JPPIPA 11(11) (2025)

Jurnal Penelitian Pendidikan IPA

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

Development of PISA-based Chemistry Test Instrument for Measuring Madrasah Aliyah Students' Scientific Literacy

Winda Ariani[1*], Nahadi[1], Heli Siti Halimatul Munawaroh[1]

[1] Department of Chemistry Education, Universitas Pendidikan Indonesia, Bandung, Indonesia.

Received: September 09, 2025
Revised: October 10, 2025
Accepted: November 25, 2025
Published: November 30, 2025

Corresponding Author:
Winda Ariani
windariani10@upi.edu

DOI: 10.29303/jppipa.v11i11.12808

Abstract: This study aims to produce a valid and reliable PISA-based chemistry test instrument on the acid–base concept to measure Madrasah Aliyah students' scientific literacy. The test instrument was developed using the ADDIE model (analyze, design, develop, implement, and evaluate). A total of 25 multiple-choice questions were tested on 60 Grade XII students majoring in natural sciences at Madrasah Aliyah in Subang. From the Rasch analysis, a Cronbach's alpha value of 0.71 indicates that the reliability of the test instrument is acceptable. In addition, the person reliability value of 0.67 indicates that the respondents' answers were quite consistent, and the item reliability value of 0.83 indicates that the reliability of the questions was good. All questions were declared valid because they met the outfit MNSQ, outfit ZSTD, and Pt-Measure Correlation criteria. The level of difficulty of the items was distributed into five very easy items, seven easy items, nine difficult items, and four very difficult items. The discriminating power of the items was classified into 24 items in the good category and one item in the sufficient category. In addition, 21 items had well-functioning distractors, while four items had distractors that did not function well.

Keywords: PISA-based test instrument; scientific literacy; ADDIE; Rasch model

Introduction

There are three essential components in the education system, namely the curriculum, instruction, and assessment (Kemendikbud, 2020). The curriculum defines the learning objectives to be achieved, instructional activities are conducted to attain these objectives, and assessment serves to determine the extent to which the objectives have been accomplished. According to Rohim et al. (as cited in Anggraini et al., 2022), assessment aims to measure what students have learned as an indicator of success in mastering specific competencies.

One of the government's efforts to improve the quality of Indonesian human resources in order to compete globally is through education. In the current era, students are required to possess 21st-century skills, commonly referred to as the 4C skills—creativity, critical thinking, collaboration, and communication—which are developed through both formal and informal educational experiences (Thornhill-Miller et al., 2023). According to Nahadi et al. (2019), these 21st-century skills can be fostered through the enhancement of students' literacy abilities.

The Ministry of Religious Affairs of the Republic of Indonesia, which oversees madrasahs across the country, has introduced an innovation in the field of assessment by implementing the Indonesian Madrasah Competency Assessment (Asesmen Kompetensi Madrasah Indonesia/AKMI). AKMI is an assessment designed to map the quality of the madrasah education system through competency-based assessment instruments in several literacy domains, namely reading literacy, numeracy literacy, scientific literacy, and social

How to Cite:
Ariani, W., Nahadi, & Munawaroh, H. S. H. (2025). Development of PISA-based Chemistry Test Instrument for Measuring Madrasah Aliyah Students' Scientific Literacy. *Jurnal Penelitian Pendidikan IPA*, *11*(11), 1291–1302. https://doi.org/10.29303/jppipa.v11i11.12808

and cultural literacy for madrasah students in Grades V, VIII, and XI (Susanti et al., 2021). The Ministry of Religious Affairs considers the implementation of AKMI to be important because madrasahs have distinctive characteristics compared to general schools, particularly in terms of the number of subjects and students' prior educational backgrounds. These unique characteristics influence the form of assessment stimuli, which need to be aligned with students' prior knowledge. Therefore, AKMI was developed based on the PISA framework by elaborating the distinctive features of madrasahs and adopting the National Assessment system developed by the Ministry of Education, Culture, Research, and Technology (Kemenag, 2022).

Research conducted by Suryadi (2024) demonstrated an improvement in AKMI results from 2022 to 2023 across all literacy domains. However, despite this improvement, the average scores of students in each literacy domain are still considered low, as illustrated in Figure 1.

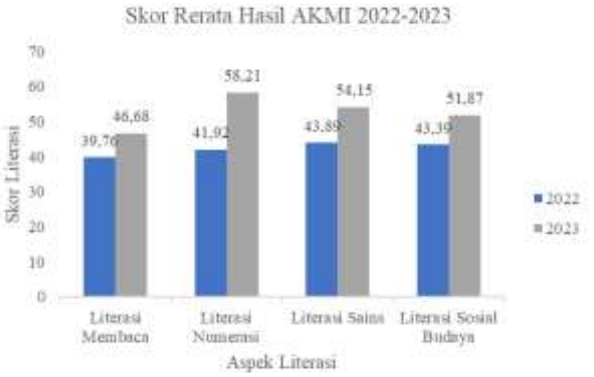


Figure 1. Average Scores of AKMI Results (2022–2023)

The AKMI results data for 2022–2023 presented in Figure 1 indicate that the scientific literacy of madrasah students remains generally very low. According to PISA, scientific literacy refers to the ability to use scientific knowledge to explain phenomena, evaluate and design scientific investigations, and interpret scientific data and evidence to make informed decisions (OECD, 2019). The Programme for International Student Assessment (PISA) is an international assessment conducted by the Organisation for Economic Co-operation and Development (OECD) to measure the reading, mathematical, and scientific literacy of 15 year old students from various countries. The PISA science framework comprises aspects of context, competencies, knowledge, and attitudes toward science, which are designed to assess students' understanding and application of science in real-life situations, thereby enabling them to think critically and make scientifically based decisions (Ovira, 2018). Despite the important role of scientific literacy in 21st-century education, many Madrasah Aliyah students still experience difficulties in mastering this competency.

Teachers play a crucial role in determining the success of the learning process. However, many teachers have not yet developed sufficient capacity to construct scientific literacy assessment instruments based on the PISA framework (Chamisijatin et al., 2022). Assessment instruments developed by Madrasah Aliyah teachers often do not adequately incorporate scientific literacy aspects, and teachers' ability to design scientific literacy items remains limited (Ardianti et al., 2022). Furthermore, the assessment formats commonly used have not fully measured students' abilities to understand and apply scientific concepts in contextualized situations (Sudirman et al., 2024). Existing test instruments generally focus on content mastery rather than on scientific literacy aspects, such as the application of science in everyday life, critical thinking skills in problem solving, and science process skills (Ridwan et al., 2013). Susanto et al. (2022) identified one of the challenges faced by teachers as the limited availability of scientific literacy evaluation instruments that can be used to train students to work on similar types of questions. PISA-oriented items emphasize problem solving and higher-order reasoning rather than rote memorization (Ovira, 2018). This is consistent with the findings of Broietti et al. (2019), which showed that chemistry-related content in PISA scientific literacy items predominantly measures students' abilities to analyze and interpret data, construct arguments, and make predictions regarding cause–effect relationships.

A study by Zhang et al. (2023), entitled "Development and Validation of an Instrument for Assessing Scientific Literacy from Junior to Senior High School," focused on the development and validation of a scientific literacy test instrument based on the PISA 2015 framework for students in Grades 6, 9, and 12. The instrument integrated content from physics, biology, chemistry, and geography to examine the progression of students' scientific literacy achievement across grade levels. Data were analyzed using the Rasch model, followed by the Bookmarking method to classify students' scientific literacy proficiency levels. However, during the instrument pilot-testing stage, Rasch model analysis was not conducted, indicating the need for further instrument validation. In addition, Anggraeni et al. (2022), in their study entitled "Students' Scientific Literacy in Chemistry Learning through Collaborative Techniques as a Pillar of 21st-Century Skills," assessed students' scientific literacy using 40 multiple-choice items developed by Moore and Foy (1997) and observation rubrics to measure senior high school students' collaboration skills. Although the development of scientific literacy assessment instruments has been widely conducted in schools, to date, no study has specifically examined the adaptation of the PISA science framework and AKMI within the context of developing chemistry test instruments for Madrasah Aliyah students. Therefore, the present study not only extends the application of the PISA framework in chemistry learning assessment at madrasahs but also contributes to the development of test instruments aligned with the requirements of AKMI.

The content of acid–base solution topics encompasses factual, conceptual, procedural, and metacognitive knowledge that is closely related to everyday phenomena (Muntholib et al., 2020). Acid–base concepts constitute a fundamental foundation of chemistry and have practical applications in various aspects of life, such as digestion systems, food preservation, acid rain, corrosion, pharmaceuticals, fertilizers, and packaged beverages (Ultay et al., 2016). Several studies have reported that students experience difficulties in understanding acid–base concepts and applying them to real-life contexts (Putri et al., 2022). Students are often unable to explain the causes of acid rain and its environmental impacts (Mufidah et al., 2024). In addition, students face challenges in describing the mechanism of action of antacids in neutralizing excess gastric acid (Saputri et al., 2022). Students' inability to connect acid–base concepts with everyday phenomena highlights the need for a context-based chemistry test instrument on acid–base topics. Based on this background, the following research questions were formulated:

1. How is the development design of a PISA framework–based chemistry test instrument on acid–base solutions for Madrasah Aliyah students?
2. What is the quality of the developed PISA framework–based chemistry test instrument based on Rasch model analysis in terms of validity, reliability, item difficulty, discriminating power, and distractor functioning?

Method

This study employed a research and development approach, with the product being a PISA framework–based chemistry test instrument on acid–base solutions for Madrasah Aliyah students. The development of the test instrument followed the ADDIE model proposed by Branch (2009), which consists of the stages of analysis, design, development, implementation, and evaluation, as illustrated in Figure 2.

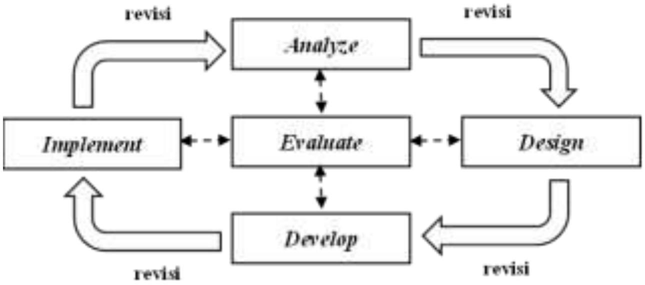


Figure 2. ADDIE Model Framework

Design of the Chemistry Test Instrument Development

During the analysis stage, document analysis was conducted on the revised 2023 Merdeka Curriculum, the PISA science framework, the AKMI science framework, and acid–base solution materials from chemistry textbooks by Raymond Chang, Petrucci, and publications issued by the Ministry of Education, Culture, Research, and Technology. In the design stage, the test format was determined, a test blueprint was developed, and test items were written. In the development stage, the content validity of the test instrument was evaluated by experts, test items were revised based on the validators' feedback, and a limited trial was conducted with students who had already received instruction on acid–base solutions. In the implementation stage, the test instrument consisting of items that had been declared valid and reliable was administered to students to measure their scientific literacy. Finally, in the evaluation stage, all phases of the development process were reviewed, including the results of expert content validation, pilot testing, and test implementation.

Research Instruments

The instruments used in this study included content validity evaluation sheets, a PISA framework–based chemistry test instrument consisting of 25 multiple-choice items, a student response questionnaire, and interview guidelines.

Participants and Research Setting

The participants involved in this study comprised five expert validators and 92 Grade XII students majoring in natural sciences from Madrasah Aliyah located in Subang Regency.

Data Analysis

The Content Validity Ratio (CVR) and Content Validity Index (CVI) methods were employed to assess the content validity of the developed instrument (Lawshe, 1975). If the CVR value of an item met or exceeded the specified threshold, the item was considered valid and suitable for use after being revised according to the validators' suggestions. The calculation of content validity using the CVR method was conducted using the following formula:

$$\mathrm{CVR} = \frac{n_e - \frac{N}{2}}{\frac{N}{2}} \qquad (1)$$

Keterangan:

CVR = content validity ratio

n_e = number of validators who rated the item as valid

N = total number of validators

Subsequently, the Content Validity Index (CVI) was calculated using the following formula:

$$\mathrm{CVI} = \frac{\Sigma \mathrm{CVR}}{\Sigma \mathrm{items}} \qquad (2)$$

The next step involved analyzing the quality of the test items using the Rasch model with the aid of the Ministep software version 5.10.2. The analysis included empirical validity testing, reliability, item difficulty, discriminating power, and distractor functioning. Empirical validity refers to the degree to which test results conform to predetermined criteria and the extent to which a test accurately measures the intended construct. According to Boone, Staver, and Yale (2014), test items in Rasch model analysis are considered valid if they meet three main criteria, as presented in Table 1.

Table 1. Criteria for Item Validity

Criteria	Analysis Results
Outfit Mean Square (MNSQ)	0.5 < MNSQ < 1.5
Outfit Z-Standard (ZSTD)	-2 < ZSTD < 2
Point Measure Correlation	0.4 < Pt-Measure Corr < 0.85

Reliability refers to the consistency of test results when administered under similar conditions. Item reliability was determined based on Cronbach's alpha coefficient and categorized according to the criteria proposed by Gliem et al. (2003), as shown in Table 2.

Table 2. Reliability Categories

Reliability Coefficient	Category
≥ 0.90	Very High
0.80 – 0.89	High
0.70 – 0.79	Acceptable
0.60 – 0.69	Questionable
0.50 – 0.59	Poor
< 0.50	Unaacptable

In Rasch model analysis, person reliability and item reliability values were obtained to determine the consistency of respondents' responses and test items. Sumintono et al. (2015) categorized these reliability values as presented in Table 3.

Table 3. Person Reliability and Item Reliability Categories

Value	Category
> 0.94	Exellent
0.91 – 0.94	Very Good
0.81 – 0.90	Good
0.67 – 0.80	Fair
< 0.67	Weak

Discriminating power indicates the ability of test items to distinguish between students with high and low abilities (Arikunto, 2009). According to Purniasari et al. (2021), discriminating power can be identified based on the standard error (S.E.) values in the Rasch model. The discriminating power of test items was classified into categories as presented in Table 4.

Table 4. Discriminating Power Categories

Discriminating Power	Category
Model S.E < 0.50	Good
0.50 ≤ Model S.E ≤ 1.00	Fair
Model S.E > 1.00	Poor

Test items are considered good if they are neither too easy nor too difficult. Items that are too easy do not stimulate higher-order thinking skills, while items that are too difficult may reduce students' motivation and lead to frustration (Arikunto, 2009). In Rasch model analysis, item difficulty is determined by comparing each item's logit measure value with the standard deviation. Sumintono et al. (2015) classified item difficulty levels as shown in Table 5.

Table 5. Item Difficulty Levels

Value	Category
Measure logit < -SD	Very Easy
-SD ≤ Measure logit ≤ 0	Easy
0 ≤ Measure logit ≤ SD	Difficult
Measure logit > SD	Very Difficult

Distractor functioning analysis was conducted to determine whether distractors effectively attracted students who had not mastered the tested subject matter. A distractor is considered effective if it is selected by at least 5% of test participants (Arikunto, 2009).

Results and Discussion

This development research aimed to produce a PISA framework–based chemistry test instrument on acid–base solutions that is capable of measuring the scientific literacy of Madrasah Aliyah students through the stages of the ADDIE model, which are described as follows.

Analyze

The activities carried out at this stage involved document analysis, including curriculum documents, the PISA and AKMI science frameworks, and instructional materials from reference textbooks. One of the chemistry learning outcomes for Phase F in the revised 2023 Merdeka Curriculum is that students are able to understand solution concepts in everyday life (Kemendikbudristek, 2023), including acid–base solutions. The learning outcomes were further elaborated into several learning objectives related to acid–base solutions. In general, the learning objectives for the acid–base topic include students' ability to distinguish the properties of acidic and basic solutions, understand various acid–base theories, comprehend the working principles of acid–base indicators, calculate the pH of acid–base solutions, write acid–base reactions, and understand the principles of acid–base titration.

Analysis of Learning Materials from Reference Textbooks

The analysis of learning materials was conducted using chemistry textbooks by Raymond Chang (2008), General Chemistry: The Essential Concepts; Petrucci et al. (2017), General Chemistry: Principles and Modern Applications; and Yuliani et al. (2022), Chemistry for Senior High School/Madrasah Aliyah Grade XII. This material analysis aimed to identify the scope and boundaries of concepts to be developed into test items. In general, the acid–base solution content developed into test items included the properties and characteristics of acidic and basic solutions; acid–base theories (Arrhenius, Brønsted–Lowry, and Lewis); pH calculations of acidic and basic solutions; acid–base indicators; types of acid–base reactions; and acid–base titration.

Analysis of the PISA and AKMI Science Frameworks

The analysis of the PISA science framework was conducted to identify the distinctive characteristics of PISA items as a guideline for item development. The results indicated that PISA cognitive assessment items must encompass aspects of context, knowledge (content, procedural, and epistemic), competencies, and cognitive levels. In addition, the items are accompanied by stimulus texts that students must comprehend in order to answer the questions. The distinctive aspects of content, context, and competencies found in PISA scientific literacy items also appear in AKMI scientific literacy items. Nevertheless, the analysis revealed differences in the distribution of scientific literacy items between the PISA and AKMI frameworks, as presented in Table 6.

Table 6. Composition of Items in the PISA and AKMI Frameworks

Aspect	PISA	AKMI
Knowledge	Conceptual (54–66%); Procedural (19–31%); Epistemic (10–22%)	Conceptual (10%); Procedural (35%); Epistemic (55%)
Context	Personal (25%); Local/National (50%); Global (25%)	Local/National (50%); Global (50%)
Competencies	Explaining phenomena scientifically (40–50%); Evaluating and designing scientific investigations (20–30%); Interpreting data and evidence scientifically to make decisions (30–40%)	Explaining phenomena scientifically (20%); Evaluating and designing scientific investigations (40%); Interpreting data and evidence scientifically to make decisions (40%)
Cognitive Level	Low; Medium; High	L1 (20%); L2 (48%); L3 (32%)

Design

At this stage, the test format was selected, a test blueprint was developed, and test items were written.

Determination of Test Format

Objective tests in the form of multiple-choice items facilitate students in completing test tasks more efficiently and allow for objective scoring (Djaen et al., 2021). In addition, multiple-choice test formats can be utilized to comprehensively measure various aspects of cognitive ability (Gurel et al., 2015). According to Muntholib et al. (2020), knowledge and competency aspects of chemical literacy that emphasize cognitive abilities can be effectively assessed using multiple-choice test formats.

Development of the Test Blueprint

The test blueprint was developed as a systematic guideline for item construction to ensure that the assessment instrument accurately reflects the intended competencies and content domains to be measured (Shofiyah et al., 2018). The blueprint was designed by integrating the core components of the PISA science framework, including context, scientific knowledge, scientific competencies, and cognitive levels. In addition, it specifies learning objectives, item indicators, and item numbering to maintain alignment between instructional goals and assessment outcomes.

Construction of Test Items

This study aimed to assess students' scientific competency achievement, which constitutes the fundamental dimension of scientific literacy. Test items were constructed to represent the three core competencies of PISA scientific literacy. The distribution of items across these competencies in the developed test instrument is presented in Table 7.

Table 7. Distribution of PISA Scientific Literacy Competencies

Scientific Competency	Items Number
Explaining phenomena scientifically	1, 6, 7, 12, 13, 14, 17, 21
Evaluating and designing scientific inquiry	2, 4, 10, 15, 16, 19
Interpreting data and scientific evidence to make informed decisions	3, 5, 8, 9, 11, 18, 20, 22, 23, 24, 25

Development

Content validation of the test items was conducted through expert judgment involving three lecturers in chemistry education and two senior high school chemistry teachers. The validation process aimed to examine the relevance, clarity, and representativeness of each item in relation to the intended indicators and measured constructs. The Content Validity Ratio (CVR) for each item was calculated and compared with the minimum acceptable CVR value proposed by Wilson et al. (2012). At a significance level of $\alpha = 0.05$ and with five validators, the minimum CVR threshold was 0.736. Accordingly, items with CVR values exceeding 0.736 were classified as valid.

Furthermore, the overall Content Validity Index (CVI) obtained was 0.92, indicating excellent content validity, as it exceeded the recommended minimum value of 0.80 (Davis, 1992; Polit & Beck, 2006). Items that did not meet the CVR criterion were revised based on qualitative feedback provided by the validators. This revision process is consistent with the assertion of Kalkbrenner (2021), who emphasized that assessment items judged to be inadequate by experts should not be eliminated outright, but rather refined to enhance alignment with the intended measurement objectives. Selected examples of item revisions are presented in Table 8.

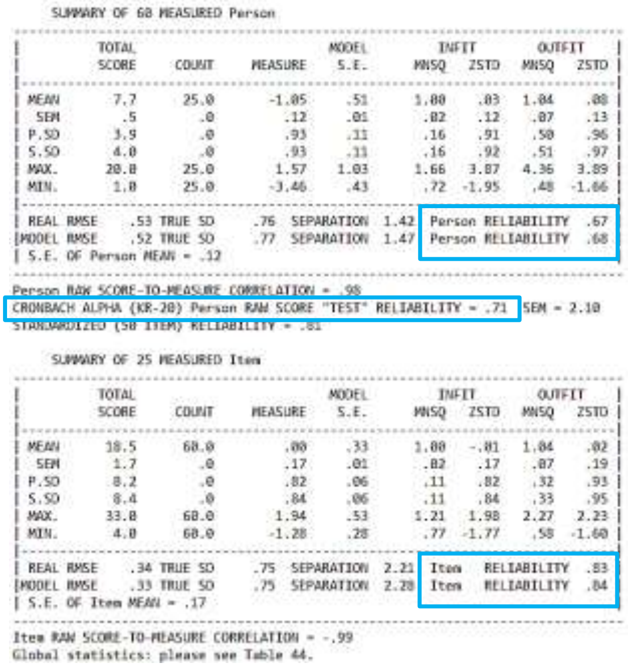
Table 8. Sample Test Items Before and After Revision

Before Revision
4. Based on the results of laboratory experiments, Ahmad concludes that … a. antacids are neutral b. antacids are acidic c. $Mg(OH)_2$ is ineffective in treating gastric pain d. $Al(OH)_3$ is the most effective in treating gastric pain e. an alkalimetric reaction occurs
8. A researcher collected 500 mL water samples from each location. The concentration of sulfuric acid in River B is … (log 2=0,3; log 4=0,6; log 5=0,7; log 8=0,9) a. 1×10^{-4} M b. 4×10^{-4} M c. 2×10^{-5} M d. 4×10^{-5} M e. 8×10^{-5} M
After Revision
4. The laboratory results obtained by Ahmad demonstrate that gastric acid solution (HCl) can be stoichiometrically neutralized by weak base components in antacids, namely magnesium hydroxide ($Mg(OH)_2$) and aluminum hydroxide ($Al(OH)_3$), until the equivalence point is reached. The quantification of gastric acid concentration through the addition of a standardized base solution is classified as which type of titration? a. oxidation–reduction b. acidimetry c. iodometry d. acid–base titration e. alkalimetry
8. A researcher collected 500 mL water samples from each location for analysis. The results indicate that the concentration of H^+ ions at the three locations is … (log 2 = 0.3; log 4 = 0.6; log 5 = 0.7; log 8 = 0.9) a. A < B < C b. A < C < B c. B < C < A d. C < B < A e. C < A < B

According to Arikunto (2009), clearly formulated questions that do not lead to multiple interpretations constitute one of the essential requirements in the construction of multiple-choice items. Additional revisions were conducted by incorporating incomplete or missing data required to answer the questions appropriately. The stimulus passages were also enriched to enhance their attractiveness and to ensure that they contained relevant and meaningful contextual information. Furthermore, revisions were made by designing answer options with comparable levels of plausibility and equivalence, thereby improving the quality of distractors, as exemplified in Item 4. Equivalent and well-functioning distractors are essential to minimize random guessing and to ensure that students select answers based on conceptual understanding rather than differences in the clarity or obviousness of the options (Haladyna et al., 2002).

The revised test instrument was subsequently piloted with 60 Grade XII students enrolled in the science track. The test consisted of 25 multiple-choice items and was administered within a 90-minute time allocation. Item analysis was conducted using the Rasch model with the Ministep software (version 5.10.2). The analyses included empirical validity, reliability, item difficulty level, item discrimination, and distractor functioning. Figure 3 presents a summary of the Rasch-based item analysis results.

Figure 3. Summary Statistics



```
     SUMMARY OF 68 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT    MEASURE   S.E.       MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      7.7       25.0      -1.05     .51       1.00    .03   1.04    .08 |
|  SEM       .5         .0        .12     .01        .02    .12    .07    .13 |
| P.SD      3.9         .0        .93     .11        .16    .91    .50    .96 |
| S.SD      4.0         .0        .93     .11        .16    .92    .51    .97 |
| MAX.     20.0       25.0       1.57    1.03       1.66   3.87   4.36   3.89 |
| MIN.      1.0       25.0      -3.46     .43        .72  -1.95    .48  -1.66 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .53 TRUE SD     .76  SEPARATION  1.42  Person RELIABILITY  .67 |
|MODEL RMSE    .52 TRUE SD     .77  SEPARATION  1.47  Person RELIABILITY  .68 |
| S.E. OF Person MEAN = .12                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .71  SEM = 2.10
STANDARDIZED (50 ITEM) RELIABILITY = .81

     SUMMARY OF 25 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT    MEASURE   S.E.       MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     18.5       68.0        .00     .33       1.00   -.01   1.04    .02 |
|  SEM      1.7         .0        .17     .01        .02    .17    .07    .19 |
| P.SD      8.2         .0        .82     .06        .11    .82    .32    .93 |
| S.SD      8.4         .0        .84     .06        .11    .84    .33    .95 |
| MAX.     33.0       68.0       1.94     .53       1.21   1.98   2.27   2.23 |
| MIN.      4.0       68.0      -1.28     .28        .77  -1.77    .58  -1.60 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .34 TRUE SD     .75  SEPARATION  2.21  Item   RELIABILITY  .83 |
|MODEL RMSE    .33 TRUE SD     .75  SEPARATION  2.28  Item   RELIABILITY  .84 |
| S.E. OF Item MEAN = .17                                                     |
-------------------------------------------------------------------------------
Item RAW SCORE-TO-MEASURE CORRELATION = -.99
Global statistics: please see Table 44.
```

Figure 3. Summary Statistics

Reliability

Based on the data presented in Figure 3, the Cronbach's alpha coefficient was 0.71. According to Gliem et al. (2003), a Cronbach's alpha value within the range of 0.70–0.79 indicates that the reliability of the test instrument is acceptable. Therefore, the interaction between respondents and the developed test items can be considered sufficiently reliable and is expected to produce consistent results when administered repeatedly. This finding is consistent with the view of Nunnally and Bernstein (1994, as cited in Gignac, 2009), who suggested that instruments with reliability coefficients above 0.70 are adequate for exploratory research. Similarly, Daud et al. (2018) classified Cronbach's alpha values ranging from 0.60 to 0.80 as indicative of good reliability.

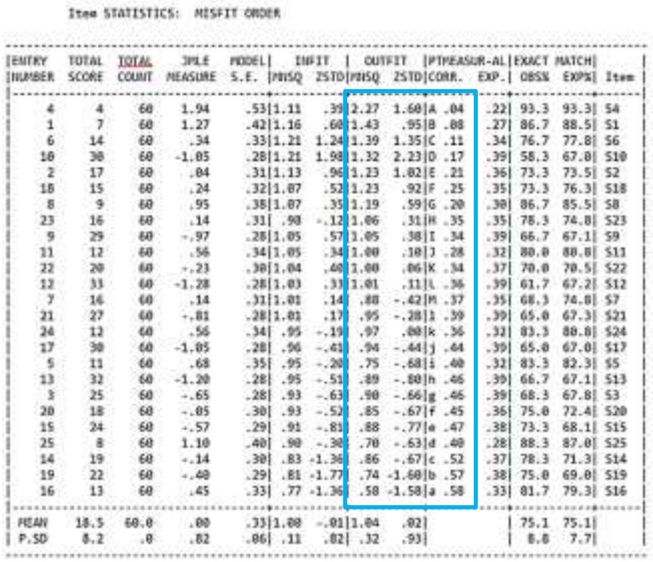
According to Sumintono et al. (2015), an item reliability value of 0.83 indicates that the item reliability is categorized as good. This finding suggests that the developed test items exhibit strong internal consistency. In addition, the item separation index of 2.21 and item strata value of 3.28 indicate that the items can be classified into three levels of difficulty, which are considered sufficiently stable. This result aligns with Bond and Fox (2015), who emphasized that high item reliability reflects the consistency of an instrument in distinguishing among varying levels of item difficulty.

The person reliability value of 0.67 indicates that respondents demonstrated moderate consistency in responding to the test items (Sumintono et al., 2015). Although the value indicates acceptable reliability, it is still considered relatively low, as good person reliability typically exceeds 0.80. The person separation index of 1.42 and person strata value of 2.23 suggest that respondents were divided into only two levels of ability, indicating limited stability. The relatively low person reliability can be attributed to two primary factors. First, the limited number of test items resulted in insufficient information to capture variations in respondents' abilities. Second, the targeting of item difficulty to respondent ability was suboptimal, as the overall item difficulty tended to exceed the average ability of the respondents. This mismatch likely led to inconsistent responses, with many students resorting to guessing.

Validity

Item validity was evaluated based on the outfit Mean Square (MNSQ), outfit Z-Standard (ZSTD), and Point-Measure Correlation (Pt-Measure Corr) values meeting the established criteria. Items that satisfied all three criteria were classified as highly valid (overfit). Items that failed to meet all criteria were considered invalid (misfit) and were subject to replacement or removal. However, items that met only one or two of the criteria were still regarded as acceptable (fit) and thus retained in the instrument (Nurdini et al., 2020; Vera et al., 2023).

Items that met the measurement requirements were retained, whereas those that failed to satisfy the criteria were revised, modified, or removed from the test instrument (Hailaya et al., 2014). As noted by Chong et al. (2023), item fit statistics serve to evaluate the extent to which each test item reflects the intended construct. Misfitting items indicate that the item may be measuring a construct different from the one intended (Baghaei & Amrahi, 2011; Boone & Staver, 2020). The item fit statistics obtained from the pilot testing are presented in Figure 5.



```
        Item STATISTICS:  MISFIT ORDER

-------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL    JMLE   MODEL|   INFIT  |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|     4      4     60    1.94     .53|1.11   .39|2.27  1.60|A .04   .22| 93.3  93.3| S4   |
|     1      7     60    1.27     .42|1.16   .60|1.43   .95|B .08   .27| 86.7  88.5| S1   |
|     6     14     60     .34     .33|1.21  1.24|1.39  1.35|C .11   .34| 76.7  77.8| S6   |
|    10     30     60   -1.05     .28|1.21  1.98|1.32  2.23|D .17   .39| 58.3  67.0| S10  |
|     2     17     60     .04     .31|1.13   .96|1.23  1.02|E .21   .36| 73.3  73.5| S2   |
|    18     15     60     .24     .32|1.07   .52|1.23   .92|F .25   .35| 73.3  76.3| S18  |
|     8      9     60     .95     .38|1.07   .35|1.19   .59|G .20   .30| 86.7  85.5| S8   |
|    23     16     60     .14     .31| .98  -.12|1.06   .31|H .35   .35| 78.3  74.8| S23  |
|     9     29     60    -.97     .28|1.05   .57|1.05   .38|I .34   .39| 66.7  67.1| S9   |
|    11     12     60     .56     .34|1.05   .34|1.00   .10|J .28   .32| 80.0  80.8| S11  |
|    22     20     60    -.23     .30|1.04   .40|1.00   .06|K .34   .37| 70.0  70.5| S22  |
|    12     33     60   -1.28     .28|1.03   .33|1.01   .11|L .36   .39| 61.7  67.2| S12  |
|     7     16     60     .14     .31|1.01   .14| .88  -.42|M .37   .35| 68.3  74.8| S7   |
|    21     27     60    -.81     .28|1.01   .17| .95  -.28|l .39   .39| 65.0  67.3| S21  |
|    24     12     60     .56     .34| .95  -.19| .97   .00|k .36   .32| 83.3  80.8| S24  |
|    17     30     60   -1.05     .28| .96  -.41| .94  -.44|j .44   .39| 65.0  67.0| S17  |
|     5     11     60     .68     .35| .95  -.20| .75  -.68|i .40   .32| 83.3  82.3| S5   |
|    13     32     60   -1.20     .28| .95  -.51| .89  -.80|h .46   .39| 66.7  67.1| S13  |
|     3     25     60    -.65     .28| .93  -.63| .90  -.66|g .46   .39| 68.3  67.8| S3   |
|    20     18     60    -.05     .30| .93  -.52| .85  -.67|f .45   .36| 75.0  72.4| S20  |
|    15     24     60    -.57     .29| .91  -.81| .88  -.77|e .47   .38| 73.3  68.1| S15  |
|    25      8     60    1.10     .40| .90  -.30| .70  -.63|d .40   .28| 88.3  87.0| S25  |
|    14     19     60    -.14     .30| .83 -1.36| .86  -.67|c .52   .37| 78.3  71.3| S14  |
|    19     22     60    -.40     .29| .81 -1.77| .74 -1.60|b .57   .38| 75.0  69.0| S19  |
|    16     13     60     .45     .33| .77 -1.36| .58 -1.58|a .58   .33| 81.7  79.3| S16  |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN   18.5   60.0     .00     .33|1.00  -.01|1.04   .02|           | 75.1  75.1|      |
| P.SD    8.2     .0     .82     .06| .11   .82| .32   .93|           |  8.8   7.7|      |
-------------------------------------------------------------------------------------
```

Figure 5. Item Fit Order

Therefore, based on the results of the Rasch model analysis presented in Figure 5, it can be concluded that all test items are valid and suitable for use. Seven items were classified as highly valid (overfit), namely Items 3, 13, 14, 15, 16, 17, 19, and 20, while the remaining 17 items were categorized as valid (fit).

Item Discrimination

Item discrimination refers to the effectiveness of an item in differentiating respondents according to the measured ability levels (Afriani et al., 2023). In the Rasch model, item discrimination is identified through the Model Standard Error (Model S.E.) values in the Item Measure output. Based on the analysis presented in Figure 5, information regarding the discrimination power of each test item was obtained. The results indicated that 24 items demonstrated good discrimination power, while one item showed moderate discrimination. Thus, the developed test items were able to effectively distinguish between respondents with high and low ability levels, as summarized in Table 9.

Table 9. Results of Item Discrimination Analysis

Discrimination Power	Item Number	%
Good (Model S.E < 0.50)	1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25	96%
Moderate (0.50 ≤ Model S.E ≤ 1.00)	4	4%

Item discrimination analysis aims to determine the extent to which a test item can distinguish between respondents with high and low ability levels (Bagiyono, 2017). According to Sakinah (2017, as cited in Purniasari et al., 2021), a test instrument can be considered to have good discrimination power if more than 50% of the items fall within the good discrimination category. Therefore, the developed test instrument is regarded as having good item discrimination.

Item Difficulty

Figure 5 also presents a standard deviation value of 0.82. The difficulty level of each item was determined by comparing the item measure (logit) values with the standard deviation. Based on the analysis, the distribution of item difficulty levels is presented in Table 10.

Table 10. Results of Item Difficulty Analysis

Difficulty Level	Item Numbers	%
Very Easy (Measure logit < -0.82)	9, 10, 12, 13, 17	20%
Easy (-0.82 ≤ Measure logit ≤ 0)	3, 14, 15, 19, 20, 21, 22	28%
Difficult (0 ≤ Measure logit ≤ 0.82)	2, 5, 6, 7, 11, 16, 18, 23, 24	36%
Very Difficult (Measure logit > 0.82)	1, 4, 8, 25,	16%

The 25 developed test items were classified into four levels of difficulty: very easy, easy, difficult, and very difficult. The test instrument is considered adequate because it includes a balanced distribution of difficulty levels. A test instrument is regarded as good when it contains a proportional range of item difficulty (Rusiyah et al., 2020). The variation in item difficulty is expected to accurately measure respondents' actual competencies and effectively differentiate their ability levels.

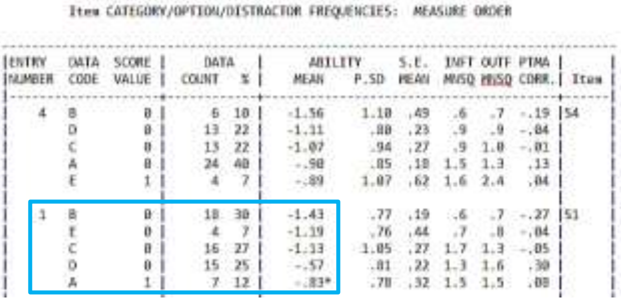
Distractor Functioning

Distractor functioning was examined using the Distractor Frequencies output. The results of the analysis are presented in Table 11.

Table 11. Results of Distractor Functioning Analysis

Distractor Functioning	Item Numbers	%
Functioning	3, 4, 5, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25	84%
Not Fungctioning	1, 2, 6, 10	16%

Overall, all five answer options were selected by respondents across the test items. The Rasch model analysis revealed that 21 items had well-functioning distractors, while four items—Items 1, 2, 6, and 10—contained non-functioning distractors.



```
        Item CATEGORY/OPTION/DISTRACTOR FREQUENCIES:  MEASURE ORDER

------------------------------------------------------------------------------
|ENTRY   DATA  SCORE |     DATA    |     ABILITY     S.E.  INFT OUTF PTMA |      |
|NUMBER  CODE  VALUE |  COUNT    % |   MEAN    P.SD  MEAN  MNSQ MNSQ CORR.| Item |
|--------------------+-------------+----------------------------------------+------|
|    4   B         0 |      6   10 |  -1.56    1.10  .49    .6   .7  -.19 |S4    |
|        D         0 |     13   22 |  -1.11     .80  .23    .9   .9  -.04 |      |
|        C         0 |     13   22 |  -1.07     .94  .27    .9  1.0  -.01 |      |
|        A         0 |     24   40 |   -.90     .85  .18   1.5  1.3   .13 |      |
|        E         1 |      4    7 |   -.89    1.07  .62   1.6  2.4   .04 |      |
|                    |             |                                        |      |
|    1   B         0 |     18   30 |  -1.43     .77  .19    .6   .7  -.27 |S1    |
|        E         0 |      4    7 |  -1.19     .76  .44    .7   .8  -.04 |      |
|        C         0 |     16   27 |  -1.13    1.05  .27   1.7  1.3  -.05 |      |
|        D         0 |     15   25 |   -.57     .81  .22   1.3  1.6   .30 |      |
|        A         1 |      7   12 |   -.83*    .70  .32   1.5  1.5   .08 |      |
```

Figure 6. Distractor Functioning

As shown in Figure 6, in Item 1 the correct answer (Option A) was more frequently selected by respondents with lower ability levels, whereas respondents with higher ability levels tended to choose an incorrect distractor, namely Option D. A similar pattern was observed in Items 2, 6, and 10. This finding is inconsistent with the statement of Firman (2000), who asserted that a well-functioning distractor should be selected more frequently by the lower-ability group than by the higher-ability group. In addition to these four items, several distractors did not meet Arikunto's (2009) criterion of being selected by at least 5% of respondents, specifically in Items 6, 7, 10, 12, 16, and 22. Consequently, these distractors require revision to prevent ambiguity and reduce the likelihood of random guessing during test completion.

Implementation

The validated and reliable test instrument, consisting of 25 multiple-choice items, was implemented with 32 Grade XII students enrolled in the science track at a Madrasah Aliyah located in Subang Regency. The data obtained from the implementation were used to measure students' scientific literacy achievement. Based on the AKMI framework (Ministry of Religious Affairs, 2022), students' score percentages were interpreted into five levels of scientific literacy proficiency: needs assistance (≤ 30%), basic (31–60%), competent (61–80%), proficient (81–90%), and creative extension required (91–100%).

Evaluation

During the evaluation stage, revisions were made based on expert judgments obtained from the content validity assessment to ensure alignment between the test instrument, the instructional content, and the intended measurement objectives. In addition, Rasch model analysis of the pilot test results was conducted to identify valid and reliable items suitable for subsequent use. Improvements focused on replacing non-functioning distractors and refining language to enhance clarity and communicative effectiveness. Well-designed distractors are expected to improve the overall quality of the developed test instrument.

Conclusion

The PISA framework-based chemistry test instrument, consisting of 25 multiple-choice items on acid–base concepts, was found to be valid, as it met the criteria for outfit MNSQ, outfit ZSTD, and Point-Measure Correlation values. Overall, the instrument demonstrated acceptable reliability, with a Cronbach's alpha of 0.71, good item reliability of 0.83, and moderate person reliability of 0.67. Although the person reliability value was adequate, it could be improved by administering the instrument to a larger and more heterogeneous sample or by increasing the number of test items. The distribution of item difficulty levels comprised five very easy items, seven easy items, nine difficult items, and four very difficult items. Item discrimination analysis revealed that 24 items exhibited good discrimination power, while one item demonstrated moderate discrimination. In terms of distractor functioning, 21 items contained well-functioning distractors, whereas four items included non-functioning distractors. Therefore, further development of PISA framework-based test instruments in other chemistry topics is feasible, particularly through the use of varied test formats.

Acknowledgments

We would like to express our sincere gratitude to all parties who have contributed to support the implementation of this research.

Author Contributions

Conceptualization, methodology, resources, and writing—review and editing, W.A, N, and H.S.H.M.; validation, N and H.S.H.M.; formal analysis, investigation, data curation, writing—original, draft preparation, and visualization, W.A. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by Beasiswa Indonesia Bangkit Kementerian Agama Republik Indonesia.

Conflicts of Interest

The authors declare no conflict of interest.

References

Afriani, E., Susilaningsih, E., Haryani, S., & Prasetya, A. T. (2023). Analisis Kompetensi Minimum Siswa pada Materi Hidrolisis Garam melalui Pengembangan Instrumen Tes Bermuatan AKM dengan Konteks Saintifik Daily Life. *Chemistry in Education, 12*(2), 162-170. https://doi.org/10.15294/chemined.v12i2.69259

Anggraeni, C., Permanasari, A., & Heliawati, L. (2022). Students' Scientific Literacy in Chemistry Learning through Collaborative Techniques as a Pillar of 21st-Century Skills. *Journal of Innovation in Educational and Cultural Research, 3*(3), 457-462. https://doi.org/10.46843/jiecr.v3i3.162

Anggraini, K. E., & Setianingsih, R. (2022). Analisis Kemampuan Numerasi Siswa SMA dalam Menyelesaikan Soal Asesmen Kompetensi Minimum (AKM). *Jurnal Mathedunesa*, *11*(3), 837-849. https://doi.org/10.26740/mathedunesa.v11n3.p837-849

Ardianti, R., Surahman, E., & Sujarwanto, E. (2022). Pengembangan Instrumen Penilaian Literasi Sains pada Bahasan Usaha dan Energi di Madrasah Aliyah. *Journal for Physics Education and Applied Physics*, *4*(1), 9-14. https://doi.org/10.37058/diffraction.v4i1.5359

Arikunto, S. (2006). *Prosedur Penelitian Suatu Pendekatan Praktik*. Jakarta: PT Rineka Cipta.

Arikunto, S. (2009). *Dasar-dasar Evaluasi Pendidikan Edisi*. Jakarta: PT Bumi Aksara.

Baghaei, P. (2008). The Rasch Model as a Construct Validation Tool. *Rasch measurement transactions, 22*(1), 1145-1162.

Bagiyono. (2017). Analisis Tingkat Kesukaran dan Daya Pembeda Soal Ujian Pelatihan Radiografi Tingkat 1. *Widyanuklida*, *16*(1), 1–12. http://jurnal.batan.go.id/index.php/widyanuklida/article/view/4068.

Branch, R. M. (2009). *Instructional Design: The ADDIE Approach*. London: Springer.

Broietti, F. C. D., Nora, P. S., & Costa, S. L. R. (2019). Dimensions of Science Learning: A Study on PISA Test Questions Involving Chemistry Content. *Acta Scientiae,* *21*(1), 95-115. https://doi.org/10.17648/acta.scientiae.v21iss1id4947

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). New York: Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. London: Springer.

Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch Analyses in the Human Sciences*. London: Springer International Publishing.

Chamisijatin, L., Pantiwati, Y., & Zaenab, S. (2022). Pendampingan Peningkatan Mutu Satuan Pendidikan melalui Penyusunan Tiga Instrumen Utama di SMP Muhammadiyah 02 Kota Batu. *Jurnal Abdimas (Journal of Community Service)*, *4*(2), 249–260. https://doi.org/10.36312/sasambo.v4i2.673

Chang, R. (2008). *General Chemistry: The Essential Concept*. (5th ed.). New York: Mc Graw Hill Company.

Chong, J., Mokshein, S. E., & Mustapha, R. (2023). Instrument Validation based on the Six Aspects of Messick Validity Framework using the Rasch Rating Scale Model. Publication at: https://www.researchgate.net/publication/366956609

Daud, K.A.M., Khidzir, N.Z., Ismail, A.R, & Abdullah, F.A. (2018). Validity and Reliability of Instrument to Measure Social Media Skills among Small and Medium Entrepreneurs at Pengkalan Datu River. *International Journal of Development and Sustainability, 7*(3), 1026-1037.

Davis, L. L. (1992). Instrument Review: Getting the Most from a Panel of Experts. *Applied Nursing Research*, *5*(4), 194–197. https://doi.org/10.1016/S0897-1897(05)80008-4

Djaen, N., Rahayu, S., Yahmin, & Muntholib. (2021). Chemical Literacy of First-Year Students on Carbon Chemistry. *Jurnal Pembelajaran Kimia, 6*(1), 41-62. https://doi.org/10.17977/um026v6i12021p041

Firman, H. (2000). *Penilaian Hasil Belajar dalam Pengajaran Kimia*. Bandung: Jurusan Pendidikan Kimia FPMIPA UPI.

Gignac, G.E. (2009). Psychometrics and the Measurement of Emotional Intelligence. In: Parker, J., Saklofske, D., & Stough, C. (Eds.), *Assessing Emotional Intelligence* (pp. 9-40). The Springer Series on Human Exceptionality. Boston: Springer. https://doi.org/10.1007/978-0-387-88370-0_2

Gliem, J.A., & Gliem, R.R. (2003). Calculating. Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales (Eds). *Midwest Research to Practice Conference in Adult, Continuing, and Community Education* (pp. 82-88). Ohio : Ohio State University.

Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2015). A Review and Comparison of Diagnostic Instruments to Identify Students' Misconceptions in Science. *Eurasia Journal of Mathematics, Science and Technology Education,* *11*(5), 989–1008. https://doi.org/10.12973/eurasia.2015.1369a

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Hailaya, W., Alagumalai, S., & Ben, F. (2014). Examining the Utility of Assessment Literacy Inventory and its Portability to Education Systems in the Asia Pacific Region. *Australian Journal of Education, 58*(3), 297–317. https://doi.org/10.1177/0004944114542984

Kalkbrenner, M. T. (2021). A Practical Guide to Instrument Development and Score Validation in the Social Sciences: The Measure Approach. *Practical*

Assessment, Research, and Evaluation, *26*(1) , 1-18. https://doi.org/10.7275/svg4-e671

Kemenag. (2022). *Framework Asesmen Kompetensi Madrasah Indonesia (AKMI) 2022*. Jakarta: Direktorat KSKK Madrasah, Ditjen Pendis Kemenag RI.

Kemendikbud. (2020). *AKM dan Implikasinya pada Pembelajaran*. Jakarta: Pusat Asesmen dan Pembelajaran, Badan Penelitian, Pengembangan dan Perbukuan, Kementrian Pendidikan dan Kebudayaan.

Kemendikbudristek. (2023). Capaian Pembelajaran untuk SMA/MA/Program Paket C pada Kurikulum Merdeka. Jakarta: Badan Standar, Kurikulum, dan Asesmen Pendidikan, Pusat Kurikulum dan Pembelajaran.

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

Maharani, S. D., Suganda, V. A., Laihat, M., Harini, B., Pulungan, M., & Safitri, M. L. O. (2022). *Asesmen Pembelajaran di Sekolah Dasar*. Palembang: Bening Media Publishing.

Mufidah, N., & Ardhana, I. A. (2024). Science Literacy Profile of Eleventh-Grade High School Students in Acid-Base. *Jurnal Pembelajaran Kimia, 9*(1), 40-45. http://dx.doi.org/10.17977/um026v9i12024p%25p

Muntholib, M., Khusmawardani, E., Utomo, Y., Muchson, M., & Yahmin, Y. (2020). Development and Implementation of Multiple-choice Chemical Literacy Survey in Acid-Base Chemistry. *AIP Conference Proceedings* of *The 3rd International Conference on Mathematics and Science Education (ICoMSE)*. https://doi.org/10.1063/5.0000547

Muntholib, M., Ibnu, S., Rahayu, S., Fajaroh, F., Kusairi, S., & Kuswandi, B. (2020). Chemical Literacy: Performance of First Year Chemistry Students on Chemical Kinetics. *Indonesian Journal of Chemistry, 20*(2), 468-482. https://doi.org/10.22146/ijc.43651

Nahadi, N., Siswaningsih, W., Purnawarman, P., Lestari, T., Febriani, A. E., & Rohmawati T. (2022). Development of Minimum Competency Assessment (AKM) on Chemical Materials. *Moroccan Journal of Chemistry*, *10*(3), 452-463. https://doi.org/10.48317/IMIST.PRSM/morjchem-v10i3.33067

Nurdini, N., Suhandi, A., Ramalis, T. R., A. Samsudin, A., Fratiwi, N. J., & Costu, B. (2020). Developing Multitier Instrument of Fluids Concepts (MIFO) to Measure Student's Conception: A Rasch Analysis Approach. *Journal of Advanced Research in Dynamical and Control Systems*, 12(6), 3069–3083. https://doi.org/10.5373/jardcs/v12i6/s20201273

OECD. (2019). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing.

OECD. (2023). *PISA 2025 Science Framework (Draft)*. Paris: OECD Publishing.

Ovira, E. D. (2018). Pengembangan dan Validasi Tes Kimia dengan *Framework* PISA pada Materi Kelas XI Semester 1. *MENARA Ilmu, 12*(1), 33-41. https://doi.org/10.33559/mi.v12i80.641

Petrucci, R. H., Herring, F. G., Madura, J. D., & Bissonnette, C. (2017). *General Chemistry : Principles and Modern Applications. (11th ed.)*. Kanada: Pearson.

Purniasari, L., Masykuri, M., & Ariani, S. R. D. (2021). Analisis Butir Soal Ujian Sekolah Mata Pelajaran Kimia SMAN 1 Kutowinangun Tahun Pelajaran 2019/2020 Menggunakan Model Iteman dan Rasch. *Jurnal Pendidikan Kimia*, *10*(2), 205–214. https://doi.org/10.20961/jpkim.v10i2.48244

Putri, A. A. A., Hussain, H., & Ramdhani. (2022). Pengembangan Instrumen Tes Literasi Sains pada Dimensi Pengetahuan Materi Asam Basa. *Jurnal Inovasi Pendidikan Matematika dan IPA, 2*(4), 536-547. https://doi.org/10.51878/SCIENCE.V2I4.1797

Polit, D. F., & Beck, C. T. (2006). The Content Validity Index: Are You Sure You Know What's Being Reported? Critique and Recommendations. *Research in Nursing & Health, 29*(5), 489–497. https://doi.org/10.1002/nur.20147

Ridwan et al. (2013). Development of a Student Scientific Literacy Test Instrument on the Topic of Diversity of Living Creatures. *Journal of Biology Education and Learning, 4*(1), 71–78.

Rusiyah., Eraku, S. S., & Supadmi, S. (2020). Analisis Soal Ujian Akhir Semester Mata Pelajaran Geografi Dengan Menggunakan Pemodelan Rasch. *Jurnal Geografi Dan Pembelajaran Geografi*, *5*(1), 11. https://doi.org/10.31851/swarnabhumi.v5i1.4136

Saputri, E. N., Wigati, I., & Laksono, P. J. (2022). Kemampuan Literasi Kimia pada Aspek Kompetensi Sains pada Materi Asam Basa. Prosiding Seminar Nasional Pendidikan Kimia, 223–231.

Shofiyah, N., & Sartika, S. B. (2018). *Buku Ajar Mata Kuliah Asesmen Pembelajaran*. Sidoarjo: Umsida Press.

Sudirman, Rusilowati, A., & Susilaningsih, E. (2024). Development of Multiple-Choice Test Instruments to Improve Scientific Literacy in Madrasah Aliyah (MA). *International Journal of Scientific Research and Management,* 12(6), 3465-3475. https://doi.org/10.18535/ijsrm/v12i06.el04

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Bandung: Trim Komunikata.

Suryadi, A. (2024). Pemanfaatan Hasil AKMI untuk Pembelajaran Berkelanjutan di Madrasah. *Jurnal Pelita Manajemen Pendidikan*, *1*(1), 1-8. https://doi.org/10.65226/jpmp.v1i1.1

Susanti, L. D., Pahrudin, A., & Yetri, Y. (2021). Analisis Pelaksanaan Asesmen Kompetensi Madrasah Indonesia (AKMI). *Journal of Interdisciplinary Science and Education*, *1*(2), 17-24. https://doi.org/10.70371/jise.v1i2.23

Susanto, E., Susanta, A., & Rusdi, R. (2022). Pelatihan Penyusunan Instrumen Tes Matematika *Online* Berbasis PISA bagi Guru Matematika SMP Bengkulu. *Jurnal Pengabdian Masyrakat*, *2*(3), 114−120. https://doi.org/10.47065/jpm.v2i3.330

Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J.-M., Morisseau, T., Bourgeois-Bougrine, S., Vinchon, F., El Hayek, S., Augereau-Landais, M., Mourey, F., Feybesse, C., Sundquist, D., & Lubart, T. (2023). Creativity, Critical Thinking, Communication, and Collaboration: Assessment, Certification, and Promotion of 21st Century Skills for the Future of Work and Education. *Journal of Intelligence*, *11*(54), 1-32. https://doi.org/10.3390/jintelligence11030054.

Ultay, N., & Calik, M. (2016). A comparison of different teaching designs of 'acids and bases' subject. *Eurasia Journal of Mathematics, Science & Technology Education, 12*(1), 57-86. https://doi.org/10.12973/eurasia.2016.1422a

Vera, R., Kaniawati, I., & Utama, J. A. (2023). Using the Rasch Model to Develop a Measure of Students' Problem Solving Ability in Optical Instruments. *Jurnal Pendidikan MIPA, 24*(2), 419-431. http://dx.doi.org/10.23960/jpmipa/v24i2.pp419-431

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the Critical Values for Lawshe's Content Validity Ratio. *Measurement and Evaluation in Counseling and Development*, *45*(3), 197-210. https://doi.org/10.1177/0748175612440286

Yuliani, G., Dianhar, H., & Suhendar, A. (2022). *Kimia untuk SMA/MA Kelas XII*. Jakarta: Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.

Zhang, L., et al. (2023). Development and Validation of an Instrument for Assessing Scientific Literacy from Junior to Senior High School. *Disciplinary and Interdisciplinary Science Education Research, 5*(21), 1-15. https://doi.org/10.1186/s43031-023-00093-2