JETIA
JOURNAL OF EDUCATIONAL TECHNOLOGY INNOVATION AND APPLICATIONS
IISTR



Critical Thinking Assessment in Chemistry Education

A Review of Tool Development and Validation

[1]Hassan Aliyu*, [2]Amina M. Chado, [2]Umar Idris Sarkin Bauchi, [2]Labake Ajoke Fadipe

Affiliations:
[1] Sokoto State University, Sokoto, Nigeria
[2] Federal University of Technology, Minna, Nigeria
*Correspondence: nagoronyo@gmail.com

ABSTRACT

Critical thinking (CT) is essential for scientific literacy in chemistry education, yet effectively assessing it remains challenging due to limitations in traditional methods and domain-general instruments. This systematic review synthesizes 19 studies (2015–2024) to map trends in developing and validating chemistry-specific CT assessment tools. Findings reveal the 4D model (Define, Design, Develop, Disseminate) as the predominant framework for instrument development, emphasizing practical implementation. The most valid and reliable instruments—such as Kriswantoro et al.'s (2021) Critical Thinking Assessment Model Integrated with Science Process Skills (CTSPS) and Susetyo et al.'s (2021) Critical Thinking Test for Salt Hydrolysis—demonstrated robust psychometric properties (e.g., Aiken V > 0.80, Cronbach's α ≥ 0.82, Rasch reliability > 0.85). Acid-base chemistry emerged as the primary context for CT enhancement (32% of studies), followed by chemical equilibrium (26%) and thermochemistry (16%), due to their inherent complexity and demand for higher-order reasoning. The review highlights persistent gaps, including fragmented validation practices and topic-specific instrument limitations. It advocates for standardized methodologies, broader topic coverage, and cross-cultural validation to advance CT assessment in chemistry education.

Keywords
Assessment Instruments
Chemical Concepts
Chemistry Education
Critical Thinking
Development and Validation

Article history
Received 20 July 2026
Revised 16 September 2026
Accepted 11 October 2026

journal.iistr.org/index.php/JETIA
DOI: 10.56741/IISTR.jetia.002526

JETIA
Journal of Educational Technology Innovation and Applications

Vol. 2, No. 03, pp. 295-312

Copyright: ©2026 by the authors. Licensee IISTR, Yogyakarta, Indonesia. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA) license.

Introduction

Critical thinking (CT) stands as a cornerstone of scientific literacy and a fundamental learning outcome in contemporary chemistry education. As defined by Facione (1990), CT encompasses the ability to "analyse, evaluate, and synthesize information to guide decision-making," equipping students to effectively navigate complex chemical phenomena, ranging from intricate reaction mechanisms to sophisticated thermodynamic systems. Its undeniable significance is globally recognized and indicated by influential educational frameworks, such as the OECD's Education 2030 initiative, which explicitly prioritizes CT as an indispensable skill for solving multifaceted real-world challenges [2]. In the context of chemistry, CT transcends mere rote memorization, enabling learners to critically appraise experimental designs, meticulously interpret empirical data, and adeptly reconcile macroscopic observations with microscopic, molecular-level reasoning [3].

Despite the widespread consensus on the paramount importance of CT, assessing these skills effectively within chemistry education remains a formidable challenge. Traditional assessment methods frequently overemphasize procedural knowledge, often at the expense of cultivating and evaluating higher-order thinking [4]. Concurrently, generic, domain-general CT instruments, such as the Watson-Glaser Critical Thinking Appraisal or the Ennis-Weir Critical Thinking Essay Test, often fall short due to their inherent lack of contextual validity for chemistry-specific tasks [5], [6]. Consequently, the past decade has witnessed a significant surge in the development of domain-specific instruments, meticulously designed to address critical thinking within particular chemical concepts. Examples include tools targeting acid-base equilibrium [7], thermochemistry [8], and reaction kinetics [9]. These specialized instruments are designed to capture nuanced indicators of CT—such as sophisticated argument analysis, complex problem-solving abilities, and robust systems thinking—all within authentic and relevant chemical contexts [10].

However, this very proliferation of new instruments has, ironically, led to fragmented methodologies and inconsistent validation practices across the field. The heterogeneity observed in current research is considerable. For instance, studies vary widely in their developmental frameworks, with some predominantly employing models like the 4D or ADDIE models (e.g., [11], [12]), while others use less common or unique approaches. There's also significant variation in psychometric rigor, ranging from reliance on basic content validity indices to more advanced statistical techniques like Rasch analysis or Confirmatory Factor Analysis (e.g., [7], [13], [5]). Furthermore, the contextual scope of these instruments differs considerably, with some being highly specialized and tied to specific topics like acid-base chemistry, and others aiming for

broader competencies (e.g., [14], [15]). This widespread methodological diversity complicates the process of selecting appropriate assessment tools for educators and researchers, and concurrently obscures valuable insights into establishing best practices for the design of effective CT assessments in chemistry.

To comprehensively address these existing gaps and provide clarity in the field, this systematic review rigorously synthesizes findings from 19 studies published between 2015 and 2024, all focused on critical thinking instrument development and validation within chemistry education. Our objectives are strategically threefold: first, to systematically map the methodological trends prevalent in instrument design, including the dominant research and development models and validation frameworks utilized; second, to critically evaluate the psychometric robustness of existing instruments, focusing on the quality and extent of their reported validity and reliability evidence; and third, to identify the specific chemical concepts and contexts that researchers have found most conducive to the manifestation and assessment of critical thinking skills. By critically appraising the current state of the field through this structured approach, this review ultimately aims to provide actionable insights that will guide future research and practical applications, steering them toward the development of more reliable, scalable, and pedagogically aligned critical thinking assessments in chemistry.

The aim of this study is to systematically review existing literature to identify prevalent models for instrument development, pinpoint the most valid and reliable critical thinking assessment instruments, and determine the common chemical concepts used in chemistry education research. Specifically, the objectives of the study are

1. To determine the prevailing instructional design models employed by researchers in the validation of instruments designed to assess critical thinking skills within chemistry education.
2. To evaluate the critical thinking assessment instruments that demonstrate the highest levels of validity, as reported in the relevant literature for chemistry education.
3. To evaluate the critical thinking assessment instruments that demonstrate the highest levels of reliability, as reported in the relevant literature for chemistry education.
4. To determine the specific chemical concepts that are most frequently utilized by researchers as contexts for developing instruments aimed at enhancing students' critical thinking skills in chemistry.

Methods

This systematic review synthesized studies on the development and validation of instruments used to assess critical thinking (CT) skills in chemistry education. Our methodology

followed a structured approach to ensure both transparency and reproducibility. This involved clearly defining the inclusion and exclusion criteria, detailing the data extraction process, and outlining the methods used for analysis.

A. Inclusion and Exclusion Criteria

1. Focused on CT assessment in chemistry

Instruments specifically designed to measure CT skills in chemistry contexts (e.g., acid-base, thermochemistry, reaction rates) were prioritized. Studies addressing broader STEM fields were excluded unless chemistry-specific outcomes were delineated. Reported validation metrics: Only studies with explicit validity (e.g., content validity index, factor analysis) and reliability (e.g., Cronbach's alpha) data were included. For example, [8] & [7] were retained due to their reporting of CVR (≥0.49) and Cronbach's alpha (≥0.65). Used empirical methods: Studies employing R&D models (e.g., ADDIE, 4D, Borg and Gall) or validation frameworks (e.g., Rasch analysis, CFA) were included. Theoretical papers or qualitative studies without instrument validation were excluded.

2. Targeted secondary or higher education

Studies involving K–12 or undergraduate students were included, while those focusing on primary education or non-formal settings were excluded.

3. Data Extraction and Synthesis

The study synthesized data from several key categories to ensure a high-quality review. This involved analyzing the authors and publication years of relevant works published between 2015 and 2025 to discern trends. We also examined the objectives of these studies, highlighting common goals such as integrating critical thinking with specific scientific domains like acid-base or thermochemistry. We also categorized each study's methodology by its development model and tabulated validation metrics to rigorously assess their soundness.

4. Final Data for the Review

This comprehensive review systematically synthesizes findings from 19 rigorously selected articles, all published between 2015 and 2025 in highly reputable, peer-reviewed journals. This specific timeframe was chosen to ensure the inclusion of the most current and relevant empirical evidence. These foundational studies are critical because they directly address the complex interplay of validity and reliability within instruments designed to assess critical thinking skills among chemistry students. The empirical evidence presented in these articles provides a robust basis for judging the extent to which these assessment tools accurately measure what they intend to measure (validity) and consistently produce the same results under similar conditions (reliability), thereby offering a crucial foundation for

effective educational practice. Furthermore, a significant criterion for inclusion was the explicit focus of the authors on the real-life application of chemical concepts. This ensures that the review not only covers the theoretical assessment of critical thinking but also highlights pedagogical approaches that connect abstract chemical principles to tangible, everyday phenomena, ultimately fostering a deeper and more meaningful understanding for students.

Results and Discussion

The findings derived from this systematic review are comprehensively presented in Table 1, which serves as a central repository for key information pertaining to each included study. The organized presentation within Table 1 facilitates a clear and comparative analysis of the methodologies and outcomes across the diverse range of studies, allowing for a structured identification of prevailing trends, exemplary instruments, and frequently utilized chemical contexts.

Table 1. Results of the systematic review of the development and validation of Instruments/ Tools for Assessing Critical Thinking Skills of Chemistry Students

Sno	Objective	Conclusion	Ref.
1	Measure CT and literacy	Effective for CT Skills	[9]
2	Develop Chem_PRO module	Validated effective tool	[16]
3	Assess CT disposition	Reliable valid tool	[14]
4	Measure CT automotive chemistry	Valid reliable instrument	[15]
5	Enhance CT welding testing	Effective hands-on tasks	[17]
6	Develop CT instrument for atomic nucleus	Valid reliable for physics in physics and chemistry education.	[18]
7	Create materials for equilibrium	Validate effective, valid, and suitable	[19]
8	Assess analytical thinking & literacy on colligative properties	Highly valid & feasible	[12]
9	Develop CT process skills	Validated valid, reliable, and effective CT tool.	[13]
10	Develop CT test instrument with a problem-solving context	Validated effective tool	[5]
11	Design thermochemistry CT test	Valid CVR .723 α .652	[8]
12	Evaluate CT and self-efficacy	Effective for combined constructs	[7]
13	Enhance students' CT skills	Valid, practical, & effective tool	[20]
14	Measure analytical thinking literacy	Practical evaluation tool	[21]
15	Enhance CT and metacognition	Promising needs further research	[11]
16	Measure CT and literacy	Valid & reliable tool	[6]
17	Measure CT acid-base solutions	Needs further development	[22]
18	Measure CT process skills	Valid & suitable assessment	[23]
19	Assess CT in chemistry	Effective CT tool	[18]

The studies summarized in Table 1 present a diverse portfolio of developed assessments and instructional modules designed to measure and cultivate critical thinking and chemical literacy across various chemistry domains. These developments encompass acid–base chemistry ([14], [22]), chemical equilibrium ([19], [21], [6]), thermochemistry [8], colligative properties [12],

buffer solutions [11], electrolyte solutions [23], and specialized vocational contexts, including automotive chemistry [15] and welding [17]. Pedagogical designs varied considerably, featuring computer-based testing frameworks [16], the Chem_PRO module [24], blended learning environments operationalized via the 4D model [20], and investigation-based toolkits [11]. In terms of format, several researchers adopted multi-tier or essay-based approaches, such as the three-tier CTD-AB diagnostic instrument [14], a 9-essay test integrating critical thinking with self-efficacy [7], and a 15-essay assessment for colligative properties [12]. Collectively, these contributions signify substantive methodological advancement in chemistry education assessment; nevertheless, the extension of such instruments to under-represented topics and higher-order cognitive competencies remains an ongoing priority.

The psychometric integrity of these instruments was rigorously scrutinized through both classical test theory and modern latent-trait frameworks, with sample sizes ranging from 31 to over 1,300 participants. Reliability estimates were generally robust: [9] reported Rasch-based item and subject reliabilities of 0.91 and 0.85, respectively; [24] documented a Cronbach's alpha of 0.924; [7] and [6] obtained coefficients of 0.82 and 0.85; [22] secured reliability values of 0.74 and 0.82 for their essay and problem-solving tests; and [5] reported an alpha of 0.823 alongside a content validity ratio (CVR) exceeding 0.622. Validity evidence was equally compelling across multiple studies: [13] achieved an Aiken index of 0.923, RMSEA of 0.021, and GFI of 0.97 with 289 participants, while [21] and [6] confirmed Aiken's V values exceeding 0.86. Content validity was affirmed through expert judgments, with [16] attaining 93.75%–99.54% across various validity facets, [12] reporting 93.33% overall feasibility, [15] receiving "very good" expert ratings, [17] achieving a 95 content-suitability score, and [19] obtaining validation scores between 4.17 and 4.33. Practicality and user acceptability were also favourable; [20] recorded 78% practicality, [25] attained 75.7% user approval despite an unexpectedly low student mean score of 10.1 out of 64, and teacher/student ratings in [19] ranged from 3.80 to 4.28.

A. What models are predominantly used by researchers for the validation of instruments for assessing critical thinking skills in chemistry education?

Based on a comprehensive analysis of the 19 studies presented in Table 1 of the systematic review, the 4D model (Define, Design, Develop, Disseminate) emerges as the predominantly used framework for developing and validating critical thinking (CT) assessment instruments in chemistry education. This model's structured, iterative approach proved particularly well-suited for creating context-specific tools targeting higher-order thinking skills.

The 4D Model (Define, Design, Develop, Disseminate) emerged as the most frequently employed instructional design framework, utilized in five distinct studies due to its clear, sequential phases and explicit "Disseminate" stage, which likely appeals to educational

researchers focused on practical implementation and broader impact of their validated instruments. This model underpinned the development of various educational tools and instruments. For instance, [9] used it to create an integrated instrument for assessing critical thinking (CT) and chemical literacy related to reaction rates. Similarly, [7] developed an integrated assessment for CT skills and self-efficacy in acid-base chemistry. The model also facilitated the creation of more comprehensive instructional materials, such as those by [20], who developed blended learning-based chemistry devices (syllabus, lesson plans, materials, instruments) aimed at enhancing CT. In earlier work, [6] applied the 4D model to develop an integrated CT and chemical literacy instrument for chemical equilibrium. Finally, [11] utilized the model to create investigation-based instructional tools (lesson plans, model books, instruments) for buffer solutions, specifically targeting CT and metacognition.

The ADDIE Model (Analysis, Design, Development, Implementation, Evaluation) was the second most frequently utilized instructional design model, appearing in four studies. While sharing systemic and phased similarities with the 4D model, ADDIE's upfront emphasis on "Analysis" and explicit "Implementation" and "Evaluation" phases render it particularly robust for comprehensive educational product development. This model underpinned the creation of diverse instruments and materials. For example, [16] employed ADDIE to develop a Science Learning Module featuring CBT Assessments designed to foster critical thinking, even though their specific topic was the human respiratory system rather than chemistry. In the realm of chemistry and related sciences, [25] utilized ADDIE to develop a critical thinking skill instrument for atomic nucleus material, applicable in higher education physics or chemistry contexts. Furthermore, [12] developed an assessment instrument for analytical thinking and chemical literacy focused on colligative properties using this model. Similarly, [21] applied ADDIE to create an integrated assessment targeting analytical thinking and chemical literacy within the context of chemical equilibrium.

The Borg and Gall Research and Development (R&D) Model, known for its extensive preliminary research, iterative prototyping, and multi-stage field testing, was employed in two studies, highlighting its suitability for large-scale instrument development. [17] utilized this model to develop a performance-based assessment instrument for critical thinking in a welding quality testing course, demonstrating its application beyond traditional chemistry contexts while still focusing on critical thinking assessment. Similarly, [18] leveraged the Borg and Gall model to create a measurement tool specifically for critical thinking skills related to hydrocarbons, thermochemistry, and chemical equilibrium.

The 3D Model (Define, Design, Develop), a more streamlined variant of the 4D model that purposefully omits the "Disseminate" phase, was identified in one notable study. This model's

focused approach is often favoured when the primary objective is the creation and validation of an educational product, without an immediate emphasis on large-scale diffusion or widespread adoption. In the context of this review, [19] effectively employed the 3D model to develop comprehensive teaching materials specifically aimed at enhancing critical thinking skills within the domain of chemical equilibrium. This application highlights the model's utility in targeted curriculum development where the emphasis lies squarely on the meticulous definition of learning objectives, the iterative design of instructional content, and the rigorous development of materials to achieve specific pedagogical goals, such as fostering higher-order thinking in a complex chemistry topic. The successful implementation of the 3D model in this research indicates its efficacy for projects centred on the creation of robust and effective educational resources.

Beyond the more prevalent instructional design models, several studies employed unique or less common frameworks designed to their specific research objectives. For instance, [26] uniquely combined Design and Development Research (DDR) with the Sidek Module Development Model to create their innovative Chem_PRO design thinking module. [14] utilized the 3D+II Model (Defining, Designing, Developing, Implementing) for the creation of their CTD-AB disposition instrument, showcasing a model that explicitly integrates an implementation phase. Furthermore, [23] adopted McIntire's Test Development Model for their integrated assessment instrument, highlighting the use of specialized psychometric approaches. Lastly, [8]; [5] both employed a Development and Validation Approach, indicating a systematic method for instrument creation and rigorous testing, even if the specific model was not explicitly named. These diverse approaches indicate the flexibility and adaptability researchers employ when developing instruments and materials in educational contexts.

Although a range of robust development models were applied across the reviewed studies, including the well-established ADDIE and Borg & Gall frameworks, it was the 4D Model (Define, Design, Develop, Disseminate) that demonstrated the highest prevalence. This model's structured, four-phase approach, particularly its emphasis on the "Disseminate" stage, seems to resonate strongly with educational researchers. This is likely due to its utility in guiding the systematic design and rigorous validation of critical thinking assessment instruments, while simultaneously facilitating their practical implementation and wider acceptance within the chemistry education community.

B. What are the most valid and reliable instruments for assessing critical thinking skills in chemistry education?

A meticulous examination of the systematic review data, as detailed in Table 1, revealed several instruments exhibiting exemplary validity and reliability evidence for the assessment of

critical thinking (CT) skills within chemistry education. The selection of these instruments was predicated on a combination of rigorous methodological approaches during their development and the presentation of compelling validation metrics, such as notably high content validity indices, confirmatory factor analysis outcomes, and robust Rasch reliability coefficients.

Table 2. Instruments that demonstrated the strongest validity and reliability evidence for assessing critical thinking (CT) skills in chemistry education

Study	Validity Evidence	Reliability Evidence	Methodological Rigor
[13]	CFA, Aiken Index (0.923)	α=0.886; Composite Rel. >0.88	Large sample (N=289); Robust D&D model
[5]	CVR >0.622; Empirical validity	α=0.823	Expert validation (7 experts)
[7]	Aiken's V >0.80; Factor Analysis (KMO=0.858)	α=0.82	Integrated CT + self-efficacy
[9]	Rasch validity (PCM-1PL)	Item Rel. 0.91 (Rasch)	4D model; Expert & Rasch validation
[6]	Aiken's V >0.86; IRT/factor analysis	α=0.85	Two-tier MCQs; Large sample (N=158)

One particularly robust instrument identified was the Critical Thinking Assessment Model Integrated with Science Process Skills (CTSPS), developed by [13]. This instrument demonstrated exceptional psychometric properties, affirming its suitability for assessing critical thinking in chemistry education. Its validity was substantiated by an outstanding Content Validity Index (Aiken) of 0.923, indicating strong expert agreement on its relevance and comprehensiveness. Further reinforcing its structural integrity, Construct Validity was confirmed through Confirmatory Factor Analysis (CFA), yielding an RMSEA of 0.021 and a GFI of 0.97, both values signifying an ideal model fit. This suggests that the instrument effectively measures the underlying theoretical constructs it purports to assess. In terms of reliability, the CTSPS also performed admirably, reporting a high Cronbach's Alpha of 0.886, which signifies excellent internal consistency among its items. Additionally, its Composite Reliability ranged from 0.88 to 0.90, further confirming the instrument's consistent measurement across different dimensions of critical thinking and science process skills. These strong validation metrics collectively establish the CTSPS as a highly credible and dependable tool for researchers and educators alike.

Another noteworthy instrument demonstrating robust psychometric properties is the Critical Thinking Test with Problem-Solving Context (Salt Hydrolysis), developed by [5]. This specialized test focuses on evaluating critical thinking skills within the specific domain of salt hydrolysis, grounding abstract concepts in practical problem-solving scenarios. The validity of this instrument was thoroughly established, with its Content Validity Ratio (CVR) exceeding 0.622 for all items, a strong indicator that the test items are highly relevant and representative of the construct being measured. Furthermore, empirical validity was confirmed, suggesting that the test effectively measures what it purports to measure in real-world applications and aligns

with external criteria. In terms of reliability, the instrument achieved an excellent Cronbach's Alpha of 0.823. This high coefficient indicates strong internal consistency, meaning that the individual items within the test reliably measure the same underlying critical thinking abilities, providing consistent results across different administrations. Collectively, these validation and reliability metrics indicate the quality and dependability of [5]'s instrument as a valuable tool for assessing critical thinking within a specific chemistry context.

The Integrated Critical Thinking (CT) & Self-Efficacy Test for Acid-Base Chemistry, developed by [7], also stands out for its well-established psychometric properties. This instrument was specifically designed to simultaneously assess students' critical thinking skills alongside their self-efficacy beliefs within the challenging domain of acid-base chemistry. Its validity was rigorously affirmed, with Aiken's V values exceeding 0.80, indicating high content validity and strong agreement among experts regarding the relevance and appropriateness of the test items. Further strengthening its structural integrity, Factor Analysis yielded a Kaiser-Meyer-Olkin (KMO) value of 0.858, which is considered excellent and confirms strong construct validity. This suggests that the test effectively measures the intended underlying theoretical constructs of critical thinking and self-efficacy. In terms of reliability, the instrument demonstrated strong internal consistency, evidenced by a Cronbach's Alpha of 0.82. This high reliability coefficient confirms that the test items consistently measure the same attributes, ensuring dependable and reproducible results for educators and researchers evaluating student understanding and confidence in acid-base chemistry.

The Integrated Critical Thinking (CT) & Chemical Literacy Test for Reaction Rates, developed by [9], represents a very recent and well-validated contribution to assessment instruments in chemistry education. This instrument was specifically designed to evaluate both critical thinking skills and chemical literacy within the dynamic context of chemical reaction rates. Its validity was rigorously confirmed using the Rasch model, a modern psychometric approach that provides robust insights into item and person fit. Specifically, the Infit and Outfit Mean Square (MSQ) values fell consistently within the ideal range of 0.7–1.3, indicating that the items functioned as expected and measured the intended construct effectively without any unexpected patterns in responses. This precise alignment indicates the instrument's strong measurement integrity. Furthermore, the reliability of this instrument is exceptionally high: the Item Reliability was an impressive 0.91, suggesting that the test items consistently perform well in measuring critical thinking and chemical literacy. Similarly, the Person Reliability stood at 0.85, indicating that the instrument consistently differentiates between individuals at different proficiency levels. These strong Rasch-based reliability coefficients provide compelling evidence of the instrument's

consistency and precision, making it a highly dependable tool for assessing these crucial skills in students studying reaction kinetics.

The Integrated Critical Thinking (CT) & Chemical Literacy Test for Chemical Equilibrium, developed by [6], consistently demonstrates strong psychometric properties, making it a reliable instrument for assessing these vital skills in students. The instrument's validity was robustly established, evidenced by an Aiken's V value exceeding 0.86, which indicates excellent content validity and strong consensus among experts regarding the relevance and appropriateness of the test items for assessing the intended constructs. Further supporting its structural integrity, factor analysis confirmed the unidimensionality of the instrument, affirming that it primarily measures a single, coherent construct (critical thinking and chemical literacy within the context of chemical equilibrium) rather than multiple unrelated ones. In terms of reliability, the instrument achieved a high Cronbach's Alpha of 0.85. This strong coefficient signifies excellent internal consistency, assuring that the test items consistently measure the same underlying abilities and that the instrument produces dependable and reproducible results. These compelling validation and reliability metrics position the [6] instrument as a trustworthy and effective tool for educators and researchers to gauge student proficiency in critical thinking and chemical literacy within the challenging topic of chemical equilibrium.

These particular instruments consistently demonstrate superior psychometric quality due to several shared characteristics that set them apart. Firstly, they all employed a multi-method validation approach, combining expert judgment through metrics like Content Validity Ratio (CVR) and Aiken's V with robust statistical tests such as Confirmatory Factor Analysis (CFA) and the Rasch model. This comprehensive validation strategy provides strong, convergent evidence for their validity. Secondly, each instrument reported exceptionally high reliability scores, with all Cronbach's Alpha values at or exceeding 0.82, significantly surpassing the commonly accepted threshold of 0.70. This indicates excellent internal consistency, assuring that the instruments consistently measure the intended constructs. Furthermore, their context-specific design is a key strength; they are not generic critical thinking tests but are meticulously crafted around core chemistry topics like acid-base chemistry, chemical equilibrium, reaction rates, and thermochemistry. This contextualization enhances their relevance and applicability in chemistry education. Finally, their development adhered to rigorous methodologies, utilizing established instructional design models such as the 4D Model, ADDIE Model, and Borg & Gall R&D Model. This systematic and often iterative development process ensures that the instruments are not only theoretically sound but also practically effective and well-tested.

While the aforementioned instruments demonstrate significant strengths, it's important to acknowledge certain caveats that influence their broader applicability and interpretation. A

primary concern for some studies, such as [14] with a sample size of N=31, is the limited sample size. Such smaller samples can potentially affect the generalizability of the findings, meaning the results might not be broadly applicable to larger, more diverse student populations. Consequently, caution should be exercised when extrapolating conclusions from these specific contexts. Furthermore, many of these highly valid instruments exhibit considerable topic specificity. For example, the CTD-AB instrument is precisely prepared for acid-base chemistry, while the DS-CT instrument is designed for thermochemistry. While this narrow focus allows for exceptionally high validity and in-depth assessment within their respective domains, it inherently limits their direct applicability to other areas of chemistry. Researchers and educators seeking to assess critical thinking across a broader spectrum of chemistry topics would, therefore, need to consider developing or validating additional instruments or adapting these with further rigorous testing.

Among the rigorously evaluated instruments, those by [13];[5];[7] demonstrate the most robust statistical evidence for assessing critical thinking in chemistry education. Notably, [13]'s CTSPS model stands out as the strongest overall, primarily due to its innovative integration of critical thinking with science process skills and its exceptional validation and reliability metrics. For educators and researchers with specific contextual needs, such as evaluating critical thinking related to reaction rates or chemical equilibrium, the specialized tools developed by [9];[6] come highly recommended for their proven efficacy within their respective domains.

C. Which chemical concepts are predominantly used by researchers for enhancing students' critical thinking skills?

Based on the systematic review data meticulously organized in Table 1, a discernible pattern emerges regarding the chemical concepts predominantly leveraged by researchers as contexts for developing instruments and interventions aimed at enhancing critical thinking (CT) skills in chemistry education. These topics are not merely chosen at random but rather because they often present inherent complexities, require multi-step reasoning, or involve abstract principles that lend themselves well to the application of higher-order thinking skills. Researchers consistently gravitated towards core areas of chemistry that necessitate strong analytical and problem-solving abilities. This strategic selection ensures that the developed assessment tools and pedagogical interventions are grounded in challenging yet fundamental chemical knowledge, thereby fostering deeper conceptual understanding alongside critical thinking development. The consistent use of these particular concepts across multiple studies further indicates their perceived efficacy as fertile ground for cultivating and measuring critical thinking in chemistry students.

Table 3. Chemical concepts were predominantly used by researchers as contexts for developing instruments and interventions to enhance critical thinking (CT) skills in chemistry education

Sno	Concept	CT Skill Alignment	Representative Studies
1	Acid-Base	Requires problem-solving, prediction, and real-world application (e.g., pH, titrations).	[14];[7];[16];[22];[17];[6]
2	Equilibrium	Demands systems thinking, reasoning about dynamic processes, and predicting shifts.	[19];[6];[21];[21];[18]
3	Thermochemistry	Integrates quantitative analysis, energy transfer, and data interpretation.	[8];[18];[13]
4	Reaction Rates	Involves experimental design, variable analysis, and evidence evaluation.	[9]

Acid-Base Chemistry emerged as the most frequently utilized chemical concept, serving as the context for critical thinking instrument and intervention development in six distinct studies, accounting for a significant 32% of all reviewed research. This prevalence highlights the inherent suitability of acid-base chemistry for fostering higher-order thinking, likely due to its conceptual complexity, reliance on abstract principles, and practical implications. For instance, [14] developed an instrument specifically to assess critical thinking disposition within acid-base chemistry, while [7] created an integrated test focusing on both critical thinking and self-efficacy in this domain. Broader applications also surfaced, with [16] utilizing the "acid, base, and salt" topic for a design thinking module. Early work by [22] also centred on developing critical thinking assessment instruments for acid-base solutions. Even studies with seemingly indirect links, such as [17]'s work on welding quality testing, occasionally touched upon underlying acid-base phenomena. Furthermore, the broad scope of [6]'s research on critical thinking in chemical equilibrium naturally encompassed acid-base equilibrium, reinforcing this topic's pervasive role. The consistent focus on acid-base chemistry indicates its value as a rich and challenging domain for cultivating and measuring critical thinking skills in chemistry students.

Chemical Equilibrium closely followed as another prominent context, featuring in five studies and representing 26% of the reviewed literature. This concept's significant presence is likely due to its inherent complexity, requiring students to grasp dynamic processes, apply Le Chatelier's principle, and perform intricate calculations, all of which demand robust critical thinking skills. Researchers leveraged this topic to develop a variety of educational tools. For instance, [19] focused on creating teaching materials specifically designed to enhance critical thinking skills within the context of equilibrium. [6] developed an integrated test to assess both critical thinking and chemical literacy related to equilibrium, while [21] similarly focused on analytical thinking in this domain. Furthermore, the broader topic often includes sub-concepts, as seen with [11], who developed instructional tools for buffer solutions, a key application of chemical equilibrium

principles. Even earlier research, such as [18], incorporated equilibrium alongside thermochemistry when developing critical thinking measurement tools. The consistent emphasis on chemical equilibrium across these studies indicates its critical role as a fertile ground for cultivating and evaluating higher-order thinking in chemistry education.

Thermochemistry also emerged as a significant context for critical thinking skill development, appearing in three studies and accounting for 16% of the reviewed literature. This consistent focus likely stems from thermochemistry's demanding nature, which requires students to interpret energy changes, apply Hess's Law, understand spontaneity, and perform calculations involving heat transfer. These concepts inherently push students beyond mere recall, compelling them to engage in analysis, synthesis, and evaluation—core critical thinking skills. For instance, [8] specifically developed a domain-specific critical thinking test designed to thermochemistry, indicates its suitability for targeted assessment. Similarly, [13] integrated critical thinking with science process skills within a thermochemistry context, demonstrating the topic's capacity to facilitate the development of multifaceted cognitive abilities. Even earlier work by [18] focused on developing a critical thinking instrument that included thermochemistry alongside other fundamental concepts like hydrocarbons and equilibrium. The repeated selection of thermochemistry across these studies affirms its value as a fertile and challenging domain for cultivating and accurately measuring students' critical thinking prowess in chemistry.

Beyond the most frequently explored topics, several other chemical concepts also served as valuable contexts for developing critical thinking instruments and interventions, each appearing in two studies. Reaction rates emerged as one such area, notably leveraged by [9] to create an integrated assessment of critical thinking and chemical literacy. This choice is logical given the conceptual and mathematical challenges involved in understanding how reaction speeds are influenced by various factors, requiring students to analyse, interpret, and predict. Similarly, colligative properties were utilized in two studies, with [12] developing an instrument to assess analytical thinking and chemical literacy within this specific domain. Colligative properties, involving the behaviour of solutions, demand students to apply principles to unfamiliar scenarios and make quantitative predictions. Finally, both salt hydrolysis and buffer solutions each served as contexts in two distinct studies, often building upon foundational acid-base and equilibrium knowledge. [5], for instance, developed a critical thinking test grounded in problem-solving scenarios related to salt hydrolysis, a topic that requires nuanced understanding of acid-base principles and their application. [11] focused on buffer solutions, a concept that demands an in-depth grasp of equilibrium and its practical implications for maintaining pH. The selection of these diverse yet interconnected topics demonstrates researchers' efforts to develop context-rich

assessments that challenge students' critical thinking across a spectrum of fundamental and applied chemistry concepts.

A clear set of key trends emerges from the analysis of chemical concepts chosen as contexts for critical thinking (CT) instrument development. Primarily, the selection of topics is driven by their inherent complexity, compelling students to engage in higher-order thinking. Concepts such as chemical equilibrium and acid-base chemistry, which necessitate multi-step reasoning, grappling with ambiguity, and understanding real-world relevance, force students to actively analyse, infer, and evaluate information rather than merely recalling facts. Furthermore, a strong quantitative focus is evident in the chosen topics. Disciplines like thermochemistry and reaction rates inherently integrate mathematical principles and data analysis, thereby demanding critical thinking for the accurate interpretation of graphs, the execution of complex calculations, and the logical derivation of conclusions. Beyond theoretical understanding, many instruments are rooted in contextual problem-solving. Concepts such as salt hydrolysis and buffer solutions, along with applications in areas like automotive chemistry (as seen in studies like [27], link theoretical chemical principles to applied, real-world scenarios. This approach not only enhances the relevance of chemistry but also develops students' abilities to transfer knowledge to practical challenges in fields like environmental science or engineering. Finally, a significant trend is the integration of CT with chemical literacy, with 10 out of 19 reviewed studies explicitly combining these two critical skills. This dual focus emphasizes the importance of not only being able to think critically about chemical phenomena but also possessing the ability to explain them clearly, analyse related data effectively, and communicate scientific understanding comprehensively.

Conclusion

This systematic review indicates significant advancements and persistent challenges in developing and validating critical thinking assessment instruments for chemistry education. The 4D model dominates the field as the primary framework. Researchers emphasize structured development for real-world application. Instruments like the CTSPS model and integrated chemical literacy tests set benchmarks for psychometric rigor through multi-method validation, including Rasch analysis and confirmatory factor analysis, and they achieve high reliability coefficients exceeding 0.80. However, validation practices remain fragmented, with researchers inconsistently reporting content validity indices. Acid-base chemistry and chemical equilibrium serve as the predominant testing contexts because they demand multi-step reasoning and systems thinking, yet this narrow focus leaves thermochemistry and reaction rates underrepresented, which limits generalizability. The review proposes three actions. First, researchers must adopt unified protocols combining expert judgment like Aiken's V with robust

statistical techniques such as confirmatory factor analysis or Rasch modelling. Second, future studies must diversify contexts to include organic chemistry and electrochemistry. Third, educators should use validated instruments like Susetyo's problem-solving test, adapt them for local curricula, and conduct thorough pilot studies. Finally, cross-cultural collaboration remains essential to refine assessments for diverse educational settings, ensuring psychometric soundness and pedagogical effectiveness in fostering chemistry-specific critical thinking skills globally.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] P. A. Facione, *The California Critical Thinking Skills Test (CCTST): Form A*. Millbrae: CA: California Academic Press, 1990.

[2] OECD, "Training for staff in early childhood education and care must promote practices that foster children ' s learning, development and well-being," Organisation for Economic Cooperation and Development. Accessed: Jun. 23, 2019. [Online]. Available: https://www.oecd.org/education/training-for-staff-in-early-childhood-education-and-care-must-promote-practices-that-foster-children-s-learning-development-and-well-being.htm

[3] V. Talanquer and A. R. Szozda, "An Educational Framework for Teaching Chemistry Using a Systems Thinking Approach," *J. Chem. Educ.*, vol. 101, p. 1785–1792, 2024, doi: 10.1021/acs.jchemed.4c00216.

[4] M. M. Cooper, "The Crosscutting Concepts: Critical Component or 'third Wheel' of Three-Dimensional Learning?," *J. Chem. Educ.*, vol. 97, no. 4, pp. 903–909, 2020, doi: 10.1021/acs.jchemed.9b01134.

[5] B. Susetyo, W. Siswaningsih, and F. R. Oktavi, "Development of Critical Thinking Test Instruments with Problem Solving Context on The Salt Hydrolysis Material," *J. Pendidik. Sains Indones.*, vol. 9, no. 2, pp. 243–255, 2021, doi: 10.24815/jpsi.v9i2.19862.

[6] S. Sadhu and E. W. Laksono, "Development and validation of an integrated assessment for measuring critical thinking and chemical literacy in chemical equilibrium," *Int. J. Instr.*, vol. 11, no. 3, pp. 557–572, 2018, doi: 10.12973/iji.2018.11338a.

[7] N. Nuryuliana and A. K. Prodjosantoso, "Development of Integrated Assessment Instrument to Measure Critical Thinking Skills and Self-Efficiency in Acid-Base Concept," *JTK (Jurnal Tadris Kim.*, vol. 6, no. 2, pp. 223–235, 2021, doi: 10.15575/jtk.v6i2.14872.

[8] E. Nursa'Adah, D. Sari, and I. Affifah, "Domain-specific critical thinking skills: Designing and analysis the instrument test on thermochemistry," *J. Phys. Conf. Ser.*, vol. 1731, no. 1, p. 012025, 2021, doi: 10.1088/1742-6596/1731/1/012025.

[9] N. W. Ariefiani and E. W. Laksono, "Development of An Integrated Assessment Instrument to Measure Students' Critical Thinking and Chemical Literacy Skills for Rate of Reaction Topic," *J. Penelit. Pendidik. IPA (Journal Res. Sci. Educ.*, vol. 10, no. 1, pp. 7726–7734, 2024, doi: 10.29303/jppipa.v10i10.9069.

[10] Y. F. Lay and A. L. Chandrasegaran, "The contribution of teacher preparation on grade 8 students' science achievement in TIMSS: A comparative study between Malaysia and Singapore," *J. Balt. Sci. Educ.*, vol. 17, no. 4, pp. 576–589, 2018.

[11] M. Danial, M. Sawal, and Nurlaela, "Development of Chemistry Instructional Tools and Its Effect on Critical Thinking Skills, Metacognition, and Concept Mastery of Students," *J. Phys. Conf. Ser.*, vol. 1028, no. 1, 2018, doi: 10.1088/1742-6596/1028/1/012041.

[12] Yandriani, R. U. Rery, and M. Erna, "Developing and Validating the Assessment Instruments to Measure Students' Analytical Thinking Ability and Chemical Literacy on Colligative Properties," *J. Phys. Conf. Ser.*, vol. 1788, no. 1, 2021, doi: 10.1088/1742-6596/1788/1/012027.

[13] Kriswantoro, B. Kartowagiran, and E. Rohaeti, "A Critical Thinking Assessment Model Integrated with Science Process Skills on Chemistry for Senior High School," *Eur. J. Educ. Res.*, vol. 10, no. 1, pp. 285–298, 2021, doi: https://doi.org/10.12973/eu-jer.8.4.935.

[14] A. Wahyudi, R. Richardo, I. Eilks, and C. Kulgemeyer, "Development of Three Tier Open-Ended Instrument to Measure Chemistry Students' Critical Thinking Disposition Using Rasch Analysis," *Int. J. Instr.*, vol. 16, no. 3, pp. 191–204, 2023, doi: 10.29333/iji.2023.16311a.
[15] D. S. Sari, Y. Widiyawati, I. Nurwahidah, and T. Setiawan, "STEM critical thinking assessment for measuring students' critical thinking skills in the automotive chemistry course," *J. Penelit. Pendidik. IPA*, vol. 9, no. 7, pp. 5289–5295, 2023, doi: 10.29303/jppipa.v9i7.4750.
[16] H. D. Ratnady, M. Tindangen, E. T. Maasawet, A. Hakim, Y. Hudiyono, and A. Hardoko, "Development of Science Learning Modules with Computer-Based Test (CBT) Assessments to Train Critical Thinking Skills of Students at SMP Negeri 7 Bontang," *J. Penelit. Pendidik. IPA (Journal Res. Sci. Educ.*, vol. 10, no. 9, pp. 6564–6573, 2024, doi: 10.29303/jppipa.v10i9.8391.
[17] J. Adri, Refdinal, Ambiyar, and A. S. Abdullah, "Critical Thinking Skills in Performance-Based Assessment: Instrument Development and Validation," *J. Tech. Educ. Train.*, vol. 14, no. 1, pp. 91–100, 2022, doi: 10.30880/jtet.2022.14.01.008.
[18] Kartim and A. A. K. Budiastra, "The Development of Measurement Tools for Senior High School Students' Critical Thinking Skills in Chemistry," *Int. J. Adv. Res. Manag. Soc. Sci.*, vol. 4, no. 5, pp. 108–16, 2015.
[19] R. Siahaan, M. Sitorus, and S. Silaban, "The development of teaching materials oriented to critical thinking skills for chemistry class XI high school," *J. Pendidik. Kim.*, vol. 13, no. 1, pp. 60–68, 2021, doi: 10.24114/jpkim.v13i1.24145.
[20] S. Hadisaputra, M. S. Ihsan, Gunawan, and A. Ramdani, "The development of chemistry learning devices based blended learning model to promote students' critical thinking skills," *J. Phys. Conf. Ser.*, vol. 1521, no. 2020, pp. 042083-1-, 2020, doi: https://doi.org/10.1088/1742-6596/1521/4/042083.
[21] E. Ad'hiya and E. W. Laksono, "Development and validation of an integrated assessment instrument to assess students' analytical thinking skills in chemical literacy," *Int. J. Instr.*, vol. 11, no. 4, pp. 241–256, 2018, doi: 10.12973/iji.2018.11416a.
[22] W. Sumarni, K. I. Supardi, and N. Widiarti, "Development of assessment instruments to measure critical thinking skills," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 349, no. 1, 2018, doi: 10.1088/1757-899X/349/1/012066.
[23] S. R. D. Astuti, Suyanta, E. W. Lfx, and E. Rohaeti, "An integrated assessment instrument: Developing and validating instrument for facilitating critical thinking abilities and science process skills on electrolyte and nonelectrolyte solution matter," *AIP Conf. Proc.*, vol. 1847, 2017, doi: https://dx.doi.org/10.1063/1.4983909.
[24] M. N. Aris, N. H. Ibrahim, and A. N. D. Halim, "Innovate to Educate: The Development and Validation of the Design Thinking Chemistry Module," *Int. J. Acad. Res. Progress. Educ. Dev.*, vol. 13, no. 1, pp. 2379–2389, 2024, doi: 10.6007/ijarped/v13-i1/21103.
[25] P. Dwijananti, A. Ruwaida, B. N. Mindyarto, and Wiyanto, "Development of a critical thinking skill instrument for physics and chemistry students in higher education," *J. Phys. Conf. Ser.*, vol. 1918, no. 5, 2021, doi: 10.1088/1742-6596/1918/5/052076.
[26] N. Md. Aris, N. H. Ibrahim, and N. D. Abd. Halim, "Innovate to Educate: The Development and Validation of the Design Thinking Chemistry Module," *Int. J. Acad. Res. Progress. Educ. Dev.*, vol. 13, no. 1, 2024, doi: 10.6007/ijarped/v13-i1/21103.
[27] D. A. Sari, E. Ellizar, and M. Azhar, "Development of problem-based learning module on electrolyte and nonelectrolyte solution to improve critical thinking ability," *J. Phys. Conf. Ser.*, vol. 1185, no. 1, 2019, doi: 10.1088/1742-6596/1185/1/012146.

Authors



Dr. Hassan Aliyu is affiliated with the Department of Science Education, Faculty of Education, Sokoto State University, Nigeria. His academic interests include science education, chemistry teaching, critical thinking assessment, and educational research. His scholarly work focuses on examining approaches to developing and validating assessment instruments that support higher-order thinking skills. He is interested in strengthening evidence-based educational practices and improving the quality of science learning and assessment. (email: nagoronyo@gmail.com).



Dr. Amina M. Chado is affiliated with the Department of Science Education, School of Science and Technology Education, Federal University of Technology Minna, Nigeria. Her academic interests are related to science education, instructional development, and educational assessment. Her research activities include exploring approaches to evaluating students' critical thinking abilities and improving assessment practices. She is interested in promoting innovative teaching strategies and strengthening the quality of science education. (email: chado.amina@futminna.edu.ng).



Dr. Umar Idris Sarkin Bauchi is affiliated with the Department of Science Education, School of Science and Technology Education, Federal University of Technology Minna, Nigeria. His academic interests include science teaching, educational measurement, critical thinking, and instructional assessment. His scholarly interests focus on the development and evaluation of assessment instruments for science education. He is interested in advancing effective educational practices. (email: haumar2008@gmail.com).



Prof. Dr. Labake Ajoke Fadipe is affiliated with the Department of Chemistry, School of Physical Sciences, Federal University of Technology Minna, Nigeria. His academic interests include chemistry education, chemical concepts, critical thinking, and scientific assessment. His scholarly interests emphasize the relationship between chemistry knowledge and higher-order cognitive skills. He is interested in supporting the development of valid and reliable assessment instruments and promoting evidence-based approaches. (email: labsfad@yahoo.com).

.