficult	$X > 0.14$	Very difficult
$0.0 < X < 0.21$	Difficult	$0.0 < X < 0.14$	Difficult
$-0.21 < X < 0.0$	Easy	$-0.14 < X < 0.0$	Easy
$X < -0.21$	Very easy	$X < -0.14$	Very easy

Table 8. Level of Difficulty and Accuracy of Question Items

Item Question	Question Difficulty Level		Level of Question Accuracy	
	Measure	Decision	S.E Model	Decision
S1	0.15	Difficult	0.06	Carefully
S2	-0.11	Easy	0.07	Carefully
S3	-0.08	Easy	0.07	Carefully
S4	-0.23	Very easy	0.08	Carefully
S5	0.54	Very difficult	0.06	Carefully
S6	-0.16	Very easy	0.08	Carefully
S7	-0.13	Easy	0.08	Carefully
S8	-0.03	Easy	0.08	Carefully
S9	-0.01	Easy	0.08	Carefully
S10	0.19	Very difficult	0.08	Carefully
S11	0.18	Difficult	0.06	Carefully
S12	-0.15	Easy	0.07	Carefully
S13	0.06	Difficult	0.06	Carefully
S14	0.02	Difficult	0.06	Carefully
S15	-0.04	Easy	0.06	Carefully
S16	-0.13	Easy	0.08	Carefully
S17	-0.13	Easy	0.08	Carefully
S18	0.18	Very difficult	0.08	Carefully
S19	0.01	Difficult	0.08	Carefully
S20	0.22	Very difficult	0.08	Carefully
S21	0.02	Difficult	0.06	Carefully
S22	-0.01	Easy	0.06	Carefully
S23	-0.39	Very easy	0.09	Carefully
S24	0.14	Difficult	0.06	Carefully
S25	-0.09	Easy	0.07	Carefully

The difficulty level of a question can be observed in the item measure output, where the acceptance range for the Logit value of item difficulty is -2.0 < X < +2.0. The Standard Deviation value for Item Measure for Part I is 0.21, and for Part II, it is 0.14. We can determine the range for the difficulty level of questions based on the criteria in Table 7, and the difficulty and precision of question items can be seen in Table 8.

Based on the table above, each question item can be categorized as having a different level of difficulty, either very difficult, difficult, easy, and very easy. An item is said to be good if the difficulty value is not less than -2 and not more than +2 (Lukman & Muhammad, 2022). The higher the index value on the difficulty level of a question item, the more difficult the question is, and vice versa. The accuracy of the items in measuring the ability measured is good, where is the score SE < 0.5, with the value moving from 0.06 − 0.09.

Figure 7. (a)Person Map Item (b) Item Map Person

In Part I of the instrument, there is one question categorized as very difficult, which is item number 5, and the easiest question is found at number 23. The logit values for both items and persons can be observed in the output tables for item measure and person measure. The average difficulty (M+) is evident on the logit scale for question items at numbers 14, 21, and 22, with logit values of 0.02, 0.02, and -0.01, respectively. Item number 5 is considered the most challenging by students, likely due to misconceptions among students based on field observations. This insight can prompt educators to better prepare materials before the teaching process. The average abilities of students can be seen on the item map person for 12P, 15P, 47L, 50L, 77L, with logit values of 0.03, and 0.6L, 13L, 14L, 17P, 31P, 41P, 48P, 49P, 52P, 58P, 70P, 79L, 81P, 89L, with logit values of -0.12. Meanwhile, students with the ability to answer high-difficulty questions are those with codes from 04P to 88P, with a logit value of 0.74. A higher logit value indicates better difficulty levels, and vice versa. Students with high logit values not only have high abilities but also excel in answering challenging questions correctly.

The difficulty levels of the 10 items in Part II, which consists of rearranging and matching question types, vary from very difficult to very easy. Observing the logit scale, item number 20 appears to be the most difficult with a logit value of 0.22, while item number 6 is the easiest with a logit value of -0.16. The average difficulty is found in items 19, 8, and 9 with logit values of 0.01, -0.01, and -0.03, respectively.

The average logit values for the abilities of individuals can be observed for individuals 08L to 66L, ranging from 0.06 to -0.23. Students with the ability to answer high-difficulty questions are those with codes 24L, 41P, 45P, with logit values of 2.55, 1.87, and 1.87, respectively. On the other hand, a student with a low

ability is identified with code 43L and a logit value of -0.73. A higher logit value indicates a better difficulty level for question items.

Analyzing students' responses to the question items, it is generally known that items considered very difficult are those explaining ideas or concepts related to colloids, illustrating, and requiring independent decision-making or the formulation of steps to apply colloids in daily life. Items considered easy involve students classifying methods of colloid production, categorizing, and identifying types of colloids in daily life. Based on the logit scale, it can be formulated that items categorized as very difficult and difficult are mostly manageable by students, achieving a satisfactory score (Susongko, 2016; Tabatabaee-Yazdi et al., 2018). For moderate and easy items, most students can answer them and obtain maximum scores. This indicates that the characteristics and abilities of individuals for these question items are functioning well (Kazemi et al., 2020; Muslihin et al., 2022; Ningrum et al., 2019).

Analysis of the Discriminating Power of Question Items

Analysis of the differentiating power of the Rasch model via the Winstep application can be done by looking at the item separation values in the output*Summary Statistic.* This separation explains how many groups of items and people there are when taking a test. Separation for Part I questions is 2.85 and Part II is 1.36.

Table 9. Item Grouping (Distinguishing Power)

Question Type	Grouping
Part I	4
Part II	2

The question items in Part I are better at grouping item abilities compared to the question items in Part II. The questions in Part I can differentiate 4 groups of abilities, while the questions in Part II are only able to differentiate 2 abilities. The greater the separation value, the better the quality of the instrument in distinguishing person and item abilities.

Question Item Distractor (Distractor)

Distractors represent the distribution of selected and non-selected answers by respondents. Distractors can be considered as incorrect answer choices designed to mislead. Distractor analysis is conducted only for multiple-choice questions, ensuring that the other answer options serve the purpose of misleading respondents into providing the correct answer. Distractors are considered good if they are chosen by approximately 5% of respondents. In this study, the detection of distractors applies only to questions 1 through 5, which are multiple-choice questions consisting of five answer options. Distractor data can be observed in the output table item measure in the data count and % sections. The results of this distractor analysis can be seen in Table 10.

Table 10. Results of Distractor Analysis on Multiple Choice Questions

Number Question	Correct answer	Data %	Which Works Well	Category
S1	D. 60 %	A. 12 B. 12 C. 7 E. 10	4	Very good
S2	B.78%	A. 9 C. 7 D. 1 E. 5	3	Good
S3	D.73%	A. 6 B.7 C.7 E.5	4	Very good
S4	A. 83 %	B.5 C.5 D.5 E.1	3	Good
S5	C.30%	A.31 B. 10 D. 19 E. 11	4	Very good

The questions developed have distractors with good and very good categories. The order of questions that have good distractors starts from sequence 5, 1, 3, 2, and 4.

Score Interpretation

The scoring results were obtained from the Winstep application on the output table item measure data section Count or percentage.

Table 11. Results of Score Interpretation Analysis of Mastery of Question Items

No	Average	Present	Information
1	224	60%	Enough
2	292	78%	Good
3	284	76%	Good
4	326	84%	Good
5	112	30%	Not good
6	250	66%	Good
7	246	65%	Fair
8	230	61%	Fair
9	228	61%	Fair
10	195	52%	Fair
11	216	57%	Fair
12	300	80%	Good
13	248	66%	Good
14	260	69%	Good
15	276	73%	Good
16	246	65%	Fair
17	246	65%	Fair
18	196	52%	Not good
19	224	60%	Fair
20	189	50%	Not good
21	260	69%	Good
22	268	71%	Good
23	340	90%	Very good
24	228	61%	Good
25	288	77%	Good

As a result of students' mastery of the question items, in general the question items are suitable for use. The mastery of students who are able to answer questions correctly is found in question number 23 with a percentage of 90% in the easy question category. Meanwhile, the least mastery is in question number 5 with a percentage of 30% which is categorized as very difficult if viewed based on the level of difficulty. Question number 5 with a small percentage is due to misconceptions among students in the field. This is not a drawback in this development, but it can encourage teaching staff in schools to better prepare materials before teaching.

Conclusion

Based on the results of the conducted research, the literacy assessment instrument using Digital Assessment Tools can be deemed suitable for use as a measure of students' abilities in the evaluation process. This conclusion is supported by meeting logical validity values according to experts, including content validity, linguistic and construction validity, as well as media validity, as evidenced by Aiken's V scores above 0.8. The instrument is also considered suitable for further testing. Out of 25 questions, 24 were declared valid and fit the Rasch model. The questions exhibit a diverse range of difficulty levels, from very easy to very difficult, with a Standard Deviation of 0.21 for Part I and 0.14 for Part II. The discriminative power of each question was analyzed for both parts. Part I, with a separation of 2.85, grouped items into 4 categories, and Part II, with a separation of 1.36, grouped items into 2 categories. Distractors for multiple-choice questions are categorized as good and very good, indicating that they function as intended.

Acknowledgments

The author would like to thank profusely, to AKSI-ADB Malikussaleh University for its funding in the Research Grant for Young Researcher scheme.

Author Contributions

R.Z, M.LZ, and I.R.L as conceptualization; M.L.Z and R.Z as methodology and analysis; A.I.P and M.R.F as data collection; M.T.K and I.R.L as visualization and validation analysis; R.Z and M.L.Z as writing-review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

Funding for this research is AKSI-ADB Malikussaleh University.(Andriani et al., 2024)

References

Alvina, S., Mellyzar, M., Zahara, S. R., Masrina, M., & Afrianti, S. (2022). The Influence of POGIL and MFI Models on Science Literacy and Science Process Skills for Junior High School. *Jurnal Penelitian Pendidikan IPA*, *8*(4), 1907–1915. Retrieved from https://garuda.kemdikbud.go.id/documents/detail/3150349

Amany, A. (2020). Quizizz sebagai media evaluasi pembelajaran daring pelajaran matematika. *Buletin Pengembangan Perangkat Pembelajaran*, *2*(2), 1–11. Retrieved from https://journals.ums.ac.id/index.php/bppp/article/view/13811

Amelia, R. N. (2021). Identifikasi Item Fit Dan Person Fit Dalam Pengukuran Hasil Belajar Kimia. *Jurnal Ilmiah WUNY*, *3*(1), 13–26. Retrieved from https://rb.gy/i1smlv

Andriani, R., Mellyzar, M., Lukman, I. R., Muttakin, M., Pasaribu, A. I., & Fadli, Mhd. R. (2024). A Review of Digital Assessment in Education: Tools, Feature, and Effectiveness. *Proceedings of Malikussaleh International Conference On Education Social Humanities and Innovation (Miceshi)*, 0023–0023. Retrieved from https://proceedings.unimal.ac.id/miceshi/article/view/490

Brentani, E., & Golia, S. (2007). Unidimensionality in the Rasch model: how to detect and interpret.

Statistica, *67*(3), 253–261. https://doi.org/10.6092/issn.1973-2201/3508

Capuno, J. G. C. (2023). Quizziz: A Game-based Formative Assessment Tool for Enhancing Students Self-Regulated Learning. *International Journal of Social Learning (IJSL)*, *3*(3), 329–340. https://doi.org/10.47134/ijsl.v3i3.206

Dzikrullah, M. I., & Syafi'i, A. (2021). Quizizz As Interactive and Gamified Assessment Platform: Voices form Indonesian EFL Secondary Learners. *Jurnal Educative: Journal of Educational Studies*, *6*(2), 140–152. https://doi.org/10.30983/educative.v6i2.4916

Indahri, Y. (2021). Asesmen Nasional sebagai Pilihan Evaluasi Sistem Pendidikan Nasional. *Aspirasi: Jurnal Masalah-Masalah Sosial*, *12*(2), 195–215. https://doi.org/10.46807/aspirasi.v12i2.2364

Iqbal, W. M. G., Fadhilah, R., & Hadiarti, D. (2018). Pengembangan alat evaluasi berbasis wondershare quiz creator pada materi koloid kelas XI di SMA Koperasi Pontianak. *Ar-Razi Jurnal Ilmiah*, *6*(1), 11–19. https://doi.org/10.29406/arz.v6i1.937

Kazemi, S., Ashraf, H., Motallebzadeh, K., & Zeraatpishe, M. (2020). Development and validation of a null curriculum questionnaire focusing on 21st century skills using the Rasch model. *Cogent Education*, *7*(1). https://doi.org/10.1080/2331186X.2020.1736849

Lukman, A., & Muhammad, H. H. (2022). Analisis Tingkat Kesulitan Butir dan Kemampuan Matematika Siswa Berdasarkan Hasil Ujian Sekolah. *Jurnal Ilmiah Wahana Pendidikan*, *8*(23), 611–616. https://doi.org/10.5281/zenodo.7421825

Lukman, I. R., Mellyzar, M., Alvina, S., & Saa'dah, N. (2022). Development of a Chemical Literacy Assessment on Colloid (CLAC) Instrument to Measure Chemical Literacy. In *Proceedings of Malikussaleh International Conference on Multidisciplinary Studies (MICoMS)* (Vol. 3, pp. 00010-00010). https://doi.org/10.29103/micoms.v3i.50

Mudrikah, S. (2021). Upaya Menumbuhkan Budaya Paperless Melalui Pemanfaatan Ispring Quiz Maker Di SMK YPPM Boja. *Panrita Abdi-Jurnal Pengabdian Pada Masyarakat*, *5*(1), 89–99. https://doi.org/10.20956/pa.v5i1.9221

Muslihin, H. Y., Suryana, D., Ahman, A., Suherman, U., & Dahlan, T. H. (2022). Analysis of the Reliability and Validity of the Self-Determination Questionnaire Using Rasch Model. *International Journal of Instruction*, *15*(2), 207–222. https://doi.org/10.29333/iji.2022.15212a

Ningrum, E., Evans, S., & Soh, S.-E. (2019). Validation of the Indonesian version of the Safety Attitudes Questionnaire: A Rasch analysis. *PLOS ONE*, *14*(4), e0215128. https://doi.org/10.1371/journal.pone.0215128

Oktharia, E., Rudibyani, R. B., & Sofia, E. (2017). Pengembangan Instrumen Asesmen Pengetahuan untuk Mengukur Penguasaan Konsep Siswa. *Jurnal Pendidikan Dan Pembelajaran Kimia*, *6*(1), 74–86. Retrieved from https://rb.gy/cjm86v

Pratiwi, V. (2017). Pengembangan Alat Evaluasi Pembelajaran Berbasis ICT MenggunakanWondershare Quiz Creator pada Materi Penyusunan Aset Tetap. *Prosiding Seminar Pendidikan Ekonomi Dan Bisni*, *3*(1), 1–7. Retrieved from https://jurnal.fkip.uns.ac.id/index.php/snpe/article/view/10698

Prawesti, A. J., Koesdyantho, A. R., & Widyaningrum, R. (2021). Asesment Hasil Belajar Peserta Didik Kelas ICT dalam Pembelajaran Online di SD Negeri Kroyo Sragen. *Jurnal Sinektik*, *4*(2), 142–151. https://doi.org/10.33061/js.v4i2.5337

Purnamawati, P., Arfandi, A., & Nurfaeda, N. (2019). The level of use of information and communication technology at vocational high school. *Jurnal Pendidikan Vokasi*, *9*(3), 249–257. https://doi.org/10.21831/jpv.v9i3.27117

Purwanto, N. (2009). *Prinsip-prinsip dan Teknik Evaluasi Pengajaran*. PT Remaja Rosdakarya.

Rahayu, K. A. (2021). The implementation of authentic assessment in English instruction. *Jurnal Penelitian Dan Pengembangan Pendidikan*, *5*(1), 122–12. https://doi.org/10.23887/jppp.v5i1.31723

Ramadhani, K. P., & Ardi, H. (2022). Penggunaan aplikasi quizizz sebagai media pembelajaran dan asesmen pada materi bahasa Inggris. *ABDI HUMANIORA: Jurnal Pengabdian Masyarakat Bidang Humaniora*, *3*(1), 1–14. https://doi.org/10.24036/abdihumaniora.v3i1.119559

Sugiyono. (2016). *Methods of quantitative, qualitative and R & D research*. Alfabeta.

Sujarwanto, S., & Rusilowati, A. (2015). Pengembangan Instrumen Performance Assessment Berpendekatan Scientific Pada Tema Kalor Dan Perpindahannya. *Unnes Science Education Journal*, *4*(1), 780–787. https://doi.org/10.15294/usej.v4i1.4998

Susongko. (2016). Validation of Science Achievement Test With The Rasch Model. *Jurnal Pendidikan IPA Indonesia*, *5*(2), 268–277. https://doi.org/10.15294/jpii.v5i2.7690

Tabatabaee-Yazdi, M., Motallebzadeh, K., Ashraf, H., & Baghaei, P. (2018). Development and Validation of a Teacher Success Questionnaire Using the Rasch Model. *International Journal of Instruction*, *11*(2), 129–144. https://doi.org/10.12973/iji.2018.11210a