JPII 15 (1) (2026) 51-64

Jurnal Pendidikan IPA Indonesia

https://journal.unnes.ac.id/journals/jpii

TRANSFORMATIVE ASSESSMENT INNOVATION BASED ON MULTIPLE REPRESENTATIONS IN HOLISTIC LEARNING TO MEASURE LOGICAL AND ANALYTICAL THINKING

E. Susilaningsih*[1], I. P. Salimnah[1], S. Nurhayati[1], Harjono[1], L. E. B. Mohtar[2], S. R. S. Aris[3]

[1]Department of Chemistry Education, Graduate School, Universitas Negeri Semarang, Semarang, Indonesia
[2]Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim, Malaysia
[3]Faculty of Education, Universiti Teknologi MARA, Shah Alam, Malaysia

DOI: 10.15294/jpii.v15i1.35317

Accepted: October 20th, 2025. Approved: March 24th, 2026. Published: March 24th, 2026

ABSTRACT

The focus of chemistry learning across multiple levels of representation is essential for developing students' conceptual understanding of complex chemical phenomena. However, previous studies have predominantly emphasized the use of multiple representations as instructional strategies, while their application in assessment, particularly in acid-base chemistry, remains limited. Moreover, existing assessment instruments seldom integrate logical and analytical thinking skills simultaneously, and the use of Rasch model analysis for diagnostic assessment in this context is still relatively scarce. To address these gaps, this study aims to develop a Transformative Assessment Innovation Based on Multiple Representations integrated into Holistic Learning to measure students' logical and analytical thinking skills. This research employed a Research and Development (R&D) approach using the ADDIE model. The participants were second-semester students enrolled in the Basic Analytical Chemistry course at Universitas Negeri Semarang. The assessment instrument was developed as multi-representational items encompassing definitional, macroscopic, submicroscopic, symbolic, and stoichiometric levels, aligned with holistic learning activities. The content validity and reliability of the instrument were analyzed using the Rasch model. The results showed that the developed assessment demonstrated high validity (CVR > 0.80) and excellent reliability (Cronbach's α > 0.90). Furthermore, the implementation findings indicated that the transformative assessment effectively identified and enhanced students' logical and analytical thinking skills while providing diagnostic and reflective feedback to support continuous learning. These findings suggest that transformative assessment integrated with multi-representational and holistic learning approaches offers a novel diagnostic framework for strengthening students' higher-order thinking skills in chemistry education.

Keywords: chemistry education; transformative assessment; level representation; acid-base titration; HOTS

INTRODUCTION

21st-century skills are essential for students to face the era of globalization. Therefore, the teaching and learning process in schools needs to encourage mastery of the 4C skills, which include (Communication, Collaborative, Critical thinking and problem-solving, and Creativity and innovation), including the development of higher-order thinking skills (HOTS) (Sepriyanti et al., 2022). One of the primary challenges in chemistry learning is that students often lack logical (Rakhmawan et al., 2018) and analytical thinking skills (Prastiwi & Laksono, 2018). Understanding is a crucial aspect of the learning process. Failure to properly grasp basic concepts can hinder students, including university students, from connecting them to related concepts. This situation can lead to misunderstandings of a concept (Supar-

***Correspondence Address**
E-mail: endangsusilaningsih@mail.unnes.ac.id

man et al., 2024). If these errors occur repeatedly and consistently, they can be categorized as misconceptions. Consistency in these errors can be identified by examining several questions based on the same concept, or by comparing written test results with interview results.

HOTS can be developed through a deep understanding of concepts and their applications. This understanding can be honed by practicing skills in interpreting, translating, drawing conclusions, generating ideas, and presenting problems using previously learned concepts (Al Haq & Raicudu, 2023). However, chemistry is considered difficult for most students because it involves both factual and conceptual knowledge, as well as procedural knowledge that must be understood (Rahmawati et al., 2022). In their learning, students are required to apply mathematical concepts and equations in specific calculations. The difficulties experienced are usually caused by a lack of direct experience and the challenges in solving problems involving chemical reactions and mathematical calculations. These characteristics make chemistry a subject that requires well-developed logical and analytical thinking skills, which need to be systematically fostered and measured through appropriate assessment approaches.

Logical thinking is one of the thinking skills required in science learning (Al-Ajmi & Abdullah, 2019). Logical thinking is the ability to use certain logical principles to solve a problem (Fauziah et al., 2022). Logical thinking in the context of chemistry learning is defined as students' ability to use scientific reasoning to consistently explain changes in substances during chemical reactions based on fundamental chemical principles (Park et al., 2019). There are three indicators of logical thinking: (1) orderly thinking, where students can express information obtained and analyze phenomena from the results of simple experiments; (2) argumentation skills, namely students' skills in solving problems rationally and providing reasons for the chosen solution steps; and (3) the ability to conclude, namely determining appropriate and precise final results (Anggraini & Irawan, 2021). These indicators highlight that logical thinking is not only essential for understanding chemical concepts but also needs to be explicitly assessed through structured instruments that can capture students' reasoning processes across different levels of representation.

Analytical thinking skills involve recognizing or determining relevant information from a material, organizing it into a connected whole, and assessing it to obtain a solution to a problem (Furqan et al., 2016). There are three indicators of analytical thinking skills: (1) grouping data into parts that correspond to existing concepts, (2) organizing the parts that have been arranged into an orderly and comprehensive whole, and (3) looking for relationships between the parts that have been organized (Fitriani et al., 2021). These indicators are essential for understanding complex chemistry topics (Dewi et al., 2025), such as acid–base reactions, where students must analyze experimental observations, symbolic equations, and stoichiometric calculations simultaneously (Herunata et al., 2021).

Science learning provides students with opportunities to learn about, respond to, appreciate, and understand science, while also developing scientific thinking habits such as critical thinking, creativity, independence, and a positive attitude toward science. Some students, in fact, still struggle to apply the knowledge they have learned, especially in making decisions, defining questions based on facts, and relating them to everyday life (Zuhaida & Imaduddin, 2019). This condition can cause students to become bored and struggle to solve problems. This condition is reflected in Indonesia's performance in the Programme for International Student Assessment (PISA) 2022, which showed a decline in students' achievement across reading, mathematics, and science compared to 2018, with an average decrease of 12–13 points (OECD, 2023). These findings indicate that students' analytical and logical thinking skills have not been optimally developed and assessed. Therefore, there is an urgent need for assessment approaches in science learning that not only measure learning outcomes but also diagnose students' thinking processes and support the development of higher-order cognitive skills.

Understanding chemistry becomes easier when presented in terms of levels of representation (Susilaningsih et al., 2018). The levels of representation used are definition, macroscopic, submicroscopic, symbolic, and stoichiometry (Safitri et al., 2019). Each representational level plays a distinct role in supporting students' conceptual understanding, ranging from observable phenomena to abstract symbolic and quantitative reasoning. In the context of acid–base titration, these representations are essential for measuring students' understanding of concepts such as acid–base definitions, indicator color changes during titration, ionization processes at the particle level, balanced chemical equations, and stoichiometric relationships in pH calculations (Widarti et al., 2023). Therefore, chemistry learning, particularly in analytical chemistry, should focus on strengthening understanding through various levels of representation so that students can grasp chemical concepts comprehensively (Apriani et al., 2021). The implementation of levels of representation in campus learning is crucial, especially in complex chemistry materials, to help students grasp concepts comprehensively (Susilaningsih et al., 2020).

Assessment is a crucial component of education, systematically and continuously measuring students' abilities (Köksa & Ulum, 2018). The quality of education can be assessed by the final learning outcomes, which necessitate the use of quality measurement tools (Destiana et al., 2020). In the context of higher-order thinking skills, particularly analytical and logical thinking, assessment should be designed to capture how students analyze information, relate variables, and construct coherent explanations rather than merely produce correct final answers. Transformative assessment is intended not only to measure students' performance but also to reveal and stimulate changes in their reasoning structures through reflective, multi-level cognitive engagement. In chemistry education, this requires assessment tasks that challenge students to connect ideas across different levels of representation (Muin & Kurniati, 2022). To address this need, essay-based assessment integrated with multiple representations offers a promising solution (Wachidah et al., 2021; Rinta & Fitriza, 2022). Multi-representational essay items require students to interpret phenomena across definitional, macroscopic, submicroscopic, symbolic, and stoichiometric levels, thereby enabling the measurement of analytical thinking through data classification, information organization, and relationship identification, as well as logical thinking through systematic reasoning and scientific argumentation (Gunbatar et al., 2025).

Holistic learning in this study refers to an integrated instructional approach that connects multiple levels of chemical representation within a coherent cognitive framework. Rather than treating concepts as isolated units, it emphasizes alignment among learning objectives, representational tasks, and assessment criteria to capture students' comprehensive understanding (Elm, 2021). Drawing on the concept of holistic evaluation, which underscores the importance of evaluating systems across diverse scenarios and criteria, holistic learning in chemistry similarly involves identifying key representational contexts and cognitive indicators that reflect students' analytical and logical thinking (Liang et al., 2023). In this study, holistic learning serves as the pedagogical foundation for designing multi-representational assessment tasks that not only measure conceptual accuracy but also reveal the coherence, depth, and consistency of students' reasoning across representational transitions, thereby supporting transformative and reflective learning processes (Gil et al., 2023).

However, most existing studies have focused on the use of multiple representations as instructional strategies, while relatively limited attention has been given to their integration within assessment frameworks, particularly for measuring students' logical and analytical thinking skills in chemistry. Many assessment instruments remain outcome-oriented and evaluate performance at a single representational level, providing limited diagnostic insight into students' reasoning processes across multiple representations.

In educational measurement, raw scores are obtained by summing correct responses; however, they are ordinal, so score differences do not necessarily reflect equal differences in ability and are strongly influenced by sample characteristics and the distribution of item difficulty. In contrast, the Rasch model transforms ordinal scores into interval-level logit measures and separates person ability from item difficulty, producing estimates that are more objective and sample-independent (Netchinda & Duangpummet, 2025). Descriptive statistical parameters derived from Rasch modeling, such as the mean, median, and standard deviation, are substantially more stable and less biased than those based on raw scores, thereby supporting the use of the Rasch model for precise and reliable assessment of students' cognitive abilities (Susongko, 2021).

Several previous studies have explored assessment and learning strategies related to higher-order thinking skills in science and chemistry; however, their focus and scope remain limited when viewed from a multi-representational and diagnostic assessment perspective. Rokhim et al. (2023), for instance, developed a structured problem-solving assessment based on the IDEAL model to measure students' problem-solving abilities in acid-base topics, including titration. Although the instrument effectively evaluated procedural problem-solving skills, it did not explicitly integrate multiple levels of chemical representation nor distinguish logical and analytical thinking skills as separate cognitive constructs. Furthermore, the assessment emphasized problem-solving performance rather than providing a diagnostic portrayal of students' reasoning processes across representations (Rokhim et al., 2023). Studies grounded in multiple representations have predominantly focused on conceptual understanding rather than on assessment of higher-order thinking. Widarti et al. (2020) investigated students' understanding of acid-base concepts using multi-representational questions and found that students' conceptual understanding, particularly at the submicroscopic level, was generally low. Their findings highlighted the importance of multiple representations for meaningful understanding; however, the assessment was mainly descriptive and did not function as a diagnostic or transformative instrument to measure students' logical and analytical thinking skills (Widarti et al., 2020).

Research in broader science contexts has demonstrated the importance of analytical thinking as a key 21st-century competency. Setiaji et al. (2025) showed that problem-based learning supported by digital modules effectively enhanced students' analytical thinking skills and scientific attitudes in physics learning. Nevertheless, this study focused on instructional intervention rather than assessment development, and it did not address chemistry-specific representations or the psychometric robustness of the assessment instruments used (Setiaji et al., 2025).

Similarly, Al-Ajmi and Ambusaidi (2022) reported that students' scientific argumentation skills in chemistry were strongly influenced by their logical thinking abilities, indicating a close relationship between logical reasoning and higher-level scientific performance. While this study confirmed the critical role of logical thinking in chemistry learning, it did not propose an assessment model capable of systematically measuring logical and analytical thinking skills across different representational levels.

Taken together, these studies indicate that although problem-solving ability, multiple representations, analytical thinking, and logical reasoning have been widely investigated, they are often examined in isolation. Existing research typically positions multiple representations as instructional strategies rather than as an integrated diagnostic assessment framework, and assessment instruments rarely differentiate logical and analytical thinking as distinct yet related cognitive constructs. Moreover, few studies combine multi-representational assessment with Rasch modeling to generate interval-level measures and person–item mapping for precise cognitive profiling.

Consequently, this study offers a novel contribution in three key aspects. Conceptually, it reconceptualizes assessment as a transformative and reflective process that integrates multiple chemical representations to stimulate and diagnose changes in students' reasoning structures. Methodologically, it combines multi-representational essay assessment with Rasch measurement analysis, enabling objective calibration of item difficulty and person ability, as well as Wright Map-based cognitive profiling. Practically, it provides a diagnostic framework that allows instructors to identify students' logical and analytical thinking patterns across definitional, macroscopic, submicroscopic, symbolic, and stoichiometric representations.

Therefore, a clear research gap exists in developing valid and reliable transformative assessment models that systematically measure students' logical and analytical thinking skills across multiple levels of chemical representation. Accordingly, this study aims to develop and validate a transformative assessment model integrating multiple representations within holistic learning to measure students' logical and analytical thinking skills in chemistry.

METHODS

The research process implemented in this study is presented in Figure 1 (Widyastuti & Susiana, 2019).

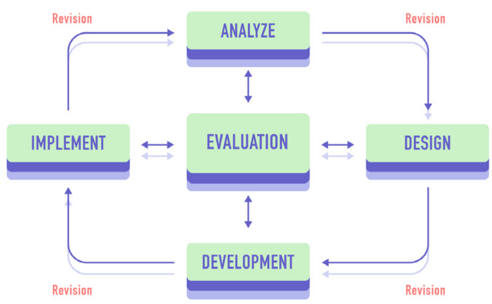

Figure 1. Research Workflow

The first stage is Analysis, which begins with preliminary observations through student interviews. These interviews aim to obtain an overview of the learning process that underlies the need for instrument development. Following the interviews, the researcher defines the problem, identifies its source, analyzes the needs, and determines appropriate solutions to address it.

The second stage is Design, which includes preparing an initial draft of the test instrument (the main product of the research), an interview questionnaire, and an instrument validation questionnaire. The initial draft of the developed test instrument consists of 6 questions, each based on multiple representations, covering definitions, macroscopic, submicroscopic, symbolic, and stoichiometric.

The third stage is Development, which is the testing phase to ensure the developed product is feasible for implementation. This process involves two steps: (1) expert appraisal/validation and (2) development testing of the test instrument.

The fourth stage is implementation, in which the test instrument was piloted on 100 second-semester students in the Basic Analytical Chemistry course. Most students were aged 18–20 and had taken a first-semester general chemistry course, indicating a similar knowledge base. Students came from a range of academic profiles (low, medium, and high) and thus represented a range of achievement levels, allowing for performance assessment at various levels. All students were informed about their participation in the study and, before completing the diagnostic test, were introduced to the research objectives and asked to provide consent for the use of their anonymous responses.

The fifth stage, evaluation, was conducted at each stage of development to identify product strengths and weaknesses, allowing for continuous improvement until an optimal instrument was achieved.

The research instrument consists of six essay items, each based on the type of acid–base titration: strong acid–strong base, strong base–strong acid, strong acid–weak base, weak base–strong acid, weak acid–strong base, and strong base–weak acid. Each item comprised five subquestions corresponding to the representational levels: definitional, macroscopic, submicroscopic, symbolic, and stoichiometric. In total, the instrument included 30 sub-questions. Definitions in chemistry, for example, can be represented through verbal or written descriptions. Macroscopic representations involve directly observable phenomena, such as experiments or everyday events. Microscopic representations encompass phenomena invisible to the naked eye and are used to explain macroscopic phenomena. Symbolic representations include chemical symbols for elements, molecules, and chemical reactions. Stoichiometry is used to represent basic concepts, such as coefficients in chemical equations, mole and mass calculations, and conversions between various units of matter. This approach was chosen to encourage students to think more deeply and understand concepts holistically. Each question is worth 5 points if answered correctly. Two expert chemistry lecturers have validated this instrument to ensure its suitability for research use.

This study was conducted with second-semester chemistry students enrolled in the Basic Analytical Chemistry course at Universitas Negeri Semarang. The diagnostic test was administered in a lecturer-led classroom setting, with students given 90 minutes to complete six essay items comprising 30 sub-questions related to their understanding of acid–base titration chemistry. Data were collected and analyzed thematically. It should be emphasized that the study involved students from a single institution, which is appropriate for development-oriented research. However, this contextual limitation may restrict the generalizability of the findings to other institutions, different chemistry topics, or other educational levels. Therefore, although the instrument demonstrated strong validity and reliability within this setting, further research is needed to examine its applicability across more diverse educational contexts.

Content validity is demonstrated through expert judgment. Content validity assessment criteria encompass three main aspects: (1) assessment of content suitability, (2) construct suitability, and (3) appropriate and communicative use of language.

Item analysis using the Rasch model is performed with the Winsteps application, including: item reliability (Summary Statistics), validity (Item Fit Order), Wright Map, item difficulty level (Item Measure), person measure, person fit order, and scalogram. Conceptual understanding abilities were mapped to logical and analytical thinking abilities using predetermined indicators.

RESULTS AND DISCUSSION

The analysis of the research results focused on evaluating a transformative assessment instrument that integrates multiple representations within holistic chemistry learning. Unlike conventional chemistry assessments that predominantly emphasize final answers or single-representation performance, the developed instrument was explicitly designed to diagnose students' logical and analytical thinking processes across five levels of chemical representation: definitional, macroscopic, submicroscopic, symbolic, and stoichiometric. This integrative structure constitutes a key novelty of the study, as it enables the assessment not only of conceptual accuracy but also of the coherence and consistency of students' reasoning across interconnected representational levels.

The application of the Rasch Model represents a significant contribution to chemistry education research (Talens et al., 2025). While previous studies have employed classical test theory to evaluate assessment instruments, this study utilized Rasch analysis to obtain objective, sample-independent measures of item difficulty and student ability. According to Bond and Fox (2013), the Rasch approach allows researchers to convert ordinal scores into interval-level measures, enabling a more precise examination of item characteristics and the distribution of learner abilities. The analysis covered item reliability, person reliability, item and person fit indices, item difficulty level, person ability measure, Wright map distribution, and Scalogram interpretation. These indicators collectively serve to verify whether the assessment system aligns with the principles of construct validity and measurement stability. Another important contribution of this study lies in the explicit separation and simultaneous measurement of logical thinking and analytical thinking as distinct but related cognitive constructs. Most existing chemistry assessments integrate higher-order thinking skills in a general manner, without differentiating specific thinking dimensions. The results of this study demonstrate that the transformative assessment instrument can discriminate between students' performance on these two thinking skills, thereby providing more nuanced information for instructional decision-making and reflective learning practices (Bond & Fox, 2013).

The research results include feasibility testing, instrument quality analysis, and the profiling of students' logical and analytical thinking abilities using the developed transformative assessment instrument. The testing process involved three stages: a small-scale trial, a large-scale trial, and implementation. The small-scale trial involved 15 undergraduate chemistry students who had previously completed the Basic Analytical Chemistry course. The large-scale trial was conducted with 60 chemistry students who also met the same prerequisite. Finally, the implementation stage involved 100 second-semester chemistry students (divided into three groups) enrolled in the Basic Analytical Chemistry course. These participant criteria ensured that all respondents had the necessary foundational knowledge to meaningfully engage with the instrument's representational content.

The product tested consisted of six multiples-representation essay questions, each comprising four questions addressing the definitional, macroscopic, submicroscopic, symbolic, and stoichiometric aspects. A total of 30 questions were analyzed. Validation aimed to determine whether the developed instrument was feasible and capable of measuring logical and analytical thinking abilities. A summary of the expert validation results is shown in Table 1.

Table 1. Recapitulation of Expert Validation of Test Instruments

Validator	**Score**	**Category**
1	76	very feasible
2	75	very feasible
3	77	very feasible

Validation results showed that experts gave an average score of 76 out of 80, indicating that the instrument is highly valid. Expert comments and suggestions were used to refine the test instrument. The validity of the test items can be demonstrated using the Content Validity Ratio (CVR).

A test instrument is considered reliable if participants tested at different times consistently show a consistent ranking position or are in the same group (Widoyoko, 2014). In the Rasch model analysis, reliability is demonstrated through three main indicators: item reliability, student reliability, and Cronbach's Alpha (KR-20) value. Participant reliability reflects the consistency of answers, item reliability indicates the quality of the questions, while Cronbach's Alpha (KR-20) describes the interaction between students and the items. The results of the reliability calculations for each testing stage are shown in Table 2.

Table 2. Reliability Summary

Test Phase	Item Reliability	Person Reliability	Cronbach Alpha (KR-20)
Small Scale	0.70–0.72 (Moderate)	0.88–0.88 (Good)	0.91 (Excellent)
Large Scale	0.87–0.88 (Good)	0.90–0.90 (Good)	0.95 (Excellent)
Implementation	0.89–0.90 (Good)	0.95–0.96 (Good)	0.92 (Excellent)

The results indicate that the assessment instrument achieved a reliability coefficient above the minimum threshold ($r \geq 0.70$), signifying high consistency across trials (Arikunto, 2015). This consistency suggests that both the items and participants' responses were stable and reproducible, implying that the instrument can effectively differentiate student performance across dimensions of logical and analytical thinking.

Analysis using the Rasch Model showed that the developed assessment instrument possessed excellent validity and reliability. Based on the implementation results, Item Reliability ranged from 0.89 to 0.90, and Person Reliability ranged from 0.95 to

0.96, with a Cronbach's Alpha (KR-20) of 0.90, categorized as "very good." These values indicate that each test item consistently measured the constructs of logical and analytical thinking. The Wright Map further revealed a balanced distribution between item difficulty and student ability. This balance is critical because it demonstrates that the instrument was neither too easy nor too difficult for the target population, enabling accurate discrimination across varying levels of students' thinking skills (Nurussaniah et al., 2025). As a result, students with higher logical and analytical abilities were consistently positioned above more challenging items, while students with lower abilities clustered around easier items. This pattern confirms that the instrument has strong discriminating power and supports meaningful interpretation of students' cognitive profiles. This finding aligns with Bond & Fox (2013), who emphasized that Rasch Measurement Theory provides more stable reliability because it accounts for the interaction between individual responses and item characteristics. All items were found to be fit after revision, demonstrating strong construct alignment and conceptual representation.

The item fit analysis showed that all questions met the validity criteria during the implementation stage. Initially, one item (1a1) was identified as a misfit; however, after revision and re-evaluation, the item demonstrated acceptable fit statistics. This result indicates that all items functioned as intended and were interpreted by students in accordance with the underlying constructs being measured.

Based on the Item Measure, the instrument consisted of 9 very difficult, 27 difficult, 20 moderate, and 10 easy items. This proportional distribution of item difficulty is essential because it allows the assessment to capture students' thinking skills across a wide ability continuum. Easier items primarily elicit basic logical reasoning and foundational analytical processes, while more difficult items require higher-order analytical integration across multiple chemical representations. As a result, the instrument can distinguish not only between high- and low-performing students but also between different levels of cognitive complexity in logical and analytical thinking.

This balance between high-order and low-order items reflects the principle of construct representation, ensuring that each aspect of the targeted thinking skills is adequately measured. According to Linacre, variation in item difficulty within Rasch measurement is necessary to ensure that each ability level is psychometrically represented and meaningfully interpreted (Linacre, 2009). In this study, this variation confirms that the assessment model successfully accommodates students' ability heterogeneity and avoids ceiling or floor effects that could obscure diagnostic interpretation.

The level of item suitability is used as an indicator of each item's validity in the test instrument. Based on the analysis results, almost all items are in the fit category according to the Mean Square Outfit (MNSQ), Z-Standard (ZSTD), and Point Measure Correlation (Pt Measure Corr) criteria. An item is declared invalid (misfit) if it does not meet these three criteria. However, if the MNSQ and Pt Measure Corr values do not match but the ZSTD value remains within the specified range, the item is still considered valid and retained. Invalid items are revised and retested until they meet the validity criteria.

The results of the item fit analysis at each testing stage are shown in Table 3.

Table 3. Level of Suitability of Question Items

Test Stages	Valid	Invalid
Small-Scale Trial	2a1, 2b1, 3a1, 3b1, 4a1, 4b1, 5a1, 5b1, 5c1, 5d1, 1a2, 2a2, 2b2, 3a2, 3b2, 4a2, 4b2, 5a2, 5b2, 5c2, 5d2, 1a3, 2a3, 2b3, 3a3, 3b3, 4a3, 4b3, 5a3, 5b3, 5c3, 5d3, 1a4, 2a4, 2b4, 3a4, 3b4, 4a4, 4b4, 5a4, 5b4, 5c4, 5d4, 1a5, 2a5, 2b5, 3a5, 3b5, 4a5, 4b5, 5a5, 5b5, 5c5, 5d5, 1a6, 2a6, 2b6, 3a6, 3b6, 4a6, 4b6, 5a6, 5b6, 5c6, 5d6	1a1
Large-Scale Trial	1a1, 2a1, 2b1, 3a1, 3b1, 4a1, 4b1, 5a1, 5b1, 5c1, 5d1, 1a2, 2a2, 2b2, 3a2, 3b2, 4a2, 4b2, 5a2, 5b2, 5c2, 5d2, 1a3, 2a3, 2b3, 3a3, 3b3, 4a3, 4b3, 5a3, 5b3, 5d3, 1a4, 2a4, 2b4, 3a4, 3b4, 4a4, 4b4, 5a4, 5b4, 5c4, 5d4, 1a5, 2a5, 2b5, 3a5, 3b5, 4a5, 4b5, 5a5, 5b5, 5c5, 5d5, 1a6, 2a6, 2b6, 3a6, 3b6, 4a6, 4b6, 5a6, 5b6, 5c6, 5d6	5c3
Implementation	1a1, 2a1, 2b1, 3a1, 3b1, 4a1, 4b1, 5a1, 5b1, 5c1, 5d1, 1a2, 2a2, 2b2, 3a2, 3b2, 4a2, 4b2, 5a2, 5b2, 5c2, 5d2, 1a3, 2a3, 2b3, 3a3, 3b3, 4a3, 4b3, 5a3, 5b3, 5c3, 5d3, 1a4, 2a4, 2b4, 3a4, 3b4, 4a4, 4b4, 5a4, 5b4, 5c4, 5d4, 1a5, 2a5, 2b5, 3a5, 3b5, 4a5, 4b5, 5a5, 5b5, 5c5, 5d5, 1a6, 2a6, 2b6, 3a6, 3b6, 4a6, 4b6, 5a6, 5b6, 5c6, 5d6	-

Note: Item codes consist of three components: the first number represents the sub-question, the letter indicates the representation level (a = definitional, b = macroscopic, c = submicroscopic, d = symbolic, e = stoichiometric), and the last number denotes the item number. For example, code 1a2 refers to the first sub-question at the definitional representation level in item 2.

During the small-scale trial, only one item (1a1), which corresponds to a definitional-level subquestion in item 1, did not meet the fit criteria. This result indicates that initial misinterpretation occurred at the definitional representation level. After revision, the same item met the fit criteria in the large-scale trial, demonstrating that the revision effectively improved item clarity and construct alignment.

In the large-scale trial, one submicroscopic-level item (5c3) was identified as a misfit, suggesting that students had trouble interpreting particle-level representations in the context of item 5. This finding is consistent with previous studies that have reported that submicroscopic representations pose greater cognitive challenges for students. Following refinement, all items met the fit criteria during implementation.

Overall, the progression from isolated misfit items in the early testing stages to complete item fit at the implementation stage indicates that the transformative assessment instrument underwent systematic improvement and refinement. This result confirms that the instrument validly measures students' logical and analytical thinking skills across definitional, macroscopic, sub-microscopic, symbolic, and stoichiometric representations.

Person-fit analysis in the Rasch model is used to assess the match between students' abilities and their response patterns when answering questions. Response discrepancies can be caused by various factors, such as lucky guesses, careless errors, or improper cooperation (cheating). According to Sumintono and Widhiarso (2015), inappropriate responses appear at the top of the person-fit table.

The results of the person-fit analysis at all stages showed that all students had response patterns appropriate to their ability level (fit) (Sumintono & Widhiarso, 2015). Thus, the test data can be considered valid and can be further interpreted.

Item difficulty analysis was used to classify questions into easy, medium, or difficult categories. Higher logit values indicate greater difficulty. In the implementation phase, 9 questions were very difficult, 27 were difficult, 20 were medium, and 10 were easy. This distribution indicates that the questions were distributed proportionally across the participants' ability levels, enabling the instrument to fairly and accurately measure a broad range of logical and analytical thinking.

Person measure analysis was used to determine students' problem-solving abilities based on logit scores. During the implementation phase, 27 students were categorized as very high, 42 as high, and 31 as average. This demonstrates a consistent variation in ability and the instrument's effectiveness in distinguishing logical and analytical thinking levels among participants.

The Wright Map shows the relationship between the distribution of student abilities and the difficulty level of test items on a logit scale. According to Sumintono and Widhiarso (2015), items and participants within 2 standard deviations are considered to have a normal distribution. In implementation, the most difficult questions were 1b, 1c, and 5c, while the easiest was 4a. The highest-ability students were 13P, 23L, 25L, and 43P, while the lowest-ability students were 32L. This demonstrates a balance between ability and item difficulty, indicating that the developed assessment has high targeting accuracy.

Table 4. Summary of Item and Person Measures Based on Testing Stages

Testing Stage	Item Measure	Person Measure
Small-Scale Trial	3b3, 3b4 (difficult) 5a1, 5a2, 5a3, 5a4, 5a5, 5a6, 5b1, 5b2, 5b3, 5b4, 5b5, 5b6 (easy)	04P (high ability) 13L (low ability)
Large-Scale Trial	3b3, 3b2 (difficult) 5a1, 5a2, 5a3, 5a4, 5a5, 5a6, 5b1, 5b2, 5b3, 5b4, 5b5, 5b6 (easy)	17P, 22P, 28P, 45P (high ability) 39P (low ability)
Implementation	3b4 (difficult) 3a2, 3a5 (easy)	006P, 095P (high ability) 090P (low ability)

The scalogram was employed to examine the relationship between students' ability levels and the difficulty level of test items. This analytical approach indicates that individual performance assessment should not rely solely on raw scores; it should also consider the relative difficulty of each item. When identical logit values are observed among several participants, further analysis is necessary to determine whether there are indications of academic dishonesty, such as collaboration during testing. If a student's ability logit value is higher than the logit value of the items answered incorrectly, it may indicate a lack of carefulness or attentional error during the test.

Although the student possesses strong cognitive ability, factors such as accuracy and attention to detail substantially influence the outcomes. Conversely, if two or more participants exhibit identical response patterns across all items, this may suggest potential collaboration (cheating) during the assessment process. Such conditions underscore the importance of strict supervision during test administration and the need for follow-up verification to ensure that assessment results accurately reflect individual abilities, free from external interference. In contrast, if a student's ability logit is lower than the logit of correctly answered items, this may reflect random guessing or "lucky guess" behavior.

Differential Item Functioning (DIF) analysis was conducted to determine whether any items exhibited differential difficulty across respondent groups. The DIF analysis results from the Winsteps output are presented in Figure 2, which illustrates the stability of item functions across respondent groups and ensures that each item has equal levels of fairness and consistency in measuring logical and analytical thinking skills.

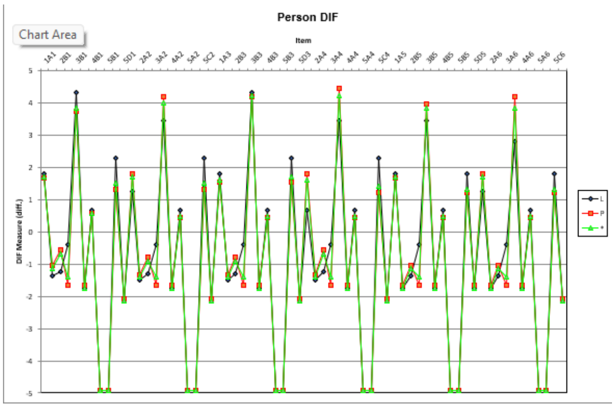

Figure 2. Item DIF

Overall, Figure 2 shows that the difficulty estimates for most items across respondent groups follow a similar pattern and remain closely aligned. The distance between group curves for the same item is relatively small and does not exceed the commonly accepted DIF threshold (±0.5 logit). This indicates that most items function equivalently across groups and do not favor or disadvantage any group.

Several items exhibit peaks at higher logit values, indicating greater difficulty. However, these peaks are consistently observed across all respondent groups, suggesting that the increased difficulty is inherent to the item content rather than being influenced by group membership. Likewise, items with lower logit values consistently appear easier for all groups, demonstrating stable item behavior.

The absence of substantial divergence between group lines across items indicates that none of the items demonstrates meaningful DIF. This finding confirms that the transformative assessment instrument maintains fairness and measurement invariance across respondent groups. Consequently, differences in student performance can be attributed to actual variations in logical and analytical thinking abilities rather than bias introduced by item characteristics.

From a psychometric perspective, these results support the instrument's robustness, as items consistently measure the intended constructs across diverse respondents. The DIF analysis, therefore, strengthens the validity claim for the developed transformative assessment and confirms its suitability for diagnostic use in chemistry education contexts.

This study developed a transformative assessment model based on multiple representations in basic analytical chemistry learning, designed to measure students' logical and analytical thinking skills in depth. This approach was designed not only to evaluate the final learning outcomes but also to trace the thinking processes involved at each level of chemical representation. Through the application of five levels of representation definition, macroscopic, microscopic, symbolic, and stoichiometric, students were guided to understand the interconnections among chemical concepts comprehensively. This model reinforces the perspective that conceptual understanding in chemistry develops only when learners connect empirical phenomena with the theoretical models they construct (Taber, 2019).

During implementation, students demonstrated increased cognitive and reflective activity, particularly when interpreting titration results based on indicator changes and pH calculations. Thus, the developed transformative assessment proved effective in shifting the learning orientation from product-based to reflective scientific thinking. This approach is relevant to the demands of the 21st century, which require comprehensive mastery of high-order thinking skills.

The distribution of indicators of logical and analytical thinking abilities, which describe the involvement of each aspect in each question item, is presented in Table 5.

Table 5. Distribution of Indicators

Aspects	Logical Thinking Indicator	Analytical Thinking Indicator
Definition	Orderly thinking: systematically presenting information about the concepts of strong acids and strong bases.	Grouping data on the concepts of acids and bases based on the degree of ionization.
Macroscopic	Orderly thinking: analyzing indicator color changes in titration.	Organizing color observations into stages of the titration process.
Microscopic	Argumenting: rationally explaining the neutralization reaction of H^+ and OH^-.	Finding the relationship between the ionization of substances and the reaction products (water and salt).
Symbolic	Argumenting: justifying the selection of an appropriate chemical reaction equation.	Organizing symbols and chemical reactions into a complete equation.
Stoichiometry	Drawing conclusions: determining the final product (pH) from calculations.	Finding the relationship between moles, volume, and pH at each stage of the titration.

Assessment of this instrument was conducted by giving a maximum score of 5 for each aspect of multiple representation. Based on the results of the implementation test with 100 respondents, the average score for each multiple representation aspect was obtained, along with a description of the results for each question item, as presented in Table 6. These results provide an overview of students' achievement in logical and analytical thinking when understanding concepts through various forms of chemical representation.

Table 6. Assessment Results

Aspects	1	2	3	4	5	6	Average
Definition	4,34	4,5	4,67	4,49	4,51	4,27	4,46
Macroscopic	4,25	4,11	3,87	3,94	4,16	4,32	4,11
Microscopic	4,35	4,34	4,06	3,58	4,33	4,30	4,16
Symbolic	4,35	4,43	3,97	4,24	4,43	4,37	4,30
Stoichiometry	4,35	4,43	3,67	4,34	4,42	4,34	4,27

To provide a more comprehensive understanding of the assessment results, a comparison of the average scores for each aspect of multiple representation was carried out as shown in Figure 3.

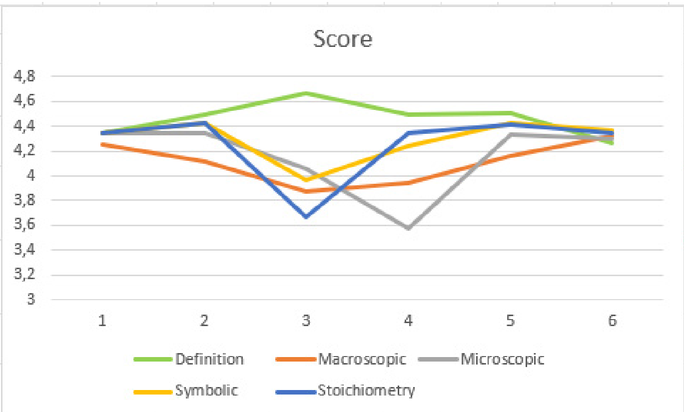

Figure 3. Assessment Results

Overall, the definition aspect achieved the highest score, with an average of 4.46. This was because students found it relatively easy to understand and remember basic concepts through verbal and symbolic explanations, which they had frequently encountered in previous lessons. The ability to explain the definitions of acids and bases, as well as the concept of ionization, had become part of their strong and well-internalized conceptual knowledge.

Conversely, the macroscopic and microscopic aspects received the lowest scores. This was because these two aspects required students to relate real-world phenomena to processes that are not directly visible at the particle level. Students often had difficulty visualizing the chemical changes that occur during reactions, especially when having to relate them to experimental phenomena or laboratory observations.

A more detailed review of each item revealed that the definition aspect in item 3 (explaining the difference in ionization strength of HCl and NH_4OH) demonstrated the highest score, as students were able to clearly identify the difference between a strong acid and a weak base based on the concept of degree of ionization. Conversely, the lowest average score was found in the microscopic aspect of item 4 (describing the formation of NH_4Cl and water at the particle level), indicating that students still have difficulty representing neutralization reactions visually and conceptually at the particle level.

This finding indicates the need to strengthen multi-representation-based learning that integrates conceptual understanding, microscopic visualization, and analytical skills, enabling students to think more logically and analytically when understanding chemical phenomena holistically.

Analysis of logical thinking indicators revealed that students performed best in orderly thinking and scientific argumentation. At the symbolic and stoichiometric levels, students were able to compose reaction equations and determine pH values systematically and logically. However, at the microscopic level, some students still had difficulty explaining reaction mechanisms, indicating a need to reinforce microscopic concepts in learning. This finding is consistent with Johnstone's theory, which states that the main difficulty in learning chemistry lies in the transition among macroscopic, microscopic, and symbolic representations. The reflection-based assessment process effectively helped students recognize and gradually correct their own thinking errors (Johnstone, 1993). This result also supports Cooper et al., who asserted that representation-based reasoning enhances logical coherence in solving chemical problems (Cooper, 2023). Therefore, this assessment effectively measures and fosters systematic, logical thinking among students.

Students' analytical thinking ability was assessed through three main indicators: data classification, information organization, and identification of variable relationships. Based on the person measure in Rasch analysis, 27 students were categorized as very high, 42 as high, and 31 as moderate, showing a balanced ability distribution. Students in the high category generally demonstrated consistency in linking stoichiometric calculation results with macroscopic phenomena such as indicator color changes during titration. These findings are consistent with Facione's concept of analytical reasoning, which involves dissecting argument structures and evaluating evidence to reach rational conclusions (Facione, 2015). In this assessment, students were trained to identify relationships between acid-base theory and empirical phenomena observed in the laboratory. The multi-representation-based approach enabled them to build cohesive and logical mental models. This result aligns with Gilbert & Treagust (2009)who emphasized that integrating symbols, graphs, and text in learning enhances students' analytical thinking coherence.

The Person–Item Map analysis showed that the distribution of student abilities and item difficulty levels was ideally balanced. High-logit items (+1.5 to +2.5), such as 3b4 and 5c, required a deep analytical understanding of ionization concepts and pH calculations, while low-logit items (–1 to 0), such as 4a, measured basic abilities in identifying reaction types. This pattern illustrates that the assessment had high targeting accuracy, corresponding to students' actual zones of proximal development. This finding supports Vygotsky's theory that optimally challenging tasks promote the emergence of new cognitive abilities. In practice, this map can serve as a diagnostic tool for instructors to design learning interventions tailored to students' ability levels (Vygotsky, 1978).

The Wright Map analysis demonstrated that the distribution of item difficulty and student ability was precisely aligned, indicating high measurement precision and consistency with students' cognitive development zones. Items located at higher logit levels (e.g., 3b4 and 5c) required students to integrate multiple representations, particularly linking sub-microscopic explanations of ionization processes with symbolic representations and stoichiometric pH calculations. Only students with higher analytical ability were able to successfully respond to these items, indicating that they effectively discriminated higher-order reasoning skills. In contrast, items with lower logit values (e.g., 4a) assessed fundamental identification and definitional understanding of acid–base concepts and were accessible to most students, reflecting basic logical reasoning skills.

Importantly, the Wright Map revealed that a small proportion of students were located below several medium- and high-difficulty items, indicating specific gaps in representational reasoning, especially at the sub-microscopic and stoichiometric levels. This finding provides empirical evidence that difficulties in acid–base titration learning are not merely conceptual but are closely related to students' limited ability to translate between different levels of chemical representation.

Beyond demonstrating psychometric quality, the Wright Map functions as a diagnostic and reflective tool. It allows instructors to identify which representational levels pose the greatest challenges for students and to design targeted instructional interventions accordingly. This condition reflects Vygotsky's (1978) notion of the zone of proximal development, in which learning is optimized when tasks slightly exceed students' current competence. Furthermore, by making students' cognitive positions visible, the assessment supports Mezirow's transformative learning perspective, in which reflection on mismatches between existing understanding and task demands promotes deeper conceptual change (Mezirow, 2018). In conclusion, the transformative assessment developed in this study transcends traditional testing by integrating psychometric precision with cognitive transformation, thereby fostering the development of logical and analytical thinking as essential competencies for 21st-century scientific literacy.

CONCLUSION

This study developed a novel transformative assessment instrument that integrates multiple chemical representations within a holistic learning framework to measure students' logical and analytical thinking skills. Unlike conventional outcome-oriented assessments, this instrument functions as a diagnostic and reflective tool, enabling the examination of students' reasoning processes across definitional, macroscopic, sub-microscopic, symbolic, and stoichiometric representations. This approach contributes to chemistry education research by repositioning assessment as a means of cognitive transformation rather than merely as a means of learning evaluation. These findings indicate that the instrument effectively differentiates students' analytical thinking across representational levels. From a pedagogical perspective, the transformative assessment provides diagnostic information that can be used to identify misconceptions and guide targeted instructional interventions, particularly in abstract chemistry topics. The Wright Map helps instructors align learning activities with students' zones of proximal development, thereby promoting reflective, meaningful learning. This study is limited to a single chemistry topic and institutional context, and it focuses primarily on logical and analytical thinking skills. Future research should apply this assessment model to broader chemistry topics, diverse educational settings, and additional higher-order thinking dimensions. Overall, the findings support the potential of transformative, multi-representational assessment as a robust approach for advancing assessment practices in chemistry education.

REFERENCES

Al-Ajmi, B., & Abdullah, A. (2019). The level of scientific argumentation skills in chemistry subject among grade 11 students: The role of logical thinking. *Science Education International*, *33*(1), 66–74.

Al-Ajmi, B., & Ambusaidi, A. (2022). The Level of Scientific Argumentation Skills in Chemistry Subject among Grade 11th Students: The Role of Logical Thinking. *Science Education International*, *33*(1), 66-74.

Al Haq, F. N. H., & Raicudu, M. I. R. (2023). Pemahaman Konsep Peserta Didik Kelas VII Pada Materi Segiempat. *Prosiding Seminar Nasional Pendidikan Matematika, Universitas Mulawarman*, *3*, 82–89.

Anggraini, D., & Irawan, E. (2021). Analisis Kemampuan Berpikir Logis Siswa Kelas VII pada Tema Pencemaran Lingkungan. *Jurnal Tadris IPA Indonesia*, *1*(2), 228–238.

Apriani, R., Harun, A. I., Erlina, E., Sahputra, R., & Ulfah, M. (2021). Pengembangan modul berbasis multipel representasi dengan bantuan teknologi augmented reality untuk membantu siswa memahami konsep ikatan kimia. *Jurnal IPA & Pembelajaran IPA*, *5*(4), 305-330.

Arikunto, S. (2015). *Dasar-Dasar Evaluasi Pendidikan*. Bumi Aksara.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Cooper, G. (2023). Examining Science Education in ChatGPT : An Exploratory Study of Generative Artificial Intelligence. *Journal of Science Education and Technology*, *32*(3), 444–452.

Destiana, D., Suchyadi, Y., & Anjaswuri, F. (2020). Pengembangan instrumen penilaian untuk meningkatkan kualitas pembelajaran produktif di sekolah dasar. *Jurnal Pendidikan Dan Pengajaran Guru Sekolah Dasar (JPPGuseda)*, *03*(September), 119–123.

Dewi, C. A., Rahayu, S., Muntholib, M., & Parlan, P. (2025, July). Students' chemical literacy through ORDIEC strategies (orientation, review, designing, investigation, evaluation and conclusion): Cases of acid-base topics. In *AIP Conference Proceedings* (Vol. 3206, No. 1, p. 070011). AIP Publishing LLC.

Elm, J. (2021). Toward a holistic understanding of the formation and growth of atmospheric molecular clusters: A quantum machine learning perspective. *The Journal of Physical Chemistry A*, *125*(4), 895–902.

Facione, P. A. (1998). *Critical Thinking : What It Is and Why It Counts*. CA: Measured Reasons and the California Academic Press.

Fauziah, F., Suyanto, E., Wahyudi, I., & Herlina, K. (2022). Effect of Learning Model Application of Guided Laboratory Inquiry to Logical Thinking Ability of Students on Heat to Electricity Conversion Material. *Jurnal Pendidikan Matematika Dan IPA*, *13*(1), 114–126.

Fitriani, F., Fadly, W., & Faizah, U. N. (2021). Analisis keterampilan berpikir analitis siswa pada tema pewarisan sifat. *Jurnal Tadris IPA Indonesia*, *1*(1), 55-67.

Furqan, M., Karyanto, P., & Salma, S. (2016). The Implementation of E-Module Based on Problem-Based Learning to Improve the Analytical Thinking Abilities and Reduce Misconception of X MIA 1 of SMA Negeri 1 Banyudono Academic Year of 2014/2015. *Proceeding Biology Education Conference: Biology, Science, Enviromental, and Learning, 12*(1), 410–414.

Gil, V. S., Bran, A. M., Franke, M., Schlama, R., Luterbacher, J. S., & Schwaller, P. (2023). Holistic chemical evaluation reveals pitfalls in reaction prediction models. *Philippe.Schwaller@epfl.Ch*, 1–17.

Gilbert, J. K., & Treagust, D. F. (2009). *Models and Modeling in Science Education*. Springer.

Gunbatar, S. A., Sirin, G. T., Ilkyaz, O. C., & Mutlu, Y. (2025). Exploring the artificial intelligence interaction profiles of participants with different levels of teaching experience for lesson planning in the context of acids and bases. *Chemistry Education Research and Practice*, *26*(4), 977–995.

Herunata, H., Rosyida, I., Sulistina, O., & Wijaya, H. W. (2021, March). Correlational analysis of conceptual understanding, chemical representation, and representational competence on chemistry equilibrium. In *AIP Conference Proceedings* (Vol. 2330, No. 1, p. 020016). AIP Publishing LLC.

Johnstone, A. H. (1993). The development of chemistry teaching: A changing response to changing demand. *Journal of Chemical Education*, *70*(9), 701–705.

Köksa, D., & Ulum, Ö. G. (2018). Language assessment through Bloom's Taxonomy. *Journal of Language and Linguistic Studies*, *14*(2), 76–88.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., New-, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-navas, D., Hudson, D. A., … Zhang, T. (2023). *Holistic Evaluation of Language Models*. Transactions on Machine Learning Research.

Linacre, J. M. (2009). *A User's guide to Winsteps-ministep: Rasch-model computer programs. Program Manual 3.68. 0*. IL.

Mezirow, J. (2018). *Transformative learning theory. In K. Illeris (Ed.), Contemporary theories of learning*. Routledge.

Muin, W. A., & Kurniati, L. (2022). Improving critical thinking skills in writing essay. *International Journal of Education, Information Technology and Others (IJEIT)*, *5*(3), 102–106.

Netchinda, S., & Duangpummet, P. (2025, July). Development of a Measurement Instrument to Assess Students' Conceptual Understanding in Three Levels of Chemical Representation of Galvanic Cells by Using the Rasch Model. In *2025 10th International STEM Education Conference (iSTEM-Ed)* (pp. 1-6). IEEE.

Nurussaniah, N., Setyosari, P., Kuswandi, D., & Ulfa, S. (2025). Psychometric validation of an analytical skills test in physics using the Rasch model. *Journal of Baltic Science Education*, *24*(3), 522–537.

OECD. (2023). *PISA 2022 Results (Volume I): The State of Learning and Equity in Education: Vol. I*. OECD Publishing.

Park, C., Kim, S., Choi, H., & Paik, S. (2019). Exploring Learning Progression of Logical Thinking in Acid and Base Chemical Reactions. *Journal of the Korean Chemical Society*, *63*(5), 376–386.

Prastiwi, M. N. B., & Laksono, E. W. (2018). The ability of analytical thinking and chemistry literacy in high school students learning The ability of analytical thinking and chemistry literacy in high school students learning. *Journal of Physics: Conference Series*, 1–8.

Rahmawati, Y., Zulhipri, Hartanto, O., Falani, I., & Iriyadi, D. (2022). Students' conceptual understanding in chemistry learning using PhET interactive simulations. *Journal of Technology and Science Education*, *12*(2), 303–326.

Rakhmawan, A., Firman, H., Redjeki, S., & Mulyani, S. (2018). Contribution of Logical Thinking Ability to Student s ' Achievement in Three Level of Representations in Chemical Dynamic Materials. *Jurnal Penelitian Dan Pembelajaran IPA*, *4*(2), 116–126.

Rinta, C., & Fitriza, Z. (2022). Pengembangan Tes Diagnostik Esai Terstruktur Kimia Materi Titrasi Asam Basa Sebagai Instrumen Analisis Hasil

Belajar Siswa. *Edukimia*, *4*(2), 84–89.

Rokhim, D. A., Atikah, A., Vitarisma, I. Y., Rahayu, S., & Muntholib, M. (2023). Assessment of high school students' ability to solve structured problems with ideal model on acid–base. *Orbital: The Electronic Journal of Chemistry*, *14*(4), 276–284.

Safitri, N. C., Nursa'adah, E., & Wijayanti, I. E. (2019). Analisis Multipel Representasi Kimia Peserta Didik pada Konsep Laju Reaksi. *EduChemia: Jurnal Kimia Dan Pendidikan*, *4*(1), 1–12.

Sepriyanti, N., Nelwati, S., Kustati, M., & Afriadi, J. (2022). The effect of 21st-century learning on higher-order thinking skills (HOTS) and numerical literacy of science students in Indonesia based on gender. *Jurnal Pendidikan IPA Indonesia*, *11*(2), 314–321.

Setiaji, B., Adiningsih, E. T., Pebriana, I. N., Supahar, Kuswanto, H., Wilujeng, I., & Wiyatmo, Y. (2025). Facilitating students' analytical thinking skill and scientific attitude in distance learning using local wisdom-based physics PBL e-module. *Revista Mexicana de Física E, 22, 020216*, *December*, 1–11.

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

Suparman, A. R., Rohaeti, E., & Wening, S. (2024). Student Misconception In Chemistry : A Systematic Literature Review. *Pegem Journal of Education and Instruction*, *14*(2), 238–252.

Susilaningsih, E., Nuswowati, M., & Natasukma, M. M. (2020). Profile of misconception in particulate level of acid basic subjects. *IOP Conference Series: Materials Science and Engineering, 830, 042082*.

Susilaningsih, E., Wulandari, C., Supartono, Kasmui, & Alighiri, D. (2018). The use of multi representative learning materials : definitive , macroscopic , microscopic , symbolic , and practice in analyzing students ' concept understanding. *Journal of Physics: Conference Series*, 983, 012165.

Susongko, P. (2021). The comparison of descriptive statistical parameter estimation stability using raw scores and Rasch model. *Journal of Physics: Conference Series, 1918(4), 042026.*

Taber, K. S. (2019). Alternative conceptions and the learning of chemistry. *Israel Journal of Chemistry*, *59*, 450–469.

Talens, C. U., Ong, I., Chiu, M. H., & Sevilla, F. B. (2025). Assessing conceptual understanding in high school chemistry: Particulate nature of gas concept test (PNGCT) with Rasch modeling. *Chemistry Teacher International*, *88*, 1–13.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wachidah, L. R., Laila, Y., Irmawati, A., & Amin, S. (2021). Implementasi Penggunaan Tes Essay dalam Evaluasi Pembelajaran Daring pada Siswa Kelas VII SMP Negeri 1 Tlanakan. *Ghâncaran: Jurnal Pendidikan Bahasa Dan Sastra Indonesi*, 16–27.

Widarti, H. R., Nazriati, N., & Yunitasari, I. (2020). Understanding of acid–base concept based on multiple representations on cross-class level in senior high school. *AIP Conference Proceedings, 2215(1), 020026*.

Widarti, H. R., Nazriati, N., & Yunitasari, I. (2023). Acid-base misconception using diagnostic test based multiple representation on different grades of university students Acid-Base Misconception Using Diagnostic Test Based Multiple Representation on Different Grades of University Students. *AIP Conference Proceedings, 2569*(1), 030002, 030002(January).

Widoyoko, E. P. (2014). *Evaluasi Program Pembelajaran*. Pustaka Pelajar.

Widyastuti, E., & Susiana, S. (2019, March). Using the ADDIE model to develop learning material for actuarial mathematics. In *Journal of Physics: Conference Series* (Vol. 1188, No. 1, p. 012052). IOP Publishing.

Zuhaida, A., & Imaduddin, M. (2019). Analisis Level Literasi Laboratorium Kimia dari Calon Guru IPA Tahun Pertama. *Jurnal Inovas Pendidikan Kimia*, *13*(2), 2349 – 2359.