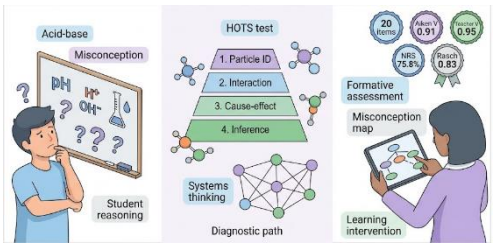
***JKPK (JURNAL KIMIA DAN PENDIDIKAN KIMIA)**, Vol. 11, No. 1, 2026* *pp. 109-129*
Chemistry Education Study Program, Universitas Sebelas Maret *ISSN 2503-4146*
https://jurnal.uns.ac.id/jkpk *ISSN 2503-4154 (online)*

The HOTS-Characteristic Diagnostic Tests Using the System Thinking Approach to Detect Acid-base Misconceptions in High School Chemistry

Septy Nur Fadhilah*, Asih Widi Wisudawati

Chemistry Education, Faculty of Tarbiyah and Teacher Training , Sunan Kalijaga State Islamic University , Yogyakarta , Indonesia

ABSTRACT



Systems thinking is an approach to understanding the complex relationships between components in a system by combining analytical and systematic thinking. Misconceptions in the thinking process can hinder understanding of chemical concepts, so a diagnostic instrument that can comprehensively identify errors in thinking is needed. Many diagnostic instruments have been developed, but none have used a systems thinking approach. This study aims to develop a diagnostic test instrument based on the systems thinking approach with Higher Order Thinking Skill (HOTS) characteristics. The development follows Wilson's four-block model with a sequential exploratory mixed method, covering qualitative and quantitative stages. The instrument consists of 20 graded items covering four aspects of systems thinking: particle identification, component interaction, cause-and-effect relationships through calculation, and particle interaction inference. Validation shows a high Aiken V index from experts (0.91) and teachers (0.95), indicating content, construct, and language suitability. Student responses yielded an average NRS score of 75.8% (good category), with a note of time constraints. Rasch analysis showed good reliability (0.83) and lower person reliability (0.57) due to homogeneous responses on several basic items. Theoretically, this study confirms the importance of integrating HOTS and system thinking in the development of chemistry diagnostic instruments. The resulting instrument can be used by teachers as a formative assessment to detect students' misconceptions in depth and design more targeted learning interventions.

Keywords: *Assessment instruments; diagnostic tests; higher-order thinking; systems thinking approach .*
**Corresponding Author: fadhilahsepty28@gmail.com*
How to cite: S. N. Fadhilah and A. W. Wisudawati, "The HOTS-characteristic diagnostic tests using the system thinking approach to detect acid-base misconceptions in high school chemistry," *Jurnal Kimia dan Pendidikan Kimia (JKPK)*, vol. 11, no. 1, pp. 109–129, 2026. [Online]. Available: https://doi.org/10.20961/jkpk.v11i1.111442
Received: 2025-11-24 *Accepted: 2026-03-08* *Published: 2026-04-30*

INTRODUCTION

Indonesia's educational quality remains a major concern, particularly in science literacy and higher-order cognitive performance. The latest Programme for International Student Assessment (PISA) data in 2022 showed that Indonesia's ranking increased by approximately five to six positions compared with 2018. Score-based performance, however, showed a decline across literacy, numeracy, and science domains. Indonesia's literacy score decreased from 371 in 2018 to 359 in 2022, numeracy decreased from 379 to 366, and science decreased from 396 to 383 [1]. These findings indicate that improvement in ranking does not necessarily reflect stronger learning quality, especially when students' conceptual understanding and reasoning abilities remain below the expected level.

© 2026 The Authors. This open-access article is distributed under a (CC-BY-SA License)

Educational challenges in the current learning context require students to develop broader insight, stronger reasoning, and better problem-solving abilities [2]. Higher Order Thinking Skills (HOTS) are therefore important because they encourage students to analyze information, evaluate evidence, make decisions, and solve problems based on systematic reasoning. Science learning should not only lead students to memorize facts, but also help them understand how scientific knowledge is constructed through observation, questioning, reasoning, and explanation. Weaknesses in these thinking processes may cause students to form incomplete or incorrect concepts, which can later develop into misconceptions.

Misconceptions occur when students construct ideas that are inconsistent with accepted scientific concepts or expert explanations [3]. These misconceptions may originate from incomplete reasoning, prior knowledge, teacher explanations, or limited learning resources [4],[5]. Assessment tools should therefore be able to reveal not only whether students answer correctly or incorrectly, but also how students think when solving problems. Classroom assessment practices, however, are still often dominated by questions that measure lower cognitive levels, such as remembering and understanding. This condition limits teachers' ability to identify deeper conceptual difficulties and to design learning interventions based on students' actual thinking processes.

HOTS-oriented assessment instruments can help teachers evaluate students' analytical, evaluative, and creative thinking more systematically. HOTS-based questions can support learning by encouraging reasoning, decision making, and problem solving [6]. Such instruments are also useful for detecting learning difficulties and misconceptions because students' answers can reflect the way they interpret concepts and construct explanations [7]. The use of HOTS-based assessment therefore has strategic value not only as an evaluation tool, but also as a diagnostic source for improving the quality of chemistry learning.

Diagnostic assessment provides a more specific approach for identifying students' strengths and weaknesses in learning. Cognitive diagnostic assessment is designed to reveal students' knowledge structures, reasoning patterns, and errors in processing information [8]. A well-designed diagnostic test should not merely indicate that students fail to master a topic, but should also provide information about the form of misunderstanding that occurs and the possible reasoning path behind it [9]. Previous research on acid-base learning showed that misconceptions may emerge from memorization-based learning and insufficient conceptual clarification [10]. Diagnostic assessment can therefore help teachers identify the level and location of students' difficulties and use the information to improve learning design, student support, and curriculum refinement [11].

Chemistry learning is particularly vulnerable to misconceptions because many chemical concepts are abstract and require

the integration of macroscopic, submicroscopic, and symbolic representations. Acid-base material is one of the topics frequently associated with conceptual difficulty and misconception. Students often struggle to understand acid-base concepts because the topic involves terminology, mathematical calculations, symbolic equations, equilibrium relationships, and titration processes [12]. Difficulties in understanding basic acid-base concepts may also affect students' ability to understand more advanced concepts, particularly those involving pH calculation, ionization, weak acid-base equilibrium, and titration analysis [13].

Systems thinking has become an important approach in chemistry education because it helps students understand relationships among components within complex systems. This approach encourages students to view chemical phenomena not as isolated facts, but as interconnected processes involving components, interactions, cause–effect relationships, dynamic behavior, and broader impacts [14]. Systems thinking in chemistry education supports holistic understanding by enabling students to analyze the composition, structure, behavior, and effects of chemical systems [15]. Mahaffy and Krief describe systems thinking in chemistry education as a way to connect molecular-level knowledge with sustainability, society, and environmental systems [16]. This perspective is relevant for acid-base learning because students must connect particle-level interactions, symbolic calculations, and observable chemical phenomena.

Several diagnostic instruments have been developed to identify chemistry misconceptions, including two-tier, three-tier, four-tier, essay-based, and HOTS-oriented diagnostic tests. These instruments have contributed to identifying incorrect answers, confidence levels, reasoning patterns, and critical thinking performance. A three-tier multiple-choice diagnostic test integrating critical thinking indicators has been developed to measure students' critical thinking skills on acid-base material through validity, reliability, difficulty level, discriminating power, and questionnaire analysis [17]. Another study developed an Internet-Based Test (IBT) two-tier HOTS instrument using Google Form for acid-base titration material [18]. These studies show that diagnostic assessment has developed toward more complex formats and technology-supported implementation.

Existing studies have developed diagnostic instruments to identify chemistry misconceptions and related thinking skills, including three-tier multiple-choice tests integrating critical thinking indicators and Internet-Based Test (IBT) two-tier HOTS instruments for acid-base material [17],[18]. These instruments have contributed to identifying incorrect answers, reasoning patterns, confidence levels, and critical thinking performance, although the confidence component is more strongly represented in multi-tier diagnostic formats [17]. Diagnostic assessment has also been recognized as a useful approach for

identifying students' conceptual difficulties and supporting instructional improvement [8],[10],[11]. Many existing instruments, however, still focus on final conceptual errors rather than tracing the sequence of students' reasoning across interconnected chemical concepts. This limitation becomes important in acid-base calculation topics because students are required to connect terminology, particle interactions, equilibrium relationships, symbolic equations, and quantitative reasoning [12],[13]. Systems thinking offers a relevant framework to address this limitation because it emphasizes relationships among components, interactions, cause–effect mechanisms, dynamic behavior, and broader system effects in chemistry learning [14]–[16].

This study addresses the identified gap by developing a HOTS-characteristic diagnostic test instrument based on a systems thinking approach to detect students' misconceptions in acid-base calculation material. The proposed instrument is designed to guide students through a structured reasoning sequence, including identifying particles or components in a system, recognizing interactions among components, connecting cause-and-effect relationships through calculations, and inferring particle interactions based on calculation results. This structure combines diagnostic assessment, higher-order thinking demands, and systems thinking sequences to provide a more comprehensive diagnosis of students' conceptual understanding and misconception patterns. The instrument is expected to help teachers identify not only whether students experience misconceptions, but also how these misconceptions are formed through their reasoning processes.

This study is guided by three main research questions concerning the development and evaluation of a HOTS-characteristic diagnostic test instrument based on a systems thinking approach for identifying students' misconceptions in acid–base calculation material. The first question examines how the instrument can be systematically developed to detect students' misconceptions effectively. The second question investigates the quality of the instrument based on expert validation, teacher assessments, and student responses analyzed using Aiken's V and response analysis. The third question explores the extent to which the empirical quality of the instrument is supported by Rasch Model analysis.

METHODS

1. Research Design

This study uses a mixed research method or mixed methods, by combining qualitative and quantitative research methodology. Research design in this mixed research uses sequential models, especially the exploratory sequential design, to follow the four building blocks of measurement tool development Wilson. This study follows a cycle in the development of assessment, using a cycle of four building blocks, including; (a) construct map, (b) item design, (c) outcome space, and (d) the measurement model [19]. Figure 1 illustrates the four building blocks by Wilson used in this study.

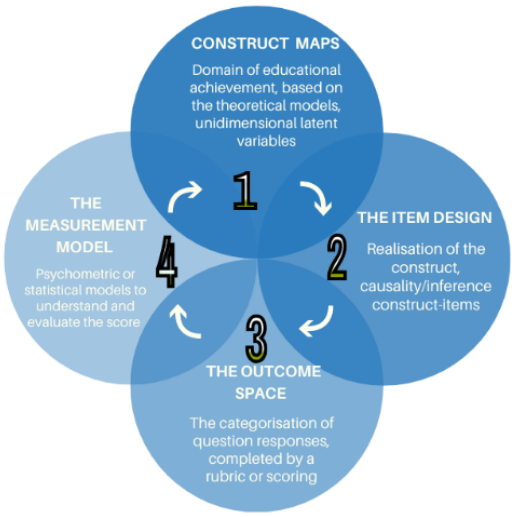


Figure 1. Four building blocks to describe measurement tools [19]

2. Qualitative Stage

The qualitative stage was conducted during the construct map phase of Wilson's four-block model. This stage aimed to identify the conceptual domain, misconception potential, and systems thinking sequence that would become the basis for developing the diagnostic instrument. The qualitative process involved a literature review, curriculum analysis, and semi-structured interviews with chemistry teachers. These activities were used to formulate item indicators that reflected the characteristics of diagnostic tests, Higher Order Thinking Skills (HOTS), and systems thinking sequences.

The literature review focused on studies related to chemistry misconceptions, diagnostic assessment, HOTS-oriented assessment, and systems thinking in chemistry education. Curriculum analysis was conducted on senior high school chemistry content, particularly acid-base material, to identify learning outcomes, core competencies, and indicators with a high potential for conceptual difficulty. Semi-structured interviews with chemistry teachers were then used to explore empirical experiences regarding students' difficulties and misconceptions in acid-base learning. Interview data were coded thematically to identify recurring patterns of conceptual difficulty, which were then used as the basis for item design and construct map development.

3. Quantitative Stage

The quantitative stage was conducted during the outcome space and measurement model phases of Wilson's four-block model. This stage aimed to evaluate the quality of the developed diagnostic instrument through expert judgment, teacher assessment, student responses, and empirical testing. Content and construct validity were first examined based on expert and teacher evaluations. The results of these evaluations were analyzed using Aiken's V formula to determine the validity level of the instrument [20]. The Aiken's V validity index ranges from 0 to 1, with the classification criteria presented in Table 1.

$$V = \frac{\Sigma s}{[n(c-1)]}$$

Description:

V = coefficient of validity of the contents
s = (validator assigned value) – (minimum possible validation value)
n = number of highest rated
c = highest rating

Table 1. Validity based on Aiken's V scale

Aiken's V Scale	Categories
≤ 0,4	Less Valid
0,4 – 0,8	Valid
≥ 0,8	Very Valid

Student responses were analyzed using the Numeric Rating Scale (NRS) to evaluate the readability, practicality, and perceived quality of the developed instrument. The percentage of student responses was calculated using the following formula:

$$\%NRS = \frac{\Sigma_{i=1}^{n} NRS}{Maximum\ NRS} \times 100\%$$

Notes:

%NRS = Percentage of student response
$\Sigma_{i=1}^{n}$ NRS = Total student response score on each question item
Maximum NRS = n x best choice score

If the number of good criteria > 50%, then the student's response is positive. Whereas if the criteria are good < 50% then it is said that the student's response is negative. The percentages of student response categories are presented in Table 2.

Table 2. Percentage of student response categories

%NRS	Categories
6% ≤ %NRS < 25%	Very less
25% ≤ %NRS < 43%	Less
43% ≤ %NRS < 62%	Enough
62% ≤ %NRS < 81%	Good
81% ≤ %NRS < 100%	Very good

The empirical testing phase represented the fourth stage of Wilson's four-block model, namely the measurement model. This stage was conducted to obtain empirical evidence regarding the psychometric quality of the developed diagnostic instrument. Students' responses were analyzed using the Rasch Model because this approach allows item difficulty and student ability to be estimated on the same logit scale. Rasch analysis also provides more detailed diagnostic information than descriptive scoring because it can identify whether items function consistently, whether item difficulty is appropriately distributed, and whether the instrument can distinguish different levels of student ability.

Table 3. Criteria for rasch model analysis

Rasch Analysis	Criteria
Unidimensionality	>60% (special) 40-60% (good) 20-40% (enough) ≥20% (minimum) <20% (less) <15% (*unexpected variance*)
Unexplained Variance	<3% (very good) 3-5% (good) 5-10% (enough) 10-15% (minimum) >15% (bad) <15% (*unexpected variance*)
Differential Item Functioning (DIF)	Probability value <5% (0,05)
Item Fit Order	0,5 < MNSQ < 1,5 -2,0 < ZSTD < +2,0 0,4 < Pt Mean Corr < 0,85
Item measure	Measure logit > +0,73 (very difficult) 0,0< measure logit < +0,73 (difficult) 0,0>measure logit> -0,73 (easy) Measure logit < -0,73 (very easy)
Reliability	>0,94 (special) 0,91-0,94 (very good) 0,81-0,90 (good) 0,67-0,80 (enough) >0,67 (weak)
Cronbach's Alpha	0,00<r<0,50 (bad) 0,50<r<0,60 (ugly) 0,60<r<0,70 (enough) 0,70<r<0,80 (good) 0,80<r<1,00 (very good)
Separation	>5 (special) 3-4 (very good) 2-3 (good) ≥1,5 (accepted) <1,5 (not accepted)

[21]

The Rasch analysis included unidimensionality, unexplained variance, Differential Item Functioning (DIF), item fit order, item measure, reliability, Cronbach's alpha, and separation indices. Unidimensionality and unexplained variance were used to examine whether the instrument measured a dominant construct. DIF analysis was used to identify potential item bias across student groups. Item fit order was used to evaluate the consistency of each item with the Rasch model, while item measure was used to classify item difficulty. Reliability, Cronbach's alpha, and separation indices were used to evaluate response consistency and the discriminating capacity of the instrument. These indicators were interpreted using established Rasch criteria to determine whether the items were appropriate, required revision, or needed cautious interpretation [21]. The Rasch results provided an empirical basis for judging the quality of the instrument before it was interpreted as a diagnostic tool. The criteria used to interpret each Rasch indicator are summarized in Table 3.

4. Research Ethics

Research ethics were applied to protect participants' rights, privacy, and welfare throughout the study. The participants involved in this research included expert validators, chemistry teachers, and students. Ethical procedures were implemented by obtaining consent from all parties involved before data collection. Expert validators provided verbal consent, while school participants, including teachers and students, were involved after official permission had been obtained from the school.

Participant confidentiality was maintained by using anonymous codes in all research data and reporting processes. The collected data were used only for research purposes and were not manipulated or used in ways that could harm the participants. The principle of beneficence was applied by ensuring that the study contributed to the development of chemistry learning and diagnostic assessment. The principle of justice was maintained by providing equal opportunities for students in the selected school categories to participate in the study.

5. Research Subjects and Sampling Techniques

Research subjects consisted of three groups: chemistry teachers for needs analysis, expert lecturers and chemistry teachers for instrument validation, and students for readability testing and empirical trials. The empirical trial involved senior high school students in the Special Region of Yogyakarta who had completed acid-base learning material. School selection was initially planned using stratified random sampling based on students' summative achievement scores, with schools classified into high, medium, and low cognitive strata. Permission refusal from one intended school caused the final empirical trial to include only schools from medium and low cognitive strata. This condition limited the representation of students with high cognitive ability and should be considered when interpreting the generalizability of the findings.

Student selection within the selected schools was conducted using purposive sampling with guidance from chemistry teachers. A total of 102 students participated in the study. This sample size was considered sufficient for Rasch analysis because a minimum of 30–50 respondents is generally acceptable, while more than 100 respondents is recommended to obtain more stable item and person parameter estimates [21]. The sample therefore supported the analysis of unidimensionality, item fit, reliability, and Differential Item Functioning (DIF). Data collection was conducted from April to May 2025.

RESULTS AND DISCUSSION

1. Instrument development (construct maps)

The developed instrument followed the principles of Higher Order Thinking Skills (HOTS) assessment. Each item was designed using a contextual stimulus based on everyday phenomena. The questions were arranged in graded levels according to the systems thinking approach to detect students' misconceptions. The characteristics of systems thinking and the four aspects used in the instrument are presented in Table 4.

Table 4. Characteristics and three aspects of the order of system thinking

Characteristics of system thinking	Four aspects of system thinking steps used	Bloom's taxonomy
1. Think in terms of parts	Identify the components or particles involved in a phenomenon, characterizing properties at the molecular level	Understanding
2. Seeing things in the bigger picture		
3. Find out the effect caused by an action	Identify interactions between system components (limited to chemical reactions)	Application and analysis
4. Identify how a relationship might affect the system		
5. Understand the concept of dynamic behavior	Attributing the cause and effect of the reaction in the form of a calculation equation	Application and analysis
6. Understand how the structure of the system shapes the behavior of the system	Deducing the interactions between particles from the calculations of the particles involved in the system	Application and analysis
7. Seeing things from a different point of view		

Each basic competency and question indicator is developed into 2 forms of questions, so that the number of questions developed is 20 items. The designed instrument combines four different levels for each principle of systems thinking. The questions are presented from the first competence to the fourth competence gradually. The structuring of the stages of the question is in line with Arnold and Wade in [22], which defines system thinking as a set of analytical skills used to improve the ability to identify and understand, predict and design to be able to produce the desired effect in a system. Meanwhile, questions with HOTS characters are located in the third level of questions. At the third level, the questions follow the Bloom C4-C6 taxonomy with

several forms of questions: ordinary multiple choice, complex multiple choice (True/False), short answer, and blurb. The developed problem indicators include analyzing the concentration of H^+ ions in solutions, analyzing the concentration of OH^- ions in solutions, analyzing both H^+ and OH^- ion concentrations in solutions, analyzing the properties of substances based on their pH values, correlating the weak base equilibrium constant (pKb) with the ionization constant of water (pKw), determining the base dissociation constant (Kb) of specific chemical solutions, correlating solution volume and concentration in the dilution process, predicting solution concentration based on titration data, summarizing pH values obtained from titration data, and analyzing the degree of ionization of a given sample. A total of 20 questions were developed based on these indicators, and one example of the developed questions is presented in Figure 2.

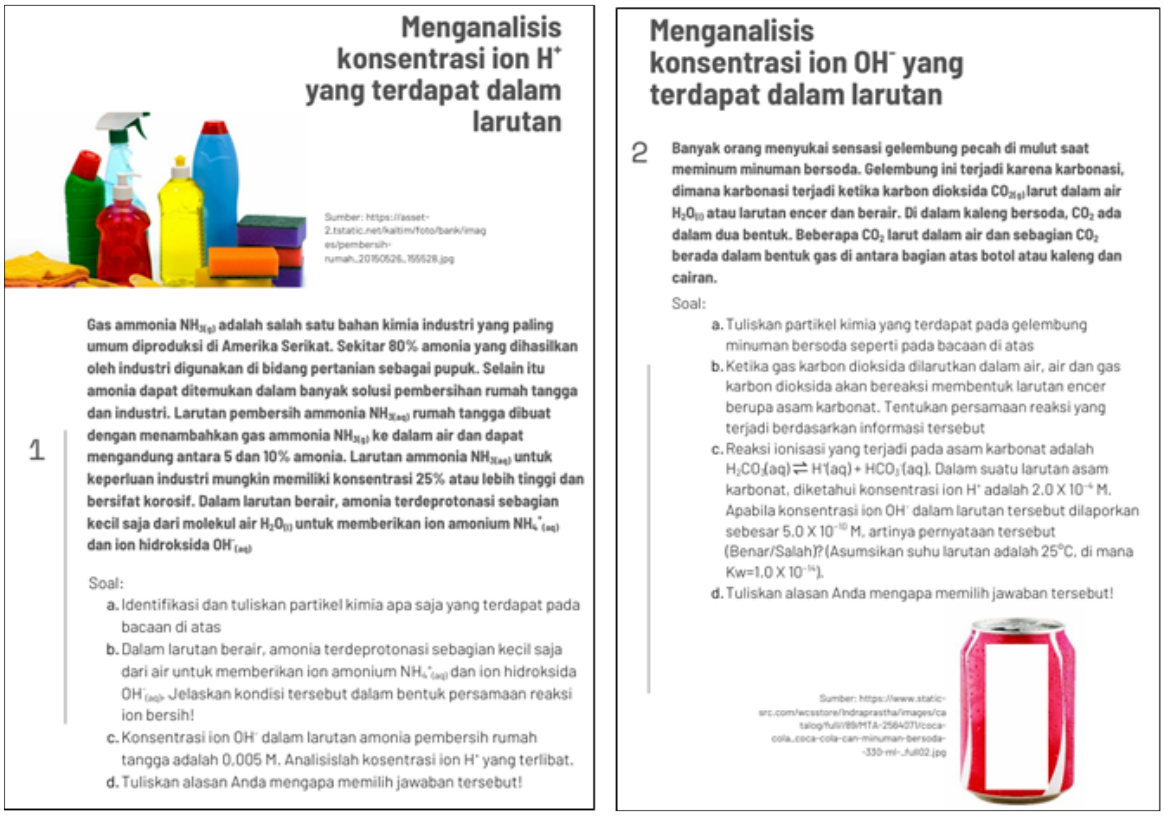


Figure 2. The developed items

The developed 20-item instrument was divided into two question packages, namely Package A and Package B, to support effective implementation during testing. Each package represented 10 basic and construct competencies. Content validity was then examined by two expert validators using Aiken's V calculation. The validators consisted of a material expert who assessed the accuracy and relevance of the chemistry content and an evaluation expert who assessed the quality of the instrument as an assessment tool. The results of expert validation are presented in Table 5.

Table 5. Aiken V calculations for expert validity

Test items	Experts		S1	S2	Σs	n(c-1)	V	Judgment
	1	2						
1-20	57,9	55,1111	37,9	35,1111	73,0	80,0	0,9125	Very valid

The results of expert validation with Aiken V index of 0.91 indicate that the instruments developed have high consistency in material, construct, and language aspects. This value indicates that the item has been in accordance with the principles of HOTS and system thinking, so theoretically the instrument is able to measure cognitive skills to be achieved. However, this figure also indicates that there are some items that still need improvement, especially in the item that has not fully displayed the characteristics of HOTS. In addition, some stimuli also need to be improved to deepen the information presented, this is important because a weak stimulus can reduce the power of the instrument in triggering systemic reasoning.

Instruments that have been validated by experts, then made improvements according to the advice and input provided by the validator. In addition, the assessment was also carried out by two teachers in helping to help assess how the quality of the instrument can be applied practically in schools. The results of teacher assessments calculated using the Aiken validity index are presented in Table 6.

Table 6. Teacher response assessment results

Indicators	Reviewers		S1	S2	Σs	n(c-1)	V	Judgment
	Teacher 1	Teacher 2						
1-17	48	51	31	34	65	68	0,9558	Very valid

The teacher's assessment showed that a higher Aiken score of 0.95 reinforced the practical validity of the instrument. Teachers assess the instrument is very feasible to use in the classroom, both in terms of readability and suitability to the learning context. Small differences between expert and teacher assessments indicate that the instrument is academically adequate, but the application in the classroom needs to emphasize more practical aspects. This confirms the importance of triangulation between theoretical and practical validity, in which the instrument is not only academically appropriate, but also relevant for the teacher as the main user in the classroom.

The Aiken V value, both the validity carried out by experts and teacher assessments, shows three aspects of validation, namely, material (chemical content), construct, and language. Questions that have been made further improvements, then tested the validity of the empirical test as well as to get student responses to the instrument. The number of students involved is 102 respondents, but in the analysis of data on empirical tests involving only 80 respondents. This is because there is one study group or about 22 students who cannot complete the test due to the limited time available in the implementation of the piloting project at that time. Thus, the results obtained

in the study group are considered not to reach the expected target in the implementation of the trial. However, the results of the student response questionnaire were still used to obtain responses to the quality of the instrument in a descriptive manner. Student responses obtained showed differences compared to the responses of experts and teachers before. The percentage of student responses was calculated using the Numeric Rating Scale (NRS), and the results are presented in Table 7.

Table 7. Student response results

Indicator	Max score	%	Results	%average	Results
Indicator-1	510	75,09803922	Good	75,88235294	Good
Indicator-2	510	60	Moderate		
Indicator-3	510	83,7254902	Excellent		
Indicator-4	510	83,33333333	Excellent		
Indicator-5	510	77,84313725	Good		
Indicator-6	510	71,76470588	Good		
Indicator-7	510	77,25490196	Good		
Indicator-8	510	78,03921569	Good		

Student responses with an average NRS score of 75.8% (category “Good”) give a more complex picture. Students rated the instrument positively in terms of readability and question form, but gave a low score on the time indicator of 60%. Students tend to give a fairly small score on indicator - 2, namely the statement “the time given is in accordance with the test item (question) given”. According to them, the time or duration given is not suitable to answer the whole question. However, it differs from the legibility aspect of the instruments in that they score very well against the use of appropriate and comfortable font shapes and sizes for reading. This reveals the gap between the design of the instrument and the real conditions of students in the classroom. The design of the instrument demands multilevel analysis so it takes longer in completing it, which makes students feel the existing duration is not enough. Empirically, these results affirm that content and construct validity do not automatically guarantee practical validity at the student level. Instruments that are complex without time adjustment can cause obstacles, even though the quality of the matter is high.

The difference in assessment between experts, teachers, and students reflects multi-level validity: experts emphasize theoretical suitability, teachers emphasize practicality, and students highlight experience of use. All three complement each other, but also show the weak points of the instrument. High validity by experts and teachers can be biased if not balanced with the direct experience of students who feel burdened in time. Thus, the teacher can use the instruments developed as a diagnostic tool, but it is necessary to adjust the working time or divide the questions into several sessions so that the students do not feel overwhelmed. Furthermore, the results of empirical tests performed by analysis using Rasch Model. The quality of the instrument can be seen

from the results of analysis using rasch model which includes unidimensional test, bias analysis or Differential Item Function (DIF), item Fit test, item Measure level analysis, and reliability test. The analysis was conducted with data sourced from 80 student respondents, with each respondent in a question package of 40 respondents. The analysis of each question item is carried out separately between question points a, b, c and d; this is to ensure that each question item can measure the competence to be achieved.

2. Unidimensionality

Unidimensionality analysis was conducted to ensure that the instrument measured a single dominant construct. In Rasch analysis, unidimensionality indicates that the variance in item responses is mainly explained by one underlying ability or construct [23]. A one-factor assumption is considered acceptable when the principal dimension explains the dominant proportion of variance and the unexplained contrasts remain low. High unexplained contrasts may indicate the presence of additional dimensions, meaning that the instrument may measure more than one underlying construct.

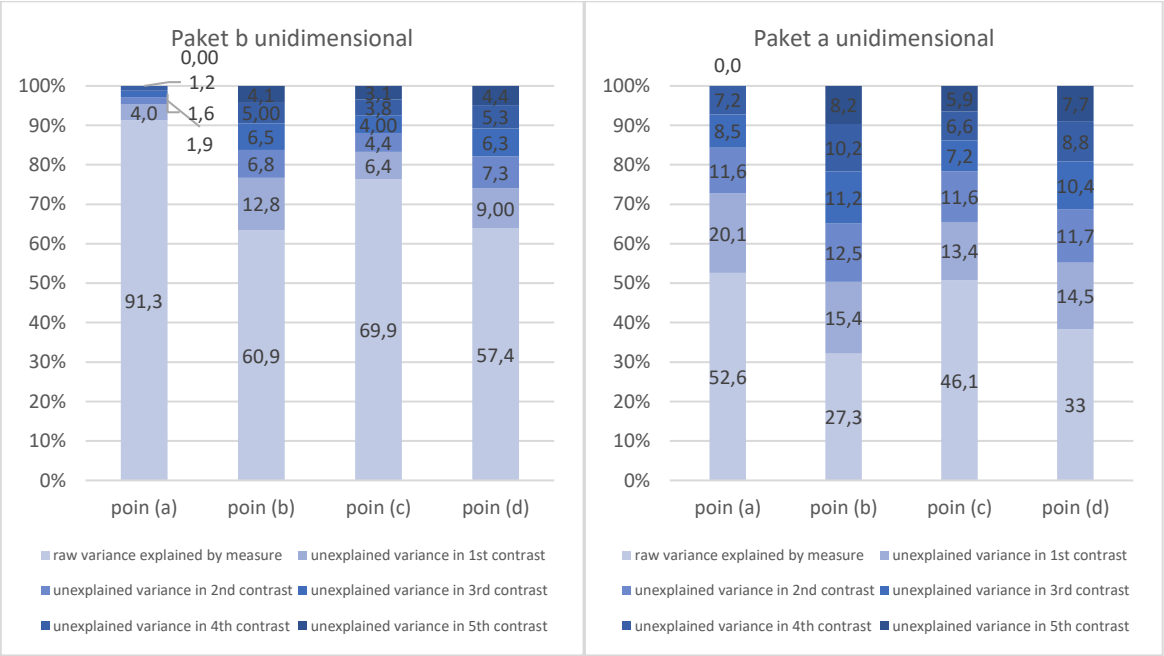


Figure 3. Summary of unidimensionality test results

The unidimensionality results showed different patterns across Package A and Package B, as summarized in Figure 3. Package B generally demonstrated stronger unidimensionality than Package A. Point (a) in Package B showed the strongest unidimensionality, with 91.3% of variance explained by the measure and very low unexplained variance. Package A in point (a) also showed acceptable unidimensionality, with 52.6% of variance explained by the measure, although the unexplained variance in the first contrast remained relatively high.

Point (b) showed a more varied pattern. Package A had the lowest unidimensionality value, with 27.3% of

variance explained by the measure, indicating the possible presence of additional dimensions. Package B showed stronger unidimensionality, with 60.9% of variance explained by the measure. Point (c) showed a relatively good pattern, with 46.1% of variance explained in Package A and 69.9% in Package B. These results indicate that point (c) was more consistent in measuring the intended construct, particularly students' ability to analyze particle interactions through calculation-based reasoning. Point (d) showed 33.0% of variance explained in Package A and 57.4% in Package B, indicating that Package B remained more structurally consistent than Package A.

Differences between Package A and Package B suggest that Package B items more consistently measured the intended constructs, whereas several Package A items showed potential subdimensions. This condition may be related to differences in stimulus complexity or variations in students' strategies when connecting symbolic representations with macroscopic phenomena. Some items in Package A may have triggered multiple cognitive pathways because they required students to move across different levels of Bloom's taxonomy and systems thinking sequences. Theoretically, this finding supports the idea that students may shift between macroscopic, submicroscopic, and symbolic representations when solving chemistry problems. The instrument therefore needs to guide students more explicitly from component identification, particle interaction, cause-and-effect reasoning, to system-level inference to strengthen its unidimensional structure.

3. DIF (Differential Item Functioning)

Differential Item Functioning (DIF) analysis was conducted to identify whether certain items showed potential bias across student groups. DIF analysis is important in Rasch-based instrument validation because it helps determine whether items function fairly for different groups of respondents with comparable ability levels [21]. An item is considered to contain DIF when students from different groups with similar ability have different probabilities of answering the item correctly [21]. This study examined DIF based on differences between the two school groups involved in the empirical trial. The summary of DIF analysis is presented in Figure 4.

Bias analysis showed that point (a) did not indicate item bias in either Package A or Package B because all analyzable items had probability values above 0.05. Several items could not be analyzed due to no response variation, which is consistent with the basic function of point (a) as an initial component-identification item. Item 5a in Package B approached the significance threshold (PROB = 0.0941), so it should be reviewed further, although it was not categorized as biased based on the 0.05 criterion [21].

DIF was detected in several items across points (b), (c), and (d). Point (b) showed DIF in item 9b in Package A (PROB = 0.0418) and items 4b and 5b in Package B (PROB = 0.0201 and 0.000). Point (c) showed DIF in item 8c in Package A (PROB = 0.0212) and items 4c and 8c in Package B (PROB = 0.0089 and 0.0277). Point (d) showed no DIF in Package A, but Package B showed DIF in items 6d and 8d, with

probability values of 0.0122 and 0.0136, respectively. These results indicate that several items may function differently across school groups and therefore require closer examination.

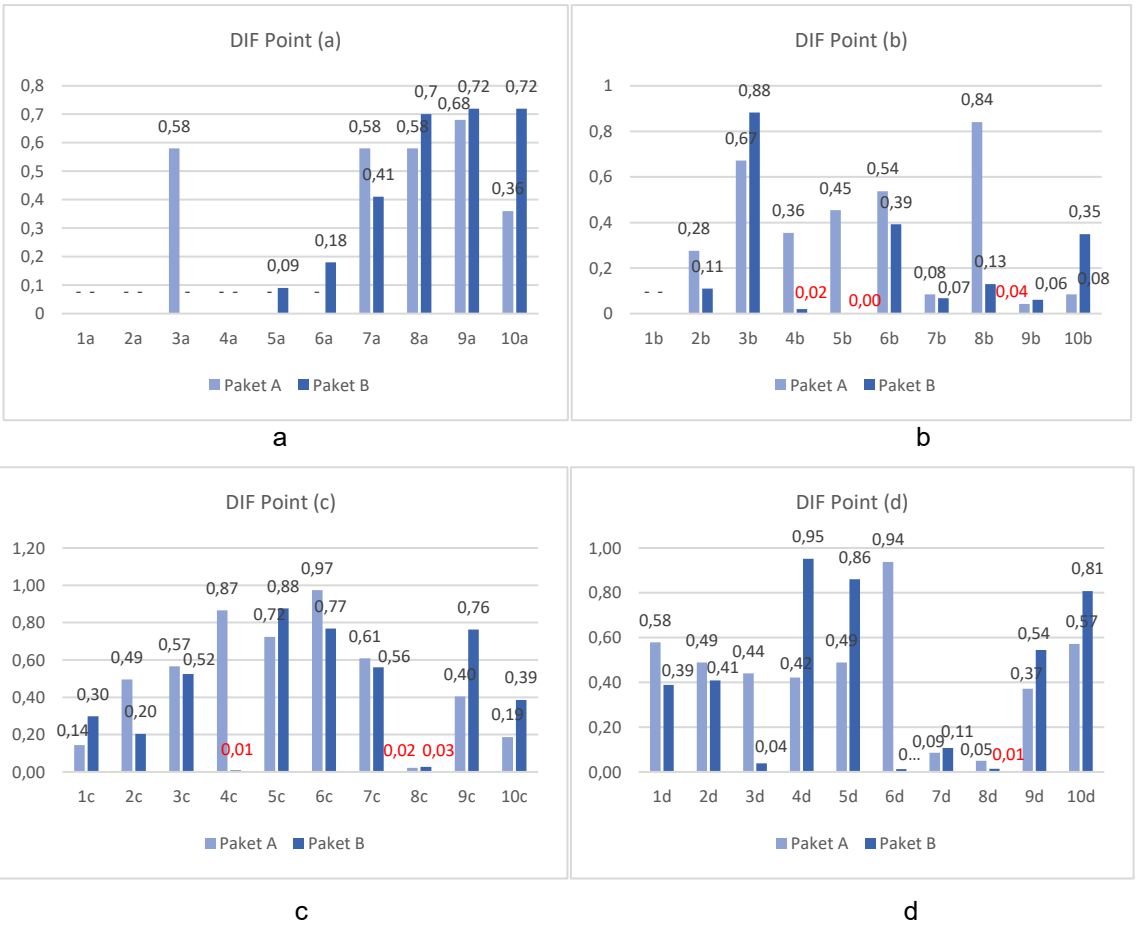


Figure 4. Summary of DIF analysis results (bias item)

Potential item bias may be related to differences in students' learning experiences, teacher instructional strategies, and exposure to HOTS-based questions. DIF should therefore be interpreted not only as a statistical issue, but also as an indication of variation in learning contexts and reasoning patterns. This interpretation is relevant because HOTS-oriented assessment requires reasoning, decision making, and problem solving [6],[7], while systems thinking requires students to connect components, interactions, and cause effect relationships within a system [14],[15]. These findings indicate that several items need review to improve fairness, clarity, and construct alignment across student groups.

1. Item Fit Order

Item fit order analysis evaluated item suitability using Outfit MNSQ, Outfit ZSTD, and Pt-Measure Corr. Items were considered fit when they met at least two criteria: 0.5 < MNSQ < 1.5, −2.0 < ZSTD < +2.0, and 0.4 < Pt-Measure Corr < 0.85 [21]. The results are presented in Table 8.

Table 8. Item fit analysis results

	Package A			Package B				Package A			Package B		
Item	Outfit MNSQ	Outfit ZSTD	Pt-Measure Corr	Outfit MNSQ	Outfit ZSTD	Pt-Measure Corr	Item	Outfit MNSQ	Outfit ZSTD	Pt-Measure Corr	Outfit MNSQ	Outfit ZSTD	Pt-Measure Corr
1a	-	-	-	-	-	-	1b	-	-	-	-	-	-
2a	-	-	-	-	-	-	2b	0,92	-0,05	0,32	0,71	0,24	0,15
3a	4,41	2,15	0,12	-	-	-	3b	1,08	0,50	0,29	1,44	1,05	0,66
4a	-	-	-	-	-	-	4b	0,93	-0,24	0,38	1,46	0,74	0,29
5a	-	-	-	9,90	9,91	0,28	5b	0,94	0,04	0,23	3,80	4,07	0,30
6a	-	-	-	0,11	-0,69	0,61	6b	0,72	-0,26	0,30	0,20	-0,58	0,50
7a	2,10	1,17	0,42	0,12	-0,70	0,91	7b	0,74	-1,69	0,60	0,31	-1,35	0,85
8a	0,23	-0,82	0,63	0,08	-0,73	0,91	8b	0,86	-0,25	0,32	0,36	-0,63	0,83
9a	0,40	-1,79	0,87	0,06	-0,84	0,89	9b	1,80	2,35	0,13	0,27	-1,28	0,87
10a	0,59	-0,23	0,85	0,06	-0,84	0,89	10b	0,67	-2,26	0,66	0,27	-0,39	0,81
1c	2,42	2,26	0,57	0,23	-0,35	0,53	1d	1,34	1,16	0,67	0,61	-0,28	0,53
2c	0,42	-1,21	0,78	0,22	-0,39	0,46	2d	0,88	-0,21	0,68	1,08	0,37	0,43
3c	0,60	-0,58	0,72	0,72	0,16	0,54	3d	1,00	0,14	0,68	0,97	0,23	0,54
4c	1,97	1,04	0,64	9,09	6,13	0,43	4d	0,91	-0,17	0,67	0,88	0,13	0,52
5c	0,56	-0,44	0,71	0,17	-0,76	0,78	5d	0,88	-0,21	0,68	0,59	-0,40	0,68
6c	1,45	1,02	0,59	0,59	-0,33	0,72	6d	0,91	-0,15	0,69	0,75	-0,30	0,64
7c	0,69	-0,74	0,75	0,71	0,10	0,63	7d	0,95	0,06	0,66	0,62	-0,05	0,68
8c	1,47	1,16	0,65	0,60	-0,18	0,67	8d	1,71	1,54	0,61	0,72	-0,19	0,64
9c	0,61	-0,97	0,78	0,19	-0,68	0,53	9d	1,00	0,11	0,70	1,48	0,77	0,55
10c	0,25	-0,32	0,67	0,08	-0,73	0,31	10d	0,43	-0,65	0,73	4,64	1,89	0,16

Item fit results showed varied patterns across item points and packages. Several items in point (a) were not calculated because they had no response variation, which is understandable because point (a) was designed as the initial stage of the systems thinking sequence for identifying particles or components in the system. Although some point (a) items did not fully meet the statistical fit criteria, they were retained as entry-level diagnostic items for mapping students' initial conceptual understanding. This decision is consistent with the function of diagnostic assessment in identifying students' conceptual difficulties and supporting instructional improvement [10],[11].

Point (b) generally showed better fit patterns, with most items in both packages meeting at least two Rasch fit criteria. Item 9b in Package A and items 5b and 1b in Package B require further review because they did not meet sufficient fit criteria or had no response variation. Point (c) also showed mostly acceptable fit results, although item 1c in Package A and items 4c and 10c in Package B need revision. Misfit in point (c) may occur because calculation-based items require students to connect conceptual understanding, symbolic procedures, and quantitative reasoning in acid-base contexts [12],[13].

Point (d) showed stronger fit results in Package A, where all items met at least two criteria, while Package B had one misfitting item, namely item 10d. This point represents the most complex stage because students were required to infer particle interactions from calculation results. Misfit at this stage may reflect differences in students' ability to integrate macroscopic, submicroscopic, and symbolic representations, which is central to systems thinking in chemistry [15],[16]. Overall, psychometric mismatch in several items does not necessarily require item removal because some basic items remained pedagogically relevant for diagnostic purposes. Revision should therefore focus on

improving item clarity, stimulus quality, and construct alignment rather than eliminating all statistically imperfect items.

2. Item Measure

The item measure values were used to identify the easiest and most difficult items for respondents, as summarized in Table 9. Item Measure analysis shows that the items at point (A) in both package A and package B represent uneven levels of difficulty, with some items being too easy and some quite difficult. A good instrument is able to show how the classification of problems can be spread evenly ranging from problems that are considered very difficult, difficult to easy, to very easy [24]. Thus, it is not optimal enough to measure the ability of students as a whole. However, this is in accordance with the design of the problem as the initial stage of mastering the concept. In contrast, the distribution of item difficulty in points (b) of packages A and B is fairly even. Thus, it is effective to measure the expected ability in this point question (b). For item points (c) shows that package A has an even distribution of difficulty levels, while package B tends to be less evenly distributed because of the lack of easy category items, both of which place item 10c as the most difficult item. Likewise, in point (d), the distribution of difficulty levels is fairly evenly distributed in package A, but less evenly in package B because of the lack of easy category items.

Uneven distribution indicates that learners tend to master basic concepts well, but difficulties on complex calculations (e.g. on titration or degree of ionization). This indicates a systemic misconception: students can answer the definition, but fail to connect the cause-and-effect in the system. The instrument needs to balance the distribution of difficulty in order to be able to map capabilities across levels. An even distribution of difficulties is important for completely mapping the trajectory of systemic thinking.

Table 9. Summary of item Measure results

Package A		**Package B**		**Package A**		**Package B**		**Package A**		**Package B**		**Package A**		**Package B**	
Item	JMLE Measure	Item	JMLE Measure	Item	JMLE Measure	Item	JMLE Measure	Item	JMLE Measure	Item	JMLE Measure	Item	JMLE Measure	Item	JMLE Measure
10a	2,64	9a	3,90	9b	2,01	10b	3,04	10c	3,62	10c	6,31	10d	1,77	10d	4,14
9a	1,09	10a	3,90	7b	0,91	8b	2,41	9c	0,83	9c	4,32	7d	0,67	9d	2,98
3a	-1,24	8a	2,66	10b	0,91	9b	1,87	7c	0,61	7c	2,27	3d	0,62	7d	1,81
7a	-1,24	7a	0,53	3b	0,42	7b	1,62	8c	0,61	8c	1,63	8d	0,62	8d	1,02
8a	-1,24	6a	-4,67	4b	0,03	3b	0,55	6c	-0,24	6c	0,06	6d	-0,03	6d	0,40
1a	-3,62	5a	-6,32	8b	-0,71	5b	0,36	1c	-0,45	4c	-1,03	9d	-0,03	5d	-1,64
1a	-3,62	1a	-7,70	2b	-0,89	4b	-2,72	2c	-0,66	5c	-2,11	1d	-0,43	3d	-1,63
4a	-3,62	2a	-7,70	5b	-1,10	6b	-2,72	3c	-0,87	3c	-3,12	4d	-1,00	4d	-1,98
5a	-3,62	3a	-7,70	6b	-1,59	2b	-4,41	5c	-1,33	1c	-3,92	2d	-1,19	2d	-2,54
6a	-3,62	4a	-7,70	1b	-4,34	1b	-5,69	4c	-2,12	2c	-4,40	5d	-1,19	1d	-2,74

3. Reliability

Reliability analysis was conducted to examine the consistency of the developed diagnostic instrument. Person reliability indicates the consistency of students' responses, whereas item reliability indicates

the stability of item difficulty estimates. Cronbach's alpha was used to evaluate the overall interaction between persons and items [25]. Separation indices were also examined to determine the extent to which the instrument could distinguish different levels of student ability and item difficulty [26].

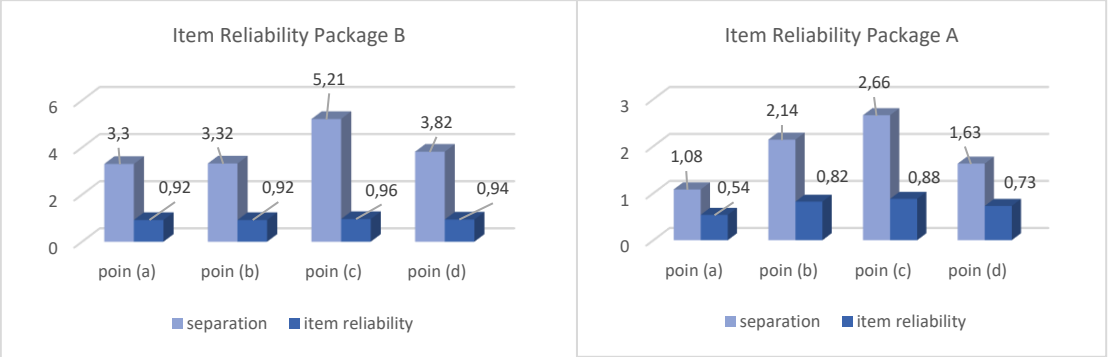


Figure 5. Summary of reliability test results

Reliability results showed different patterns across item points and packages. Point (a) in Package A showed low reliability and weak separation, indicating that the items were less able to distinguish students' abilities. This condition may be related to the basic nature of point (a), which was designed to measure initial component identification in the systems thinking sequence. Package B showed better reliability and separation in point (a), indicating more stable item functioning and better ability to classify students' responses.

Point (b) in Package A also showed limited ability to distinguish students' abilities because the score distribution was relatively narrow. Package B demonstrated better performance, with more acceptable reliability and separation values. Point (c) showed stronger reliability patterns in both packages, indicating that items involving cause-and-effect reasoning through calculation were more effective in differentiating students' ability levels. Point (d) also showed acceptable reliability, especially in Package B, which had stronger item separation and more consistent item difficulty estimation.

The reliability and separation results indicate that the instrument was generally stable at the item level, although person reliability varied across packages and item points. Lower person reliability may reflect homogeneous student responses, especially on basic items, or differences in students' learning backgrounds across the medium and low school strata. This finding suggests that the instrument can distinguish item difficulty levels more consistently than student ability levels. Broader empirical testing involving students from high, medium, and low cognitive strata is therefore needed to obtain a more complete picture of the instrument's discriminating capacity. The separation values and corresponding strata for Package A and Package B are summarized in Table 10.

Table 10. Summary of separation strata results

Package		Separation of value		Results/separation of strata
A	Point (a)	Respondent	0,00	0
		Item	1,08	2
	Point (b)	Respondent	0,55	1
		Item	2,14	3
	Point (c)	Respondent	1,73	3
		Item	2,66	4
	Point (d)	Respondent	1,53	2
		Item	1,63	2
B	Point (a)	Respondent	1,77	3
		Item	3,30	5
	Point (b)	Respondent	1,49	2
		Item	3,32	5
	Point (c)	Respondent	1,73	3
		Item	5,06	7
	Point (d)	Respondent	1,66	2
		Item	3,82	5

Overall, the reliability results support the use of the developed instrument as a diagnostic tool, particularly for identifying variation in students' reasoning across the systems thinking sequence. Items requiring higher cognitive processes, such as calculation-based reasoning and particle interaction inference, showed greater potential to differentiate students' thinking patterns. This result is consistent with the purpose of HOTS-based diagnostic assessment, which aims to reveal not only students' final answers but also differences in reasoning quality and misconception patterns.

4. Comparative Analysis of Instruments

Comparison with previous diagnostic instruments shows that the developed instrument has both advantages and limitations. The three-tier diagnostic instrument developed by Irfiana et al. [17] integrated critical thinking indicators to identify students' thinking skills and misconceptions in acid-base material. The IBT-based two-tier HOTS instrument developed by Suri and Andromeda [18] also showed practical value through technology-supported implementation and HOTS-oriented item design. These studies indicate that diagnostic instruments in acid-base learning have developed toward more complex formats, including multi-tier structures, critical thinking indicators, and digital testing platforms [17],[18].

Previous instruments, however, still tend to focus on identifying final conceptual errors, answer justification, confidence level, or item quality rather than mapping the sequence of students' reasoning. This limitation is important because acid-base concepts require students to connect symbolic calculations, equilibrium relationships, and conceptual interpretation [12],[13]. The instrument developed in this study addresses this limitation by integrating HOTS indicators with systems thinking sequences. HOTS-oriented assessment requires students to engage in analysis, reasoning, decision making, and problem

solving [6],[7], while systems thinking emphasizes relationships among components, interactions, cause–effect mechanisms, and system-level behavior [15],[16]. This integration allows the instrument to identify not only what misconceptions occur, but also how students' reasoning trajectories are formed.

Empirical analysis using the Rasch Model also provides an additional contribution compared with diagnostic instruments that rely mainly on classical test analysis. Rasch analysis allows the evaluation of item fit, item difficulty, reliability, separation, unidimensionality, and potential item bias, thereby providing stronger evidence of instrument quality [21],[23]. The findings of this study indicate that the developed instrument is generally valid and reliable, although several challenges remain, including school-group bias, homogeneous responses on basic items, and uneven distribution of item difficulty. These limitations suggest that further testing with more diverse student groups, including high-cognitive-level school strata, is needed to strengthen the generalizability and sensitivity of the instrument in detecting variations in students' systems thinking trajectories.

CONCLUSION

The development of HOTS-based diagnostic instruments with a systems thinking approach is the goal of this study, as an answer to the challenge of assessment that has only focused on the basic ability to remember and understand. The resulting instrument consists of 20 questions with a multilevel structure designed to guide students through the stages of systemic thinking, from component identification to cause-and-effect analysis and mathematical interference.

The validation results showed that the instruments developed were considered feasible by experts and teachers, and received positive responses from students despite records regarding the duration of the work. Analysis using Rasch model reinforces the quality of the instrument with evidence of construct validity, reliability, and the ability to distinguish the trajectory of students ' abilities, although some basic questions show homogeneous responses that it is natural to measure mastery of the initial concept. The main contribution of this instrument is to provide a more meaningful formative assessment tool for teachers in identifying misconceptions and designing learning interventions. This instrument has the potential to be used as an evaluation model that balances the demands of HOTS with systemic thinking skills in its integration, thus supporting more contextual and problem-solving-oriented Chemistry learning. Further research is suggested to be able to use the instrument on a large scale implementation to strengthen generalization and uncover more complex misconceptions experienced by students.

REFERENCES

[1] "Education GPS - Indonesia - Student Performance (PISA 2022)." Accessed: Dec. 16, 2023. [Online]. Available: https://gpseducation.oecd.org/CountryProfile?primaryCountry=IDN&treshold=10&topic=PI

[2] I. N. Ramadhani and W. Sukmawati, "Analisis pemahaman literasi sains berdasarkan gender dengan tes diagnostik three-tier multiple choice," Ideas: Jurnal Pendidikan, Sosial, dan Budaya, vol. 8, no. 3, p. 781, Aug. 2022, doi: https://doi.org/10.32884/ideas.v8i3.860.

[3] N. Mukhlisa, "Miskonsepsi pada peserta didik," Speed Journal: Journal of Special Education, vol. 4, no. 2, pp. 66-76, Jan. 2021, doi: https://doi.org/10.31537/speed.v4i2.403

[4] D. Ngurah and L. Laksana, "Miskonsepsi dalam materi IPA sekolah dasar," JPI (Jurnal Pendidikan Indonesia), vol. 5, no. 2, pp. 166-175, Oct. 2016, doi: https://doi.org/10.23887/jpi-undiksha.v5i2.8588.

[5] O. Deni, E. Nugroho, and M. A. Prayitno, "Analisis miskonsepsi peserta didik dalam memahami konsep kimia dengan menggunakan tes diagnostik TTMC," Jurnal Education and Development, vol. 9, no. 1, pp. 72-72, Jan. 2021, doi: https://doi.org/10.37081/ed.v9i1.2300.

[6] P. Dewi, Elvinawati, and R. Elvia, "Pengembangan butir soal HOTS untuk menguji kemampuan berpikir tingkat tinggi siswa di MA Negeri 2 Kota Bengkulu," Alotrop, vol. 5, no. 2, pp. 141-148, Aug. 2021, doi: https://doi.org/10.33369/atp.v5i2.17119.

[7] B. Anggara, "Pengembangan soal Higher Order Thinking Skills sebagai tes diagnostik miskonsepsi matematis siswa SMA," Algoritma: Journal of Mathematics Education, vol. 2, no. 2, pp. 176-191, Dec. 2020, doi: https://doi.org/10.15408/ajme.v2i2.18387.

[8] I. K. Mustika, "Optimalisasi tes diagnostik berbasis IT dalam meningkatkan mutu pembelajaran Bahasa Bali pada Kurikulum Merdeka di SMA Negeri 1 Seririt," Kalangwan Jurnal Pendidikan Agama, Bahasa dan Sastra, vol. 12, no. 2, pp. 13-22, Sep. 2022, doi: https://doi.org/10.25078/kalangwan.v12i2.1674.

[9] I. N. Diartha, W. Wildan, and M. Muntari, "Penilaian kinerja (performance assessment) dalam pembelajaran kimia," Jurnal Pijar MIPA, vol. 11, no. 1, pp. 65-69, Mar. 2016, doi: https://doi.org/10.29303/jpm.v11i1.64.

[10] N. L. Azizah, L. Mahardiani, and D. S. Yamtinah, "Analisis miskonsepsi dengan tes diagnostik two-tier multiple choice dan in-depth interview pada materi asam basa," Jurnal Pendidikan Kimia, vol. 11, no. 2, pp. 168-177, Dec. 2022, doi: https://doi.org/10.20961/jpkim.v11i2.60345.

[11] T. Fan, J. Song, and Z. Guan, "Integrating diagnostic assessment into curriculum: A theoretical framework and teaching practices," Language Testing in Asia, vol. 11, no. 1, Dec. 2021, doi: https://doi.org/10.1186/s40468-020-00117-y.

[12] S. Fajrin, A. Haetami, D. Muhamad, and A. Marhadi, "Identifikasi kesulitan belajar kimia siswa pada materi pokok larutan asam dan basa di kelas XI IPA 2 SMA Negeri 1 Wolowa Kabupaten Buton," Jurnal Pendidikan Kimia FKIP Universitas Halu Oleo, vol. 5, no. 1, 2020, doi: https://doi.org/10.36709/jpkim.v5i1.13106.

[13] R. Tri Astuti and H. Marzuki, "Analisis kesulitan pemahaman konsep pada materi titrasi asam basa siswa SMA," Orbital: Jurnal Pendidikan Kimia, vol. 1, no. 1, pp. 22-27, Feb. 2018, doi: https://doi.org/10.19109/ojpk.v1i1.1862.

[14] P. G. Mahaffy, E. J. Brush, J. A. Haack, and F. M. Ho, "Journal of Chemical Education call for papers - Special issue on reimagining chemistry education: Systems thinking, and green and sustainable chemistry," J. Chem. Educ., vol. 95, no. 10, pp. 1689-1691, Oct. 2018, doi: https://doi.org/10.1021/acs.jchemed.8b00764.

[15] V. Talanquer and A. R. Szozda, "An educational framework for teaching chemistry using a systems thinking

approach," J. Chem. Educ., vol. 101, no. 5, pp. 1785-1792, May 2024, doi: https://doi.org/10.1021/acs.jchemed.4c00216.

[16] P. G. Mahaffy, A. Krief, H. Hopf, G. Mehta, and S. A. Matlin, "Reorienting chemistry education through systems thinking," Nat. Rev. Chem., vol. 2, no. 4, 2018, doi: https://doi.org/10.1038/s41570-018-0126.

[17] A. Irfiana, W. Sumarni, J. Kimia, F. Matematika, D. Ilmu, and P. Alam, "Desain instrumen tes three-tier multiple choice bermuatan critical thinking skills untuk mengukur keterampilan berpikir kritis siswa SMA terkait materi asam basa," Chemistry in Education, vol. 11, no. 2, pp. 101-110, Dec. 2022, doi: https://doi.org/10.15294/chemined.v11i2.55071.

[18] S. A. Suri and A. Andromeda, "Development of IBT based two-tier Higher Order Thingking Skills (HOTS) test instruments on acid-base titration materials for SMA/MA students," Entalpi Pendidikan Kimia, vol. 3, no. 1, pp. 58-65, Feb. 2022, doi: https://doi.org/10.24036/epk.v3i1.249.

[19] M. Wilson, Constructing Measure: An Item Response Modeling Approach. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2005.

[20] L. R. Aiken, Perspective of Individual Difference: Assessment of Intellectual Functioning, 2nd ed. New York, NY, USA: Plenum Press, 1996.

[21] B. Sumintono and W. Widhiarso, Aplikasi Pemodelan Rasch pada Assessment Pendidikan. Cimahi, Indonesia: Trim Komunikata, 2015.

[22] E. Betley, E. J. Sterling, and A. L. Porzecanski, "The value of systems thinking in a rapidly changing world," Lessons in Conservation, vol. 11, pp. 5-8, 2021, doi: https://doi.org/10.5531/cbc.linc.11.1.1.

[23] T. Strachan, U. H. Cho, T. Ackerman, S.-H. Chen, J. De La Torre, and E. H. Ip, "Evaluation of the linear composite conjecture for unidimensional IRT scale for multidimensional responses," Appl. Psychol. Meas., vol. 46, no. 5, pp. 347-360, Jul. 2022, doi: https://doi.org/10.1177/01466216221084218.

[24] S. Mulyanti and S. Rahmania, "Pengembangan instrumen tes penguasaan konsep senyawa alkil halida: Analisis validitas model Rasch," Jurnal Zarah, vol. 10, no. 1, pp. 21-27, Jun. 2022, doi: https://doi.org/10.31629/zarah.v10i1.4161.

[25] N. Ngadi, "Analisis model Rasch untuk mengukur kompetensi pengetahuan siswa SMKN 1 Kalianget pada mata pelajaran perawatan sistem kelistrikan sepeda motor," Jurnal Pendidikan Vokasi Otomotif, vol. 6, no. 1, pp. 1-20, Dec. 2023, doi: https://doi.org/10.21831/jpvo.v6i1.63479.

[26] M. Ibnu, B. Indriyani, H. Inayatullah, and Y. Guntara, "Aplikasi Rasch Model: Pengembangan instrumen tes untuk mengukur miskonsepsi mahasiswa pada materi mekanika," in Seminar Nasional Hari Pendidikan Nasional FKIP Untirta 2019, 2019, pp. 205-210.