Integrated Science Education Journal
Vol. 7, No. 2, March 2026, pp. 265~274
ISSN: 2716-3725, DOI: 10.37251/isej.v7i2.2385

Review Article

Mapping the Landscape of Critical Thinking Assessment in STEM Education: A Systematic Review of Psychometric Properties, Contextual Implementation, and Future Directions

Yuleks Juru Mudi[1,*], Kana Hidayati[1], Muhammad Nursa'ban[1], Widowati Pusporini[1]
[1] Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

Article Info

Article history:

Received Dec 01, 2025
Revised Jan 27, 2026
Accepted Feb 17, 2026
OnlineFirst Mar 25, 2026

Keywords:

Assessment
Critical Thinking
Psychometrics
STEM Education
Systematic Review

ABSTRACT

Purpose of the study: This study aims to systematically map the landscape of critical thinking assessment in STEM education, with a particular focus on psychometric characteristics, contextual implementation, and emerging research trends.

Methodology: A Systematic Literature Review (SLR) was conducted following the PRISMA protocol using the Scopus database as the primary source. A total of 58 studies published between 2018 and 2025 were analyzed through bibliometric mapping using VOSviewer and thematic synthesis.

Main Findings: The findings indicate a substantial increase in research on critical thinking assessment in STEM education since 2020, aligning with growing global attention to 21st-century competencies. However, most studies continue to position assessment primarily as a tool for evaluating learning outcomes or the effectiveness of pedagogical interventions, such as project-based, problem-based, and inquiry-based learning. Only a limited number of studies systematically examine the psychometric quality of assessment instruments, including evidence of construct validity, reliability, and multidimensional structure. This pattern reveals a clear gap between assessment practices in STEM education and established standards for educational measurement, which may lead to weak or potentially misleading conclusions about students' critical thinking abilities.

Novelty/Originality of this study: This review integrates bibliometric and thematic analyses to identify conceptual and methodological gaps in the existing literature and proposes a coherent direction for the development of critical thinking assessments that are both psychometrically robust and contextually relevant within STEM education.

Corresponding Author:
Yuleks Juru Mudi,
Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta,
Jl. Colombo No. 1, Karangmalang, Yogyakarta, 55281, Indonesia
Email: yuleksmudi2000@gmail.com

1. INTRODUCTION

Critical thinking assessment has become an integral component of STEM education, particularly due to its role in evaluating students' readiness to address complex problems and respond to the demands of the 21st century [1], [2]. However, despite its widespread application across various STEM learning contexts, a fundamental issue remains insufficiently addressed, namely the weak psychometric quality of many critical

thinking assessment instruments currently in use. Numerous studies employ critical thinking assessment primarily as a tool for evaluating learning outcomes, while only a limited number explicitly examine instrument quality through evidence of construct validity, reliability, or well-defined dimensional structures [3]-[5]. This situation may lead to findings that are methodologically weak or potentially misleading when assessment results are used as the basis for evaluating the effectiveness of STEM learning.

Over the past decade, research on critical thinking in STEM education has demonstrated a notable shift in focus. Studies have moved beyond an exclusive emphasis on general critical thinking skills toward the integration of more contextualized and discipline-specific competencies aligned with the characteristics of individual STEM fields [4], [6]. This shift reflects broader changes in educational paradigms as well as increasing workforce demands for higher-order reasoning skills. Nevertheless, advancements in pedagogical approaches have not been fully accompanied by corresponding developments in measurement practices, particularly with respect to conceptual consistency and the psychometric quality of the assessment instruments employed [2], [7].

The urgency of this issue becomes more apparent when viewed in relation to large-scale assessment results. International reports consistently indicate that students' critical thinking abilities remain at relatively low levels, including within the context of science literacy in Indonesia [8], [9]. Such findings are frequently used as the basis for designing instructional interventions and informing educational policy. However, the validity of inferences drawn from assessment results is highly dependent on the quality of the instruments employed. Assessment tools that lack adequate evidence of validity and reliability pose a substantial risk of producing inaccurate score interpretations and leading to inappropriate educational decisions [10]-[12].

Although critical thinking is widely recognized as an essential competency in STEM education, conceptual and methodological challenges in its assessment persist. Divergent perspectives on the definition of critical thinking, along with variations in assessment approaches ranging from standardized tests and analytic rubrics to performance-based assessments, complicate efforts to establish a consistent understanding across studies [1], [3]. In addition, ongoing debates regarding the use of domain-general and domain-specific approaches highlight the importance of STEM disciplinary contexts in shaping how critical thinking is conceptualized and measured [13], [14].

Several previous systematic literature reviews have examined critical thinking in STEM education, primarily from the perspectives of instructional strategies and pedagogical design [4], [6]. However, most of these reviews continue to position critical thinking assessment as a supporting component of learning rather than as a primary object of inquiry from a measurement perspective. As a result, key aspects of instrument quality, including evidence of validity, reliability, and dimensional structure, are often addressed only in a limited or purely descriptive manner, without in-depth psychometric analysis [5], [7], [15].

Based on these identified gaps, this systematic review aims to comprehensively map critical thinking assessment practices in STEM education by placing measurement quality as the central focus. This review integrates two complementary dimensions, namely psychometric characteristics and contextual implementation. The psychometric dimension encompasses evidence of validity, reliability, and the dimensional structure of assessment instruments, while the contextual implementation dimension examines how these instruments are applied across different STEM disciplines, educational levels, and cultural contexts [16], [17].

This study employs a systematic literature review methodology guided by the PRISMA framework to identify, select, and synthesize relevant research on critical thinking assessment in STEM education. The analysis combines bibliometric mapping and thematic synthesis to organize findings according to disciplinary domains, types of assessment instruments, and pedagogical contexts. Through this approach, the review seeks to provide a stronger conceptual and methodological foundation for the development and use of critical thinking assessment instruments that are valid, reliable, and contextually relevant within STEM education.

2. RESEARCH METHOD

This study employs a Systematic Literature Review (SLR) approach to comprehensively analyze and map research on critical thinking assessment in STEM education. The SLR approach was selected because it enables a structured, transparent, and replicable process of literature searching, screening, evaluation, and synthesis, thereby enhancing the methodological rigor and credibility of the review. To ensure accuracy and consistency in reporting, this review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework, which consists of four main stages: identification, screening, eligibility assessment, and study inclusion. The application of the PRISMA framework serves not only as a procedural guide but also as a quality control mechanism to ensure that the reviewed studies are relevant to the research focus and meet established scientific standards.

The literature search was conducted using the Scopus database as the sole data source. Scopus was selected because it is widely recognized as one of the largest and most comprehensive bibliographic databases indexing high-quality, peer-reviewed journals across multiple disciplines, including education, science,

technology, and psychometrics. In addition, Scopus provides rich and consistent metadata, which strongly supports bibliometric analysis and the systematic mapping of research trends.

The review period was limited to publications from 2018 to 2025 in order to capture recent developments and emerging trends in critical thinking assessment within the context of STEM education. The search strategy was designed to encompass variations of key terms representing STEM education, critical thinking skills, and assessment or measurement approaches. The search string used in this study was as follows: ("STEM education" OR "science technology engineering mathematics" OR "STEM literacy") AND ("critical thinking" OR "higher-order thinking" OR "critical reasoning") AND (assess* OR measur* OR evaluat* OR instrument OR tool OR scale OR questionnaire* OR psychometric OR validit* OR reliab*).

To ensure source quality and accessibility, the search was restricted to peer-reviewed journal articles written in English and available in open-access format. This restriction was applied to allow transparent examination and verification of the reviewed studies by readers.

Articles retrieved from the literature search were further screened using predefined inclusion and exclusion criteria. The establishment of these criteria aimed to ensure that only relevant, recent, and methodologically appropriate studies were included in the analysis. The inclusion criteria comprised peer-reviewed journal articles published between 2018 and 2025 that explicitly addressed STEM education and critical thinking assessment. Eligible studies were required to be written in English and available in open-access format. In contrast, non-journal publications such as conference proceedings, books, book chapters, editorials, and articles with restricted access were excluded from the review. A detailed overview of the inclusion and exclusion criteria is presented in Table 1.

Table 1. Inclusion and Exclusion Criteria

No.	Category	Inclusion Criteria	Exclusion Criteria
1	Type of Publication	Peer-reviewed journal articles	Non-journal publications (conference, book, website)
2	Publication Year	2018–2025	Before 2018
3	Keywords	"STEM education" AND "critical thinking"	Unrelated or missing keywords
4	Language	English	Non-English articles
5	Access Type	Open access (Gold/Green/Hybrid)	Restricted-access publications

The initial literature search yielded 273 articles from the Scopus database. After removing duplicate records and applying the publication year restriction (2018–2025), the number of articles was reduced to 255. The screening process was then conducted in a stepwise manner in accordance with the PRISMA guidelines. During the initial screening stage, several types of publications were excluded, including book chapters (25), books (2), review articles (5), conference proceedings (91), editorials (1), non-English articles (1), studies addressing unrelated topics (36), and articles that were not available in open-access format (37). Following this screening process, a total of 58 articles met all inclusion criteria and were deemed eligible for further analysis.

The 58 selected articles were analyzed in depth using two complementary analytical approaches: bibliometric analysis and thematic synthesis. Bibliometric analysis was employed to map the development of the research field, publication trends, the distribution of STEM disciplines, and overall research patterns related to critical thinking assessment within the review period. This approach primarily contributed to addressing RQ1, which focuses on the evolution and trends of research in this area. Subsequently, thematic synthesis was conducted to examine the substantive content of each article. This process involved extracting information related to the types of critical thinking assessment instruments used, the purposes of instrument application, educational levels, STEM disciplinary contexts, and reported psychometric characteristics, including evidence of validity, reliability, and construct structure. The thematic analysis played a central role in addressing RQ2 and RQ3, particularly with respect to patterns of instrument implementation, methodological strengths and limitations, and directions for future research.

To ensure the quality of the review, only articles published in peer-reviewed journals and meeting all predefined inclusion criteria were included in the analysis. In addition, data extraction and categorization were conducted systematically based on predetermined analytical categories, thereby reducing potential bias and enhancing the consistency and transparency of interpretation. An overview of the entire process of study identification, screening, and selection is presented in Figure 1 using a PRISMA flow diagram.

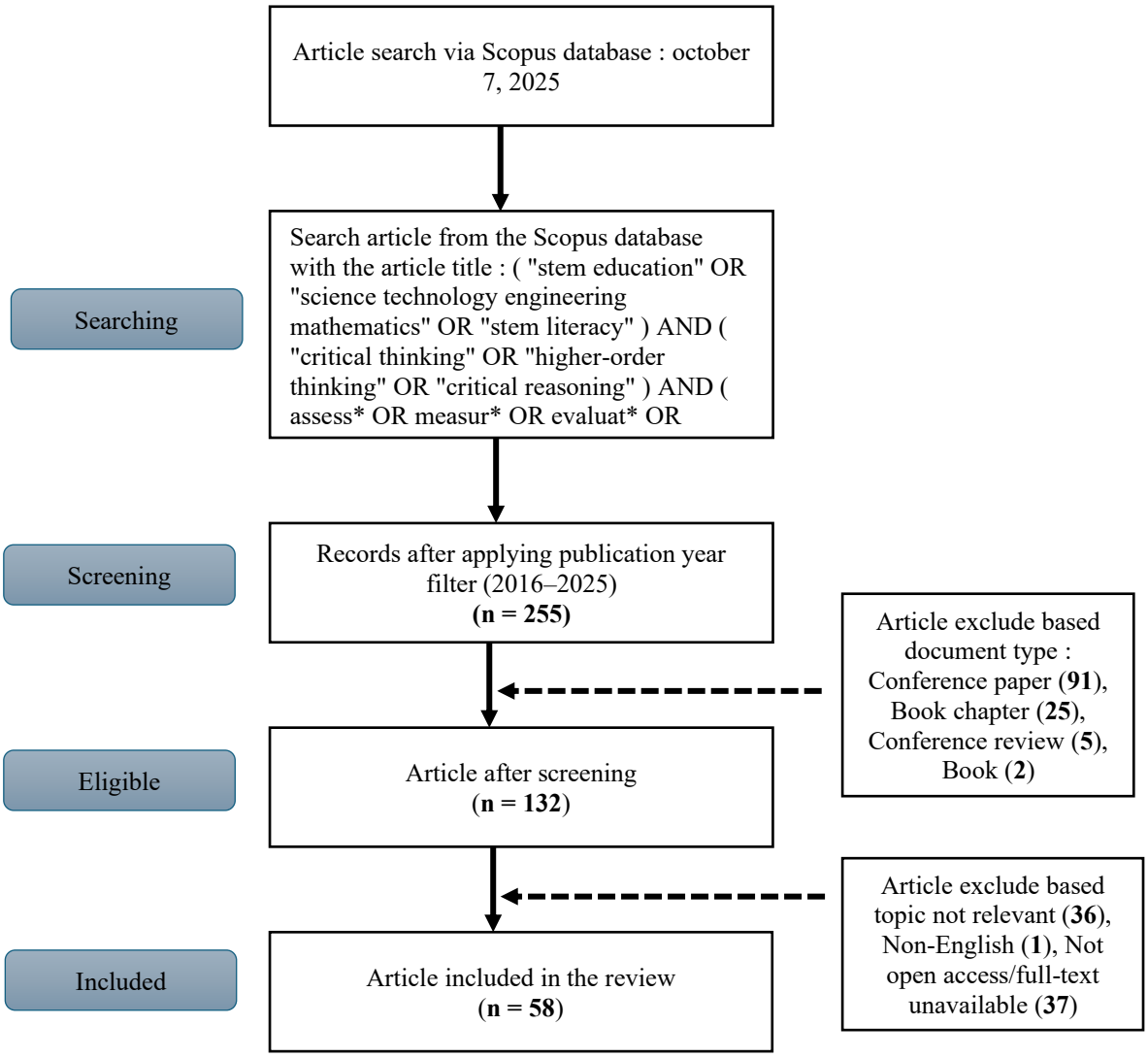

Figure 1. PRISMA flow diagram showing the study identification, screening, and selection process for the systematic literature review.

3. RESULTS AND DISCUSSION

3.1. Research Trends and Thematic Development

This systematic review is based on 58 articles retrieved from the Scopus database and focuses on mapping critical thinking assessment in STEM education. The analysis covers annual publication trends, the distribution of contributions by country and patterns of author collaboration, as well as the development of emerging research themes over the period 2018–2025 [1], [2].

As shown in Figure 2, publications addressing critical thinking assessment in STEM education exhibit a consistent upward trend. During the early period from 2018 to 2019, the number of publications was relatively limited, indicating that this topic had not yet become a central focus of research [4]. A noticeable increase began to emerge between 2020 and 2022, coinciding with growing global attention to 21st-century skills, particularly critical thinking as an essential competence in STEM education [16].

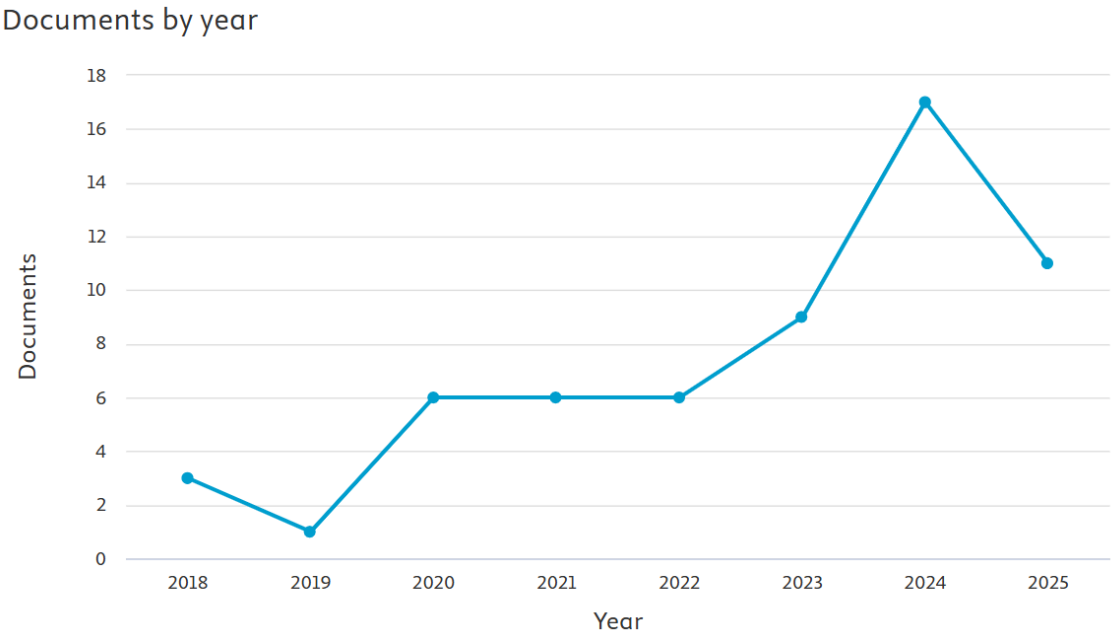

Figure 2. Publication trends of studies on critical thinking assessment in STEM education (2018–2025), based on 58 articles indexed in the Scopus database.

The most significant increase occurred in 2023 and reached its peak in 2024. This development indicates a shift from predominantly pedagogical exploration toward a more serious interest in the development and evaluation of critical thinking assessment. These findings align with the growing demand for competency-based learning accountability and the use of empirical evidence in STEM education [17]. Keyword analysis based on the co-occurrence map (Figure 3) shows that dominant themes remain centered on "critical thinking" and "STEM education", with strong linkages to instructional approaches such as problem-based learning and project-based learning [3], [18]. This pattern suggests that existing research continues to adopt a predominantly pedagogical orientation, in which critical thinking assessment is treated primarily as a learning outcome rather than as a construct that requires systematic development and rigorous psychometric testing [5], [19].

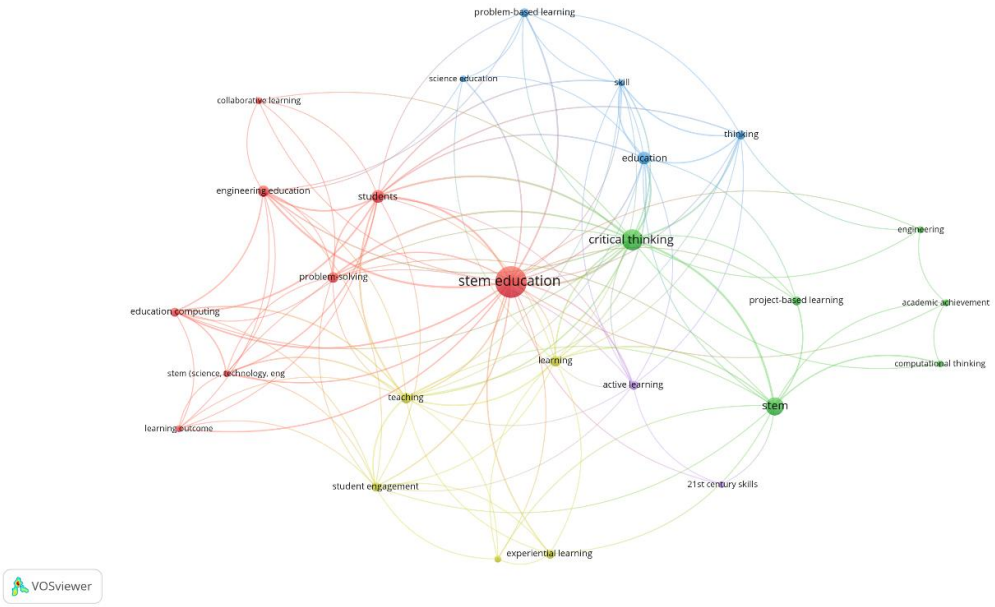

Figure 3. Keyword co-occurrence network of critical thinking assessment in STEM education generated using VOSviewer.

Table 2. Keyword by outhors

Rank	Keyword	Total Link Strenght
1	Critical Thingking	82
2	Stem Education	74
3	Students	64
4	Education	49
5	Thingking	48
6	Teaching	40
7	Skill	36
8	Learning	34
9	Stem	25
10	Problem Based Learning	24

The international collaboration network (Figure 4) indicates that the United States, Malaysia, Indonesia, China, and Turkey are the main contributors in this research area. However, cross-regional collaboration remains limited, suggesting that cross-cultural validation of critical thinking assessment instruments has not yet become a major focus of research [13], [14].

Figure 4. International collaboration network of publications on critical thinking assessment in STEM education (Scopus data, visualized using VOSviewer).

Overall, the findings for RQ1 indicate that despite a significant increase in the number of publications, research remains largely concentrated on instructional practices, while the strengthening of the psychometric foundations of critical thinking assessment in STEM education has received comparatively limited attention.

3.2 Contextual Implementation of Critical Thinking Assessment

The implementation of critical thinking assessment in STEM education demonstrates a wide diversity of approaches, while simultaneously revealing substantial methodological fragmentation. The majority of studies employ assessment primarily as a tool for evaluating learning outcomes rather than as a central object of systematic instrument development [1], [20]. Innovative instructional approaches such as problem-based learning, project-based learning, and inquiry-based learning are frequently reported as effective in enhancing students' critical thinking skills [4], [6]. However, the instruments used in these contexts are predominantly observation rubrics, essay tests, and perception-based questionnaires, which are rarely accompanied by adequate evidence of construct validity and reliability [3].

At the higher education level, critical thinking assessment has increasingly been integrated into reflective and technology-mediated tasks, including digital media projects and learning portfolios [21], [22], [23]. Although these approaches offer the potential to capture authentic cognitive processes, most assessments remain subjective in nature and have not undergone systematic psychometric validation procedures [11], [12]. Another emerging issue concerns teachers' assessment literacy. Studies indicate that educators who possess a strong understanding of formative assessment principles are better able to evaluate key critical thinking indicators, such as interpretation, inference, and argument evaluation [24], [25]. Conversely, constraints related to time, institutional support, and assessment training often lead to reliance on conventional testing formats that are insufficiently sensitive to the complexity of critical thinking [2].

Recent developments in educational technology and artificial intelligence have also begun to influence critical thinking assessment practices. Several studies report the use of AI-assisted tools for argument analysis and automated feedback generation [26], [27], [28]. Nevertheless, most of these approaches lack rigorous examination of algorithmic validity, data reliability, and ethical considerations, raising important concerns regarding the credibility and accountability of AI-based assessment systems [11], [16].

To clarify the patterns of critical thinking assessment implementation and the characteristics of instruments used across various STEM contexts, a cross-study comparative synthesis is presented in Table 3.

This table summarizes the types of instruments, assessment approaches, and reported psychometric evidence across the reviewed articles, thereby providing a structured overview of dominant assessment practices in the field. The findings related to RQ2 consistently reveal a clear gap between assessment practices implemented in educational settings and the psychometric standards that should underpin the measurement of critical thinking abilities.

Table 3. Synthesis of Instrument Types, Assessment Approaches, and Psychometric Evidence in Studies of Critical Thinking Assessment in STEM Education (2018–2025)

Main Study Focus	Type of Instrument Used	Assessment Approach	Reported Psychometric Evidence	Typical Study Patterns in the Corpus
Evaluation of STEM instructional effectiveness (PBL, PjBL, IBL, robotics, makerspaces)	Performance rubrics, observation sheets, project products	Authentic, performance-based	Generally limited to content validity; reliability and construct validity are rarely reported	STEM PBL studies, robotics-based learning, makerspaces
Measurement of perceptions, attitudes, and readiness for critical thinking	Self-report questionnaires	Perceptual	Internal consistency reliability (e.g., Cronbach's alpha) reported in a limited number of studies	STEM attitude studies, teacher and pre-service teacher readiness
Development of STEM-based critical thinking instruments	Scenario-based tests, contextualized items	Cognitive	Construct validity and reliability reported in a small subset of studies	Instrument development studies and limited systematic reviews
Reflective and digital assessment	Videos, podcasts, e-portfolios, digital projects	Formative, reflective	Psychometric evidence is generally not reported	Video essays, digital escape rooms, online STEM learning
Technology- and AI-based assessment	AI-assisted assessment, automated feedback	Formative–adaptive	Focus on pedagogical innovation; validity and reliability are rarely examined	AI studies, LLM applications, digital literacy
Cross-context and cross-national studies	Various	Mixed	Cross-cultural validation is very limited	STEM reviews in Indonesia, Asia, and regional contexts

Note: This table is constructed based on the information explicitly reported in the reviewed articles. The absence of psychometric evidence in the table reflects the lack of reporting of such procedures in the original studies, rather than their confirmed absence in practice [11], [12], [29].

3.3 Future Directions of Critical Thinking Assessment Research in STEM Education

The comparative synthesis of the 58 reviewed articles indicates that research on critical thinking assessment in STEM education needs to shift from the dominance of pedagogical approaches toward strengthening measurement foundations. As summarized in Table 2, only a limited number of studies explicitly report evidence of construct validity, internal reliability, or dimensional structure testing of assessment instruments [1], [2].

Instruments developed without adequate validity and reliability testing pose a risk of producing weak or misleading conclusions [12]. This risk becomes particularly serious when assessment results are used to support claims about the effectiveness of STEM learning interventions or to inform educational decision-making and policy development [11]. Therefore, future research agendas should prioritize instrument development as a central focus. STEM-based critical thinking assessment instruments need to be systematically designed by explicitly linking cognitive indicators to authentic STEM contexts and tested using modern psychometric approaches, such as confirmatory factor analysis and Rasch or item response theory (IRT) modeling [30]-[34].

In addition, performance-based and reflective assessments should be integrated with systematic validation procedures to enable cross-context comparisons and longitudinal tracking of students' critical thinking development [35]-[39]. The integration of qualitative and quantitative approaches also represents an important strategy for capturing the complexity of critical thinking in a more comprehensive manner [40]-[44]. The use of technology and artificial intelligence in adaptive assessment further offers new opportunities for innovation. However, additional research is required to ensure algorithmic transparency, fairness, and measurement accountability in AI-assisted assessment systems [45]-[49].

Finally, cross-cultural validation and international collaboration need to be strengthened, as most critical thinking instruments are developed within specific cultural contexts and have not been widely tested in developing countries, including Indonesia [50], [51]. Overall, the findings of this review emphasize that the future of critical thinking assessment in STEM education depends on a balanced integration of pedagogical innovation and psychometric rigor. With a strong measurement foundation, critical thinking assessment can function as a credible scientific instrument to support the quality and sustainability of STEM education.

4. CONCLUSION

This systematic review provides a comprehensive contribution to the field of critical thinking assessment in STEM education by integrating bibliometric analysis, thematic synthesis, and an examination of psychometric evidence across existing studies. The findings reveal a clear upward trend in research productivity, indicating growing scholarly attention to critical thinking within STEM contexts. However, this expansion remains predominantly centered on pedagogical practices, with relatively limited emphasis on the development of psychometrically sound assessment instruments. Furthermore, the review demonstrates that current assessment practices are still largely dependent on performance rubrics, essay-based tests, and self-report measures, while rigorous reporting of construct validity and reliability is often insufficient. This indicates a critical gap between classroom assessment practices and the measurement standards required to accurately evaluate students' critical thinking abilities. In addition, the study identifies a significant misalignment between the theoretical understanding of critical thinking as a multidimensional construct encompassing cognitive, metacognitive, and affective domains and its practical measurement, which frequently fails to capture this complexity. As a result, many assessment outcomes risk being partial or even misleading, particularly when used to evaluate the effectiveness of STEM learning interventions.

The implications of these findings are substantial across research, practice, and policy domains. From a research perspective, there is a pressing need to shift the focus toward the development of robust, psychometrically validated instruments, where construct validity, reliability, and dimensionality are treated as central components. The adoption of advanced analytical approaches such as confirmatory factor analysis, Rasch modeling, and item response theory is essential to ensure measurement precision and comparability. In educational practice, the findings highlight the importance of strengthening teachers' assessment literacy, particularly in designing instruments that are sensitive to students' reasoning processes and higher-order thinking. Innovative pedagogies such as project-based learning, digital reflection, and AI-assisted feedback must be supported by accurate and meaningful assessment systems to genuinely foster critical thinking development. At the policy level, the study underscores the need for a systemic commitment to integrating high-quality assessment into STEM curricula, including support for cross-cultural validation, international collaboration, and the ethical use of assessment technologies. This is especially relevant in developing contexts, where balancing local relevance with global standards is crucial. Overall, advancing critical thinking assessment in STEM education requires the integration of psychometric rigor, contextual relevance, and sustainable innovation, transforming assessment from a procedural requirement into a credible and impactful tool for enhancing meaningful learning and developing reflective, innovative, and responsible learners.

ACKNOWLEDGEMENTS

The authors would like to thank all researchers whose studies were included in this systematic review. No external funding was received for this study.

AUTHOR CONTRIBUTIONS

Conceptualization, Yuleks Juru Mudi and Kana Hidayati; Methodology, Yuleks Juru Mudi; Data curation, Yuleks Juru Mudi; Formal analysis, Yuleks Juru Mudi; Investigation, Yuleks Juru Mudi; Writing – original draft preparation, Yuleks Juru Mudi; Writing – review and editing, Kana Hidayati, Muhammad Nursa'ban, and Widowati Pusporini; Supervision, Kana Hidayati and Muhammad Nursa'ban. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The author(s) declare no conflict of interest.

USE OF ARTIFICIAL INTELLIGENCE (AI)-ASSISTED TECHNOLOGY

The authors declare that no artificial intelligence (AI) tools were used in the generation, analysis, or writing of this manuscript. All aspects of the research, including data collection, interpretation, and manuscript preparation, were carried out entirely by the authors without the assistance of AI-based technologies.

REFERENCES

[1] I. Azmi, M. F. N. L. Abdullah, Z. Alwadood, and F. Calaminos, "Assessing critical thinking in mathematics education: A systematic review and analysis using the prisma framework," *Int. J. Essent. Competencies Educ.*, vol. 4, no. 1, pp. 54–69, 2025, doi: 10.36312/ijece.v4i1.1858.

[2] C. H. Ng and M. Adnan, "The needs of competency assessment in STEM education: A systematic literature review," *Int. J. Mod. Educ.*, vol. 6, no. 23, pp. 455–469, 2024, doi: 10.35631/ijmoe.623031.

[3] A. Alias, L. E. Mohtar, S. K. Ayop, and F. R. Rahim, "A systematic review on instruments to assess critical thinking & problem-solving skills," *EDUCATUM*, vol. 9, no. Sp, pp. 38–47, 2022, doi: 10.37134/ejsmt.vol9.sp.5.2022.

[4] N. W. A. Hakim and C. A. Talib, "Measuring critical thinking in science: Systematic review," *Asian Soc. Sci.*, vol. 14, no. 11, 2018, doi: 10.5539/ass.v14n11p9.

[5] O. L. Liu, L. Frankel, and K. C. Roohr, "Assessing critical thinking in higher education: Current state and directions for next-generation assessment," *ETS Res. Rep. Ser.*, vol. 2014, no. 1, pp. 1–23, 2014, doi: 10.1002/ets2.12009.

[6] S. Ramadani, P. Sinaga, and W. Liliawati, "Critical thinking in science learning: Systematic literature review," *J. Pemberdaya. Masy.*, vol. 4, no. 1, pp. 183–196, 2025, doi: 10.46843/jpm.v4i1.378.

[7] S. M. Brookhart, *How to assess higher-order thinking skills in your classroom*. Alexandria, Virginia: ASCD, 2010.

[8] B. I. Sappaile, A. Rahman, I. Ilwandri, T. A. Santosa, I. Ichsan, and J. N. T. Papia, "The effect of the STEM learning model on student's critical thinking in Indonesia: Meta-analysis," *Edumaspul J. Pendidik.*, vol. 7, no. 1, pp. 1425–1436, 2023, doi: 10.33487/edumaspul.v7i1.6129.

[9] T. Tanti, A. Astalini, D. A. Kurniawan, D. Darmaji, T. O. Puspitasari, and I. Wardhana, "Attitude for physics: The condition of high school students," *Jurnal Pendidikan Fisika Indonesia,* vol. 17, no. 2, pp. 126-132, 2021, doi: 10.15294/jpfi.v17i2.18919.

[10] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, 2014.

[11] M. T. Kane, "Validating the Interpretations and Uses of Test Scores," *J. Educ. Meas.*, vol. 50, no. 1, pp. 1–73, Mar. 2013, doi: 10.1111/jedm.12000.

[12] S. Mollah, "The factors, forms, causes, positive and negative impacts of the digital divide on educational practices from both educators' and learners' perspectives: A systematic review," *Integrated Science Education Journal*, vol. 7, no. 1, pp. 86-103, 2026, doi: 10.37251/isej.v7i1.2314.

[13] Y. B. Bhakti, R. Arthur, and Y. Supriyati, "Development of an assessment instrument for critical thinking skills in physics: A systematic literature review," in *Journal of Physics: Conference Series*, 2023. doi: 10.1088/1742-6596/2596/1/012067.

[14] D. T. Tiruneh, M. De Cock, A. G. Weldeslassie, J. Elen, and R. Janssen, "Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism," *Int. J. Sci. Math. Educ.*, vol. 15, no. 4, pp. 663–682, 2017, doi: 10.1007/s10763-016-9723-0.

[15] R. H. Ennis, "The nature of critical thinking: An outline of critical thinking dispositions and abilities," University of Illinois, Urbana, IL, 2011.

[16] J. W. Pellegrino, "A new era for STEM assessment: Considerations of assessment, technology, and artificial intelligence BT - Uses of Artificial Intelligence in STEM Education," X. Zhai and J. Krajcik, Eds., Oxford University Press, 2024, pp. 17–37. doi: 10.1093/oso/9780198882077.003.0002.

[17] G. Reynders, J. M. Lantz, S. M. Ruder, C. L. Stanford, and R. Cole, "Rubrics to assess critical thinking and information processing in undergraduate STEM courses," *Int. J. STEM Educ.*, vol. 7, no. 1, pp. 1–15, 2020, doi: 10.1186/s40594-020-00208-5.

[18] T. Tanti, A. Astalini, D. Darmaji, D. A. Kurniawan, and R. Fitriani, "Student perception review from gender: Electronic moduls of mathematical physics," *JPI (Jurnal Pendidikan Indonesia)*, vol. 11, no. 1, pp. 125-132, 2022, doi: 10.23887/jpiundiksha.v11i1.35107.

[19] S. Deo and K. Hölttä-Otto, "Critical thinking assessment in engineering education: A Scopus-based literature review," 2023, doi: 10.1115/1.4064275.

[20] Q. Wang and A. H. Abdullah, "Enhancing students' critical thinking through mathematics in higher education: A systemic review," *SAGE Open*, vol. 14, no. 3, 2024, doi: 10.1177/21582440241275651.

[21] T. Tanti, W. Utami, D. Deliza, and M. Jahanifar, "Investigation in vocation high school for attitude and motivation students in learning physics subject", *Journal Evaluation in Education (JEE),* vol. 6, no. 2, pp. 479-490, 2025, doi: 10.37251/jee.v6i2.1452.

[22] R. L. Nurjanah, "The Presentation of Students' Critical Thinking Skill in Writing Essays with Microlearning Strategy and E-portfolios Integration," *J. English Lang. Teach. Linguist.*, vol. 10, no. 3, p. 459, 2025, doi: 10.21462/jeltl.v10i3.1766.

[23] P. Caratozzolo, V. Lara-Prieto, S. Hosseini, and J. Membrillo-Hernández, "The use of video essays and podcasts to enhance creativity and critical thinking in engineering," *Int. J. Interact. Des. Manuf.*, vol. 16, no. 3, pp. 1231–1251, 2022, doi: 10.1007/s12008-022-00952-8.

[24] P. Hull and T. Vígh, "Teachers' Assessment Literacy: A descriptive literature review," *Hungarian Educ. Res. J.*, pp. 1–16, 2024, doi: 10.1556/063.2024.00317.

[25] R. A. Shafii and J.-L. Berger, "Teacher Assessment Literacy, Formative Assessment Practices, and Their Perceived Efficacy in Tanzania: A Scoping Review," *Stud. Educ. Eval.*, vol. 86, pp. 101496, 2025, doi: 10.1016/j.stueduc.2025.101496.

[26] M. Sabri and A. Wais, "Artificial Intelligence dalam Pengukuran dan Penilaian Pendidikan," *J. Eval. Pendidik.*, vol. 16, no. 2, pp. 95–107, 2025, doi: 10.21009/jep.v16i2.60738.

[27] F. A. Yanti, R. W. Wardana, B. Buyung, E. Heryensi, and N. Khamis, "Automatic Assessment- Based Artificial

Intellegent to Measure Students Environmental Literacy," *Indones. J. Learn. Adv. Educ.*, vol. 7, no. 3, pp. 461–480, 2025, doi: 10.23917/ijolae.v7i3.11240.
[28] C. Zhao, "AI-assisted Assessment in Higher Education: A Systematic Review," *J. Educ. Technol. Innov.*, vol. 6, no. 4, pp. 39–58, 2024, doi: 10.61414/jeti.v6i4.209.
[29] S. Ntumi and K. Twum Antwi-Agyakwa, "A Systematic Review of Reporting of Psychometric Properties in Educational Research," *Mediterr. J. Soc. Behav. Res.*, vol. 6, no. 2, pp. 53–59, 2022, doi: 10.30935/mjosbr/11912.
[30] T. G. Bond and C. M. Fox, *Applying the Rasch model: Fundamental measurement in the human sciences*, 3rd ed. New York, NY: Routledge, 2015.
[31] R. J. De Ayala, *The theory and practice of item response theory*. New York, NY: Guilford Press, 2009.
[32] Q. Pan, F. Reichert, Q. Liang, J. de la Torre, and N. Law, "Measuring Digital Literacy Across Ages and Over Time: Development and Validation of A Performance-Based Assessment," *Educ. Inf. Technol.*, vol. 30, no. 15, pp. 22065–22100, 2025, doi: 10.1007/s10639-025-13592-8.
[33] M. Alfaleh, "Sustainable AI-Driven Assessment in Higher Education: A Systematic Review of Fairness, Transparency, Pedagogical Innovation, and Governance," *Sustainability*, vol. 18, no. 2, pp. 785, 2026, doi: 10.3390/su18020785.
[34] M. Liu, "Ensuring Fairness in AI-Assisted EFL Writing Assessment," *J. Educ. Educ. Res.*, vol. 16, no. 3, pp. 97–102, 2025, doi: 10.54097/p137vr68.
[35] C. M. Evans, "Applying a culturally responsive pedagogical framework to design and evaluate classroom performance-based assessments in Hawai'i," *Applied Measurement in Education*, vol. 36, no. 3, pp. 269-285, 2023, doi: 10.1080/08957347.2023.2214655.
[36] N. Hartini, E. Prihatin, Y. Rahyasih, E. Herawan, D. Nurbani, S. Dzakirah, and S. Jiayin, "Can Artificial Intelligence Automate the Microteaching Evaluation?," *Educational Process: International Journal*, vol. 19, pp. 2025607, 2025, doi: 10.22521/edupij.2025.19.607.
[37] H. R. Romandoni, F. Nurhasanah, and S. Maharani, "Integration and evaluation of computational thinking in mathematics education: A systematic review of research 2016-2025," *Mosharafa: Jurnal Pendidikan Matematika*, vol. 14, no. 4, pp. 903-918, 2025, doi: 10.31980/mosharafa.v14i4.3548.
[38] S. K. Mahmud, and M. Kurt, "Enhancing inclusive sustainability-oriented learning in higher education using adaptive learning platforms and performance-based assessment," *Sustainability*, vol. 18, no. 3, pp. 1489, 2026, doi: 10.3390/su18031489.
[39] N. Q. Nguyen, and L. P. Nguyen, "Formative assessment for knowledge cultivation in university EFL classrooms: a systematic review from higher education," *Language Testing in Asia*, vol. 16, no. 1, pp. 22, 2026, doi: 10.1186/s40468-026-00430-y.
[40] Y. Liu, "Paradigmatic compatibility matters: A critical review of qualitative-quantitative debate in mixed methods research," *Sage Open*, vol. 12, no. 1, 2022, doi: 10.1177/21582440221079922.
[41] D. L. Dinsmore, and L. K. Fryer, "Critical thinking and its relation to strategic processing," *Educational Psychology Review*, vol. 35, no. 1, pp. 36, 2023, doi: 10.1007/s10648-023-09755-z.
[42] J. A. Paul, M. Sinha, and J. D. Cochran, "Instruments to assess students' critical thinking—A qualitative approach," *Decision Sciences Journal of Innovative Education*, vol. 21, mo. 3, pp. 123-143, 2023, doi: 10.1111/dsji.12295.
[43] R. Cerchione, M. Morelli, R. Passaro, and I. Quinto, "A critical analysis of the integration of life cycle methods and quantitative methods for sustainability assessment," *Corporate Social Responsibility and Environmental Management*, vol. 32, no. 2, pp. 1508-1544, 2025, doi: 10.1002/csr.3010.
[44] W. M. Lim, "What is qualitative research? An overview and guidelines," *Australasian marketing journal*, vol. 33, no. 2, pp. 199-229, 2025, doi: 10.1177/14413582241264619.
[45] C. Aloisi, "The future of standardised assessment: Validity and trust in algorithms for assessment and scoring," *European Journal of Education*, vol. 58, no. 1, pp. 98-110, 2023, doi: 10.1111/ejed.12542.
[46] M. Thelwall, and K. Kousha, "Technology assisted research assessment: algorithmic bias and transparency issues," *Aslib Journal of Information Management*, vol. 77, no. 1, pp. 175-190, 2025, doi: 10.1108/AJIM-04-2023-0119.
[47] S. C. Nouis, V. Uren, and S. Jariwala, "Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making: A qualitative study of healthcare professionals' perspectives in the UK," *BMC Medical Ethics*, vol. 26, no. 1, pp. 89, 2025, doi: 10.1186/s12910-025-01243-z.
[48] S. A. Birahim, "Contesting the algorithm: advancing a right to challenge AI decisions under the GDPR for algorithmic fairness," *Transforming Government: People, Process and Policy*, vol. 19, no. 4, pp. 895-913, 2025, doi: 10.1108/TG-05-2025-0148.
[49] M. Alfaleh, "Sustainable AI-Driven assessment in higher education: A systematic review of fairness, transparency, pedagogical innovation, and governance," *Sustainability*, vol. 18, no. 2, pp. 785, 2026, doi: 10.3390/su18020785.
[50] J. G. Trigueiro, M. V. da Costa, M. A. F. Barreto, M. Zwarenstein, and R. E. Fontenele Lima de Carvalho, "Cross-cultural adaptation and evidence of validity of the interprofessional collaboration scale (IPC-BR) for Brazil," *Journal of Interprofessional Care*, vol. 39, no. 2, pp. 257-266, 2025, doi: 10.1080/13561820.2025.2451957.
[51] M. Alavi, D. Le Lagadec, and M. Cleary, "Challenges of cross-cultural validation of clinical assessment measures: A practical introduction," *Journal of Advanced Nursing*, vol. 82, no. 1, pp. 941-949, 2026, doi: 10.1111/jan.16906.