Jurnal Pendidikan MIPA
e-ISSN: 2685-5488 | p-ISSN: 1411-2531
URL: https://jpmipa.fkip.unila.ac.id/index.php/jpmipa
Jounal's DOI: http://doi.org/10.23960/jpmipa
Vol. 27, No. 3, 2026
Article ID: 1417
Pages: 1480-1500

The Evolution of Four-Tier Diagnostic Tests in Physics Education: A Systematic Literature Review

Arian Trianov Solistio[1,2], Achmad Samsudin[1,*], Winny Liliawati[1], Andi Suhandi[1], & Fitria Arifiyanti[1]
[1]Department of Physics Education, Universitas Pendidikan Indonesia, Indonesia
[2]Doctoral School of Education, University of Szeged, Hungary

*Corresponding email: achmadsamsudin@upi.edu

Received: 09 June 2026 *Accepted: 04 August 2026* *Published: 26 August 2026*

Abstract: This review examines Four-Tier Diagnostic Tests (FTDTs) in physics education, focusing on purposes, designs, models, scoring, measurement evidence, administration modes, and conceptual-change roles. Following PRISMA, Scopus and ERIC were searched for empirical studies published between 2021 and 2025. The final corpus comprised 23 studies: 14 journal articles and nine conference papers. Data were analyzed using descriptive statistics and framework-based narrative synthesis. Eleven studies used classical FTDTs, seven used response-enhanced FTDTs, and five used analytically advanced FTDTs. Most retained the answer–confidence–reason–confidence structure. In contrast, others added open or semi-open reasoning, scaled confidence, mental-model analysis, Rasch or Partial Credit Model scoring, and cognitive diagnostic modeling. Rule-based categorical diagnosis was the most common scoring approach. Among nine development studies, content review and reliability were common, but classification accuracy, fairness, and cross-context stability were limited. Research focused mainly on senior high school, especially mechanics and fluids. Paper-based administration was the most common single mode, although digital formats were also common. FTDTs primarily supported diagnostic profiling and intervention evaluation, whereas fewer studies examined patterns of conceptual change. No study clearly used individual profiles to adapt instruction. FTDT research has progressed toward richer responses, advanced scoring, and conceptual-change evaluation. Future studies should strengthen measurement evidence, standardize scoring, expand cross-context validation, and develop digital systems that link diagnosis to targeted remediation, monitoring, and sustained instructional support.

Keywords: four-tier diagnostic test, physics education, misconceptions, diagnostic assessment, conceptual change.

Article's DOI: https://doi.org/10.23960/jpmipa.v27i3.pp1480-1500

INTRODUCTION

Physics education is not limited to the mastery of formulas and physical laws. Students must also develop scientific understanding by connecting new knowledge with concepts they already hold (Isra & Mufit, 2023; Samsudin et al., 2021). However, this process is often difficult. Many physics concepts are abstract, mathematically demanding, represented in different forms, and applied in situations that differ from those presented during instruction (Neumann & Hopf, 2017). These difficulties become greater when scientific explanations conflict with students' existing ideas and everyday intuitions. As a result, students may solve numerical problems correctly without fully understanding the physical concepts involved (Espinoza, 2004).

From a constructivist perspective, students do not enter the classroom without prior knowledge. They bring personal experiences, previous learning, and intuitive explanations that influence how they interpret new information (Jung, 2020). Prior knowledge can support learning when it is consistent with accepted scientific concepts. However, it can also lead to misconceptions when students' explanations conflict with scientific understanding (Guerra-Reyes et al., 2024; Smith et al., 1994). Misconceptions are not simply a lack of knowledge or temporary errors. They are explanations that students consider reasonable and may hold with high confidence (Allen, 2014). Because these ideas are connected to students' existing cognitive structures, they may remain even after formal instruction (Clement, 1982). Effective physics instruction, therefore, requires both suitable teaching strategies and diagnostic methods to identify students' conceptual states before remediation is planned.

Misconceptions cannot be identified accurately by examining only whether an answer is correct or incorrect. Conventional multiple-choice tests are efficient for assessing large groups, but they provide limited information about whether a correct response reflects genuine understanding, memorization, uncertainty, or guessing (Treagust, 1986). Tiered diagnostic assessments were developed to address this limitation. One of these instruments is the Four-Tier Diagnostic Test (FTDT). The FTDT contains four components: a conceptual answer, confidence in the answer, a conceptual reason, and confidence in the reason (Caleon & Subramaniam, 2010; Kaltakci-Gurel et al., 2017). This structure provides more detailed evidence about students' conceptual states than conventional multiple-choice assessment (Fratiwi et al., 2017). Depending on the decision rules used, FTDT responses can distinguish scientific understanding, partial understanding, lack of knowledge, guessing, error, and high-confidence misconception (Tumanggor et al., 2020).

Recent studies show that FTDTs differ in both response structure and analytical approach. Some instruments retain fixed answer and reason options, whereas others use open or semi-open reasoning and multilevel confidence scales (Diani et al., 2019; Samsudin et al., 2024a; Umar et al., 2024). Several studies have also used Rasch analysis, Partial Credit Models, Item Response Theory, mental-model classification, or cognitive diagnostic modeling to provide more detailed diagnostic information (Istiyono et al., 2023; Mi et al., 2023; Zain et al., 2025). In addition, FTDTs have been administered through printed questionnaires, online forms, mobile applications, and web-based systems (Astuti, Bhakti, Prasetya, et al., 2023; Maison et al., 2022). These differences influence the type of conclusions that can be drawn. An FTDT may only classify a student's conceptual status, but it may also examine reasoning, estimate confidence, produce attribute-mastery profiles, or track changes after instruction. A focused review must therefore examine instrument structure, scoring, measurement evidence, implementation, and pedagogical use as related dimensions.

Earlier reviews provide an important foundation, but their purposes were broader than the present study's focus. Resbiantoro et al. (2022) reviewed 72 articles published between 2005 and 2020 on the diagnosis, causes, and remediation of misconceptions in physics education. Their review compared interviews, open-ended tests, multiple-choice tests, and multi-tier diagnostic instruments. It also identified several instructional approaches for addressing misconceptions. However, FTDT was discussed as one diagnostic method among several alternatives. Consequently, the review did not examine the development of FTDT structure, scoring, validation, administration, and pedagogical use in detail.

Çelikkanlý & Kýzýlcýk (2022) focused more directly on four-tier diagnostic tests in physics education. They reviewed 69 studies published between 2010 and 2022, including 58 instrument-development studies and 11 studies that used existing four-tier tests to identify misconceptions. Their analysis covered publication characteristics, physics topics, targeted misconceptions, instrument modification, misconception criteria, validity, reliability, and number of items. They found that Indonesia contributed the largest number of studies, that the criteria for misconceptions differed across articles, and that important instrument-development procedures were often not reported. However, their review mainly described general development patterns. It did not use a study-level framework to distinguish FTDT model families, primary scoring paradigms, digital administration modes, advanced measurement evidence, or the different roles of FTDTs in conceptual-change studies.

Marjana et al. (2026) examined misconceptions in the broader field of physics through a citation-based systematic mapping review. They searched Scopus for studies published between 2015 and 2024, screened the 100 most-cited records, and included 74 studies. Their review mapped research trends, misconception profiles, diagnostic tools, and remediation strategies. Four-tier diagnostic tests were identified as one of the most frequently used diagnostic instruments. However, FTDT remained one method within the wider field of misconception research. In addition, ranking records by citation count was suitable for identifying influential publications but may have underrepresented recent studies that had not yet accumulated many citations. Citation-based mapping also provides limited evidence on how individual FTDT studies define their models, map responses to diagnostic categories, assess measurement quality, and use diagnostic results during instruction.

These limitations underscore the need for a focused, up-to-date review of FTDT research. Existing reviews have provided broad accounts of misconception diagnosis or descriptive summaries of four-tier test development. However, they have not systematically connected instrument architecture, scoring, measurement evidence, administration mode, and pedagogical role at the study level. The present study addresses this gap through a PRISMA-based selection process and a predefined extraction and coding framework. It reviews empirical FTDT studies in physics education published from January 2021 to December 2025. Each study is examined in relation to its publication characteristics, research purpose, methodological design, dominant FTDT model, scoring paradigm, measurement evidence, mode of administration, educational level, physics domain, and role in the conceptual-change process.

The review addresses the following research questions: 1) How did FTDT research in physics education develop between 2021 and 2025 in terms of publication characteristics, research purposes, and methodological designs? 2) What dominant FTDT models were used, and how did they differ in response architecture and level of diagnostic inference? 3) What primary scoring paradigms and secondary analytical procedures were used to classify, quantify, and interpret FTDT responses? 4) What measurement evidence was reported, and how were FTDTs implemented across educational levels, physics domains, and administration modes? 5) How were FTDTs used in the conceptual-change process, including initial diagnosis, intervention evaluation, conceptual-change pattern measurement, and the use of diagnostic results to guide instruction?

METHOD

Research Design and Study Selection Procedure

This study used a systematic literature review based on the PRISMA 2020 framework to ensure a transparent process of study identification, screening, eligibility assessment, and inclusion (Page et al., 2021). The review addressed five analytical domains: study characteristics and research orientation; FTDT models; scoring and analytical paradigms; measurement evidence and implementation contexts; and the use of FTDT in conceptual-change research.

Search Strategy

The literature search was conducted in Scopus and the Education Resources Information Center (ERIC). Scopus was selected for its broad multidisciplinary coverage, while ERIC was included to strengthen retrieval of research on education, assessment, teaching, and learning. The database search was last updated on July 28, 2026. Eligibility was limited to studies published between January 1, 2021, and December 31, 2025. Therefore, publications dated 2026 were not eligible. The following Boolean expression was used:

("four-tier" OR "4-tier" OR "four-level") AND ("diagnostic" OR "assessment" OR "instrument") AND ("misconception" OR "conceptual understanding") AND ("physics" OR "physics learning")

The same conceptual search string was used in both databases, with minor adjustments to the search syntax and searchable fields available in Scopus and ERIC. In ERIC, an additional filter was applied to retrieve only peer-reviewed publications. The database search identified 38 records, comprising 27 from Scopus and 11 from ERIC. Four duplicate records were identified by comparing titles, authors, publication years, and DOIs. After the duplicates were removed, 34 unique records remained for screening.

Inclusion and Exclusion Criteria

Studies were included if they reported original empirical research, used a four-tier diagnostic test to investigate physics concepts or physics learning, were published between 2021 and 2025 as journal articles or individual conference papers, and provided sufficient information on the research methods, participants, instruments, analytical procedures, and main findings to support data extraction and synthesis. Studies were excluded when they were reviews, conceptual or theoretical papers, editorials, book chapters, complete conference proceedings volumes, or other non-empirical publications. Studies using only one-tier, two-tier, or three-tier tests were excluded. Studies combining physics with other science disciplines were also excluded when physics-specific data could not be extracted separately.

No formal language restriction was applied. All records retained during screening were written in English, and no study was excluded because of language. Of the 34 unique records, one was excluded during title screening because it was a complete conference proceedings volume rather than an individual paper. Nine records were excluded during abstract screening: three review articles, three studies that did not use an FTDT, one non-empirical conceptual paper, and two studies without separately extractable physics-education evidence.

The remaining 24 reports were retrieved and assessed in full. No report was unavailable. One report was excluded after full-text assessment because its main disciplinary context was chemistry or general science rather than physics education. The final synthesis therefore included 23 empirical studies published between 2021 and 2025. The complete record-level screening decisions and reasons for exclusion are reported in Appendix A. The study-selection process is presented in Figure 1.

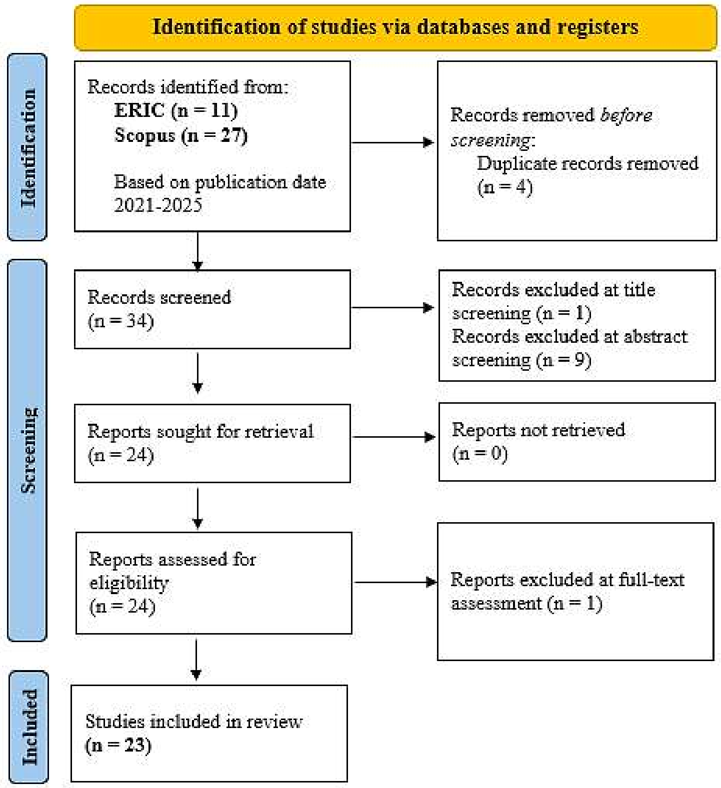

Figure 1. PRISMA 2020 flow diagram of the study-selection process

Data Extraction and Analysis

The first researcher conducted the title, abstract, and full-text screening using the predefined criteria. The second researcher verified the exclusion decisions and recorded reasons for exclusion. Uncertain cases were resolved through discussion. Data from the eligible studies were entered into a standardized extraction matrix aligned with the research questions. Each study received a unique Study ID. The extracted variables included publication characteristics, research design, participants, educational level, physics topic, research purpose, FTDT structure, scoring procedure, measurement evidence, administration mode, technology integration, conceptual-change role, key findings, and research gaps.

Validity and reliability findings were recorded as FTDT measurement evidence only when they directly referred to the diagnostic instrument, its items, or its classifications. Evidence concerning an instructional product, such as an e-module or simulation, was not treated as FTDT measurement evidence. Each study was assigned to one dominant FTDT model, namely the classical FTDT, the response-enhanced FTDT, or the analytically advanced FTDT, based on the main structure of the instrument and the level of diagnostic inference it produced.

Open or semi-open reasoning and multilevel confidence scales were coded as response enhancements. Rasch models, Partial Credit Models, Item Response Theory, mental-model analysis, and cognitive diagnostic models were considered model-defining only when they formed the main basis of diagnostic inference. Each study was also assigned one primary scoring paradigm representing the main procedure used to transform four-tier responses into conceptual categories, numerical scores, mental representations, latent-trait estimates, or cognitive-attribute profiles. Secondary procedures, including N-Gain, Reduction of

Misconception Quantity, McNemar tests, Wilcoxon tests, and conceptual-transition analysis, were coded separately.

Research purpose and conceptual-change-role codes were not mutually exclusive because a single study could serve multiple functions. A study could also report several forms of measurement evidence or technology integration. In contrast, each study received only one dominant FTDT model, one primary scoring paradigm, one educational-level category, one primary physics domain, and one administration-mode category. Administration mode was classified as paper-based, digital/online, hybrid, or not reported. Digital platforms, simulations, electronic modules, and computer-assisted instruction were not treated as changes to the FTDT model unless they directly changed the response architecture or the basis of diagnostic inference. The complete coding framework and study-level extraction matrix are provided in **Appendix B.**

Two researchers independently coded all 23 studies for the dominant FTDT model and primary scoring paradigm using the predefined coding framework. Inter-rater reliability was calculated separately for the two nominal variables using unweighted Cohen's kappa in IBM SPSS Statistics version 27. Agreement was high for both the dominant FTDT model (ê = .932; observed agreement = 95.7%; $p < .001$) and the primary scoring paradigm (ê = .939; observed agreement = 95.7%; $p < .001$). Cohen's kappa was calculated before consensus discussion. The remaining disagreements were resolved by reviewing the operational definitions and evidence reported in the full texts.

The extracted data were analyzed using descriptive statistics and framework-based narrative synthesis following Popay et al. (2006). Frequencies and percentages were used to summarize the main study characteristics and coding categories, while narrative synthesis was used to compare differences across studies. Each study was assigned one primary physics domain to avoid double counting, and studies covering distinct domains were classified as multiple-domain studies. Administration modes and the distribution of physics domains across educational levels were visualized using two-dimensional horizontal bar charts and a proportional bubble chart generated in Python with Matplotlib. Meta-analysis was not conducted because the included studies differed substantially in design, participants, FTDT formats, scoring procedures, interventions, and outcomes.

RESULT AND DISCUSSION

The findings are presented according to the five research questions.

RQ1. Research Characteristics and Methodological Orientation

FTDT studies were published throughout 2021–2025, although the annual number did not

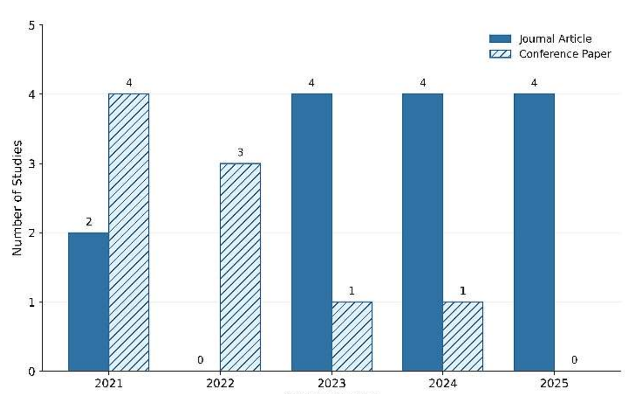

Figure 2. Annual distribution of included FTDT studies by publication type

show a consistent upward trend. Six studies were published in 2021, followed by three in 2022, five each in 2023 and 2024, and four in 2025. The publication channel changed more clearly. Conference papers were more common in 2021 and accounted for all studies published in 2022, whereas journal articles became dominant from 2023 onward. Overall, the review included 14 journal articles and nine conference papers. This pattern suggests a shift toward journal publication, but publication type alone does not indicate methodological quality.

Table 1. Study characteristics and methodological orientation

Dimension	Category	n	%
Publication type	Journal article	14	60.9
	Conference paper	9	39.1
Country	Indonesia	20	87.0
	China	1	4.3
	Greece	1	4.3
	Thailand	1	4.3
Primary research design	Developmental research	10	43.5
	Pre-experimental research	4	17.4
	Quasi-experimental research	1	4.3
	Cross-sectional descriptive research	5	21.7
	Qualitative descriptive research	1	4.3
	Mixed-methods research	2	8.7
Research purpose*	Developer	9	39.1
	Profiler	16	69.6
	Remediator	8	34.8
	Evaluator	8	34.8

Note: Percentages were calculated from the 23 included studies. Primary research-design categories were mutually exclusive, whereas research-purpose categories were not mutually exclusive because one study could serve more than one function.

Research was strongly concentrated in Indonesia, which contributed 20 of the 23 included studies. China, Greece, and Thailand each contributed one study. This pattern is consistent with the earlier review by Çelikkanlý & Kýzýlcýk (2022), which also identified Indonesia as the main contributor to four-tier diagnostic research. However, the current distribution should not be interpreted as a complete global pattern, as the search was limited to Scopus and ERIC.

Developmental research was the most common design, representing 10 studies, followed by cross-sectional descriptive research with five studies. Pre-experimental, quasi-experimental, qualitative, and mixed-methods designs were less frequent. The strong developmental orientation indicates that instrument construction, validation, scoring development, and assessment system design remain central to FTDT research. This focus is reasonable because diagnostic interpretations require adequate evidence that the instrument performs consistently and measures the intended concept (Zach, 2005).

Research design and research purpose were coded separately. Although 10 studies used a developmental design, only nine were classified as FTDT Developers. The difference occurred because one developmental study created an instructional product and used FTDT only to evaluate its outcome. Purpose coding also showed that profiling was the most common FTDT function, appearing in 16 studies. Eight studies implemented remediation, and the same eight studies used FTDT to evaluate instructional outcomes. These results indicate that diagnosis remains the primary purpose of FTDT research, although the instrument is increasingly used alongside instructional interventions and conceptual-change evaluation.

Sample size generally followed the analytical purpose of each study. Large instrument development and psychometric studies included 1,850 participants in Mi et al. (2023), 700 participants in Istiyono et al. (2023), and 300 participants in Istiyono et al. (2024). In contrast, most intervention studies involved smaller groups from one school, class, or institution. This difference suggests that evidence for instrument calibration and diagnostic classification is based on larger datasets, whereas evidence for instructional effectiveness remains largely local and short-term.

RQ2: Development of FTDT Models and Diagnostic Capabilities

Most studies retained the four-tier core, which consists of an answer, confidence in the answer, a reason, and confidence in the reason. The main differences were found in the type of response evidence collected and the level of diagnostic inference produced. Based on these differences, the 23 studies were classified into three dominant FTDT models.

Most studies retained the four-tier core, namely an answer, confidence in the answer, a reason, and confidence in the reason. The main differences were not found in the number of tiers, but in the response architecture and the level of diagnostic inference. Based on these features, the 23 studies were classified into three models: Classical FTDT, Response-Enhanced FTDT, and Analytically Advanced FTDT. Table 4 shows that development has mainly occurred by collecting richer response evidence or producing more detailed diagnostic information.

Table 2. Dominant FTDT models and diagnostic capabilities

FTDT model	Main characteristics	n	%
Classical FTDT	Uses the conventional four-tier structure and interprets responses through predefined conception categories. Digital administration or supporting item analysis does not change this classification.	11	47.8
Response-Enhanced FTDT	Extends the evidence collected through open or semi-open reasoning, student-generated responses, multilevel confidence scales, CRI, or confidence thresholds while retaining the basic categorical logic.	7	30.4
Analytically Advanced FTDT	Uses an advanced analytical architecture to produce latent, multidimensional, mental-representation, or cognitive-attribute information from four-tier responses.	5	21.7

Note: Each study was assigned a dominant FTDT model based on its primary diagnostic contribution. Administration platform and instructional intervention were coded separately

Classical FTDT was the largest model family, covering 11 studies (47.8%). The four-tier format was originally developed by extending the two-tier multiple-choice test with separate confidence ratings for the answer and reason tiers. Therefore, students' content knowledge, explanatory knowledge, and confidence in both responses can be examined together (Caleon & Subramaniam, 2010). The studies in this category retained this answer–confidence–reason–confidence structure and interpreted response combinations through predefined categories, such as scientific conception, partial understanding, lack of knowledge, misconception, and error. This structure remains useful because an incorrect answer does not always indicate a misconception. Bessas et al. (2024), for example, found that many incorrect responses reflected lack of knowledge, while Maison et al. (2022) showed that an increase in correct answers did not always mean that the underlying misconception had changed.

Model classification was based on the instrument's diagnostic architecture rather than its administration platform or supporting analysis. Astuti et al. (2023), for example, implemented the FTDT via an Android application, but its diagnostic output still adhered to conventional rule-based categories. Likewise, Putranta & Afifah (2025) used Rasch analysis to examine item quality, but the main interpretation remained categorical. Both studies were therefore classified as Classical FTDT. Some classical studies also used external evidence to explain the diagnostic results. Mustofa et al. (2024), for instance, combined FTDT results with cognitive-style classification and interviews. However, these additional methods did not change the four-tier response structure itself.

Response-Enhanced FTDT was identified in seven studies (30.4%). These instruments retained the basic four-tier logic but collected richer evidence through open or semi-open reasoning, student-generated responses, CRI, multilevel confidence scales, or confidence thresholds. Samsudin et al. (2024) used open or semi-open reasoning formats, while Mahardika et al. (2025) and Surmaini et al. (2021) combined open-ended reasoning with a certainty response index. Such features may reveal reasoning not represented in fixed response options and may provide clearer evidence for distinguishing among uncertainty, lack of knowledge, and confident misconception (Kaltakci-Gurel et al., 2017; Resbiantoro et al., 2022). However, richer responses also require clear coding rules so that students' explanations and confidence levels are interpreted consistently.

Five studies (21.7%) were classified as Analytically Advanced FTDT. These studies generally retained the four-tier response structure but used an advanced analytical architecture to produce more detailed inferences. Mi et al. (2023), for example, used a Q-matrix and cognitive diagnostic modeling to estimate students' mastery of specific conceptual attributes. Likewise, (2024) interpreted response patterns as scientific, synthetic, misconceptions, and initial mental models. Other studies used Rasch or Partial Credit Models as the main scoring architecture to represent item difficulty, student ability, and levels of conceptual understanding (Istiyono et al., 2023, 2024). These approaches extend FTDT beyond broad-conception labels, although their conclusions depend on adequate sample size, model fit, and clearly defined diagnostic outputs.

Overall, Classical FTDT remained the largest individual category, but Response-Enhanced and Analytically Advanced FTDTs together represented 12 studies (52.2%). This finding shows that recent FTDT research has followed two main paths: enriching the evidence collected from students and increasing the detail of diagnostic inference. These models should not be viewed as fixed generations or as a hierarchy of quality. Classical FTDT supports direct categorical diagnosis, Response-Enhanced FTDT provides deeper evidence about reasoning and confidence, and Analytically Advanced FTDT produces more fine-grained diagnostic profiles. FTDT development can therefore be described as core-preserving but capability-expanding.

RQ3: Primary Scoring Paradigms and Secondary Analytical Procedures

The FTDT model describes the instrument's structure and the level of diagnostic inference, whereas the scoring paradigm explains how four-tier responses were transformed into diagnostic information. Each study was therefore assigned one primary scoring paradigm. Statistical tests and conceptual-change measures applied after the initial scoring were recorded separately as secondary analytical procedures.

Table 3. Primary scoring paradigms used in FTDT studies

Primary scoring paradigm	Main function	n	%
Rule-Based Categorical Diagnosis	Converts combinations of answer correctness, reason correctness, and confidence into predefined conceptual-status categories.	11	47.8
Confidence-Index Categorical Diagnosis	Uses CRI, multilevel confidence, or confidence thresholds together with answer and reason correctness.	4	17.4
Mental-Representation Scoring	Interprets response patterns as scientific, synthetic, misconceptions, or initial mental models.	1	4.3
Psychometric Latent-Trait Scoring	Uses Rasch, PCM, or IRT as the main scoring architecture to estimate item and person parameters or ordered conceptual levels.	4	17.4
Cognitive Diagnostic Modeling	Links responses to cognitive attributes through a Q-matrix and a cognitive diagnostic model.	1	4.3
Scoring Procedure Not Sufficiently Reported	Does not provide enough information to determine how four-tier responses were converted into categories or scores.	2	8.7

Rule-Based Categorical Diagnosis was the most common paradigm, covering 11 studies (47.8%). This approach follows the basic logic of the four-tier test, in which answer accuracy, reason accuracy, and confidence are combined to distinguish different conceptual states (Desstya et al., 2025; Kiray & Simsek, 2021). However, the labels and decision rules varied across studies. Similar response patterns were described as a lack of knowledge, lack of understanding, partial understanding, no conception, or error. Bessas et al. (2024), for example, distinguished scientific conception, misconception, lack of knowledge, and guessing, whereas other studies used more detailed category systems. These differences may be instructionally useful, but they reduce direct comparison across instruments.

Confidence-Index Categorical Diagnosis was used in four studies, while one study used Mental-Representation Scoring. CRI and multilevel confidence scales help distinguish confident misconceptions from uncertainty or insufficient knowledge, which is the main purpose of confidence-based diagnosis (Hasan et al., 1999; Koto & Gusma, 2021). Confidence also provides information beyond response accuracy, although its interpretation depends on clear thresholds and students' ability to report their confidence consistently (Stankov & Lee, 2008). Likewise, Umar et al. (2024) extended the interpretation beyond conventional status categories by classifying responses into scientific, synthetic, misconception, and initial mental model categories.

Four studies used Psychometric Latent-Trait Scoring, and one used Cognitive Diagnostic Modeling. Psychometric approaches preserve ordered differences in response quality rather than treating all response combinations as equivalent nominal categories (Fletcher & Hattie, 2004; Vispoel & Kim, 2014). The Partial Credit Model, for example, was designed for items scored in ordered categories and allows category structures to vary across items (Masters, 1982). Accordingly, Istiyono et al. (2024) used PCM-based scoring to estimate item difficulty, student ability, and levels of conceptual understanding. Such interpretations require evidence of model fit, measurement error, reliability, and dimensionality (Saidi & Siew, 2019; Schumacker & Smith, 2007).

Cognitive Diagnostic Modeling produced a different type of information. Mi et al. (2023) used a Q-matrix and G-DINA modeling to link test items with specific electrostatics attributes and estimate students' attribute mastery. CDMs are multiple-classification latent-class models that link observed responses to several underlying skills or attributes rather than reducing performance to a single total score (Templin & Henson, 2006). This approach can therefore identify students who obtain similar overall scores but differ in the specific skills they have mastered, as shown by Zhai et al. (2023). The GDINA framework also supports checks of model fit, Q-matrix validity, differential item functioning, and classification accuracy (Ma & de la Torre, 2020). However, two studies did not report their scoring procedures sufficient detail, limiting reproducibility and comparability across FTDT studies.

Secondary analytical procedures were applied after the primary categories or scores had been produced. These included N-Gain, Reduction of Misconception Quantity, McNemar and Wilcoxon tests, repeated comparisons, and conceptual-transition coding. Samsudin et al. (2024), for example, combined categorical FTDT results with conceptual-change codes and N-Gain, while Adimayuda et al. (2025) used McNemar and RMQ to evaluate changes after instruction. Mahardika et al. (2025) used a reduction index and Wilcoxon testing, whereas Maison et al. (2022) compared changes in response categories across two measurements. These procedures measured the extent or direction of change, but they did not replace the primary FTDT scoring model.

Overall, FTDT scoring did not develop through a simple shift from basic to advanced methods. Instead, the studies used different primary paradigms for different diagnostic purposes and added secondary procedures when examining conceptual change or intervention effects. The central issue is therefore not the absence of analytical variety, but the lack of shared terminology, transparent decision rules, and consistent reporting. Future FTDT studies should clearly separate the original response-scoring procedure from later statistical analyses so that diagnostic results can be reproduced and compared across contexts.

RQ4: Measurement Evidence and FTDT Implementation

Measurement evidence was examined in the nine studies classified as Developers because these studies developed, adapted, calibrated, or validated an FTDT instrument, scoring model, or dedicated assessment system. Evidence concerning instructional products was excluded unless it directly referred to the FTDT instrument or its diagnostic classifications.

Table 4. Measurement evidence reported in FTDT development studies

Measurement evidence	n of 9	%
Content validity or expert review	9	100.0
Internal consistency or measurement reliability	8	88.9
Advanced psychometric or measurement modeling	6	66.7
Diagnostic classification accuracy or consistency	1	11.1
Differential item functioning or measurement fairness	1	11.1

Note: Categories are not mutually exclusive

The findings show that FTDT development did not rely only on expert judgment. Content review was reported in all Developer studies, while several also used Rasch analysis, Partial Credit Models, Item Response Theory, or cognitive diagnostic modeling. However, the depth of measurement evidence remained uneven. Mi et al. (2023) was the only study to report diagnostic classification accuracy and consistency, while Zain et al. (2025) was the only study to examine differential item functioning. Cross-validation, measurement invariance, and the stability of classifications across schools or populations were generally absent. Therefore, the main gap was not the absence of validation, but the limited evidence that FTDT classifications remain accurate and fair across different contexts. Reliability reported for an instructional product was not counted as FTDT reliability; for example, the reliability evidence in Mahardika et al. (2025) mainly concerned the developed e-module rather than the diagnostic test.

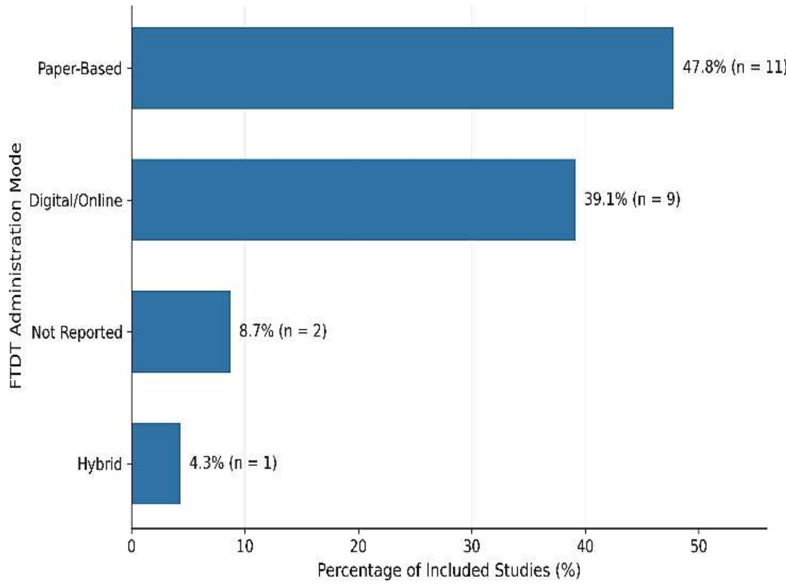

Figure 3. Distribution of FTDT administration modes

Paper-based administration was the most common single mode, appearing in 11 studies (47.8%), followed closely by digital or online administration in 9 studies (39.1%). One study used a hybrid format, while two did not report how the FTDT was administered. When digital and hybrid formats are combined, technology-mediated administration was reported in 10 studies (43.5%), indicating that digital delivery has expanded but has not replaced paper-based testing.

Most digital studies used general platforms, such as Google Forms or other online test systems, to distribute instruments and collect responses (Munggarani et al., 2021; Surmaini et al., 2021). More developed systems included the Android application by Astuti et al. (2023), which automated scoring, misconception classification, and PDF report generation, and the online system by Maison et al. (2022), which produced conceptual profiles upon submission. The Android application's automated classification and reporting functions are recorded in the extraction evidence.

These systems improved data collection and reporting, consistent with the flexibility and monitoring benefits of digital assessment (Jurâne-Brçmane, 2023). However, most digital FTDTs merely transferred conventional tests to online platforms, with little evidence of adaptive item selection, profile-based feedback, personalized remediation, learning-management-system integration, real-time instructional adjustment, or continuous monitoring. They are therefore better described as digitized diagnostic assessments than fully adaptive learning environments (Kaddouri et al., 2025).

FTDT implementation was concentrated at the senior high school level, which accounted for 17 studies (73.9%). Only two studies involved

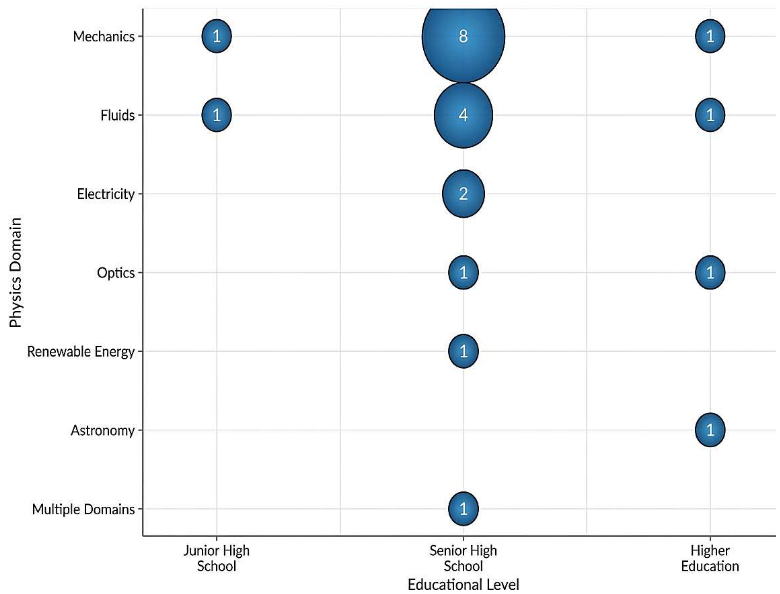

Figure 4. Distribution of physics domains across educational levels

junior high school students, while four were conducted in higher education. Mechanics was the most frequently investigated domain, with 10 studies, followed by fluids with six. Together, these two domains represented 16 studies (69.6%) and were the only domains examined across all three educational levels. Electricity and optics were each represented by two studies, whereas renewable energy, astronomy, and multiple-domain assessment appeared in only one study each.

The concentration on mechanics and fluids is understandable because motion, force, pressure, and buoyancy are closely linked to everyday experience, which may support intuitive explanations that conflict with scientific models (Clement, 1982; diSessa, 1993). Such conceptions can also be reinforced by difficulties in interpreting scientific language and causal relationships (Graham et al., 2013). However, because 73.9% of the studies involved senior high school students and 69.6% focused on mechanics or fluids, the evidence cannot yet be generalized across educational levels or physics domains. Research remains limited in the fields of electricity, optics, thermodynamics, modern physics, junior high school, and higher education.

Higher-education research was also limited to small samples of undergraduates and pre-service teachers. Azizah et al. (2022), for example, identified substantial astronomy misconceptions among physics education students, while Maison et al. (2022) found that additional instruction improved answer accuracy without substantially reducing buoyancy misconceptions. These findings support wider concerns that future teachers may retain conceptual difficulties that could later influence their classroom explanations and instructional decisions (Liu & Li, 2017)

RQ5: FTDT Across the Conceptual-Change Process

The pedagogical roles of FTDT were coded according to functions that were empirically demonstrated in each study, rather than instructional possibilities mentioned only in the implications or recommendations. Because one study could use FTDT at several stages of the conceptual-change process, the categories were not treated as mutually exclusive.

Diagnostic Baseline was the main role of FTDTs, appearing in 22 studies. These studies used FTDTs to identify misconceptions,

Table 5. Pedagogical roles of FTDT across the conceptual-change process

Conceptual-change role	n	%
Diagnostic Baseline	22	95.7
Intervention Guidance	0	0.0
Remediation or Intervention Evaluation	8	34.8
Conceptual-Change Pattern Measurement	6	26.1
Instrument or Scoring Development Only	1	4.3

Note: Categories are not mutually exclusive. Conceptual Change Measurement was coded only for studies examining transitions or changes in response patterns, not for aggregate reductions in misconception alone.

conceptual-status distributions, mental models, or cognitive-attribute mastery, either as a stand-alone diagnosis or before instruction. Only A12 focused solely on instrument and scoring development without reporting substantive conceptual findings. Although several studies administered an FTDT before an intervention, the treatment had already been selected and delivered to the whole group. Thus, using an FTDT as a pretest does not mean that its results guided instruction. This gap supports Resbiantoro et al. (2022), who argued that diagnosis does not automatically produce effective conceptual-change teaching.

Eight studies used FTDTs to evaluate interventions, including conceptual-change tasks (Samsudin et al., 2024), scaffolded simulation (Fongsamut et al., 2023), electronic modules (Mahardika et al., 2025), computer-based refutational text (Sarwono et al., 2022), additional instruction (Maison et al., 2022), PIMCA (Lampeang et al., 2021), T-MORE (Adimayuda et al., 2025) and computer-simulation-assisted instruction (Kaniawati et al., 2021). Sarwono et al. (2022) reported substantial remediation after analogy-supported refutational text. This result agrees with evidence that refutational text can help students revise intuitive ideas (Hynd et al., 1994) and that analogies are more effective when learners understand their connection to the target concept (Brown, 1992).

Similarly, Fongsamut et al. (2023) found that scaffolded simulation improved understanding, although the extent of change differed across projectile-motion misconceptions. Simulations may clarify abstract concepts, but their effectiveness still depends on instructional support (Banda & Nzabahimana, 2023). Six studies went beyond aggregate outcomes and examined category transitions, persistent misconceptions, acceptable and unacceptable changes, or changes in cognitive structure. Maison et al. (2022) showed why this distinction matters: correct responses increased after additional instruction, but the prevalence of misconceptions changed only from 24.74% to 24.32%. Much of the improvement came from reductions in lack of knowledge, false-positive responses, and false-negative responses.

Therefore, higher scores or more correct answers should not automatically be interpreted as successful conceptual change. Misconceptions may represent organized explanations shaped by prior knowledge and everyday experience (Vosniadou & Mason, 2012). They may also arise when learners place a concept in an inappropriate conceptual category, as found in ideas about light, heat, and electric current (Slotta et al., 1995). Conceptual change may therefore require reorganizing explanations rather than simply replacing an incorrect answer with a correct one (Chi et al., 1994).

Confidence and response-pattern data are also important because conceptual change involves metacognitive and motivational factors and may occur gradually across several dimensions (Duit & Treagust, 2003; Pintrich et al., 1993). Overall, the reviewed studies show that FTDTs are well established for initial diagnosis and are increasingly used to evaluate interventions and conceptual transitions. However, they have not yet formed a complete cycle in which individual diagnostic profiles guide targeted instruction, progress is monitored, and persistent misconceptions are reassessed. FTDTs should therefore be viewed not only as tools for classifying misconceptions, but also as frameworks for instructional decisions that support continuous conceptual change, curriculum improvement, and more responsive teaching (Roziqin et al., 2026). The main remaining gap is the limited translation of diagnostic evidence into differentiated and adaptive instructional action.

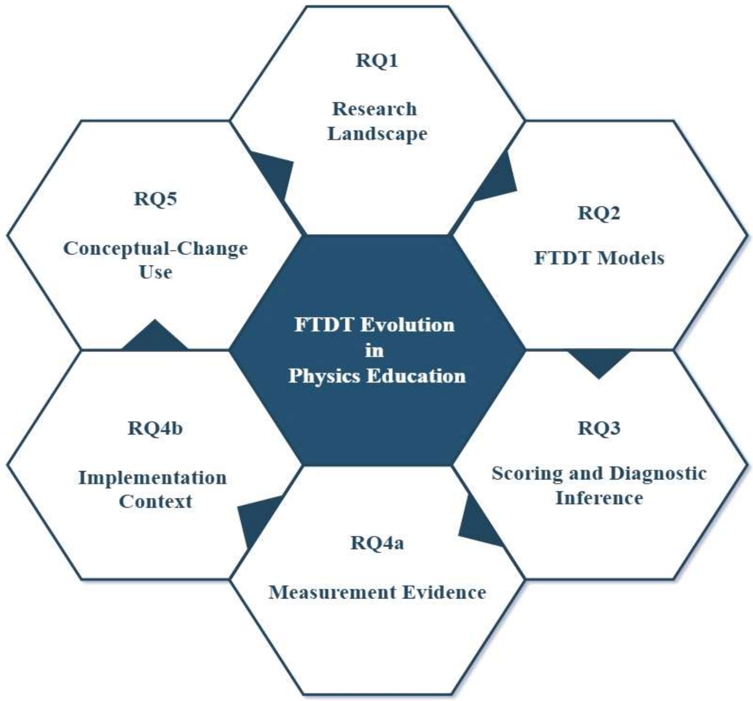

Figure 5. Integrated thematic map of FTDT research in physics education

CONCLUSION

This review shows that Four-Tier Diagnostic Tests (FTDTs) have expanded beyond misconception identification to serve as assessment tools for examining conceptual states, confidence, reasoning, and conceptual change. Their development is marked by richer response formats, advanced scoring, and pedagogical use. However, scoring frameworks remain inconsistent, measurement evidence is uneven, and diagnostic results are rarely translated into differentiated instruction. Future studies should report more than content validity and reliability by explaining scoring rules and confidence thresholds and, where appropriate, providing evidence of measurement structure, classification accuracy, and item fairness. They should also strengthen cross-context validation and develop digital systems that connect diagnosis with targeted remediation, monitoring, and sustained support. These conclusions should be interpreted cautiously because the search was limited to Scopus and ERIC and used English-language terms. Studies indexed elsewhere or published with non-English titles may have been missed. Moreover, 20 of the 23 studies were conducted in Indonesia; therefore, the findings describe the reviewed corpus rather than the complete global development of FTDT research.

DECLARATION OF GENERATIVE AI USAGE IN THE WRITING PROCESS

The authors used OpenAI's ChatGPT to improve language, clarity, organization, and consistency. All data extraction, coding, interpretation, references, and final content were checked by the authors, who take full responsibility for the manuscript.

■ REFERENCES

Adimayuda, R., Suhandi, A., Samsudin, A., Suhendi, E., Setiawan, A., & Fratiwi, N. J. (2025). Breaking Misconceptions: Technology-Integrated MORE Model for Meaningful Learning of Momentum and Impulse. *Online Learning In Educational Research*, *5*(1), 25–40. https://doi.org/10.58524/oler.v5i1.606

Allen, M. (2014). *Misconceptions in Primary Science*. McGraw-Hill Education.

Astuti, I. A. D., Bhakti, Y. B., & Prasetya, R. (2023). Android-based 4-tier physics test app to identify student misconception profiles. *International Journal of Evaluation and Research in Education*, *12*(3), 1356–1363. https://doi.org/10.11591/ijere.v12i3.25536

Astuti, I. A. D., Bhakti, Y. B., Prasetya, R., & Zulherman. (2023). Android-based 4-tier physics test app to identify student misconception profiles. *International Journal of Evaluation and Research in Education*, *12*(3), 1356–1363. https://doi.org/10.11591/ijere.v12i3.25536

Azizah, S. N., Akhsan, H., Muslim, M., & Ariska, M. (2022). Analysis of college students misconceptions in astronomy using four-tier test. *Journal of Physics: Conference Series*, *2165*(1). https://doi.org/10.1088/1742-6596/2165/1/012004

Banda, H. J., & Nzabahimana, J. (2023). The Impact of Physics Education Technology (PhET) Interactive Simulation-Based Learning on Motivation and Academic Achievement Among Malawian Physics Students. *Journal of Science Education and Technology*, *32*(1), 127–141. https://doi.org/10.1007/s10956-022-10010-3

Bessas, N., Tzanaki, E., Vavougios, D., & Plagianakos, V. P. (2024). Diagnosing students' misconception in Hydrostatic Pressure through a 4-tier test. *Heliyon*, *10*(23). https://doi.org/10.1016/j.heliyon.2024.e40425

Brown, D. E. (1992). Using examples and analogies to remediate misconceptions in physics: Factors influencing conceptual change. *Journal of Research in Science Teaching*, *29*(1), 17–34. https://doi.org/10.1002/tea.3660290104

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

Çelikkanlý, N. Ö., & Kýzýlcýk, H. Þ. (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

Chi, M. T. H., Slotta, J. D., & De Leeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. *Learning and Instruction*, *4*(1), 27–43. https://doi.org/10.1016/0959-4752(94)90017-5

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics*, *50*(1), 66–71. https://doi.org/10.1119/1.12989

Desstya, A., Sayekti, I. C., Abduh, M., & Sukartono, S. (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

Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-Tier Diagnostic Test With Certainty of Response Index on The Concepts of Fluid. *Journal of Physics: Conference Series*, *1155*, 012078. https://doi.org/10.1088/1742-6596/1155/1/012078

diSessa, A. A. (1993). Toward an Epistemology of Physics. *Cognition and Instruction*, *10*(2–3), 105–225. https://doi.org/10.1080/07370008.1985.9649008

Duit, R., & Treagust, D. F. (2003). Conceptual change: A powerful framework for improving science teaching and learning. *International Journal of Science Education*, *25*(6), 671–688. https://doi.org/10.1080/09500690305016

Espinoza, F. (2004). Enhancing mechanics learning through cognitively appropriate instruction. *Physics Education*, *39*(2), 181–187. https://doi.org/10.1088/0031-9120/39/2/007

Fletcher, R. B., & Hattie, J. A. (2004). An examination of the psychometric properties of the physical self-description questionnaire using a polytomous item response model. *Psychology of Sport and Exercise*, *5*(4), 423–446. https://doi.org/10.1016/S1469-0292(03)00036-0

Fongsamut, K., Tanasittikosol, M., & Phaksunchai, M. (2023). Effectiveness of the simulation-based learning (SBL) assisted with scaffolding approach to address students' misconceptions about projectile motion. *Physics Education*, *58*(2), 25002. https://doi.org/10.1088/1361-6552/aca57d

Fratiwi, N. J., 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*, *1848*, 50011.

Graham, T., Berry, J., & Rowlands, S. (2013). Are 'misconceptions' or alternative frameworks of force and motion spontaneous or formed prior to instruction? *International Journal of Mathematical Education in Science and Technology*, *44*(1), 84–103. https://doi.org/10.1080/0020739X.2012.703333

Guerra-Reyes, F., Guerra-Dávila, E., Naranjo-Toro, M., Basantes-Andrade, A., & Guevara-Betancourt, S. (2024). Misconceptions in the Learning of Natural Sciences: A Systematic Review. *Education Sciences*, *14*(5), 497. https://doi.org/10.3390/educsci14050497

Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the Certainty of Response Index (CRI). *Physics Education*, *34*(5), 294–299. https://doi.org/10.1088/0031-9120/34/5/304

Hynd, C. R., McWhorter, J. Y., Phares, V. L., & Suttles, C. W. (1994). The role of instructional variables in conceptual change in high school physics topics. *Journal of Research in Science Teaching*, *31*(9), 933–946. https://doi.org/10.1002/tea.3660310908

Isra, R. A., & Mufit, F. (2023). Students' conceptual understanding and causes of misconceptions on Newton's Law. *International Journal of Evaluation and Research in Education*, *12*(4), 1914–1924. https://doi.org/10.11591/ijere.v12i4.25568

Istiyono, E., Fenditasari, K., Ayub, M. R. S., Saepuzaman, D., & Dwandaru, W. S. B. (2024). *An eight-category partial credit model as very appropriate for four-tier diagnostic test scoring in physics learning*. 020013. https://doi.org/10.1063/5.0133862

Istiyono, E., Sunu Brams Dwandaru, W., 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

Jung, J. (2020). Diagnosing Causes of Pre-Service Literature Teachers' Misconceptions on the Narrator and Focalizer Using a Two-Tier Test. *Education Sciences*, *10*(4), 104. https://doi.org/10.3390/educsci10040104

Jurâne-Brçmane, A. (2023). Digital Assessment in Technology-Enriched Education: Thematic Review. *Education Sciences*, *13*(5), 522. https://doi.org/10.3390/educsci13050522

Kaddouri, M., Jmad, S., Azzimani, T., Mhamdi, K., & Abbadi, Z. (2025). Transforming Assessment Practices. In *Digital Tools and Platforms for Effective and Personalized Learning* (hal. 155–202). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-6013-3.ch006

Kaltakci-Gurel, D., Eryilmaz, A., & McDermott, L. C. (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

Kiray, S. A., & Simsek, S. (2021). Determination and Evaluation of the Science Teacher Candidates' Misconceptions About Density by Using Four-Tier Diagnostic Test. *International Journal of Science and Mathematics Education*, *19*(5), 935–955. https://doi.org/10.1007/s10763-020-10087-5

Koto, I., & Gusma, S. E. (2021). Using certainty response index to differentiate lack of knowledge and misconception about basic electrical concepts. *Journal of Physics: Conference Series*, *1731*(1), 012070. https://doi.org/10.1088/1742-6596/1731/1/012070

Lampeang, N. S., Mondolang, A. H., Tumangkeng, J. V, Makahinda, T., Umboh, I., & Poluakan, C. (2021). Use of the four-tier diagnostic test with PIMCA model on learning of microscope. In T. N., M. Y., A. A.G., W. I., A. C.U., & W. A.A.D. (Ed.), *Journal of Physics: Conference Series* (Vol. 1968, Nomor 1). IOP Publishing Ltd. https://doi.org/10.1088/1742-6596/1968/1/012039

Liu, E., & Li, M. (2016). Enhancing science teacher professional development: Lessons from a study of misconceptions of junior secondary biology teachers. In *Chinese Science Education in the 21st Century: Policy, Practice, and Research: 21 Nª~-NýVÑyf[Ye²€: ?eV{, ž[õ N—xvz* (pp. 401–412). Dordrecht: Springer Netherlands.

Ma, W., & de la Torre, J. (2020). GDINA/ : An R Package for Cognitive Diagnosis Modeling. *Journal of Statistical Software*, *93*(14). https://doi.org/10.18637/jss.v093.i14

Mahardika, A. I., Hakam, A. B., Arifuddin, M., & Halim, A. D. (2025). The Impact of Physics Modeling Learning E-Module to Overcome Misconception on Static Fluids. *TEM Journal*, *14*(2), 1808–1819. https:/ /doi.org/10.18421/TEM142-78

Maison, Asma, R., Doyan, A., & Saputri, L. (2022). How do additional instructions change the answer? Study of pre-service physics teachers' misconception about buoyancy. *Journal of Physics: Conference Series*, *2165*(1), 012048. https://doi.org/10.1088/1742-6596/2165/1/012048

Marjana, R., Achmad Samsudin, Duden

Saepuzaman, & Dadi Rusdiana. (2026). Misconception in Physics Learning: A Systematic Review of the Research Trend, Misconception Profile, Diagnosis, and Remediation Strategies. *Jurnal Pendidikan Fisika*, *14*(1), 240–262. https://doi.org/10.26618/7wdrxx78

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Mi, S., Ye, J., Yan, L., & Bi, H. (2023). Development and validation of a conceptual survey instrument to evaluate senior high school students' understanding of electrostatics. *Physical Review Physics Education Research*, *19*(1), 10114. https://doi.org/10.1103/PhysRevPhysEducRes.19.010114

Munggarani, M. E., Supriyati, Y., & Astra, I. M. (2021). Identifying high school students' misconceptions using digital four-tier diagnostic tests in distance learning. *Journal of Physics: Conference Series*, *2019*(1). https://doi.org/10.1088/1742-6596/2019/1/012016

Mustofa, H. A., Zain, Z. A., Tsania, H., Azman, M. N. A., Marmoah, S., & Masfuah, S. (2024). Analysis of students' conceptions based on cognitive style on newton's law understanding. *Jurnal Pendidikan IPA Indonesia*, *13*(2), 301–312. https://doi.org/10.15294/mx0v4952

Neumann, S., & Hopf, M. (2017). Discovering Children's Science Associations Utilizing Drawings. In *Drawing for Science Education* (hal. 111–121). Sense Publishers. https://doi.org/10.1007/978-94-6300-875-4_10

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

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond Cold Conceptual Change: The Role of Motivational Beliefs and Classroom Contextual Factors in the Process of Conceptual Change. *Review of Educational Research*, *63*(2), 167–199.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. *A product from the ESRC methods programme Version*, *1*(1), b92.

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

Resbiantoro, G., Setiani, R., & Dwikoranto. (2022). A Review of Misconception in Physics: The Diagnosis, Causes, and Remediation. *Journal of Turkish Science Education*, *19*(2), 403–427. https://doi.org/10.36681/tused.2022.128

Roziqin, R., Samsudin, A., Saepuzaman, D., Kaniawati, I., Iryanti, M., & Abdul Rahman, N. F. (2026). Mapping Misconceptions in Heat and Temperature: Rasch Analysis of Four-Tier Diagnostic Responses. *Jurnal Pendidikan MIPA*, *27*(2), 867–886. https://doi.org/10.23960/jpmipa.v27i2.pp867-886

Saidi, S. S., & Siew, N. M. (2019). Reliability and Validity Analysis of Statistical Reasoning Test Survey Instrument using the Rasch Measurement Model. *International Electronic Journal of Mathematics Education*, *14*(3). https://doi.org/10.29333/iejme/5755

Samsudin, A., Afif, N. F., Nugraha, M. G., Suhandi, A., Fratiwi, N. J., Aminudin, A. H., Adimayuda, R., Linuwih, S., & Costu, B. (2021). Reconstructing Students' Misconceptions on Work and Energy through the PDEODE*E Tasks with Think-Pair-Share. *Journal of Turkish Science Education*, *18*(1), 118–144. https://doi.org/10.36681/tused.2021.56

Samsudin, A., Zulfikar, A., Saepuzaman, D., Suhandi, A., Aminudin, A. H., Supriyadi, S., & Coºtu, B. (2024a). Correcting grade 11 students' misconceptions of the concept of force through the conceptual change model (CCM) with PDEODE*E tasks. *Journal of Turkish Science Education*, *21*(2), 212–231. https://doi.org/10.36681/tused.2024.012

Sarwono, S., Suhandi, A., & Samsudin, A. (2022). Remediate senior high school students' misconception regarding the runs out battery concept using AS-CBRText. In S. A., H. L., Y. G., I. M., K. Y.F., S. A.S., & R. L. (Ed.), *AIP Conference Proceedings* (Vol. 2468). American Institute of Physics Inc. https://doi.org/10.1063/5.0131699

Schumacker, R. E., & Smith, E. V. (2007). A Rasch Perspective. *Educational and Psychological Measurement*, *67*(3), 394–409. https://doi.org/10.1177/0013164406294776

Slotta, J. D., Chi, M. T. H., & Joram, E. (1995). Assessing Students' Misclassifications of Physics Concepts: An Ontological Basis for Conceptual Change. *Cognition and Instruction*, *13*(3), 373–400. https://doi.org/10.1207/s1532690xci1303_2

Smith, J., diSessa, A., & Roschelle, J. (1994). Misconceptions Reconceived: A Constructivist Analysis of Knowledge in Transition. *The Journal of the Learning Sciences*, *3*, 115–163. https://doi.org/10.1207/s15327809jls0302_1

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology*, *100*(4), 961–976. https://doi.org/10.1037/a0012546

Surmaini, Syafe'I, I., & Diani, R. (2021). An analysis of students' physics misconceptions in online learning using the four-tier diagnostic test with certainty of response index (CRI). *IOP Conference Series: Earth and Environmental Science*, *1796*(1). https://doi.org/10.1088/1742-6596/1796/1/012099

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. In *Psychological Methods* (Vol. 11, Nomor 3, hal. 287–305). American Psychological Association. https://doi.org/10.1037/1082-989X.11.3.287

Treagust, D. F. (1986). Evaluating students' misconceptions by means of diagnostic multiple-choice items. *Research in Science Education*, *16*, 199–207.

Tumanggor, A., Supahar, Kuswanto, H., & Ringo, E. (2020). Using four-tier diagnostic test instruments to detect physics teacher candidates' misconceptions: Case of mechanical wave concepts. *Journal of Physics: Conference Series*, *1440*. https://doi.org/10.1088/1742-6596/1440/1/012059

Umar, F. A., Samsudin, A., Kapýcý, H. Ö., Ramalis, T. R., Aminudin, A. H., Mufidah, S. N., Kunaedi, J., Astuti, I. R. W., & Dewi, F. H. (2024). Developing Dispersion and Polarization Conceptual Inventory (DiPolCI) to Identify Students' Mental Model. *Journal of Science Learning*, *7*(3), 239–247. https://doi.org/10.17509/jsl.v7i3.69058

Vispoel, W. P., & Kim, H. Y. (2014). Psychometric properties for the Balanced Inventory of Desirable Responding: Dichotomous versus polytomous conventional and IRT scoring. *Psychological Assessment*, *26*(3), 878–891. https://doi.org/10.1037/a0036430

Vosniadou, S., & Mason, L. (2012). Conceptual change induced by instruction: A complex interplay of multiple factors. In *APA educational psychology handbook, Vol 2: Individual differences and cultural and contextual factors.* (hal. 221–246). American Psychological Association. https://doi.org/10.1037/13274-009

Zach, H. (2005). Psychometrics properties of the questionnaire surveys: Validity, reliability and feasibility. *Sociologia*, *37*(3), 275–291. https://www.scopus.com/inward/record.uri?eid=2-s2.0-21344472799&partnerID=40&md5=ac8a6a66a38bbf22820a3ecbd453c20d

Zain, A. Z., Samsudin, A., & Hasanah, L. (2025). Diagnostic Instrument Test of Renewable Energy Topics: Analysis of Conceptual Change Students Using the Rasch Approach. *Online Learning In Educational Research (OLER)*, *5*(2), 231–243. https://doi.org/10.58524/oler.v5i2.675

Zhai, X., Haudek, K. C., & Ma, W. (2023). Assessing Argumentation Using Machine Learning and Cognitive Diagnostic Modeling. *Research in Science Education*, *53*(2), 405–424. https://doi.org/10.1007/s11165-022-10062-w