

COGNITIVE TEST INSTRUMENT TO MEASURE THE CHEMICAL LITERACY SKILLS OF HIGH SCHOOL STUDENTS REGARDING SALT HYDROLYSIS

Nesfitri Aufa Zakiyah¹, Yerimadesi Yerimadesi^{2*}, Desy Kurniawati¹, Alizar Alizar³

^{1,2,3}Department of Chemistry Education, Faculty of Teaching Education, University of Padang, Sumatra, Indonesia

Article Info

Article history:

Received 15/07/2026

Accepted 21/09/2026

Published 30/09/2026

Keywords:

Chemical literacy;

Cognitive test;

Rasch;

Winstep.

ABSTRACT

Scientific literacy ability has also become a focus of attention in international assessments, namely through the Programme for International Student Assessment. Indonesia ranked 68th out of 81 participating countries with a science literacy score of 383. This low achievement indicates that the scientific literacy ability of students in Indonesia still requires serious attention. Scientific literacy has several branches, one of which is chemical literacy. Difficulty in understanding chemical concepts can affect chemical literacy ability, as chemical literacy is a student's ability to understand and apply basic chemical concepts accurately and effectively in everyday life. The purpose of this study is to develop a chemical literacy test instrument using the Rasch model. The objective of this study was to develop a chemistry literacy test instrument using the Rasch model. This research is a research and development (R&D) study employing the Rasch model. The test instrument was examined for validity, reliability, difficulty level, and discrimination index. The subjects of this study were 120 high school students. The research instrument used was a set of test items that had been developed. Based on the results of the analysis, the Rasch model can be applied to the developed chemistry literacy test instrument on the topic of salt hydrolysis. The field test results for the developed instrument indicate that it is valid and can be used to measure students' chemistry literacy ability. The validity of the instrument was assessed through empirical validity analysis, with Outfit MNSQ values ranging from 0.81 to 1.20, Outfit ZSTD values ranging from -1.43 to 1.34, and point measure correlation values ranging from 0.60 to 0.72. Reliability was assessed based on a person reliability of 0.78 and an item reliability of 0.99.

This is an open-access article under the [CC BY-SA](#) license.



Corresponding Author:

Yerimadesi

Department of Chemistry Education, Faculty of Teaching Education, University of Padang, Sumatra, Indonesia

Email: yeri@fmipa.ac.id

1. INTRODUCTION

Rapid change and constant access to information made possible by the fast pace of information and communication technology define the character of the 21st century. This condition brings both new challenges and new opportunities, calling for human resources equipped with competencies distinct from those of earlier generations. Amid this global shift, education systems play a decisive role in shaping a nation's progress. Consequently, the 21st-century learning paradigm has moved away from merely transmitting factual knowledge, focusing instead on

cultivating foundational competencies such as critical thinking, creativity, collaboration, and communication, in line with the broader set of 21st-century skills that also includes problem-solving and cooperative working ability (Atik & Marieke, 2023).

International assessments have likewise placed growing emphasis on scientific literacy competence, most notably through the Programme for International Student Assessment, administered by the Organization for Economic Co-operation and Development (OECD). According to the 2022 PISA results, Indonesia ranked 68th among 81 participating countries, recording a scientific literacy score of 383 well below the OECD average of 485 (OECD, 2023). Such a low outcome signals that Indonesian students' scientific literacy skills still demand serious attention (Hasan, 2024). This finding underscores the pressing need for a valid, reliable, and sensitivity-adjusted literacy instrument that can serve as a more targeted diagnostic tool than an overall scientific literacy score, one capable of pinpointing students' specific weaknesses in chemistry literacy and thereby informing improvements to chemistry instruction at the senior secondary level.

Chemical literacy, one of the recognized branches of scientific literacy, involves understanding the properties of particles, chemical reactions, and chemical laws and theories, along with their application to everyday situations. Strong chemical literacy therefore presupposes a solid grasp of basic scientific concepts, particularly within chemistry itself (Shwartz et al, 2006). When students struggle to understand chemical concepts, their chemical literacy inevitably suffers, given that chemical literacy is essentially the capacity to apply fundamental chemical concepts accurately and effectively in daily life (Burhanuddin et al, 2023).

This concern is reinforced by results from the 2025 Academic Ability Test released by the Ministry of Primary and Secondary Education (Kemendikdasmen). Chemistry, as an elective subject, recorded a notably low national average of only 34.92 out of 100 among 361,634 participants, while in West Sumatra the average stood at just 34.66 across 12,564 examinees (Kemendikdasmen, 2025). A similar pattern emerges in TKA data specific to students in Payakumbuh City and Lima Puluh Kota Regency, pointing to persistently weak chemical literacy in these areas. This gap between theoretical chemistry instruction and its practical application highlights the need for an evaluation instrument capable of comprehensively capturing students' chemical literacy.

A sound chemical literacy assessment, as outlined by Shwartz et al (2006) should address four interrelated dimensions: content, context, Higher-Order Learning Skills (HOLS), and attitude. Instruments limited to the content dimension alone cannot adequately capture the full range of students' chemical literacy, which makes the availability of more comprehensive instruments essential. A review of four related prior studies reveals several gaps that this study seeks to address. First, the instruments developed by Sartika & Yusmaita, (2020) and by Sholeha dkk, (2023) lacked a balanced spread of item difficulty items in the former skewed toward the difficult-to-very-difficult range, while none in the latter reached the difficult category, limiting their ability to discriminate among high-ability students. Second, Rizki & Yusmaita, (2021) applying an Item Response Theory approach through the Rasch model, still produced several items that only partially satisfied statistical feasibility criteria, indicating a need for further refinement. Third, the study by Eliza, W., & Yusmaita, (2021) had not yet reached the stage of empirical testing with students, leaving reliability, difficulty level, and discrimination index data unavailable. In response to these gaps, this study adopts the Rasch model both as the development framework and as the

method of data analysis for the test instrument produced.

Salt hydrolysis, as a topic, closely mirrors everyday phenomena and industrial practice, encompassing concepts such as ionic equilibrium, acid-base properties, and ionic reactions in aqueous solution. Mastering this material requires students to move fluidly across three levels of chemical representation-macroscopic, submicroscopic, and symbolic. This layered complexity often proves challenging, particularly when students must connect an observable phenomenon, such as a shift in solution pH, to what is happening at the molecular level. Learning salt hydrolysis, then, cannot rely on formula memorization alone; it demands deep conceptual understanding paired with solid chemical literacy skills.

The Rasch model was selected for this study largely for its capacity to expose the underlying quality of test items, ensuring that a high score genuinely reflects high ability rather than a testing artifact. It is also particularly effective at differentiating item difficulty in line with students' actual ability levels. As (Jamaludin et al, 2021) note, Rasch model analysis offers a measurement-instrument development approach that safeguards instrument quality, including alignment between item difficulty and test-taker ability. Adopting the Rasch model further allows researchers to flag poor-fitting or biased items, enabling revision or removal of ineffective items from the instrument. Beyond this, the model performs calibration across three components at once: the measurement scale, the persons (test takers), and the items themselves (Surya, 2024).

Developing a test instrument on salt hydrolysis is therefore essential to supporting chemical-literacy-oriented learning. Several benchmarks must be met for such an instrument to be considered sound or valid; chief among them, an instrument is valid only insofar as it accurately measures what it is intended to measure (Kemendikdasmen, 2025), and the resulting test must clearly reflect the extent to which a learning objective has been achieved. This study relies on the Rasch model as a guiding framework to refine the developed instrument, which is designed to satisfy all four dimensions of chemical literacy chemical content, contextual understanding, higher-order thinking skills, and the affective dimension (Jayanti, S. D., & Sutrisno, 2025).

The feasibility of the developed instrument will be examined through the Rasch Model, one of the more contemporary measurement approaches, offering a deeper lens for evaluating item quality by allowing researchers to estimate student ability and item difficulty simultaneously on a shared scale (Purniasari dkk, 2021) While the four previous studies cited above all drew on Shwart et al, (2006) four-dimension chemical literacy framework, each addressed a different topic colloids, chemical bonding, basic chemical laws, and the periodic system of elements. This study instead takes up salt hydrolysis, a topic that is inherently more demanding because it requires the simultaneous integration of three representational levels: the macroscopic (observable pH changes), the submicroscopic (molecular-and ion-level processes), and the symbolic (ionic equilibrium reaction equations). It is also more directly tied to everyday phenomena and industrial application than a purely abstract concept such as the periodic system or basic chemical laws. Because students must link observed phenomena to underlying molecular processes rather than simply recall formulas, the cognitive demand is correspondingly higher requiring deeper HOLS oriented items than those used for previous topics.

This study further incorporates an explicit psychometric calibration stage through the Rasch Model, adding value by estimating test-taker ability and item difficulty simultaneously on an interval scale rather than relying solely on raw or ordinal scores, as in classical approaches. Because the Rasch model can detect misfitting or biased items, ineffective items can be revised or

removed in a way that classical validity - reliability testing alone cannot achieve. Simultaneous calibration across three components the measurement scale, persons, and items yield more comprehensive quality control over the instrument. The resulting instrument is thus positioned to offer stronger measurement properties than earlier chemical literacy instruments, with item difficulty empirically tested against student ability rather than derived purely from theoretical cognitive taxonomy. It continues to address all four dimensions of chemical literacy (content, context, HOLS, attitude) found in prior work, while adding a further layer of psychometric validation through Rasch. Building on this rationale, the present study aims to develop a question-based test instrument to measure students' chemical literacy using the Rasch model.

2. METHOD

2.1. Types of Research

This study adopted a Research and Development (R&D) approach to construct a chemical literacy assessment instrument, following the Rasch-based development framework outlined by Liu, (2020) with the Rasch Model applied as the psychometric method for evaluating item quality and measurement properties.

2.2. Research Subject

Participants comprised high school students drawn from two schools with "Good (A)" accreditation: 60 from MAN 1 Kota Payakumbuh and 60 from SMAN 1 Harau. Participants were selected through purposive sampling, a technique in which individuals are deliberately chosen based on specific characteristics or qualities relevant to the study, also known as judgment sampling (Etikan et al., 2016) Such deliberate selection can be guided by various considerations, including access to individuals who share a particular experience, membership in a relevant group, or the capacity to contribute meaningfully to the research (Hossan, D., Dato' Mansor, Z., & Jaharuddin, 2023).

2.3. Research Instrument

This study used a chemical literacy test instrument that had been developed, based on the Rasch model. This instrument was reviewed by 5 validators: Prof. Dr. Desy Kurniawati, S.Pd., M.Si, a Postgraduate Chemistry Education lecturer in the field of Analytical Chemistry; Budi Oktavia, M.Si., Ph.D, a Postgraduate Chemistry Education lecturer in the field of Analytical Chemistry; Umar Kalmar Nizar, S.Si., M.Si., Ph.D, a Postgraduate Chemistry Education lecturer in the field of Physical Chemistry; Dr. Riga, S.Pd., M.Si, a Postgraduate Chemistry Education lecturer in the field of Organic Chemistry; and Dr. Fauzana Gazali, S.Pd., M.Pd, a Postgraduate Chemistry Education lecturer in the field of Education, using the 10 stages developed by Liu, (2020). The Rasch model measurement stages carried out in this study were: (1) determining the purpose and target population of the test instrument, (2) defining the construct, (3) identifying the performance of the defined construct, (4) conducting a trial/pilot test, (5) performing Rasch model analysis, (6) examining item fit statistics, (7) examining the Wright map, (8) checking invariance

properties (repeating steps 4 through 7), (9) establishing validity and reliability, and (10) compiling guidelines for using the test instrument. Rasch is currently a general term encompassing measurement theory, statistical models, and a set of related algebraic techniques (Tesio et al, 2024).

The developed test instrument was first validated by the validators to examine its feasibility before being used in the field trial. The test instrument consists of 8 essay questions arranged based on the aspects of chemical literacy. Scoring was conducted after students completed the cognitive test consisting of the questions that had been developed. The results of the test answers were analyzed according to the scoring rubric: if a student was able to answer all questions completely, the score obtained was 3. If the student was only able to answer partially, the score obtained was 2. Students who were unable to answer the test questions received a score of 0. These results were then accumulated to obtain the final score from the administration of the developed test instrument. Scoring was carried out by the researcher alone, without the involvement of other raters.

2.4. Data Analysis Technique

The data were first tested for logical validity using CVR. This method was chosen because CVR has several advantages, such as being practical, easy to use, having a concise computer calculation, being equipped with a critical value as a reference, and placing emphasis on expert agreement regarding the importance of an item (Lawshe, 1975). After logical validation was carried out, it was followed by empirical validation. Empirical validation began with a limited small-scale trial. Valid trial results would then proceed to a large-scale trial. The collected data were processed using the Winsteps application. Winsteps is a computer program specifically designed to analyze data based on the Rasch model. This software provides various features that facilitate users in conducting analysis, ranging from data input and model selection to the interpretation of analysis results (Lasahinda, A., Pulukadang, R., & Kumesan, 2025). The data obtained were tested for validity, reliability, discrimination index, and difficulty level to examine the feasibility of the developed test instrument.

3. RESULTS AND DISCUSSION

This study yielded a product consisting of eight essay questions measuring chemical literacy in the topic of salt hydrolysis at the SMA/MA level, subsequently analyzed using the Rasch model. Among its strengths, Rasch modeling can detect guessing behavior, estimate scores for missing data, flag response errors, and operate independently of the raw number of correct responses (Hasanah, H., & Aini, 2025). It also calculates item difficulty in a manner independent of the specific sample or assessment used, enabling stronger comparability across assessments and over time (Hope et al, 2025) and offers a more robust, longitudinal view of item quality that benefits multi-site and multi-year assessment efforts (Wrigley et al, 2012). Findings from each development stage are presented below.

3.1. Results of Determining the Objectives and Target Population for the Test Instrument

At this stage, Learning Objectives were produced based on the analysis of Learning

Outcomes in accordance with the Decree of the Head of the Education Standards, Curriculum, and Assessment Agency Number 046/H/KR/2025. From the results of this CP analysis, 4 learning objectives to be applied were obtained, namely: explaining the concept of salt hydrolysis, analyzing ion equilibrium in salt solutions, identifying salts that undergo hydrolysis and determining the properties of their solutions (acidic, basic, neutral), and analyzing the pH of salt solutions resulting from hydrolysis.

3.2. Results of Construct Definition

This stage produced a test specification grid and item indicators, which were then used to develop the design of the test instrument along with its scoring rubric. The design of the chemical literacy instrument in this study was based on the aspects of Content, Context, High Order Learning Skill (HOLS), and Attitude (affective). These four aspects are the chemical literacy aspects proposed by Shwartz (Shwartz et al, 2006) An example of a developed item design is presented in Table 1.

Table 1. Design of Chemical Literacy Test Items on Salt Hydrolysis Material

Chicken Eggshell Waste as a Feed Supplement for Laying Hens

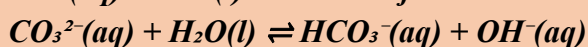
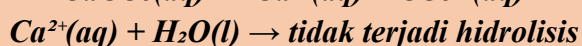
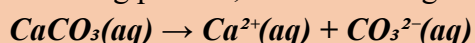


Chicken eggshell

Read the following passage carefully!

Mr. Hendra, a laying hen farmer in Pesisir Selatan Regency, experienced a problem with thin and easily cracked eggshells despite high production, which was found to be caused by a calcium deficiency in his chickens' feed.

A researcher suggested utilizing chicken eggshell waste as a cheap and environmentally friendly feed supplement because it is rich in calcium carbonate (CaCO_3), which is a salt of the weak acid H_2CO_3 and the strong base $\text{Ca}(\text{OH})_2$. In the washing process, the salt undergoes the following:



The Ca^{2+} ion does not undergo hydrolysis because it originates from the strong base $\text{Ca}(\text{OH})_2$, while the CO_3^{2-} ion, which is the anion of the weak acid H_2CO_3 , undergoes hydrolysis and produces OH^- ions, causing the

Context

Content

solution to be basic. This basic condition needs to be neutralized before the eggshell powder is given to the chickens so that it does not disrupt the livestock's digestive system.

HOLS

Understanding the chemistry behind livestock feed processing is important for modern farmers so they can produce healthy, high-quality, and environmentally friendly livestock products by making wise use of waste..

Affective

Answer the questions below correctly!

Based on the passage above, analyze the relationship between the type of ions resulting from the ionization reaction of CaCO_3 and the properties of the solution formed, and explain how these properties can affect the safety of using eggshell powder as chicken feed!

The scoring rubric drew on a five-level chemical literacy framework: scientific illiteracy, nominal, functional, conceptual, and multi-dimensional literacy. However, only the first four levels (through conceptual literacy) were applied here, following Bybee (Shwartz et al, 2006) given that salt hydrolysis demands considerable conceptual depth and analytical skill-one of several chemistry topics whose profound conceptual content calls for genuinely deep understanding.

3.3. Results of Identifying the Performance of the Defined Construct

At this stage, the results of the validation of the test instrument were obtained before it was used in the field trial. Validation was carried out by 5 experts as validators. Content Validity Ratio (CVR) analysis was then used to examine the results of the validation sheets that had been completed by several experts. Lawshe's CVR (Content Validity Ratio) is one of the methods widely used to measure content validity. The results of the expert validation are presented in the following Table 2.

Table 2. Results of the Instrument Validation Test

Test Item	Validator (V)	Chemical Literacy Aspects				CVR	Category
		Context	Content	HOLS	Affectiv		
S1	V1	1	1	1	1	1	Valid
	V2	1	1	1	1		
	V3	1	1	1	1		
	V4	1	1	1	1		
	V5	1	1	1	1		
S2	V1	1	1	1	1	1	Valid
	V2	1	1	1	1		
	V3	1	1	1	1		
	V4	1	1	1	1		
	V5	1	1	1	1		
S3	V1	1	1	1	1	1	Valid
	V2	1	1	1	1		
	V3	1	1	1	1		

	V4	1	1	1	1		
	V5	1	1	1	1		
S4	V1	1	1	1	1	1	Valid
	V2	1	1	1	1		
	V3	1	1	1	1		
	V4	1	1	1	1		
	V5	1	1	1	1		
S5	V1	1	1	1	1	1	Valid
	V2	1	1	1	1		
	V3	1	1	1	1		
	V4	1	1	1	1		
	V5	1	1	1	1		
S6	V1	1	1	1	1	1	Valid
	V2	1	1	1	1		
	V3	1	1	1	1		
	V4	1	1	1	1		
	V5	1	1	1	1		
S7	V1	1	1	1	1	1	Valid
	V2	1	1	1	1		
	V3	1	1	1	1		
	V4	1	1	1	1		
	V5	1	1	1	1		
S8	V1	1	1	1	1	1	Valid
	V2	1	1	1	1		
	V3	1	1	1	1		
	V4	1	1	1	1		
	V5	1	1	1	1		

Based on Table 2, it can be seen that all questions (question 1 to question 8) have already fallen into the valid category. This was obtained from the results of item revisions based on the validators' critiques and suggestions; based on the final validation, a CVR value of 1 was obtained. The questions can be considered valid because they have a CVR value above the critical value of 0.763.

3.4 Results of the Trial/Field Test

In the small-group trial, students' average scores fell into the moderate category. Items 3, 4, and 7 tended to yield lower scores, though 3 of the 9 participating students still answered them well, while items 2, 5, and 6 produced comparatively higher scores. Items 1 and 8 were answered well by all students in this trial. Because the trial involved children spanning different cognitive levels, these results suggested the instrument was ready for the larger study.

The subsequent field trial, involving 120 SMA/MA students, generated response data that were processed in Winsteps to establish each item's validity, reliability, difficulty level, and discrimination index.

3.5. Results of Rasch Model Analysis

The data analyzed using the Rasch model yielded empirical Validity, Reliability, Difficulty Level, and Discrimination Index Data, processed using the Winsteps application.

3.5.1. Validity Test Results

The results of the item validity analysis for the chemical literacy questions on the salt hydrolysis material were obtained using the Winsteps application. The data obtained can be seen in Figure 1.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PTMEASUR-AL CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
7	269	120	-.92	.16	1.25	1.81	1.20	1.34	A .60	.65	58.3	62.7	S7
5	217	120	.29	.15	1.10	.85	1.12	.98	B .64	.69	49.2	57.0	S5
8	283	120	-1.29	.17	1.11	.85	1.02	.18	C .67	.63	69.2	65.3	S8
6	230	120	.01	.15	1.08	.69	1.07	.62	D .62	.68	65.0	57.2	S6
1	306	120	-2.00	.18	.86	-.99	.95	-.16	d .62	.58	72.5	71.6	S1
3	123	120	2.31	.15	.89	-.91	.89	-.90	c .71	.68	60.0	58.2	S3
4	124	120	2.29	.15	.84	-1.36	.89	-.90	b .70	.68	59.2	58.2	S4
2	260	120	-.69	.16	.80	-1.64	.81	-1.43	a .72	.66	75.0	61.1	S2
MEAN	226.5	120.0	.00	.16	.99	-.1	.99	.0			63.5	61.4	
P.SD	65.0	.0	1.48	.01	.16	1.2	.13	.9			8.0	4.7	

Figure 1. Item Fit Order of the Chemical Literacy Instrument on Salt Hydrolysis Material

Based on Figure 1, the image above shows the results of the item validity test using the Winsteps application. Validity relates to the accuracy of an instrument in measuring what it is supposed to measure (Zayrin et al, 2025). The analysis results show that all items, from question 1 (S1) to question 8 (S8), had Outfit MNSQ values ranging from 0.81 to 1.20, all of which still fall within the acceptance criteria range of 0.5–1.5. This indicates that there were no significant extreme responses deviating from the model's predictions across all instrument items. The Outfit ZSTD values for all items ranged from -1.43 to 1.34, which also still met the criteria of -2.0 to +2.0, thereby reinforcing the conclusion that no items were significantly misfit. Although all questions were declared fit, item S7 showed an Outfit MNSQ value of 1.20 with a ZSTD value of 1.34, the highest among all items. Although item S7 still meets the fit criteria, its relatively high Outfit value indicates variation in response patterns, which may be attributed to differences in students' interpretation of the item context or possible misconceptions regarding the salt hydrolysis material. However, the exact cause of this variation requires further analysis, such as distractor analysis or student interviews.

3.5.2. Reliability Test Results

The results of the reliability analysis of the chemical literacy test instrument on the salt hydrolysis material. The data obtained can be seen in Figure 2.

SUMMARY OF 120 MEASURED Person								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	15.1	8.0	1.09	.62	.98	-.15	.99	-.09
SEM	.4	.0	.14	.01	.06	.12	.06	.12
P.SD	4.1	.0	1.48	.08	.70	1.31	.70	1.26
S.SD	4.2	.0	1.49	.08	.71	1.32	.71	1.27
MAX.	22.0	8.0	4.12	.90	3.94	3.55	3.37	3.40
MIN.	6.0	8.0	-2.06	.56	.16	-2.48	.12	-2.37

REAL RMSE	.70	TRUE SD	1.31	SEPARATION	1.88	Person RELIABILITY		.78
MODEL RMSE	.63	TRUE SD	1.35	SEPARATION	2.15	Person RELIABILITY		.82
S.E. OF Person MEAN = .14								

Person RAW SCORE-TO-MEASURE CORRELATION = 1.00								
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .82 SEM = 1.77								
STANDARDIZED (50 ITEM) RELIABILITY = .97								

SUMMARY OF 8 MEASURED Item								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	226.5	120.0	.00	.16	.99	-.09	.99	-.03
SEM	24.6	.0	.56	.00	.06	.45	.05	.35
P.SD	65.0	.0	1.48	.01	.16	1.20	.13	.93
S.SD	69.4	.0	1.59	.01	.17	1.28	.13	.99
MAX.	306.0	120.0	2.31	.18	1.25	1.82	1.20	1.34
MIN.	123.0	120.0	-2.00	.15	.80	-1.64	.81	-1.43

REAL RMSE	.16	TRUE SD	1.47	SEPARATION	8.98	Item RELIABILITY		.99
MODEL RMSE	.16	TRUE SD	1.47	SEPARATION	9.28	Item RELIABILITY		.99
S.E. OF Item MEAN = .56								

Item RAW SCORE-TO-MEASURE CORRELATION = -1.00								
Global statistics: please see Table 44.								
UMEAN=.0000 USCALE=1.0000								

Figure 2. Summary Statistics of the Chemical Literacy Instrument Analysis on Salt Hydrolysis Material

Based on Figure 2, the analysis yielded a person reliability of 0.78, an item reliability of 0.99, and a Cronbach's Alpha of 0.82. Reliability, broadly, reflects how consistently a measuring instrument can be trusted: an instrument is considered reliable when repeated measurement of the same phenomenon produces relatively stable results (Rahmayanti, 2024). The person separation index of 1.88–2.15 obtained here indicates that the instrument can distinguish respondents into three ability strata, while the notably high item separation index of 8.98–9.28 reflects a wide and consistent spread of item difficulty. Together, these results point to good reliability and confirm the instrument's suitability for measuring the intended variable.

3.5.3 Discrimination Index Test Results

Difficulty level is the degree of difficulty of a test item, expressed in the form of a numerical value (Lestari, K. E., & Yudhanegara, 2017). A question is considered good if it does not fall into the category of being too easy or too difficult (Saputri H A, Zulhijrah, Larasati N Ji, 2023). The results of the difficulty level analysis in the Winsteps application can also be seen from the output table menu on the measure item. The following information is presented in Figure 3.

Item STATISTICS: MEASURE ORDER													
ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFI T	INFI T	OUTFI T	OUTFI T	PTMEASUR-AL	EXACT MATCH	EXACT MATCH	EXACT MATCH	Item
3	123	120	2.31	.15	.89	-.91	.89	-.90	.71	.68	60.0	58.2	S3
4	124	120	2.29	.15	.84	-1.36	.89	-.90	.70	.68	59.2	58.2	S4
5	217	120	.29	.15	1.10	.85	1.12	.98	.64	.69	49.2	57.0	S5
6	230	120	.01	.15	1.08	.69	1.07	.62	.62	.68	65.0	57.2	S6
2	260	120	-.69	.16	.80	-1.64	.81	-1.43	.72	.66	75.0	61.1	S2
7	269	120	-.92	.16	1.25	1.82	1.20	1.34	.60	.65	58.3	62.7	S7
8	283	120	-1.29	.17	1.11	.85	1.02	.18	.67	.63	69.2	65.3	S8
1	306	120	-2.00	.18	.86	-.99	.95	-.16	.62	.58	72.5	71.6	S1
MEAN	226.5	120.0	.00	.16	.99	-.1	.99	.0			63.5	61.4	
P.SD	65.0	.0	1.48	.01	.16	1.2	.13	.9			8.0	4.7	

Figure 3. Item Measure Results of the Analysis of the SMA/MA Chemical Literacy Instrument

Based on Figure 3, the item measure analysis produced a standard deviation of 1.48 at the SMA/MA level. Items 3 (S3, at 2.31) and 4 (S4, at 2.29) proved the most difficult, while item 1 (S1, at -2.00) was the easiest. An average measure value of 0.00 indicates that item difficulty, on the whole, centers on the midpoint of respondents' ability scale confirming that the instrument spans a wide range of difficulty, from very easy to very difficult.

3.5.4 Discrimination Index Test Results

The discrimination index reflects an item's capacity to differentiate high- from low-ability students (Arikunto, 2018) essentially, how well a test item separates lower- from higher-performing examinees (Setiyawan, R. A., & Wijayanti, 2020). This information likewise appears in Winsteps' measure-item output, summarized in Figure 4.

SUMMARY OF 120 MEASURED Person									
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFI T	INFI T	OUTFI T	OUTFI T	
MEAN	15.1	8.0	1.09	.62	.98	-.15	.99	-.09	
SEM	.4	.0	.14	.01	.06	.12	.06	.12	
P.SD	4.1	.0	1.48	.08	.70	1.31	.70	1.26	
S.SD	4.2	.0	1.49	.08	.71	1.32	.71	1.27	
MAX.	22.0	8.0	4.12	.90	3.94	3.55	3.37	3.40	
MIN.	6.0	8.0	-2.06	.56	.16	-2.48	.12	-2.37	
REAL RMSE	.70	TRUE SD	1.31	SEPARATION	1.88	Person RELIABILITY	.78		
MODEL RMSE	.63	TRUE SD	1.35	SEPARATION	2.15	Person RELIABILITY	.82		
S.E. OF Person MEAN = .14									
Person RAW SCORE-TO-MEASURE CORRELATION = 1.00									
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .82 SEM = 1.77									
STANDARDIZED (50 ITEM) RELIABILITY = .97									

Figure 4. Summary Statistics (Separation) of the Chemical Literacy Instrument Analysis on Salt Hydrolysis Material

Based on Figure 4, the discrimination index analysis results above, the person separation value obtained for SMA/MA was 2.84. With this value, the person separation for SMA/MA of 2.84, rounded to 3, can be interpreted as indicating the presence of 3 groups of individuals, namely students with high ability, students with moderate ability, and students with low ability (Sumintono, B., & Widhiarso, 2015). The percentage of questions in this study was 25% : 50% : 25%. Among the questions, 2 questions had a difficult level, 4 questions had a moderate level, and 2 questions had an easy level.

3.6. Results of the Item Fit Statistics Review

This stage consolidates the fit criteria established in the preceding analyses. Across all eight SMA/MA items, Outfit MNSQ values ranged from 0.81 to 1.20, Outfit ZSTD values from -1.43 to 1.34, and point-measure correlation values from 0.60–0.72, jointly confirming validity for every item. The reliability analysis produced a person reliability of 0.78 (adequate) and an item reliability of 0.99 (excellent), while the difficulty-level breakdown showed two difficult items, four moderate items, and two easy items. Discrimination analysis, in turn, identified three ability groups among respondents: high, moderate, and low.

3.7. Wright Map Analysis Results

The results of the discrimination index analysis in the Winsteps application can also be seen from the output table menu on the measure item. The following information is presented in Figure 5.

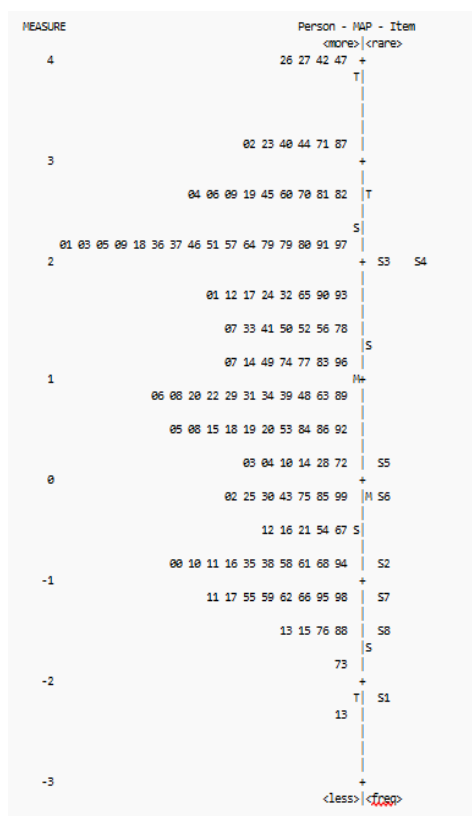


Figure 5. Wright Map Analysis of the Chemical Literacy Instrument on Salt Hydrolysis Material

Based on Figure 5, the next stage is analyzing the field trial data using the Wright map. The Figure above shows the results obtained in this study. This analysis is used to examine the distribution of student ability and the developed items. According to (Saputri H A, Zulhijrah, Larasati N Ji, 2023). Wright map analysis, also known as the person-item map, is an analysis of the distribution map of students' ability levels and the distribution of item difficulty levels. The mapping results show the order of items from the most difficult to the easiest, namely question no. 3 (S3) and question no. 4 (S4) at approximately ± 2.3 , question no. 5 (S5) at approximately ± 0.3 , question no. 6 (S6) at approximately ± 0.0 , question no. 2 (S2) at approximately ± -0.7 , question no. 7 (S7) at approximately ± -0.9 , question no. 8 (S8) at approximately ± -1.3 , and question no. 1 (S1) at approximately ± -2.0 .

The average ability value of respondents (Mean Person = 1.09 logit) is higher than the average item difficulty level (Mean Item = 0.00 logit), indicating that the instrument items are relatively easier compared to the respondents' ability in general. In addition, a gap was found in the logit range of 0.5–2.0, indicating the unavailability of items with a medium-high difficulty level. It can be seen that this instrument is fairly capable of measuring most of the respondents' ability levels; however, additional items with a high difficulty level are needed to improve measurement precision for high-ability respondents.

3.8. Results of Checking Invariance Properties (Repeating Steps 4 to 7)

After processing the data using the Winsteps application, good data validity results were obtained, the questions had good reliability for use as a variable measure, with difficulty levels varying between difficult, moderate, and easy, and the discrimination index test produced questions with 3 levels, namely easy, moderate, and difficult. At this stage, no revision or repetition of steps 4–7 was carried out because the field trial results were good.

3.9. Results of Establishing Validity and Reliability

Empirical validity was established by evaluating three Rasch fit criteria for each item: Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point-Measure Correlation (Pt Mean Corr). All eight items (S1–S8) returned Outfit MNSQ values between 0.81 and 1.20 (within the 0.5–1.5 acceptance range), Outfit ZSTD values between -1.43 and 1.34 (within the -2.0 to +2.0 limit), and Pt Mean Corr values between 0.60 and 0.72 (above the 0.3 minimum threshold). With every item meeting these fit criteria, none required elimination, confirming the chemical literacy test instrument on salt hydrolysis as empirically valid and appropriate for measuring SMA/MA students' chemical literacy.

3.10. Results of Developing Guidelines for the Use of the Test Instrument

At this stage, a guidebook/manual was produced that can be used as additional information to help users apply the test instrument effectively. At this stage, a guidebook/manual was produced that can be used as additional information to help users apply the test instrument effectively. The Rasch analysis results obtained do not merely show that this instrument is statistically "valid" and "reliable," but also provide several important implications for chemical literacy assessment

practices at the high school level.

3.10.1. Person Separation Index and Chemical Literacy as a Continuum

A person separation index of 1.88–2.15 (yielding three ability strata: low, medium, high) indicates that chemical literacy is continuous in nature rather than a dichotomous construct. This finding is consistent with Shwart et al (2006) conceptualization of chemical literacy as graded, and supports the instrument's construct validity (Messick, 1995) by distinguishing the extent to which students integrate content, context, HOLS, and attitude aspects.

3.10.2. Wright Map and Cognitive Hierarchy in Assessment Design

The gap identified in the logit range of 0.5–2.0 indicates that the instrument is more optimal in distinguishing low- to medium-ability students than in differentiating competent from highly competent students. This aligns with Bond, T. G., & Fox, (2015) principle regarding the need for a proportional item spread in Rasch-calibrated instruments. This condition is adequate for classroom diagnostic purposes but requires the addition of higher-difficulty items for selection or high-achieving student mapping purposes.

3.10.3. Item Difficulty Pattern and HOLS Cognitive Taxonomy

Items S3 and S4 (analyzing ionic equilibrium and pH) occupied the highest difficulty level, while S1 (basic concept) was the easiest, consistent with the HOLS taxonomy and Anderson, L. W., & Krathwohl, (2020) revised cognitive taxonomy. This pattern demonstrates that the instrument validly captures the gradation of cognitive complexity within the chemical literacy framework.

3.10.4. Fit Statistics and Response Validity in Essay-Based Assessment

The Outfit MNSQ values (0.81–1.20) and ZSTD values (-1.43 to 1.34), which fall within the acceptance range, indicate response consistency in accordance with students' abilities, rather than resulting from guessing. This finding is important for context-based essay items, which are prone to variation in interpretation, and simultaneously confirms the consistent application of the five-level rubric by raters, in line with the principle of response validity (Bond, T. G., & Fox, 2015; Boone, W. J., Staver, J. R., & Yale, 2014).

3.10.5. Practical Implications and Contribution to the Chemistry Assessment Literature

This instrument offers an alternative to the rote-memorization-based assessments that have long dominated practice (consistent with the low national Chemistry TKA achievement of 34.92). Its authentic discourse-based item structure, which connects the macroscopic, submicroscopic, and symbolic levels, makes this instrument a reference model for other chemistry topics, while also responding to calls in the literature (Dori, Y. J., & Sasson, 2008; Shwart et al, 2006) assessments capable of capturing higher-order thinking.

4. CONCLUSION

Theoretically, this study reinforces the Rasch model as an appropriate psychometric approach for the development of science test instruments. Practically and in terms of usability, the resulting instrument can be used by teachers to measure students' abilities more accurately and to design instruction that corresponds to students' ability levels. For future researchers, these findings demonstrate that the Rasch model can be applied as an analytical method in the development of evaluation instruments in other fields of study. The field test results of the developed instrument were declared valid. With Outfit MNSQ values ranging from 0.81 to 1.20, Outfit ZSTD values for all items ranging from -1.43 to 1.34, and point measure correlation values ranging from 0.60 to 0.72. The reliability analysis test results showed a person reliability of 0.78, categorized as fair/adequate, and an item reliability of 0.99, categorized as excellent, thus meeting the established criteria.

REFERENCES

- Anderson, L. W., & Krathwohl, D. R. (2020). Blooms Taxonomy Revised - Understanding the New Version of Bloom's Taxonomy. *Understanding the New Version of Bloom's Taxonomy*, 1(1), 1–8.
- Arikunto, S. (2018). *Dasar-dasar evaluasi pendidikan*. Rineka Cipta.
- Atik A , Marieke M, M. G. S. (2023). Uncovering important 21st-century skills for sustainable career development of social sciences graduates: A systematic review. *Educational Research Review*, 39(August 2022). <https://doi.org/10.1016/j.edurev.2023.100528>
- Bond, T. G., & Fox, C. M. (2015). Applying the Rasch Model. In *Applying the Rasch Model*. Routledge/Taylor & Francis Group. <https://doi.org/10.4324/9781410614575>
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch analysis in the human science. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 5(2), 208–211.
- Burhanuddin B, Junaidi, E., & Al Idrus, S. W. (2023). Analisis Tingkat Lierasi Kimia Siswa Kelas XI di SMAN 1 Mataram. *Chemistry Education Practice*, 6(2)(ISSN 2654-8119 (print), ISSN 2656-3940 (online)), 240–247. <https://doi.org/10.29303/cep.v6i2.5522>
- Dori, Y. J., & Sasson, I. (2008). Chemical understanding and graphing skills in an honors case-based computerized chemistry laboratory environment: The value of bidirectional visual and textual representations. *Journal of Research in Science Teaching*, 45(2), 219–250. <https://doi.org/10.1002/tea.20197>
- Eliza, W., & Yusmaita, E. (2021). Pengembangan Butir Soal Literasi Kimia pada Materi Sistem Koloid Kelas XI IPA SMA / MA Keywords : Chemical literacy instrument , Colloid system , MER , Rasch model. *JEP (Jurnal Eksakta Pendidikan)*, 5(2), 197–204.
- Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics*, 5(1), 1–4. <https://doi.org/10.11648/j.ajtas.20160501.11>
- Hasan E N. (2024). *Pesantren dalam Memecahkan Masalah Kesehatan Tubuh : Studi Kasus di Pondok Pesantren Mamba'us Sholihin* 8. 2(1).
- Hasanah, H., & Aini, F. Q. (2025). Pengembangan Instrumen Tes Keterampilan Proses Sains pada Topik Keseimbangan Kimia: Analisis dengan Model Rasch. *JagoMIPA: Jurnal*

- Pendidikan Matematika Dan IPA*, 5(1), 68–82.
<https://doi.org/10.53299/jagomipa.v5i1.1043>
- Hope, D., Kluth, D., Homer, M., Dewar, A., Jaap, A., & Cameron, H. (2025). Exploring the use of Rasch modelling in “common content” items for multi-site and multi-year assessment. *Advances in Health Sciences Education*, 30, 427–438.
- Hossan, D., Dato’ Mansor, Z., & Jaharuddin, N. S. (2023). Research Population and Sampling in Quantitative Study. *International Journal of Business and Technopreneurship (IJBT)*, 13(3), 209–222. <https://doi.org/10.58915/ijbt.v13i3.263>
- Jamaludin, J., Lay, Y. F., Khoo, C. H., & Leong, A. S. Y. (2021). Examining the Stem-Science Achievement Test (Ssat) Using Rasch Dichotomous Measurement Model. *Jurnal Pendidikan IPA Indonesia*, 10(4), 571–584. <https://doi.org/10.15294/jpii.v10i4.29405>
- Jayanti, S. D., & Sutrisno, H. (2025). Improving Chemical Literacy Ability and Self Efficacy through Context Based Approach with Prediction-Observation-Explanation Strategy. *Jurnal Penelitian Pendidikan IPA*, 11(9), 573–582. <https://doi.org/10.29303/jppipa.v11i9.11427>
- Kemendikdasmen. (2025). *Hasil Tes Kemampuan Akademik*.
- Lasahinda, A., Pulukadang, R., & Kumesan, S. (2025). Analisis Butir Soal Matematika Pada Penilaian Akhir Semester Ganjil Menggunakan Model Rasch Berbantuan Software Winstep. *Jurnal Pemikiran Dan Penelitian Pendidikan Matematika (JP3M)*, 8(2), 191–217. <https://journal.rekarta.co.id/index.php/jp3m/article/view/765>
- Lawshe, C. H. (1975). a Quantitative Approach To Content Validity. *Personnel Psychology*, 28(4), 563–575. <https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>
- Lestari, K. E., & Yudhanegara, M. R. (2017). Analisis Kemampuan Representasi Matematis Mahasiswa pada Mata Kuliah Geometri Transformasi Berdasarkan Latar Belakang Pendidikan Menengah. *Jurnal Matematika Integratif*, 13(1), 28–33. <https://doi.org/10.24198/jmi.v13.n1.11410.28-33>
- Liu, X. (2020). Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach. In *Coresource 4*. <https://doi.org/10.1002/sce.20477>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
- OECD. (2023). PISA 2022 Assessment and Analytical Framework. In *OECD Publishing*.
- Purniasari, L., Masykuri, M., & Ariani, S. R. D. (2021). Analisis Butir Soal Ujian Sekolah Mata Pelajaran Kimia Menggunakan Model Iteman dan Rasch. *Jurnal Pendidikan Kimia*, 10(2), 205–214.
- Rahmayanti, N. P. (2024). Spss Training Validity Tests and Reliability Tests for Primary Data. *BAKTI BANUA : Jurnal Pengabdian Kepada Masyarakat*, 5(2), 21–26.
- Rizki, M., & Yusmaita, E. (2021). Pengembangan Butir Soal Literasi Kimia pada Materi Ikatan Kimia Menggunakan Model Rasch. *Edukimia*, 3(2), 121–134. <https://doi.org/10.24036/ekj.v3.i2.a282>
- Saputri H A, Zulhijrah, Larasati N Ji, S. (2023). Analisis Instrumen Assesmen : Validitas, Relibitas, Tingkat Kesukara dan Daya Beda Butir Soal. *Didaktik : Jurnal Ilmiah PGSD FKIP Universitas Mandiri*, 9(5), 2986–2995.
- Sartika, A. Y. M., & Yusmaita, E. (2020). Pengembangan Asesmen Literasi Kimia pada Materi

- Hukum-Hukum Dasar Kimia dan Stoikiometri Kelas X SMA/MA. *Edukimia*, 2(3), 128–133. <https://doi.org/10.24036/ekj.v2.i3.a187>
- Setiyawan, R. A., & Wijayanti, P. S. (2020). Analisis Kualitas Instrumen Untuk Mengukur Kemampuan Pemecahan Masalah Siswa Selama Pembelajaran Daring Di Masa Pandemi. *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, 1(2), 130–139. <https://doi.org/10.46306/lb.v1i2.26>
- Sholeha, D. F., Kusasi, M., & Bakti, I. (2023). Pengembangan Instrumen Soal Higher-Order Thinking Skills (HOTS) Berbasis Google Form untuk Mengukur Hasil Belajar Peserta Didik pada Materi HID GARAM Development of Higher-Order Thinking Skills (HOTS) Questions Based on Google Form to Measure Stude. *Journal of Chemistry And Education*, 7(1), 32–48. <http://jtam.ulm.ac.id/index.php/jcae>
- Shwartz, Y., Ben-Zvi, R., & Hofstein, A. (2006). The use of scientific literacy taxonomy for assessing the development of chemical literacy among high-school students. *Chemistry Education Research and Practice*, 7(4), 203–225. <https://doi.org/10.1039/B6RP90011A>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan* (Bambang Sumintono & Wahyu Widhiarso, Ed.; 1st ed.). September 2015.
- Surya, C. and. (2024). *Content validity analysis of literacy assessment instruments*. 40(1), 81–91. <https://doi.org/10.21831/cp.v40i1.33474>
- Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2024). Interpreting results from Rasch analysis 1. The “most likely” measures coming from the model. *Disability and Rehabilitation*, 46(3), 591–603. <https://doi.org/10.1080/09638288.2023.2169771>
- Wrigley, W., Van Der Vleuten, C. P., Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71. *Medical Teacher*, 34(9), 683–697. <https://doi.org/10.3109/0142159X.2012.704437>
- Zayrin A. A., Nupus, H., Maizia, K. K., Marsela, S., Hidayatullah, R., & H. (2025). Analisis Instrumen Penelitian Pendidikan (Uji Validitas Dan Relibilitas Instrumen Penelitian). *Jurnal QOSIM Jurnal Pendidikan Sosial & Humaniora*, 3(2), 780–789. <https://doi.org/10.61104/jq.v3i2.1070>