Social, Humanities, and Educational Studies

SHEs: Conference Series 9 (1) (2026) 1063 – 1072

A Systematic Literature Review of the 4-Tier Test as an Instrument for Measuring Critical Thinking Skills and Identifying Misconceptions in Science Education

Nugraha Widya Wisista, Sarwanto, Sri Marmoah

Universitas Sebelas Maret
sarwanto@staff.uns.ac.id

Article History
accepted 1/2/2026 approved 1/3/2026 published 31/3/2026

Abstract
This study aimed to synthesise evidence on the development, validation, effectiveness, and applications of Four-Tier Test (4-Tier Test) instruments in science education through a systematic literature review. The review was conducted following the PRISMA guidelines across seven databases, namely Google Scholar, Scopus, ERIC, ScienceDirect, DOAJ, and Emerald Insight covering publications from 2015 to 2025 in both English and Indonesian. From over 1,247 initially identified records, 16 peer-reviewed studies were selected based on predetermined inclusion criteria. Findings revealed that 4-Tier Test instruments consistently demonstrated robust psychometric properties, with Aiken's V ranging from 0.75 to 0.94 and Cronbach's alpha from 0.62 to 0.90. The dual confidence rating mechanism proved effective in distinguishing genuine misconceptions from lack of knowledge and guessing across physics, chemistry, biology, and general science domains at multiple educational levels. However, research remains heavily concentrated at secondary and tertiary levels, with only two studies targeting primary school students. Methodological heterogeneity in misconception classification criteria further limits cross-study comparability. It is concluded that 4-Tier Tests represent a significant advancement in diagnostic assessment, yet their full potential remains constrained by inconsistent development standards and insufficient representation at the primary school level, a gap that warrants priority in future research.
Keywords: *4-Tier Diagnostic Assessment, Critical Thinking Skills, Misconceptions, Science Education, Systematic Literature Review*

Abstrak
Penelitian ini bertujuan untuk mensintesis bukti-bukti mengenai pengembangan, validasi, efektivitas, dan penerapan instrumen Four-Tier Test (Tes Empat Tingkat) dalam pendidikan sains melalui tinjauan literatur sistematis. Tinjauan dilakukan mengikuti panduan PRISMA pada tujuh database, yaitu Google Scholar, Scopus, ERIC, ScienceDirect, DOAJ, dan Emerald Insight yang mencakup publikasi tahun 2015 hingga 2025 dalam bahasa Inggris dan Indonesia. Dari lebih dari 1.247 artikel yang teridentifikasi pada pencarian awal, sebanyak 16 artikel jurnal peer-reviewed dipilih berdasarkan kriteria inklusi yang telah ditetapkan. Hasil penelitian menunjukkan bahwa instrumen Four-Tier Test secara konsisten memiliki properti psikometri yang kuat, dengan nilai Aiken's V berkisar antara 0,75 hingga 0,94 dan koefisien Cronbach's alpha antara 0,62 hingga 0,90. Mekanisme penilaian kepercayaan diri ganda terbukti efektif dalam membedakan miskonsepsi sejati dari kurangnya pengetahuan dan tebakan pada domain fisika, kimia, biologi, dan IPA umum di berbagai jenjang pendidikan. Namun, penelitian masih terkonsentrasi pada jenjang menengah dan perguruan tinggi, dengan hanya dua studi yang menyasar jenjang sekolah dasar. Heterogenitas metodologis dalam kriteria klasifikasi miskonsepsi juga membatasi komparabilitas antar studi. Disimpulkan bahwa instrumen Four-Tier Test merupakan kemajuan signifikan dalam asesmen diagnostik, namun potensi penuhnya masih terkendala oleh standar pengembangan yang tidak konsisten dan kurangnya representasi pada jenjang sekolah dasar sehingga memerlukan tindakan yang perlu diprioritaskan dalam penelitian mendatang.
Keywords: *Asesmen 4-Tier Diagnostik, Kemampuan Berpikir Kritis, Miskonsepsi, Pendidikan Sains, Tinjauan Literatur Sistematis*

Social, Humanities, and Education Studies (SHEs): Conference Series p-ISSN 2620-9284
https://jurnal.uns.ac.id/shes e-ISSN 2620-9292

INTRODUCTION

The possession of critical thinking skills and conceptual understanding is considered to be of fundamental importance to 21st-century learners. Nevertheless, the accurate assessment of these abilities remains a persistent challenge in education (Changwong et al., 2018; Facione, 2015). Conventional assessment instruments, most notably conventional multiple-choice tests, are characterised by significant limitations in their capacity to capture the intricacies of students' cognitive processes. These limitations frequently result in a failure to differentiate between genuine understanding and correct answers that are obtained through guesswork or underlying misconceptions (Retnawati et al., 2018; Widana et al., 2018).

In response to these limitations, multi-tier diagnostic tests have evolved as sophisticated alternatives progressing from two-tier to three-tier, and more recently to 4-Tier formats (Fratiwi et al., 2017; Kaltakci-Gurel et al., 2017). The 4-Tier Test signifies the most recent advancement in this progression, The 4-Tier Test encompasses four discrete levels: (1) the selection of the answer; (2) the confidence in the answer; (3) the reasoning explanation; and (4) the confidence in the reasoning. This structure facilitates the acquisition of comprehensive diagnostic information by educators. This information encompasses not only the students' existing knowledge, but also the underlying conceptions they hold, and the confidence they have in their knowledge and reasoning.

The evolution of diagnostic instruments in the domain of science education has unfolded through successive generations. Two-tier tests, which combine answers with reasons, were among the first attempts to probe students' conceptual understanding beyond surface-level knowledge. The implementation of three-tier tests has facilitated the incorporation of a confidence rating component, thereby facilitating the distinction between misconceptions and a lack of knowledge. The 4-Tier Test is a refinement of the approach by virtue of the provision of separate confidence ratings for both the answer and reasoning tiers. This results in a minimisation of the overestimation of misconceptions and underestimation of knowledge gaps that can occur with three-tier instruments (Gurel et al., 2015; Ma et al., 2025).

Rigorous research has consistently demonstrated that 4-Tier Tests offer considerable advantages over their predecessors. As shown in the studies conducted by Kaltakci-Gurel et al. (2017) and Tumanggor et al. (2020), the dual confidence rating mechanism significantly reduces guessing and enhances the accuracy of student understanding categorisation. This enhanced diagnostic capability renders 4-Tier Tests a particularly valuable tool for identifying specific areas in which instructional interventions are required

Despite this growing body of individual research, a comprehensive systematic synthesis of 4-Tier Test development, validation, and effectiveness across educational levels and science domains remains conspicuously absent. The volume of individual studies is, in fact substansial Önder Çelikkanlı & Kızılcık (2022) identified 69 studies using 4-Tier diagnostic tests in physics education alone between 2010 and 2022 yet this body of work remains fragmented, domain-specific, and unsynthesised. Critically, the only existing review specifically focused on 4-Tier tests is confined to a single subject area (physics), with no analogous synthesis available for other science domains, let alone a cross-domain, multi-level synthesis. A broader systematic review of Misconception Tier Diagnostic Technologies by Ma et al. (2025) confirmed this pattern, finding only 28 qualifying studies from SSCI-indexed journals across four decades of research. In biology education, Duran & Dikmenli (2024) similarly concluded that no direct systematic review specific to multi-tier diagnostic tests yet existed in that domain. An initial database search for the present review yielding over 1,247 records returned no systematic literature review synthesising 4-Tier test research across multiple science domains and educational levels, confirming that this gap persists to the present day. The near-exclusive focus of existing research on secondary and tertiary education further

compounds this limitation, studies such as Desstya et al. (2025) and Atmojo et al. (2024) which developed 4-Tier instruments for primary school pupils remain exceptional rather than representative. Moreover, the use of 4-Tier Tests for assessing critical thinking skills as distinct from misconception identification constitutes an additional dimension that has received scant scholarly attention (Lufita et al., 2025; Soeharto et al., 2019). To the best of the authors' knowledge, no systematic literature review has synthesised 4-Tier Test research spanning multiple educational levels, science domains and cognitive dimensions positioning the present study as the first of its kind.

The present systematic literature review aims to address these gaps by the following six points: (1) synthesising evidence on the psychometric properties (validity and reliability) of 4-Tier Test instruments across different studies and contexts; (2) analysing the effectiveness of 4-Tier Tests in diagnosing misconceptions and measuring critical thinking skills; (3) examining the applications of 4-Tier Tests across different educational levels and science domains; (4) identifying common methodological approaches in 4-Tier Test development; (5) evaluating the advantages and limitations of 4-Tier Tests compared to conventional and other multi-tier assessment instruments; and (6) providing recommendations for future research and practice in 4-Tier Test development and implementation.

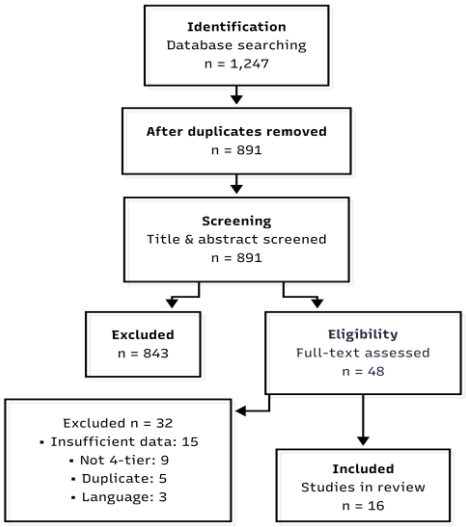
This systematic review makes a significant contribution to both the theoretical and practical aspects of educational assessment. In principle, it furnishes a comprehensive synthesis of the current state of knowledge regarding 4-Tier Test instruments, identifying patterns in their development, validation, and application. In practice, it provides educators, researchers and curriculum developers with evidence-based guidance for selecting, developing and implementing 4-Tier Tests in their specific contexts. The review also underscores the necessity for further research, with a particular emphasis on the development of 4-Tier Tests for the assessment of critical thinking in elementary education.

METHODS

This study employed a Systematic Literature Review (SLR) design following the PRISMA guidelines (Page et al., 2021). The review was guided by five research questions addressing the psychometric properties, effectiveness, methodological approaches, applications, and comparative advantages of 4-Tier Test instruments across science education contexts.

A systematic search was conducted across seven databases: Google Scholar, Scopus, ERIC, ScienceDirect, DOAJ, and Emerald Insight, covering publications from 2015 to 2025 in both English and Indonesian. One pre-2015 article from Caleon & Subramaniam (2010) was retained as a foundational exception. Search strings combined three conceptual clusters using Boolean operators: instrument-type terms (*"four-tier test"*, *"multi-tier test"*), diagnostic function terms (*"misconception"*, *"diagnostic assessment"*, *"critical thinking"*), and context terms (“*science education*”).

Articles were included if they were peer-reviewed journal articles substantively addressing 4-Tier Tests in science education, available as full-text in English or Indonesian. Conference proceedings, theses, and peripheral mentions were excluded. From over 1,247 initially identified records, approximately 48 underwent full-text review, yielding a final corpus of 16 studies (see Figure 1). Data were extracted using a structured six-column framework covering educational level, science domain, methodological and psychometric profile, key findings, and limitations. Quality was appraised using an adapted MMAT (Ma et al., 2025).



Picture 1. PRISMA Flow Diagram

RESULTS AND DISCUSSION

Table 1. Summary of Included Studies in Systematic Literature Review

Author's & Year	Level & Domain	Methods & Psychometrics	Key Finding	Limitations
Caleon & Subramaniam (2010)	Upper Secondary / Physics (Mechanical Waves)	Development & application; Multiple reliability estimates; Content + statistical validity; 4WADI instrument	The 4TMC was the first four-tier instrument reported in the literature. Nine genuine alternative conceptions were identified, expressed with moderate confidence. Students largely displayed an "illusion of knowing," with higher scores and confidence in the answer tier than the reason tier.	The instrument was developed within the Singapore 'O'-level curriculum context, limiting generalisability to other curricula. Predetermined reason options may not fully capture students' actual reasoning patterns.
Kaltakci-Gurel et al. (2017)	University / Physics (Geometrical Optics)	Multi-phase R&D (interviews → open-ended → pilot → final); Expert content validation (6 experts); Statistical validity confirmed; 40 items; N=243 PSPTs from 12 universities	The FTGOT was demonstrated to be valid and reliable for identifying pre-service physics teachers' misconceptions in geometrical optics, particularly regarding reflection, refraction, and image formation. The dual confidence rating mechanism effectively distinguished genuine misconceptions from lack of knowledge.	The sample was restricted to pre-service physics teachers in Turkey, limiting generalisation to secondary school students or other national contexts. Some items required further revision to improve distractor effectiveness.
Tumanggor et al. (2020)	Senior High School / Physics (Simple Harmonic Motion)	R&D (modified Oriondo & Dalo-Antonio model); Aiken's V = 0.75–0.77; INFIT MNSQ = 0.77–1.30; Item reliability = 0.73; 9 items	The four-tier instrument successfully detected student misconceptions, with 40.7% of students identified as holding misconceptions. The highest misconception rate occurred in the frequency subtopic (75%), followed by string length and pendulum period (60%), and spring constant and frequency relationships (58.3%).	The instrument focused exclusively on Simple Harmonic Motion without integrating prior physics content. The limited sample size and restricted sampling strategy were acknowledged as constraints requiring broader replication.

Author's & Year	Level & Domain	Methods & Psychometrics	Key Finding	Limitations
Istiyono et al. (2023)	Senior High School / Physics (Multiple Topics)	R&D (modern test theory-based); Aiken's V = 0.85–0.94; α = 0.90; 100 items	The FTDT successfully identified multiple levels of conceptual understanding, including scientific conceptions, lack of knowledge, misconceptions, false negatives, and false positives. Detailed percentages of each understanding category and misconception sequences by topic were reported.	Instrument analysis relied solely on Partial Credit Model (PCM) without examining discriminative power parameters. Testing was conducted exclusively with senior high school students in Yogyakarta, limiting national representativeness.
Amalia et al. (2025)	Senior High School / Physics (Momentum & Impulse)	ADDIE model; Rasch analysis; Construct validity (raw variance explained) = 30%; Item reliability = 0.86; α = 0.62; 19 items	Informi was demonstrated to be a valid, reliable, and effective tool for identifying student misconceptions in momentum and impulse. Most students were categorised as understanding the concept or partially understanding, with fewer in the misconception category.	The relatively small sample size (n=37) limits the generalisability of findings. The moderate person reliability index indicates that further instrument refinement is warranted.
Putranta & Afifah (2025)	Senior High School / Physics (Static Fluids)	R&D; Rasch model; Person reliability = 0.73; Item reliability = 0.96; α = 0.72; 17 valid items, 1 invalid item	The four-tier diagnostic test for Static Fluids was demonstrated to be valid and reliable. No items were classified as easy; the majority fell in the moderate-to-difficult range, indicating that static fluids represents a conceptually challenging topic for senior high school students.	Sample size and geographical scope were limited, necessitating broader replication to strengthen the generalisability of the instrument's psychometric properties.
Habiddin & Page (2019)	University / Chemistry (Chemical Kinetics)	R&D FTDICK (5-stage development); Pearson correlation validity; α = 0.78; 20 items	The FTDICK was successfully developed and validated as a reliable and valid diagnostic instrument for chemical kinetics. The instrument effectively identified first-year undergraduate students' misconceptions and incomplete understanding of chemical kinetics concepts.	Several instrument items required revision or replacement due to excessive difficulty or ineffective distractors, limiting the immediate applicability of the instrument without further refinement.
Yonata et al. (2021)	University / Chemistry (Chemical Kinetics)	R&D (Barkman 2002, 6-stage design); Content and construct validity confirmed theoretically; 19 items	The four-tier diagnostic test for Chemical Kinetics concepts was demonstrated to be theoretically valid in terms of content and construct. The instrument enables the identification of student misconception profiles and conceptual understanding of chemical kinetics at the undergraduate level.	Quantitative psychometric indices (Aiken's V, α) were not reported. The authors acknowledged that the instrument requires minor revisions prior to broader implementation.
Laliyo et al. (2021)	Senior High School / Chemistry (States of Matter)	R&D 4-D model; Fleiss κ = 0.97; α = 0.84; 20 items	The integration of the Four-Tier Multiple-Choice (4TMC) test and Partial Credit Model was demonstrated to be effective and valid for measuring students' learning progress. Low-ability students showed	The instrument may misrepresent student reasoning in attempts to connect phenomena and concepts measured in each item. No items assessed students' heuristic reasoning capabilities, representing a gap in the

Author's & Year	Level & Domain	Methods & Psychometrics	Key Finding	Limitations
			slower progress due to lack of knowledge and misconceptions in explaining the concept of changes in states of matter.	instrument's diagnostic scope.
Wu et al. (2025)	Upper Secondary / Chemistry (Isomers)	R&D; Confidence Rating Factor for both answer and reason tiers; N=385 tests analysed; 8 misconceptions identified	Students demonstrated generally suboptimal mastery of isomer conceptual understanding. Scores and confidence ratings were significantly higher in the answer tier than the reason tier. Eight misconceptions were identified and classified as moderate, serious, and typical across four content dimensions, with the greatest difficulties in number judgment and writing tasks.	The instrument was developed specifically for isomer content in Chinese upper secondary education, requiring contextual adaptation and revalidation before application in other national curricula.
Putica (2023)	Secondary School / Biochemistry (Amino Acids, Proteins & Enzymes)	Three-phase R&D; α = 0.76 (cognitive scores), α = 0.87 (confidence ratings); Test-retest r = 0.74 (cognitive), r = 0.88 (confidence); 8 items	Amino acids, proteins, and enzymes are conceptually challenging for secondary school students. The 4AAPE test proved to be a reliable and valid instrument for assessing conceptual understanding, and demonstrated utility in helping teachers develop effective remediation and prevention strategies.	With only eight items, the 4AAPE does not constitute a comprehensive diagnostic test of the full biochemistry curriculum. Many additional items could be developed to provide a more complete picture of students' conceptual understanding.
Desstya et al. (2025)	Primary School / Natural Science (IPA)	R&D 4-D model; Aiken's V = 0.79–0.92; α = 0.86; Pearson r = 0.17–0.58; 61 items	The four-tier diagnostic instrument effectively detected science misconceptions in primary school pupils. Overall, 34% of students held misconceptions, 43% showed lack of knowledge, 12% were guessing, and only 11% demonstrated clear understanding. The circulatory system topic yielded the highest misconception rate (55.55%).	The instrument is a novel tool for primary school-level science misconception diagnosis. Its use at the elementary level remains limited and requires further development and broader application to establish widespread validity.
Atmojo et al. (2024)	Primary School / Natural Science (IPA)	Qualitative (descriptive method); 20 items	The mean misconception rate among Grade V students at SDN 1 Krasak Boyolali was 29.33%, classified as overall low but still moderate in certain subtopics. The highest misconception rate occurred in the temperature-heat differentiation subtopic (36.65%), while the lowest was in the conductor-insulator subtopic (23.32%).	Student misconceptions were largely attributable to erroneous prior conceptions from everyday experience, limited learning motivation, inappropriate reasoning, difficulty connecting prior knowledge with new content, and insufficient instructional strategies.
Önder Çelikkanlı & Kızılcık (2022)	Cross-level / Physics (Systematic Review)	Systematic literature review; 69 studies analysed (58 development-focused + 11 application-	No general consensus was found regarding misconception criteria across the reviewed studies. Indonesia was identified as the country	The review was restricted to physics education only and did not encompass chemistry, biology, or other science domains, limiting the scope of conclusions

Author's & Year	Level & Domain	Methods & Psychometrics	Key Finding	Limitations
		focused); Scope: 2010–2022	with the highest number of four-tier test studies. The majority of studies focused on secondary and tertiary levels, with very few addressing primary education or non-physics science domains.	that can be drawn about cross-domain applications.
Soeharto et al. (2019)	Cross-level / General Science (Review)	Systematic literature review (PRISMA); 111 articles reviewed (33 Physics, 12 Chemistry, 15 Biology)	Student misconceptions were most prevalent in physics (33 concepts), followed by biology (15) and chemistry (12). Multi-tier tests were the most frequently used diagnostic tool (33.06%), followed by simple multiple-choice (32.23%), open-ended tests (23.97%), and interviews (10.74%).	Each diagnostic instrument type carries inherent advantages and limitations. Misconceptions are persistent and difficult to remediate, underscoring the critical importance of early identification.
Gurel et al. (2015)	Cross-level (Secondary, University, Pre-service Teachers) / General Science (Review)	Library research; 273 articles reviewed	Interviews (53%), open-ended tests (34%), multiple-choice tests (32%), and multi-tier tests (13%) were the most commonly used diagnostic methods. Combining multiple diagnostic methods was found preferable over single-method approaches.	Multi-tier diagnostic instruments remain underutilised across all science domains and require greater emphasis. Although four-tier tests address many shortcomings of simpler instruments, they demand longer testing time and may not be suitable for achievement assessment purposes.

The 16 studies included in this review span physics (n = 6), chemistry (n = 4), biology and biochemistry (n = 3), and general science reviews (n = 3) covering educational levels from primary school to university across nine countries. The distribution reflects physics as the dominant domain in four-tier test research, consistent with Önder Çelikkanlı & Kızılcık (2022), who identified 69 physics-focused four-tier studies between 2010 and 2022.

1. Psychometric Properties (RQ1)

Content validity across studies was predominantly established using Aiken's V (range: 0.75 - 0.94), consistently meeting the accepted threshold of ≥ 0.75. Internal consistency varied considerably (Cronbach's alpha ranged from 0.62 to 0.90), a pattern that reflects the multidimensional nature of four-tier instruments rather than methodological weakness. As Putica (2023) demonstrated by reporting separate alpha values for cognitive scores (α = 0.76) and confidence ratings (α = 0.87). Studies employing Rasch-based approaches by Amalia et al. (2025) and Putranta & Afifah (2025) consistently yielded higher item reliability indices (0.86 - 0.96), corroborating Ma et al. (2025) recommendation that modern test theory offers diagnostic advantages over classical methods.

2. Effectiveness in Diagnosing Misconception (RQ2)

Misconception prevalence across studies ranged from 29.33% to 40.7% at the student level, with topic-specific rates reaching as high as 75% in certain subtopics (Atmojo et al., 2024; Tumanggor et al., 2020). Critically, the dual confidence rating mechanism by Caleon & Subramaniam (2010) consistently proved effective in distinguishing genuine misconceptions from guessing and lack of knowledge across all domains reviewed, a capability confirmed by Gurel et al. (2015) as the defining advantage of four-tier over simpler diagnostic formats.

3. Methodological Approaches (RQ3)

R&D designs, predominantly 4-D and ADDIE models, dominated instrument development across the corpus. However, a critical inconsistency identified by Önder Çelikkanlı & Kızılcık (2022) persists, no consensus exists on misconception classification criteria, with different studies adopting varying thresholds and decision rules. This methodological heterogeneity directly undermines the comparability of prevalence rates across studies and represents the most pressing standardisation challenge in the field.

4. Applications Across Levels and Domains (RQ4)

Of the 16 included studies, only two specifically targeted primary school students from Desstya et al. (2025) and Atmojo et al. (2024) confirming the pronounced underrepresentation of elementary-level research documented in the Introduction. Indonesia emerges as the most prolific contributor to four-tier test literature, yet the concentration of Indonesian studies at secondary level suggests that the primary school context remains a significant research frontier even domestically (Önder Çelikkanlı & Kızılcık, 2022).

5. Comparative Advantages and Limitation (RQ5)

Four-tier tests consistently outperform simpler diagnostic formats in identifying the full spectrum of student understanding (from scientific conception to misconception to lack of knowledge) across all reviewed domains (Gurel et al., 2015; Soeharto et al., 2019). Nevertheless, two structural limitations recur: narrow conceptual scope per instrument as few as 8 items in Putica (2023) and small geographically restricted samples that constrain generalisability. These limitations do not diminish the instrument's diagnostic value but underscore the need for larger-scale, cross-contextual validation studies particularly at the primary school level where the evidence base remains thin.

CONCLUSION

This systematic literature review synthesises evidence from 16 studies on 4-Tier Test instruments across science education, confirming that these instruments consistently demonstrate robust psychometric properties with Aiken's V values ranging from 0.75 to 0.94, Cronbach's alpha from 0.62 to 0.90, and prove effective in diagnosing misconceptions across physics, chemistry, biology, and general science domains at multiple educational levels. The dual confidence rating mechanism uniquely distinguishes genuine misconceptions from lack of knowledge and guessing, an advantage confirmed across all reviewed studies. Theoretically, this review establishes that the four-tier structure operationalises both cognitive and metacognitive assessment simultaneously, aligning with Facione (2015) critical thinking framework in ways that conventional instruments cannot.

Practically, teachers are encouraged to use 4-Tier Tests as diagnostic rather than achievement instruments, while assessment developers should prioritise Rasch-based validation approaches given their demonstrated superiority over classical methods. For educational policymakers, investment in professional development is essential to ensure effective interpretation of the diagnostic information these instruments generate.

Nevertheless, this review acknowledges several limitations: the corpus is dominated by physics studies and secondary-level contexts, with only two studies targeting primary school students; methodological heterogeneity in misconception classification criteria limits cross-study comparability; and publication bias towards positive findings cannot be excluded. Future research should prioritise cross-domain and cross-level validation, standardisation of misconception criteria, and expansion of four-tier test applications to elementary education in a frontier that remains critically underexplored despite its importance for early misconception identification.

REFERENCES

Amalia, S. A., Sulisworo, D., Fratiwi, N. J., Nurdini, Novia, H., Samsudin, A., Nasbey, H., Wibowo, F. C., & Sozbilir, M. (2025). A Rasch-Validated Four-Tier Instrument to Diagnose Students' Conceptions of Momentum and Impulse. *Journal of Science Learning*, *8*(4), 401–415.

Atmojo, I. R. W., Saputri, D. Y., & Fadhilah, A. N. (2024). Misconceptions about Science Learning Materials of Class V in Elementary Schools using A Diagnostic Instrument of Four-Tier Multiple Choice. *Jurnal Pendidikan Dan Pengajaran*, *57*(3), 619–630. https://doi.org/10.23887/jpp.v57i3.73841

Caleon, I. S., & Subramaniam, R. (2010). Do Students Know What They Know and What They Don't Know? Using a Four-Tier Diagnostic Test to Assess the Nature of Students' Alternative Conceptions. *Research in Science Education*, *40*(3), 313–337. https://doi.org/10.1007/s11165-009-9122-4

Changwong, K., Sukkamart, A., & Sisan, B. (2018). Critical Thinking Skill Development: Analysis of a New Learning Management Model for Thai High Schools. *Journal of International Studies*, *11*(2), 37–48. https://doi.org/10.14254/2071-8330.2018/11-2/3

Desstya, A., Sayekti, I. C., Abduh, M., & Sukartono. (2025). Development of a Four-Tier Diagnostic Test for Misconceptions in Natural Science of Primary school Pupils. *Journal of Turkish Science Education*, *22*(2), 338–353. https://doi.org/10.36681/tused.2025.017

Duran, T., & Dikmenli, M. (2024). Use of Multi-Tier Concept Diagnostic Tests in Biology Education: A Systematic Review of the Literature. *Journal of Education in Science, Environment and Health*, 224–244. https://doi.org/10.55549/jeseh.755

Facione, P. (2015). Critical Thinking: What It Is and Why It Counts. In *Insight Assessment*.

Fratiwi, J. N., Kaniawati, I., Suhendi, E., Suyana, I., & Samsudin, A. (2017). The Transformation of Two-Tier Test Into Four-Tier Test on Newton's Laws Concepts. *AIP Conference Proceedings*, 050011. https://doi.org/10.1063/1.4983967

Gurel, K. D., Eryilmaz, A., & McDermott, C. L. (2015). A Review and Comparison of Diagnostic Instruments to Identify Students' Misconceptions in Science. *EURASIA: Journal of Mathematics, Science and Technology Education*, *11*(5). https://doi.org/10.12973/eurasia.2015.1369a

Habiddin, & Page, E. M. (2019). Development and Validation of a Four-Tier Diagnostic Instrument for Chemical Kinetics (FTDICK). *Indonesian Journal of Chemistry*, *19*(3), 720. https://doi.org/10.22146/ijc.39218

Istiyono, E., Dwandaru, W. S. B., Fenditasari, K., Ayub, M. R. S. S. N., & Saepuzaman, D. (2023). The Development of a Four-Tier Diagnostic Test Based on Modern Test Theory in Physics Education. *European Journal of Educational Research*, *12*(1), 371–385. https://doi.org/10.12973/eu-jer.12.1.371

Kaltakci-Gurel, D., Eryilmaz, A., & McDermott, C. L. (2017). Development and Application of a Four-Tier Test to Assess Pre-Service Physics Teachers' Misconceptions About Geometrical Optics. *Research in Science & Technological Education*, *35*(2), 238–260. https://doi.org/10.1080/02635143.2017.1310094

Laliyo, L. A. R., Hamdi, S., Pikoli, M., Abdullah, R., & Panigoro, C. (2021). Implementation of Four-Tier Multiple-Choice Instruments Based on the Partial Credit Model in Evaluating Students' Learning Progress. *European Journal of Educational Research*, *10*(2), 825–840. https://doi.org/10.12973/eu-jer.10.2.825

Lufita, D., Kuswanto, H., Rosana, D., Pratiwi, F. A. I., & Triananda, L. (2025). Identification of Misconception Using Diagnostic Tests: Systematic Literature Review. *Jurnal Penelitian Pendidikan IPA*, *11*(5), 10–15. https://doi.org/10.29303/jppipa.v11i5.9986

Ma, H., Yang, H., Li, C., Ma, S., & Li, G. (2025). The Effectiveness and Sustainability of Tier Diagnostic Technologies for Misconception Detection in Science Education: A Systematic Review. *Sustainability*, *17*(7), 3145. https://doi.org/10.3390/su17073145

Önder Çelikkanlı, N., & Kızılcık, H. Ş. (2022). A Review of Studies About Four-Tier Diagnostic Tests in Physics Education. *Journal of Turkish Science Education*, *19*(4), 1291–1311. https://doi.org/10.36681/tused.2022.175

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Putica, K. B. (2023). Development and Validation of a Four-Tier Test for the Assessment of Secondary School Students' Conceptual Understanding of Amino Acids, Proteins, and Enzymes. *Research in Science Education*, *53*(3), 651–668. https://doi.org/10.1007/s11165-022-10075-5

Putranta, H., & Afifah, F. (2025). Development of the Four-Tier Diagnostic Test to Identify Student Misconceptions in the Static Fluids Chapter. *Journal on Efficiency and Responsibility in Education and Science*, *18*(4), 268–281. https://doi.org/10.7160/eriesj.2025.180403

Retnawati, H., Djidu, H., Kartianom, K., Apino, E., & Anazifa, R. D. (2018). Teachers' Knowledge About Higher-Order Thinking Skills and it's Learning Strategy. *Problems of Education in the 21st Century*, *76*(2), 215–230. https://doi.org/10.33225/pec/18.76.215

Soeharto, S., Csapó, B., Sarimanah, E., Dewi, F. I., & Sabri, T. (2019). A Review of Students' Common Misconceptions in Science and Their Diagnostic Assessment Tools. *Jurnal Pendidikan IPA Indonesia*, *8*(2). https://doi.org/10.15294/jpii.v8i2.18649

Tumanggor, R. M. A., Supahar, Ringo, S. E., & Harliadi, D. M. (2020). Detecting Students' Misconception in Simple Harmonic Motion Concepts Using Four-Tier Diagnostic Test Instruments. *Jurnal Ilmiah Pendidikan Fisika Al-Biruni*, *9*(1), 21. https://doi.org/10.24042/jipfalbiruni.v9i1.4571

Widana, I. W., Parwata, Y. M. I., Parmithi, N. N., Jayantika, T. A. G. I., Sukendra, K., & Sumandya, W. I. (2018). Higher Order Thinking Skills Assessment Towards Critical Thinking on Mathematics Lesson. *International Journal of Social Sciences and Humanities (IJSSH)*, *2*(1), 24–32. https://doi.org/10.29332/ijssh.v2n1.74

Wu, M., Tian, P., Sun, D., Feng, D., & Luo, M. (2025). Evaluating Students' Conceptual Understanding of Isomers Based on a Four-Tier Diagnostic Tool in Upper Secondary Schools. *International Journal of Science and Mathematics Education*, *23*(4), 907–947. https://doi.org/10.1007/s10763-024-10494-y

Yonata, B., Suyono, & Azizah, U. (2021). Four-Tier Diagnostic Test on Chemical Kinetics Concepts for Undergraduate Students. *International Joint Conference on Science and Engineering 2021 (IJCSE 2021)*, *209*, 457–463.