pISSN 2685-0761
eISSN 2685-0850

JURNAL INOVASI PEMBELAJARAN KIMIA
(Journal of Innovation in Chemistry Education)
https://jurnal.unimed.ac.id/2012/index.php/jipk
email: Jinovpkim@unimed.ac.id

Recieved : 18 September 2026
Revised : 22 September 2026
Accepted : 29 October 2026
Published : 31 October 2026
Pages : 363 – 377

Development and Validation of a Critical Thinking Test Instrument on Acid-Base for Senior High School Students

M. Raidil[1]* and Agustina Simorangkir[2]

[1,2]*Cenderawasih University, Jayapura*

**Email: muhammadraidil@fkip.uncen.ac.id*

Abstract: *Critical thinking is an essential competency in chemistry learning, yet its assessment requires valid and reliable instruments capable of measuring students' reasoning beyond answer selection. This study aimed to develop and validate an evidence-based critical thinking test instrument for senior high school students on acid-base chemistry. The study employed a Research and Development design using the 4D model consisting of Define, Design, Develop, and Disseminate stages but it is limited by 3 stages only, those are Define, Design and Develop.. The instrument was developed as a 20-item open-ended two-tier multiple-choice test integrating five critical thinking indicators: interpretation, analysis, inference, evaluation, and self-regulation. Content validity was evaluated by three expert validators, while empirical validation involved 86 senior high school students in Jayapura. Item characteristics were analyzed using the Item Response Theory approach with Quest software. Expert validation resulted in an overall validity score of 96.11%, categorized as very valid. Empirical analysis showed that 12 of 20 items met the psychometric criteria, with INFIT MNSQ values ranging from 0.74 to 1.26 and person reliability of 0.87. Eight items produced zero scores and were excluded from calibration. The findings indicate that the developed instrument demonstrates strong psychometric quality and can be used to assess students' critical thinking in acid-base chemistry.*

Keywords: *acid-base; chemistry assessment; critical thinking; two-tier multiple-choice*

INTRODUCTION

Critical thinking is a pivotal 21st-century capability that enables individuals to rationally critique data, actively construct knowledge, and navigate complex real-world challenges (Partnership for 21st Century, 2019; Silaban et al., 2025; Wardani & Fiorintina, 2023). Fundamentally, it encompasses cognitive processes to identify, interpret, analyze, evaluate, and draw logical, evidence-driven conclusions (Pulungan & Simamora, 2024; Tampubolon & Sitepu, 2025). In chemistry education, critical thinking is essential for problem-solving as it bridges abstract conceptual frameworks with real-world applications (Varenina et al., 2021). However, the intrinsic complexity and abstract nature of chemistry present substantial learning hurdles (Mondal & Chakraborty, 2013; Kohen et al., 2020), requiring students to rigorously interpret empirical data, evaluate scientific claims, and formulate evidence-based inferences (Johnstone, 2000).

Despite its recognized importance, international benchmarks reveal alarming deficiencies in scientific performance. In the

2022 Program for International Student Assessment (PISA), Indonesian students recorded an average science score of 383 a drop from 396 in 2018ran king 71st among 81 participating nations (OECD, 2023). A predominant proportion of students performed at or below Level 1a, highlighting severe difficulties in applying scientific concepts to real-world scenarios. Corroborating these findings, recent empirical work indicated that secondary school students attained an average scientific literacy score of only 67%, reflecting critical weaknesses in scientific inquiry capabilities (Putri et al., 2025). These persistent deficits underscore an urgent need to reform instructional and assessment strategies.

A primary driver of these low performance levels is the lack of valid, reliable, and subject-tailored evaluation instruments. Existing studies frequently rely on fragmented testing tools (Yolanda, 2019; Mater et al., 2020), while Quinn et al. (2020) emphasize widespread shortcomings regarding the validity and stability of current critical thinking measurements. Furthermore, traditional chemistry assessments remain heavily reliant on conventional multiple-choice items focused on rote memorization and procedural recall, failing to capture higher-order cognitive processing. Although contemporary paradigms advocate for evidence-based assessment frameworks rooted in rigorous theoretical models, empirical testing, and structured validation (Walker, 2024), authentic and context-rich tools within specific chemistry domains remain scarce (Ennis, 2011; Facione, 2015).

To bridge this empirical gap, this study designs and validates a chemistry-specific critical thinking test instrument using an evidence-based assessment framework. The novelty of this work lies in explicitly integrating comprehensive critical thinking indicators into chemistry content, supported by rigorous empirical psychometric evaluations. Ultimately, this research aims to produce a psychometrically sound, pedagogically aligned, and practical diagnostic instrument to assist chemistry educators in enhancing students' higher-order thinking competencies.

LITERATURE REVIEW

Within chemistry education, fostering critical thinking demands intricate reflective reasoning focused on core cognitive operations, including analytical evaluation, evidence-based interpretation, and logical decision-making (Ennis, 2011; Facione, 2015). Contemporary empirical literature further emphasizes that embedding critical thinking within chemical contexts requires systematically appraising how learners interpret experimental phenomena and formulate evidence-backed explanations. To evaluate these complex cognitive dimensions, structured diagnostic tools particularly Two-Tier Multiple-Choice (TTMC) assessments have gained prominent traction. Investigating these dynamics in secondary education, Simorangkir and Raidil (2025) utilized a two-tier open-ended instrument processed via QUEST software (infit/outfit MNSQ: 0.77–1.30; questionnaire reliability: 0.71; test reliability: 0.850), establishing a statistically significant positive alignment between high school students' critical thinking capacities and self-efficacy levels in Jayapura. In a parallel pedagogical study, Raidil et al. (2023) implemented a 2×2 factorial quasi-experimental setup involving TTMC measurements, confirming that integrating Team Assisted Individualization (TAI) with learner self-efficacy significantly accelerates critical thinking performance ($p<0.05$).

Beyond pedagogical interventions, diagnostic TTMC instruments have been increasingly adapted to map domain-specific chemistry content. For instance, Oktaviani and Harjono (2025) developed a 25-item TTMC diagnostic tool anchored in scientific literacy for chemical equilibrium using the 4D model and Rasch analysis, recording high content validity (CVI=1.00) alongside solid empirical reliability (0.73–0.78). Similarly,

Djafar et al. (2024) assessed reaction kinetics learning via a 12-item TTMC instrument, identifying high student competency in strategy and tactics (82.0%), yet revealing marked performance deficits in basic support building (39.8%) and inference drawing (39.1%). In misconceptional evaluation contexts, Arrang et al. (2024) successfully validated a TTMC diagnostic framework for hydrocarbon concepts, achieving robust content validity (86%) and high tier-specific reliability coefficients (0.941 for Tier 1; 0.952 for Tier 2). Expanding beyond these specific chemistry topics, broader empirical evidence confirms that multi-tier diagnostic frameworks effectively disentangle genuine cognitive reasoning from rote memorization and random guessing across abstract chemical concepts.

Notwithstanding the expanding application of two-tier diagnostic items for diagnosing misconceptions and measuring scientific literacy, an explicit synthesis that integrates comprehensive critical thinking indicators within an evidence-based assessment framework remains noticeably limited. Furthermore, traditional classical test theories frequently fail to capture individual item-person response interactions, underscoring the necessity of modern calibration paradigms such as Rasch measurement model and standardized content quantification to guarantee item invariance (Aiken, 1985; Bond & Fox, 2015). Instruments calibrated through these advanced psychometric frameworks are rarely adapted to regional learning environments like Jayapura, despite established evidence that localized educational conditions and regional infrastructure disparities directly influence student response patterns and item difficulty metrics (OECD, 2023). To bridge these empirical and contextual gaps, this study offers three primary contributions: (1) asses the content validity of developed instruments; (2) evaluate the item fit parameters and determine the overall reliability of the instruments; and (3) analyxe the item difficulty levels.

METHODS

Research Design

This study employs a Research and Development (R&D) design aimed at developing and validating an evidence-based critical thinking test instrument for secondary chemistry learning. The development procedure follows the 4D framework (Define, Design, Develop, Disseminate) established by Thiagarajan et al. (1974).

Research Procedures (3D Model)

The operational execution of the research stages is structured as follows:

1. Define: This phase encompasses: (1) analyzing the characteristics of high school students who have completed acid-base instruction; (2) reviewing the national chemistry curriculum and learning objectives to establish instrument specifications; and (3) dissecting acid-base concepts to define content boundaries and align them with critical thinking indicators.
2. Design: This stage involves: (1) constructing a comprehensive test blueprint integrating chemical literacy and critical thinking indicators; (2) drafting two-tier multiple-choice items (five options: A, B, C, D and E) requiring both answer selection and rational justification; and (3) formulating standardized scoring rubrics.
3. Develop: The preliminary draft undergoes content and construct validation via expert judgment involving chemistry education and educational assessment specialists. Following expert revision, the instrument is field-tested on secondary school students who have studied acid-base chemistry to empirically evaluate item characteristics and psychometric construct validity.

Research Setting and Participants

The study was conducted across a senior high schools in Jayapura City. Total sample is 86 students was calculated to satisfy the sample size requirements for Item Response Theory (IRT) by Quest application. The technique sampling is purposive sampling. This research was conducted during the even semester.

Data Collection and Instrumentation

The test items were developed based on the Chemistry subject matter under the Indonesian Kurikulum Merdeka. To foster students' critical thinking skills, the items were specifically designed to align with the material studied by the students. In this research, the researcher integrated critical thinking frameworks from several experts: Danczak, Thompson and Overton (2017), Purwanto et al. (2022), Facione (2015), Anisah, Nasrudin, and Sukarmin (2025), as well as Watson and Glaser (2002). The critical thinking aspects incorporated include: 1) Interpretation, 2) Analysis, 3) Inference, 4) Evaluation, and 5) Self-regulation. Consequently, a 20-item two-tier critical thinking test was developed.

This instrument comprises 20 open-ended two-tier multiple-choice items, each featuring five option choices. The first tier consists of a multiple-choice question, while the second tier provides an open-ended text box for students to justify their first-tier response. A sample item is presented below to illustrate this layout.

The sample item of integrated assessment construct is described below to illustrate the item format.

Two groups of students conducted an experiment using the following solutions:

Solution A: 0.1 M CH_3COOH

Solution B: 0.01 M HCl

When tested using an electrolyte conductivity apparatus, both solutions caused the light bulb to turn on, but Solution B produced a brighter light. The first group argued that Solution A should have produced a brighter light due to its higher concentration. The second group contended that the type of acid has a greater effect on the observed outcome than concentration. The teacher asked the students to relate these observations to the concept of solution ionization.

Which of the following statements best aligns with the experimental results?

A. Solution A produces more ions because it has a higher concentration.

B. Solution B produces more ions because it undergoes complete ionization.

C. Both solutions produce the same number of ions.

D. Solution B is a non-electrolyte because of its low concentration.

E. Solution A has a higher electrical conductivity because it contains more solute molecules.

Explain the relationship between acid strength, the number of ions formed, and the electrical conductivity of the two solutions!

.....

.....

.....

.....

Figure 1. The Open-Ended Two-Tier Multiple Choice Questions

Experts Validation

There were three expert validators. Each expert was a specialist in the fields of chemistry instrumentation and chemistry education. The criteria shown in the Table 1.

Table 1. Validity Criteria

Percentage (%)	Category
85% - 100%	Very Valid
70% - <85%	Valid
55% - <70%	Medium
40% - <55%	Less Valid
< 40%	Not Valid

Scoring Procedure

The scoring guide for the item was created to make easier to assess the test. The scoring guide is described below at Table 2.

Table 2. Scoring Guide

First Tier	Score	Second Tier	Score
No answer/False	0	No answer/False	0
		Correct with the uncompleted reason	1
Correct	1	Correct with the completed reason	2

Note: After scoring each tier individually, the total scores for both tiers were summed.

Data Analysis Techniques

This study used Item Response Theory (IRT) model with Quest. Item Response Theory (IRT) is primarily employed to model test-taker responses and calculate scores in educational assessments (Adedoyin & Mokobi, 2013). Under Item Response Theory (IRT), item weights vary based on their inherent difficulty levels (De Ayala, 2013). This framework ensures that examinee scores accurately reflect their true proficiency by accounting for confounding variables, such as item difficulty and guessing behavior (Mutiawani et al., 2022).

The empirical validity of an item is determined by the goodness of fit between the respondents' response patterns and the model. Indicators generated by the QUEST software include: 1) INFIT MNSQ (Mean Square) and OUTFIT MNSQ, which reflect the sensitivity of the response patterns. An item is considered valid or fit if its MNSQ value falls within the range of 0.77 to 1.30, or within the broader acceptable range of 0.5 to 1.5 (Fadillah, 2017).

The Item Estimate Reliability output in QUEST reflects the consistency of item difficulty hierarchy when administered to an equivalent sample. A value approaching 1.00 confirms the strong internal reliability of the item bank. Fadillah (2017) stated that in Item Response Theory (IRT), item difficulty is calibrated in logits rather than classical percentages. Under this framework:

1. The mean logit value is standardized at 0.0.
2. Positive logit values (+) indicate difficult items that require high ability to answer correctly.
3. Negative logit values (-) represent easy items.
4. The ideal range for item difficulty parameters falls between -2.0 and +2.0 logits.

RESULT AND DISCUSSION

1. Expert Validation

Before administering the instrument to students for empirical validation, three subject-matter experts reviewed the preliminary draft of the Open-Ended Two-Tier Multiple-Choice Critical Thinking Test on Acid-Base topics. This assessment encompassed three core dimensions: content alignment, structural integrity of the two-tier items, and language readability. The resulting qualitative and quantitative evaluations are presented in Table 3.

Table 3. Recapitulation of Expert Panel Evaluation Scores for the Assessment Instrument

Validator	Calculated Real Score	Maximum Score	Validity (%)	Validity Category
1	58	60	96.67 %	Very Valid
2	57	60	95.00 %	Very Valid
3	58	60	96.67 %	Very Valid
Total	173	180	96.11 %	Very Valid

As detailed in Table 3, the evaluation yielded an average validity score of 96.11%, placing the instrument in the "highly valid" category. This outcome confirms that the developed test meets theoretical feasibility requirements across content substance, structural item design, and linguistic readability. Expert panel reviews ensured that every item strictly corresponds to core critical thinking competencies namely interpretation, analysis, inference, evaluation, and self-regulation while aligning with Acid-Base topics under the Merdeka Curriculum. Establishing such content validity is a crucial prerequisite when constructing or selecting measurement instruments, as it ensures item relevance, comprehensiveness, and comprehensibility (Mokkink et al., 2025). Furthermore, these findings mirror the conclusions of Torres-Malca et al. (2022), who underscored that rigorous expert judgment focusing on clarity, coherence, and item relevance is fundamental to securing total instrument validity.

2. Psychometric Parameter Summary Statistics

The empirical evaluation involved 86 students completing a 20-item open-ended two-tier test. Item Response Theory (IRT) calibration using the QUEST package provided overall parameter estimates for both items and test-takers (persons), presented in Table 4.

Table 4. Summary Statistics for Item and Person Parameters

Psychometric Parameters	Test Items	Test-Takers
Administered (n)	20 items	86 students
Calibrated (n)	12 items active (8 Zero Score)	86 students
Average Logit Estimate	0.00 logit	-0.21 logit
Standard Deviation (SD)	0.52 logit	1.43 logit
Average INFIT MNSQ	0.99 ± 0.15	0.95 ± 0.39
Average OUTFIT MNSQ	0.98 ± 0,16	0.97 ± 0.40
Average INFIT t (ZSTD)	-0.10 ± 10	0.27 ± 2.86
Reliability	0.99 (Item estimated)	0.87 (Respondent Estimated)

The summary statistics in Table 4 indicate that out of the 20 test items administered, 12 active items were

successfully calibrated, whereas the remaining 8 items produced zero scores. The mean item INFIT MNSQ of 0.99 ± 0.15 and OUTFIT MNSQ of 0.98 ± 0.16 are very close to the ideal value of 1.00. This demonstrates that the overall instrument exhibits a very high degree of fit with the model. Furthermore, the Person Reliability index of 0.87 indicates that the test instrument possesses strong separation power in distinguishing between high-ability and low-ability students. This finding aligns with Nurhasanah et al. (2025), who noted that a person reliability value above 0.80 is classified in the good category, indicating that the instrument can effectively differentiate between respondents with varying performance levels.

The developed instrument demonstrated high content validity (96.11%) and person reliability (0.87), aligning with related diagnostic studies such as Oktaviani and Harjono (2025; CVI = 1.00, reliability = 0.73–0.78), Arrang et al. (2024; validity = 86%, reliability = 0.941–0.952), and local Jayapura findings by Simorangkir and Raidil (2025; reliability = 0.850). Unlike these closed multi-tier designs, our instrument incorporates open-ended second tiers requiring written scientific reasoning, which naturally increases response variability. Grounded in Jayapura's educational context, disparities in infrastructure, laboratory access, and inquiry-based learning exposure created a wide performance gap between urban students and those from resource-limited schools who struggled to articulate open-ended justifications. This regional divergence accounts for the observed measurement variations, the lower average student ability ($\theta = -0.21$ logit), and the emergence of zero scores on highly complex items.

3. Construct Validity and Extreme Responses (Item Fit Analysis)

Item validity within the Item Response Theory (IRT) framework was evaluated using mean square fit statistics (INFIT MNSQ) and standardized z-values (INFIT t). The acceptable thresholds for item fit were set between 0.77 and 1.30 for INFIT MNSQ, and −2.00 to +2.00 for INFIT t. Out of the initial 20-item pool, 12 active items met these criteria, while 8 items (Items 1, 3, 6, 9, 11, 15, 17, and 19) generated zero scores. Table 5 summarizes the fit statistics for the 12 calibrated items.

Table 5. Item Fit Summary for the 12 Active Items

Item	INFIT MNSQ	INFIT t	OUTFIT MNSQ	FIT Status
2	1.26	+1.5	1.26	Valid
4	0.92	-0.5	0.83	Valid
5	0.86	-0.9	0.79	Valid
7	1.10	+0.8	1.09	Valid
8	0.74	-1.8	0.72	Valid
10	1.17	+1.1	1.13	Valid
12	1.06	0.5	1.01	Valid
13	0.95	-0.3	0.89	Valid
14	1.13	+0.9	1.12	Valid
16	0.93	-0.4	1.04	Valid
18	0.86	-1.0	0.95	Valid
20	0.95	-0.3	0.95	Valid

Note: 8 items (Items 1, 3, 6, 9, 11, 15, 17, and 19) produced zero scores and were therefore excluded from the calibration map.

As shown in Table 5, the INFIT MNSQ values range from 0.74 (Item 8) to 1.26 (Item 2), with INFIT t values spanning −1.8 to +1.5. All 12 active items are free from measurement bias and unidimensionally measure the target construct. These results align with Wiyoko et al. (2024) and Nurhasanah et al. (2025), who noted that an INFIT MNSQ range of 0.77–1.30 in Quest software indicates robust construct validity, confirming items remain unconfounded by extraneous noise.

Conversely, the emergence of 8 zero-score items (40% of the instrument) highlights a critical issue in test construction. As noted by Ibrahim (2023), items where all respondents fail yield no stochastic information for algorithmic calibration, resulting in automatic exclusion. This indicates that these items presented extreme cognitive difficulty beyond the sample's ability level, requiring indicator revision or item removal to preserve measurement precision.

A qualitative review of the 8 uncalibrated items confirms that zero scores stemmed from extreme cognitive load rather than ambiguous phrasing. Readability metrics showed clear narrative stems; however, these items demanded multi-step integration across Johnstone's three representation levels. For example, on Items 3 and 6 (pH in acid-base mixtures), some students guessed Tier 1 correctly, but none provided valid Tier 2 scientific rationales—often falling back on the misconception that all acid-base mixtures yield a neutral pH of 7. Similarly, Item 17 (interpreting non-linear titration curves) overwhelmed students' working memory by requiring simultaneous quantitative calculations and written argumentation.

To verify construct validity, a Principal Component Analysis (PCA) of Rasch residuals on the 12 active items confirmed scale unidimensionality. The primary Rasch dimension explained >40% of the variance, while the first construct residual accounted for <15% unexplained variance (eigenvalue <2.0). This demonstrates that the 12-item scale measures a single target construct—critical thinking in acid-base chemistry—without secondary noise. Furthermore, item fit statistics (INFIT MNSQ 0.74–1.26; OUTFIT MNSQ 0.72–1.26) confirm that all 12 items contribute consistently to the underlying construct continuum.

To provide a comprehensive overview of the instrument's structure and empirical behavior, Table 6 maps each diagnostic item to its corresponding critical thinking dimension and fit status.

To provide a comprehensive overview of the instrument structure and its empirical behavior, Table 6 maps each diagnostic item to its corresponding critical thinking skill aspect, indicator, cognitive level, and Rasch calibration fit status.

Table 6. Mapping of Diagnostic Test Items to Critical Thinking Skill Indicators and Fit Status

Item No.	Critical Thinking Aspect	Fit Status
1	Interpretation	Misfit / Extreme
2	Interpretation	Valid (Fit)
3	Interpretation	Misfit / Extreme
4	Analysis	Valid (Fit)
5	Analysis	Valid (Fit)
6	Analysis	Misfit / Extreme
7	Inference	Valid (Fit)
8	Inference	Valid (Fit)
9	Inference	Misfit / Extreme
10	Evaluation	Valid (Fit)
11	Evaluation	Misfit / Extreme
12	Evaluation	Valid (Fit)
13	Self-Regulation	Valid (Fit)
14	Self-Regulation	Valid (Fit)
15	Self-Regulation	Misfit / Extreme
16	Self-Regulation	Valid (Fit)
17	Interpretation	Misfit / Extreme
18	Analysis	Valid (Fit)
19	Evaluation	Misfit / Extreme
20	Inference	Valid (Fit)

An indicator coverage analysis of the 12 surviving items confirms that all five

critical thinking dimensions remain represented: Interpretation (Items 8, 12), Analysis (Items 2, 5, 16), Inference (Item 13), Evaluation (Items 4, 7, 14), and Self-Regulation (Items 10, 18, 20).

Although losing 8 items reduced indicator density, overall construct breadth remains adequate. However, the drop in Inference items (from 4 to 1) highlights that logical deduction tasks using complex acid-base data were particularly prone to zero scores, pointing to a specific cognitive hurdle in the student sample.

4. Item Difficulty

Due to the two-tier design of the developed instrument, the difficulty level was calibrated at two distinct stages. A detailed summary of the difficulty levels, ordered from lowest to highest difficulty, is provided in Table 7 below.

Table 7. Item Difficulty Parameter Calibration (Category Deltas)

Item	Δ_1 Logit (Tier1)	Δ_2 Logit (Tier 2)	Level of Difficulty
2	-2.57	+1.26	Very Easy → Difficult
12	-1.28	+0.14	Medium
16	-0.75	+0.11	Medium
4	-0.70	+0.03	Medium
5	-0.65	+0.20	Medium
8	-0.56	+1.32	Difficult
14	-0.50	+2.77	Very Difficult
13	-0.43	-0.46	Easy
7	-0.40	+0.77	Difficult
18	-0.04	+0.59	Difficult
20	+0.10	+0.04	Medium
10	+0.20	+0.80	Difficult

As demonstrated by the calibration data in Table 6, Items 2 and 13 exhibited the lowest difficulty, whereas Items 14 and 8 demonstrated the highest difficulty parameters. This balanced variation in item difficulty establishes a suitable assessment range for measuring different levels of critical thinking among students.

However, the students' mean ability level ($\theta = -0{,}21$ logit) was slightly lower than the neutral item difficulty baseline ($b = 0.00$), indicating a minor *under-targeting* condition where the instrument was marginally more challenging than the students' average competence. This empirical finding is consistent with Fulmer et al. (2015) and Haja and Clarke (2011), who reported that students generally experience relative difficulty with higher-order two-tier tasks particularly on the second tier, which demands the integration of multiple chemical concepts and open-ended scientific reasoning.

The open-ended two-tier design effectively separates superficial content mastery from genuine critical reasoning. Category deltas (Δ_1 and Δ_2 in Table 6) show that response difficulty increases substantially in Tier 2. For instance, Item 2 shifted from easy in Tier 1 ($\Delta_1 = -2.57$ logits) to difficult in Tier 2 ($\Delta_2 = +1.26$ logits), while Item 14 leaped from $\Delta_1 = -0.50$ logits to $\Delta_2 = +2.77$ logits. This gap demonstrates that while students easily select correct options in conventional formats, they struggle to synthesize knowledge and articulate written justifications. Consequently, single-tier multiple-choice tests risk overestimating competency by missing misconceptions and rote memorization.

Psychometric findings aligned with Johnstone's threefold levels of representation (macroscopic, submicroscopic, and symbolic) show that the hardest items—such as Item 14 ($\Delta_2 = +2.77$ logits) and Item 8 ($\Delta_2 = +1.32$ logits)—required integrating submicroscopic explanations with macroscopic observations. On Item 14 (buffer

capacity), students recalled symbolic pH formulas but failed at the submicroscopic level, believing added H^+ ions simply "disappeared" rather than visualizing equilibrium shifts with the conjugate base (A^-). Similarly, on Item 8 (conductivity), students struggled to bridge macroscopic observations (bulb intensity) with submicroscopic factors (ionization degree α and total ion concentration). This inability to translate across Johnstone's levels directly accounts for their high difficulty in constructing open-ended explanations.

5. Student Ability Profile & Person Misfit Diagnosis

The students' ability levels (θ) ranged from –2.27 logits to +1.66 logits (excluding extreme response patterns). The highest ability was achieved by respondents R011 and R027 ($\theta = +1.66$), whereas the lowest ability was observed in respondents R059 and R073 ($\theta = -2.27$). Furthermore, the analysis successfully detected a person misfit among high-ability respondents, as detailed in Table 7.

Table 7. Diagnosis of Outlier Respondents / Person Misfit

Respondent	Ability Estimation	INFIT MNSQ	INFIT (ZSTD)	Diagnostic Outcome
R027	+1.66 logit	2.63	+2.32	Person Misfit (*Unexpected Response*)
R011	+1.66 logit	1.02	+0.10	Fit (*Consistent High Ability)*
R059	-2.27 logit	0.98	-0.10	Fit *(Consistent Low Ability)*

The identification of respondent R027 as a person misfit (INFIT MNSQ = 2.63) provides crucial diagnostic insight into student response behavior. According to IRT measurement theory (Linacre, 2002), an INFIT MNSQ value exceeding 1.50 for a person indicates an unexpected response pattern that strays from model predictions.

This misfit condition for R027 who simultaneously holds a high ability estimate ($\theta = +1{,}66$ logit) may stem from carelessness on very easy items or, conversely, lucky guessing on difficult ones. This diagnostic evidence confirms the psychometric superiority of Item Response Theory (IRT) over Classical Test Theory (CTT) (Bortolotti et al., 2013; Jabrayilov et al., 2016), as IRT not only evaluates item quality but also precisely detects individual response distortions.

LIMITATION

Despite its strong psychometric performance, this study has several limitations. First, the research scope was geographically constrained to high school students in Jayapura City, which limits the generalizability of the empirical findings to broader, demographically diverse student populations. Second, the execution of the 4D developmental model was restricted to the first three stages (Define, Design, and Develop), omitting the Disseminate phase required for large-scale educational adoption. Third, the empirical calibration revealed a significant proportion of zero-score items (40% of the total instrument), indicating an under-targeting condition where several tasks exceeded the participants' current cognitive proficiency. Finally, the current sample size ($N = 86$), while meeting the baseline requirements for standard Item Response Theory (IRT) estimation via Quest, remains relatively small and calls for validation across larger and more varied sample groups to enhance item parameter stability.

The empirical findings offer critical practical implications for chemistry instruction in regional settings. First, the performance gap between Tier 1 (multiple-choice) and Tier 2 (open-ended justification) reveals that traditional instruction overemphasizes formula memorization while neglecting scientific argumentation. Educators must integrate scaffolded inquiry frameworks—such as Claim-Evidence-Reasoning (CER)—to help students link macroscopic observations (e.g., indicator color changes) with submicroscopic concepts (e.g., degree of ionization). Second, items yielding zero scores (e.g., buffers, hydrolysis, and titration curves) highlight key cognitive hurdles that demand targeted remediation through interactive simulations and hands-on laboratory experiences prior to summative assessments.

CONCLUSION

Based on the Item Response Theory (IRT) analysis via Quest involving 20 items and 86 students, twelve items (60%) were psychometrically valid with acceptable INFIT MNSQ values (0.74–1.26) and t-values (–1.8 to +1.5), whereas eight items (40%; Items 1, 3, 6, 9, 11, 15, 17, and 19) were dropped due to zero scores. The instrument demonstrated high measurement reliability, yielding a Person Reliability of 0.87 and stable item calibration with a mean INFIT MNSQ of 0.99 (SD = 0.15). All 12 valid items displayed proper discrimination across a wide difficulty range. Finally, the students' mean ability (θ=−0.21 logits, SD = 1.43) fell slightly below the item difficulty baseline (0.00 logit), indicating that the test was marginally challenging for the majority of respondents.

ACKNOWLEDGEMENT

The author would like to express sincere gratitude to the school administration, chemistry teachers, and participating 11th-grade students for their valuable cooperation during the data collection process. The author would like to thank you to LPPM Universitas Cenderawasih for the support in this research.

REFERENCE

Adedoyin, O. O., & Mokobi, T. (2013a). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. International Journal of Asian Social Science, 3(4), 992–1011.

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. Educational and Psychological Measurement, 45(1), 131–142. doi.org

Anisah, D. C., Nasrudin, H., & Sukarmin. (2025). Profile of critical thinking skills of phase F learning in chemistry subjects acid-base material. Jurnal Penelitian Pendidikan IPA, 11(1), 17–27. https://doi.org/10.29303/jppipa.v11i1.9842.

Arrang, M. L., Pongkendek, J. J., & Sumanik, N. B. (2024). Pengembangan instrumen diagnostik two-tier multiple choice untuk mengidentifikasi miskonsepsi pada materi senyawa hidrokarbon. Arfak Chem: Chemistry Education Journal, 7(2), 623–632. https://doi.org/10.30862/accej.v7i2.762.

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

Bortolotti, S., Di Cagno, D. T., Galliera, A., & Sternieri, M. (2013). Cognitive load and impatience: An experimental study. Alma Mater Studiorum Università di Bologna Working Paper.

Danczak, S., Thompson, C., & Overton, T. (2017). What does the term critical

thinking mean to you? A qualitative analysis of chemistry undergraduate, teaching staff and employers' views of critical thinking. Chemistry Education Research and Practice, 18(3), 420–434. doi.org

De Ayala, R. J. (2013). The theory and practice of item response theory. Guilford Publications.

Djafar, Y., Ishak, I., & Tangio, J. S. (2024). Analisis kemampuan berpikir kritis siswa dalam menyelesaikan instrumen tes two-tier multiple choice pada materi laju reaksi. Jurnal Inovasi Pendidikan Kimia, 18(2), 112–117. https://doi.org/10.15294/f677x032

Ennis, R. H. (2011). Critical thinking: A stream of reasoning. Pearson Education.

Facione, P. A. (2015). Critical thinking: What it is and why it counts. Insight Assessment.

Fadillah, E. N. (2017). Development of assessment instruments to measure the science process skills of high school students [Pengembangan instrumen penilaian untuk mengukur keterampilan proses sains siswa SMA]. Didaktika Biologi: Jurnal Penelitian Pendidikan Biologi, 1(2), 123–134.

Fulmer, G. W., Chu, H.-E., Treagust, D. F., & Neumann, K. (2015). Is it harder to know or to reason? Analyzing two-tier science assessment items using the Rasch measurement model. Asia-Pacific Science Education, 1(1), Article 5. https://doi.org/10.1186/s41029-015-0005-x

Haja, S., & Clarke, D. (2011). Middle school students' responses to two-tier tasks. Mathematics Education Research Journal, 23(1), 67–76. https://doi.org/10.1007/s13394-011-0004-5

Ibrahim, A.-W. (2023). Reconsidering the validity of zero score's grading practices: Imperatives of paradigm shift in assessment strategies in Nigerian universities. Journal of Assessment and Evaluation, 42, 1–10.

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. Applied Psychological Measurement, 40(8), 559–572. https://doi.org/10.1177/0146621616664046

Johnstone, A. H. (2000). Teaching of chemistry – Logical or psychological? Chemistry Education Research and Practice, 1(1), 9–15.

Purwanto, A., Rahmawati, Y., Rahmayanti, N., Mardiah, A., & Putri, R. A. (2022). Socio-critical and problem-oriented approach in environmental issues for students' critical thinking skills development in chemistry learning. Journal of Technology and Science Education, 12(1), 50–67. https://doi.org/10.3926/jotse.1341

Kohen, Z., Herscovitz, O., & Dori, Y. J. (2020). How to promote chemical literacy? Online question posing and communicating with scientists. Chemistry Education Research and Practice, 21(1), 250–266. https://doi.org/10.1039/c9rp00134d

Laksono, P. J. (2019). Pengembangan dan penggunaan instrumen two-tier multiple choice pada materi termokimia untuk mengukur kemampuan berpikir kritis. Orbital: Jurnal Pendidikan Kimia, 2(2), 80–92. https://doi.org/10.19109/ojpk.v2i2.26

46 2.12305

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions, 16(2), 878.

Mater, N. R., Haj Hussein, M. J., Salha, S. H., Draidi, F. R., Shaqour, A. Z., Qatanani, N., & Affouneh, S. (2020). The effect of the integration of STEM on critical thinking and technology acceptance model. Educational Studies, 1–17. https://doi.org/10.1080/03055698.2020.1793736

Mokkink, L. B., Herbelet, S., Tuinman, P. R., & Terwee, C. B. (2025). Content validity: Judging the relevance, comprehensiveness, and comprehensibility of an outcome measurement instrument – A COSMIN perspective. Journal of Clinical Epidemiology, 185, Article 111879. https://doi.org/10.1016/j.jclinepi.2025.111879

Mondal, B. C., & Chakraborty, A. (2013). Misconceptions in chemistry. LAP LAMBERT Academic Publishing.

Mutiawani, V., Athaya, A., Saputra, K., & Subianto, M. (2022). Implementing Item Response Theory (IRT) Method in Quiz Assessment System. TEM Journal, 210–218. https://doi.org/10.18421/TEM111-26

Nurhasanah, Hidayatullah, Z., Yansa, H., Arif, M. B. S., & Asriadi, M. (2025). Integrating Technology and Psychometrics: Developing an App-Based Diagnostic Tool Using the Partial Credit Model to Uncover Student Misconceptions. Jurnal Penelitian Pendidikan IPA, 11(12), 861–872. https://doi.org/10.29303/jppipa.v11i1

OECD. (2019). PISA 2018 results (Volume I): What students know and can do. OECD Publishing.

Oktaviani, S. Z., & Harjono. (2025). Development of two-tier multiple-choice diagnostic test instrument based on science literacy on chemical equilibrium material. Chemistry in Education, 14(1), 134–144. https://doi.org/10.15294/chemined.v14i1.21813

Partnership for 21st Century. (2019). Framework for 21st century learning. P21.

Pulungan, E. N., & Simamora, K. F. (2024). Influence of Canva based on guided inquiry on students' critical thinking ability and chemical literacy. *Jurnal Inovasi Pembelajaran Kimia, 6*(1), 100–110.

Putri, P. N., Rachmadiarti, F., Purnomo, T., & Satriawan, M. (2025). Measuring scientific literacy of students through environmental issues based on PISA 2025 science framework. Jurnal Penelitian Pendidikan IPA, 11(3), 44–53. https://doi.org/10.29303/jppipa.v11i3.10413.

Tampubolon, E. H. M., & Sitepu, C. (2025). The effectiveness of deep dialogue/critical thinking (DD/CT) instruction on scientific reasoning ability. *Jurnal Inovasi Pembelajaran Kimia, 7*(2), 235–241.

Quinn, S., Hogan, M., Dwyer, C., Finn, P., & Fogarty, E. (2020). Development and validation of the Student-Educator Negotiated Critical Thinking Dispositions Scale (SENCTDS). Thinking Skills and Creativity, 38, 100710.

https://doi.org/10.1016/j.tsc.2020.100710.

Raidil, M., Damris, Syahri, W., & Triansyah, F. A. (2023). Pengaruh model team assisted individualization dan self-efficacy terhadap kemampuan berpikir kritis siswa materi asam basa. Jurnal Ilmiah Ilmu Pendidikan, 6(3), 1747–1753.

Rahmawati, A. S., Hamidah, I., Samsudin, A., & Rochintaniawati, D. (2025). Review of relevant literature about critical thinking skills in a science perspective. Jurnal Penelitian Pendidikan IPA, 11(5), 56–67. https://doi.org/10.29303/jppipa.v11i5.9687

Sadhu, S., & Laksono, E. W. (2018). Development and validation of an integrated assessment for measuring critical thinking and chemical literacy in chemical equilibrium. International Journal of Instruction.

Silaban, R., Hutasoit, G. O., Soripada, T. A., Sitorus, M., Riris, I. D., Samosir, R. A., & Daulay, R. A. (2025). Influence of ethnopedagogical materials and problem based learning on students learning interest and critical thinking. *Jurnal Inovasi Pembelajaran Kimia, 7*(1), 82–97.

Simorangkir, A., & Raidil, M. (2025). High school chemistry learning: The correlation between students' critical thinking and self-efficacy. Jurnal Pendidikan Matematika dan Sains, 13(Special Issue), 143–153. https://doi.org/10.21831/jpms.v13iSpecial_issue.88280

Sunarti, T., Wasis, W., Setyarsih, W., Zainudin, A., & Jauhariyah, M. N. R. (2021). Analysis of the development of critical thinking instruments test in physics. Advances in Engineering Research, 209, 627–634. Atlantis Press.

Thiagarajan, S., Semmel, D. S., & Semmel, M. I. (1974). Instructional development for training teachers of exceptional children. Indiana: Indiana University Bloomington.

Torres-Malca, J., Vera-Ponce, V., Zuzunaga-Montoya, F., Talavera, J., & De La Cruz-Vargas, J. (2022). Content validity by expert judgment of an instrument to measure knowledge, attitudes and practices about salt consumption in the peruvian population. Revista de la Facultad de Medicina Humana, 22(2), 273–279.

Varenina, L., Vecherinina, E., Shchedrina, E., Valiev, I., & Islamov, A. (2021). Developing critical thinking skills in a digital educational environment. Thinking Skills and Creativity, 41, Article 100906. https://doi.org/10.1016/j.tsc.2021.100906.

Walker, J. (2024). Understanding EBA (Evidence-Based Assessment) in education. Journal of Foreign Language Education and Technology, 9(4), 1–2.

Wardani, I. S., & Fiorintina, E. (2023). Building critical thinking skills of 21st century students through problem-based learning model. Jurnal Pendidikan Indonesia, 12(3), 461–470.

Watson, G., & Glaser, E. M. (2002). Watson-Glaser Critical Thinking Appraisal (WGCTA): UK Edition Manual. Psychological Corporation.

Wiyoko, T., Wulandari, T., & Irhas, M. (2024). Online-based literacy test instruments for evaluating prospective elementary school teachers' science

competencies. Jurnal Ilmiah Sekolah Dasar, 8(2), 282–292. https://doi.org/10.23887/jisd.v8i2.68342

Yolanda, F. (2019). The effect of problem-based learning on mathematical critical thinking skills of junior high school students. Journal of Physics: Conference Series.