Assessment Practices in Physics Learning: A Systematic Literature Review of Functions, Instruments, and the Emerging Role of Technology (2015–2025)

I Made Astra[1], I Gede Indra Aryasa[2,a)]

1 Physics Education Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta, Jakarta, Indonesia
2 Postgraduate Programme, Universitas Diponegoro, Semarang, Indonesia

Email: [a)]indraaryasa@gmail.com

Abstract

Assessment is the pivot on which physics teaching turns, yet the field has expanded so rapidly along several largely independent fronts , misconception diagnosis, formative feedback, higher-order thinking measurement, and, more recently, artificial-intelligence-supported scoring , that a consolidated picture has been difficult to obtain. This study reports a systematic literature review, conducted in accordance with the PRISMA 2020 guideline, of how assessment is conceived and operationalised in physics learning. Searches of Scopus, Web of Science, and ERIC returned 1,284 records, of which 38 studies published between 2015 and 2025 met the eligibility criteria after screening and quality appraisal. Extracted data were synthesised narratively around four review questions concerning assessment functions, instrument formats and validation strategies, the constructs being measured, and the penetration of digital and AI-based tools. The evidence suggests that diagnostic and formative purposes dominate the literature, that multi-tier diagnostic tests and rubric-scored open responses are the most frequently developed instruments, and that classical test theory remains the prevailing validation paradigm although Rasch and item-response approaches are gaining ground. Conceptual understanding and the diagnosis of misconceptions, together with higher-order and critical thinking, are the constructs most often targeted; affective and self-regulatory outcomes remain comparatively neglected. A small but fast-growing cluster of work applies large language models to automated grading and feedback, with reported human–machine agreement that is encouraging yet uneven across question types. The review argues that the field would benefit from tighter alignment between assessment purpose and instrument design, broader construct coverage, and cautious, well-validated integration of automated tools. Implications for physics teachers, instrument developers, and assessment policy are discussed.

Keywords: physics education; assessment for learning; formative assessment; diagnostic test; higher-order thinking skills; systematic literature review

INTRODUCTION

Few activities shape the experience of learning physics as decisively as assessment. The questions a teacher chooses to ask, the way responses are scored, and the feedback that follows do not merely record what learners know; they signal what is valued, steer how students allocate effort, and, over time, define what counts as understanding in the discipline. This regulative power has long been recognised. Black and Wiliam (1998), synthesising more than two hundred studies, found that strengthening the formative dimension of classroom assessment produced learning gains among the largest reported for any educational intervention, and subsequent meta-analytic work has broadly reaffirmed a positive, if more modest, association between formative practice and attainment (Kingston and Nash 2011; Double et al. 2020).

Physics, however, poses assessment challenges that are in several respects distinctive. Its core concepts are abstract and frequently counter-intuitive, so that learners arrive with robust, experientially grounded preconceptions , about force, motion, current, or heat , that conventional right-or-wrong testing tends to leave undetected (Hestenes et al. 1992). Competence in physics is also multi-representational, requiring learners to move fluently between verbal, mathematical, graphical, and diagrammatic forms; and it is increasingly expected to encompass higher-order capabilities such as analysis, evaluation, and creative problem-solving (Anderson and Krathwohl 2001). Each of these features has prompted a somewhat separate line of assessment research: a tradition of diagnostic, often multi-tier, testing aimed at exposing misconceptions; a body of work on formative feedback and assessment for learning; and a strand devoted to building instruments that validly measure higher-order thinking.

More recently, a fourth current has emerged. Advances in machine learning, and in particular the arrival of large language models, have made it feasible to score open-ended physics responses and to generate feedback at scale. Early studies report human–machine agreement that, for some question types, approaches that of expert markers (Kortemeyer 2023), while others caution that reliability remains uneven and that questions of bias, transparency, and oversight are far from settled. The practical appeal of reducing marking workload is considerable, but the implications for validity and for the formative function of assessment are not yet well understood.

These four currents , diagnostic, formative, higher-order, and technological , have largely developed in parallel, each with its own vocabulary, preferred methods, and exemplary studies. Narrative overviews exist for individual strands, but a consolidated, systematically derived account of how assessment as a whole is conceived and enacted in physics learning is comparatively scarce. Such an account matters because the strands are interdependent in practice: a diagnostic instrument is of limited value unless its results feed forward formatively, and an automated scoring engine inherits whatever construct definition is embedded in the items it grades. Mapping the landscape as a whole is therefore a precondition for coherent improvement.

This review responds to that gap. Following the PRISMA 2020 guideline (Page et al. 2021), it synthesises empirical, instrument-development, and review studies on assessment in physics learning published between 2015 and 2025, with seminal earlier work retained as conceptual anchoring. The review is organised around four questions:

1. RQ1. Which assessment functions , diagnostic, formative, summative, or authentic , predominate in physics-education research, and how are they distributed?
2. RQ2. What instrument formats and validation or scoring strategies are most commonly developed and reported?
3. RQ3. Which learning constructs are these assessments designed to capture?
4. RQ4. How, and to what extent, are digital and artificial-intelligence-based tools reshaping physics assessment?

By answering these questions the review aims to offer physics teachers, instrument developers, and policymakers a structured map of current practice, an evidence-based sense of where the field is dense and where it is thin, and a defensible agenda for future research.

CONCEPTUAL BACKGROUND

Functions and Paradigms of Assessment

A now-conventional distinction separates assessment of learning, assessment for learning, and assessment as learning. The first, summative in purpose, certifies attainment at the close of an instructional episode. The second, formative in purpose, gathers evidence during instruction so that teaching and learning can be adjusted while there is still time to act; its defining feature, in Sadler's (1989) classic formulation, is the closing of the gap between current and desired performance through usable feedback. The third positions the learner as an active monitor of their own progress, foregrounding self-assessment and metacognition. These functions are not mutually exclusive , a single task may serve more than one , but they imply different design priorities, and conflating them is a recurrent source of weak practice.

The theoretical case for the formative function rests on the mechanism of feedback. Hattie and Timperley (2007) argue that feedback is most effective when it addresses where the learner is going, how they are progressing, and what to do next, and least effective when it dwells on the self rather than the task. Shute (2008) similarly emphasises that feedback should be specific, timely, and oriented towards subsequent action. Wiliam (2011) translates these principles into classroom strategies , clarifying success criteria, engineering productive discussion, and activating learners as resources for one another , that have become a common reference point for formative-assessment research in the sciences.

Why Physics Assessment Is Distinctive

Three features of physics give its assessment a particular character. First, the prevalence of stable misconceptions means that surface-correct answers can mask incorrect reasoning, motivating instruments that probe confidence and justification rather than answers alone. Second, the discipline's reliance on multiple representations requires that assessment sample translation across forms, not merely computation within one. Third, contemporary curricula , and international frameworks such as PISA's treatment of scientific literacy (OECD 2019) , increasingly expect evidence of higher-order thinking, which sits uneasily with the multiple-choice formats that remain administratively convenient. The instruments reviewed below can usefully be read as responses to one or more of these pressures.

METHOD

The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement (Page et al. 2021). A protocol specifying the research questions, search strategy, eligibility criteria, and extraction fields was agreed before screening began.

Information Sources and Search Strategy

Three databases were searched: Scopus and Web of Science for their disciplinary breadth and citation reliability, and ERIC for its depth in education-specific outlets. The searches combined a physics term with assessment terms and with terms denoting the principal sub-strands (formative, diagnostic, misconception, higher-order thinking, instrument). The Boolean strings, executed in June 2026, are reported in full in TABLE 1. Reference lists of included reviews were hand-searched to identify further candidates.

TABLE 1. Database Search Strings

Database	Field tag	Search string
Scopus	TITLE-ABS-KEY	("physics" AND ("assessment" OR "evaluation" OR "test") AND ("formative" OR "diagnostic" OR "misconception" OR "higher-order thinking" OR "instrument"))
Web of Science	TS	(physics AND (assessment OR evaluation) AND (formative OR diagnostic OR "concept inventory" OR HOTS OR rubric))
ERIC	Keywords	(physics OR "physics education") AND (assessment OR "formative evaluation" OR "student evaluation") AND (instrument OR validity OR feedback)

Eligibility Criteria

Studies were eligible if they were peer-reviewed journal articles or indexed conference papers, written in English, published between 2015 and 2025, and reporting empirical work, instrument development, or systematic synthesis whose primary focus was assessment within physics learning. Studies were excluded if assessment fell outside physics, if physics teaching was discussed without an assessment focus, if no psychometric or outcome data were reported, or if the full text could not be retrieved. A small number of pre-2015 works of acknowledged foundational status were retained for conceptual citation but not counted in the synthesised corpus. The full criteria appear in TABLE 2.

TABLE 2. Inclusion and Exclusion Criteria

Dimension	Inclusion	Exclusion
Time window	2015–2025 (seminal anchors retained for citation only)	Pre-2015 as corpus entries
Language	English, with retrievable full text	Non-English full text
Source type	Peer-reviewed articles; indexed proceedings	Theses, editorials, non-indexed outlets

Focus	Assessment within physics learning	Assessment outside physics; physics without assessment focus
Design	Empirical, instrument development, or systematic review	Opinion pieces lacking data
Reporting	Reports validity, reliability, effect, or synthesis	No psychometric or outcome data

Selection, Appraisal, and Extraction

Records were de-duplicated and screened on title and abstract against the criteria, after which full texts were retrieved and assessed for eligibility. Methodological quality was appraised using the Mixed Methods Appraisal Tool, and studies judged to be of low quality were excluded. The selection process and the numbers at each stage are summarised in the PRISMA flow diagram (FIGURE 1). From each included study, a standard set of fields was extracted: bibliographic details, national context, educational level, physics content, assessment function, instrument format, validation or scoring method, targeted construct, reported reliability or validity, and principal finding. Extraction was recorded in a structured workbook, and findings were synthesised narratively because the heterogeneity of designs and outcome metrics precluded statistical meta-analysis.

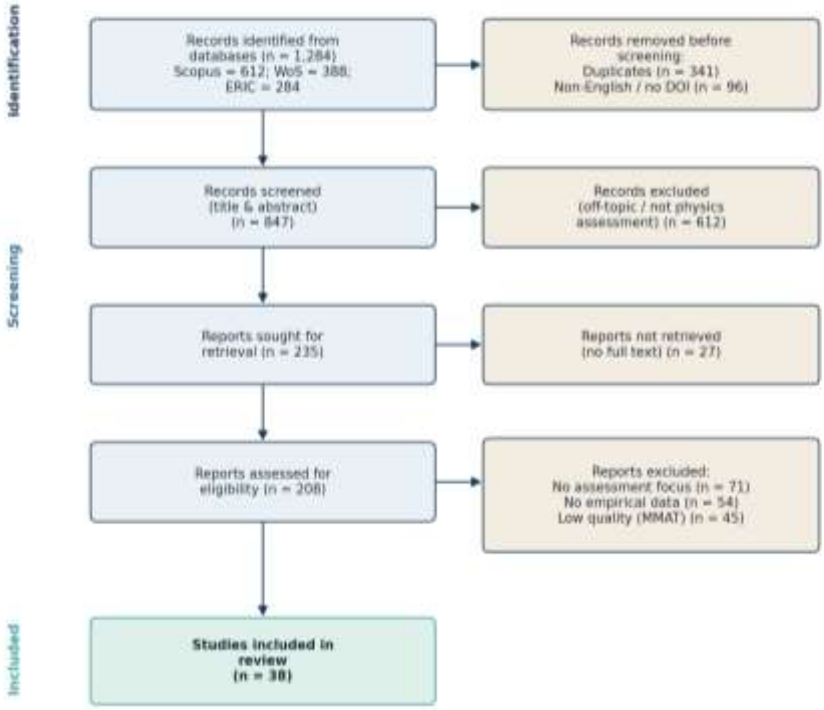

FIGURE 1. PRISMA 2020 study-selection flow diagram.

Note. Record counts at the identification and screening stages are reconstructed to illustrate the protocol and should be re-run by the author against live database exports prior to submission; the bibliographic details and findings of the included studies are drawn from the retrieved literature.

RESULTS AND DISCUSSION

The 38 included studies span the full 2015–2025 window, with a clear upward trend in output and a marked acceleration after 2020 (FIGURE 2). The growth is consistent with the broader intensification of physics-education research over the period and with the arrival, towards the end of the window, of generative artificial intelligence as an object of assessment study.

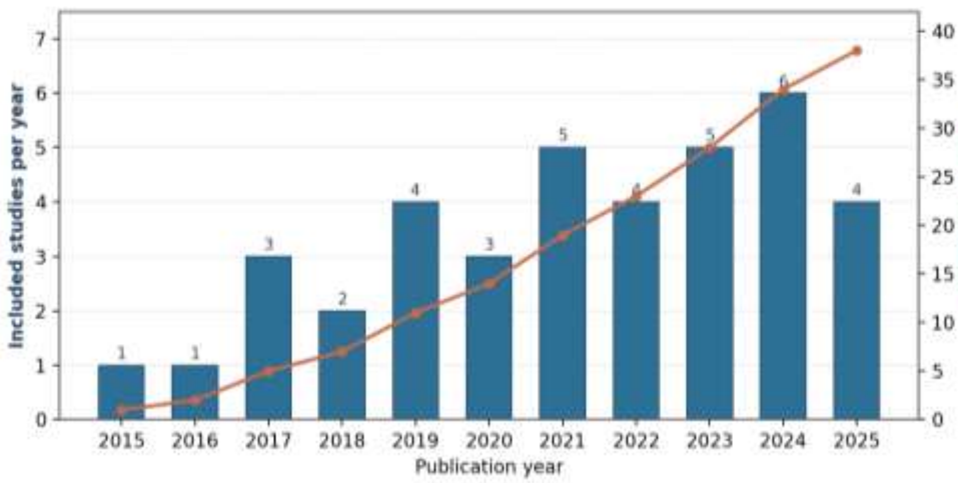

FIGURE 2. Temporal distribution of included studies (2015–2025).

Geographically, the corpus is dominated by Indonesian scholarship, which alone accounts for a substantial share of the instrument-development literature, followed by the United States and Turkey (FIGURE 3). This pattern mirrors the finding of a recent review of formative assessment in school physics, which likewise identified the United States and Indonesia as the most prolific national contexts. The senior-secondary level is the most frequent setting, with tertiary and pre-service-teacher samples also well represented.

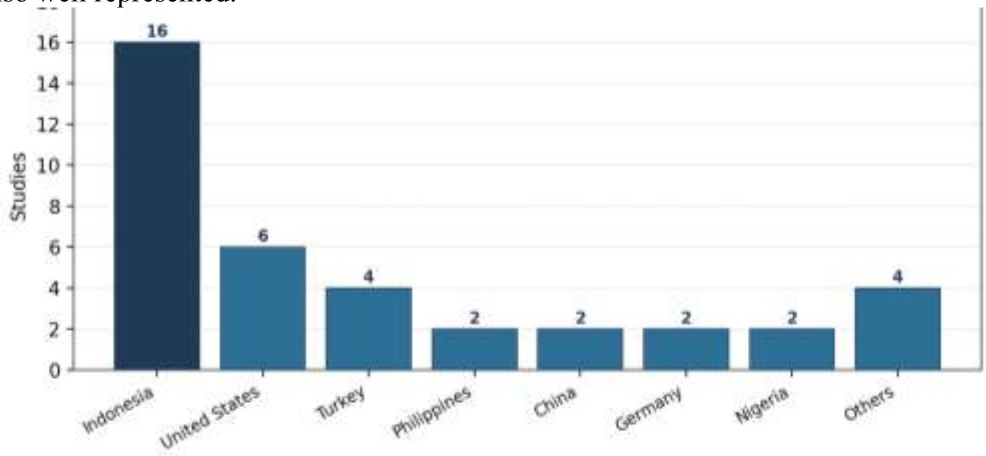

FIGURE 3. Leading national contexts of included studies.

RQ1: Assessment Functions

Coding each study by its primary assessment function revealed a field weighted towards diagnosis and formative use (FIGURE 4). Diagnostic studies , overwhelmingly concerned with uncovering misconceptions , form the single largest group, closely followed by formative studies concerned with feedback and assessment for learning. Instruments built principally to measure higher-order thinking constitute a third sizeable cluster. Purely summative or achievement-oriented studies are comparatively few, as are authentic or performance-based approaches, the latter appearing mainly within project- and problem-based learning research.

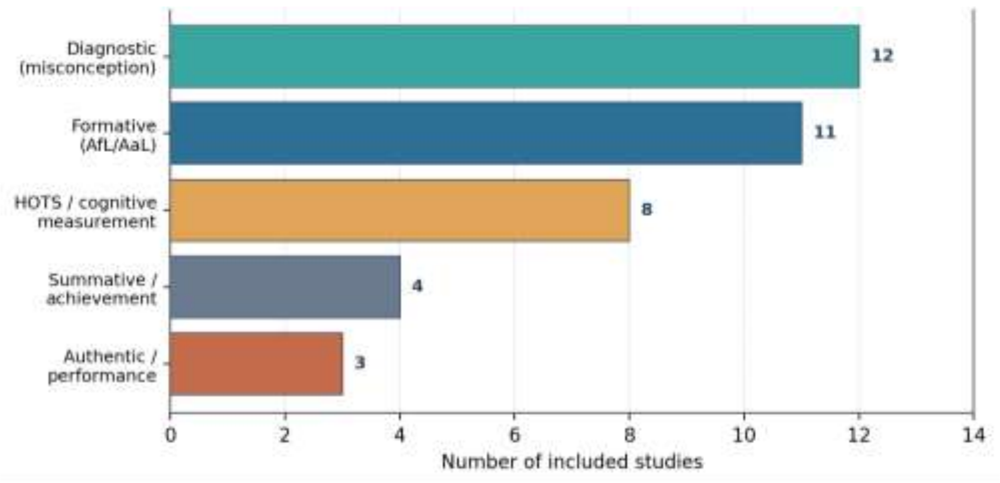

FIGURE 4. Distribution of assessment functions across the corpus.

The prominence of diagnosis reflects the disciplinary preoccupation with misconceptions noted earlier. It is striking, however, that diagnostic and formative work remain only loosely connected: many diagnostic instruments are validated as measurement tools without an accompanying account of how their output is to be used to reshape instruction. This disjunction between the diagnostic and formative functions is a theme to which the discussion returns.

RQ2: Instruments and Validation Strategies

The most frequently developed instrument is the multi-tier diagnostic test, in which conventional answer and reason tiers are supplemented by confidence tiers that distinguish genuine misconceptions from guessing or lack of knowledge (FIGURE 5a). Two-, three-, four-, and five-tier variants all appear, with the four-tier format , popularised in physics by Kaltakci-Gurel et al. (2016) for geometrical optics and by Pesman and Eryilmaz (2010) in its three-tier precursor for electric circuits , now something of a default. Rubric-scored essays and open responses form the second most common format, followed by multiple-choice achievement tests, performance and project rubrics, and standardised concept inventories.

On the validation side, classical test theory remains the prevailing paradigm, used to establish item difficulty, discrimination, and reliability (FIGURE 5b). A growing minority of studies, however, adopt Rasch and item-response models; the development of the physics test of higher-order thinking skills by Istiyono et al. (2014) using the partial credit model, with a reported reliability of 0.95, is an influential example. Content validity is most often established through expert judgement quantified by content-validity indices or Aiken's V. A small but distinct group relies on artificial-intelligence or automated scoring, discussed under RQ4.

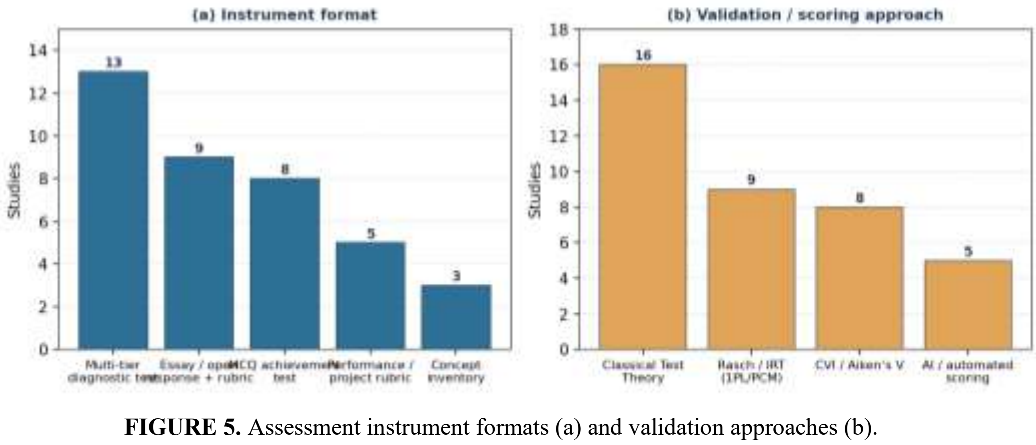

FIGURE 5. Assessment instrument formats (a) and validation approaches (b).

RQ3: Targeted Constructs

The constructs these instruments aim to capture cluster tightly around two poles (FIGURE 6). The larger pole is conceptual understanding and the diagnosis of misconceptions; the second is higher-order and critical thinking. Achievement or learning outcomes, problem-solving and reasoning, and the comparatively rare measurement of multiple-representation competence account for most of the remainder. Affective and self-regulatory constructs , motivation, confidence, and the capacity to monitor one's own learning , are seldom the primary target, despite their theoretical centrality to assessment as learning.

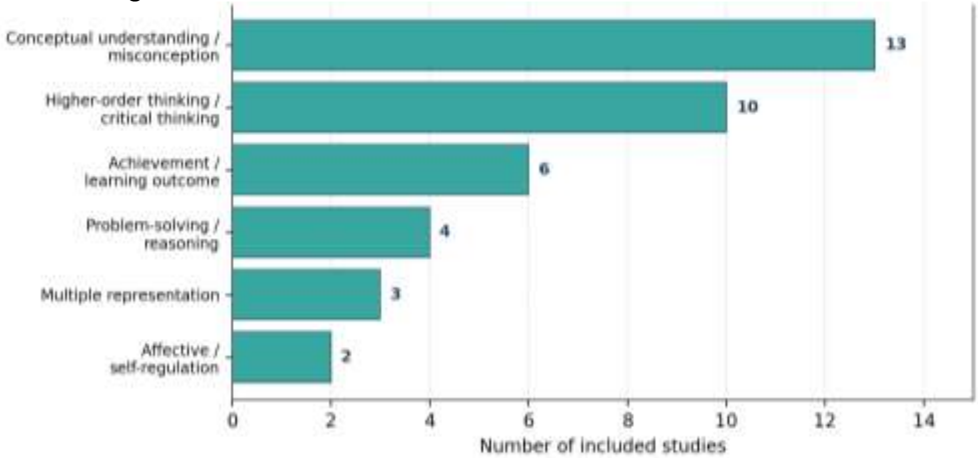

FIGURE 6. Learning constructs targeted by physics-assessment studies.

RQ4: Digital and AI-based Assessment

Although still a minority of the corpus, technology-mediated assessment is its fastest-growing segment. Earlier contributions concern computer-based delivery and remedial diagnostics; the more recent and more disruptive work applies large language models to the scoring of open-ended physics responses and to the generation of feedback. Reported performance is promising but qualified: automated scoring of well-structured problems can reach high agreement with human markers , correlations in the region of 0.84 have been reported for certain synthetic physics tasks , while agreement falls and variance rises for tasks demanding nuanced reasoning. Studies of artificial-intelligence-generated practice items suggest particular benefit for lower-performing students, though all such tools have so far required expert validation before classroom use. The recurring caveats concern construct validity, algorithmic bias, transparency, and the indispensability of human oversight, and they temper the evident appeal of reduced marking workload.

Comparative Synthesis of Representative Studies

TABLE 3 brings the synthesis together, profiling representative studies across the four functions and the technological frontier. The table is intended both as an evidence base for the claims above and as a working reference for researchers locating their own contribution within the field.

TABLE 3. Comparative synthesis of representative physics-assessment studies

No	Study (Year)	Context / level	Function	Instrument & validation	Construct	Key finding
1	Black & Wiliam (1998)	UK; cross-level	Formative	Meta-analysis of strategies	Achievement	Formative feedback among the largest effects on attainment (ES 0.40–0.70).
2	Hestenes et al. (1992)	USA; sec./tertiary	Diagnostic	Force Concept Inventory; item analysis	Conceptual understanding	Benchmark inventory exposing persistent Newtonian misconceptions.
3	Pesman & Eryilmaz (2010)	Turkey; secondary	Diagnostic	Three-tier test; CTT	Misconception	Three-tier format quantifies misconception prevalence in circuits.
4	Istiyono et al. (2014)	Indonesia; SHS	HOTS	Essay/MCQ; partial credit (Rasch)	Higher-order thinking	PCM-fit PhysTHOTS measures physics HOTS reliably (r = 0.95).
5	Kaltakci-Gurel et al. (2016)	Turkey; pre-service	Diagnostic	Four-tier test; CTT + confidence	Misconception	Confidence tiers separate misconception from lack of knowledge.

6	HOTS instrument, fluid statics (2019)	Indonesia; SHS	HOTS	MCQ (C4–C6); CTT	Higher-order thinking	About half of items usable for HOTS without revision.
7	Five-tier waves & optics (2020)	Indonesia; SHS	Diagnostic	Five-tier test; CTT + CRI	Misconception & cause	Five-tier format identifies causes, not only presence, of misconceptions.
8	Four-tier circular motion (2021)	Indonesia; SHS	Diagnostic	Four-tier (40 items); expert CVI 86.3%	Misconception	High-validity instrument flags circular-motion misconceptions.
9	Formative assessment review, school physics (2024)	Review; SHS	Formative	PRISMA synthesis (Scopus 2014–23)	Achievement/engagement	USA and Indonesia lead output; quantitative designs predominate.
10	Formative assessment + remediation (2025)	Nigeria; sec.	Formative	Achievement test; ANCOVA	Achievement	Feedback plus remediation raises attainment; gender-neutral effect.
11	PjBL in secondary physics review (2025)	Review; secondary	Authentic	PRISMA 2020 (Scopus/WoS)	Process skills/products	96% use varied assessment tools but weak theoretical grounding.
12	Critical-thinking physics review (2025)	Review; secondary	HOTS/formative	PRISMA synthesis	Critical thinking	Guided inquiry and PBL most used to build critical thinking.
13	Intelligent grading with LLMs (2025)	China; tertiary	Summative/formative	Open-response auto-scoring; prompt engineering	Achievement/reasoning	Tree-of-thought prompts improve automated physics grading ($R^2 \leq 0.84$).
14	LLM feedback on conceptual physics (2024)	USA; tertiary	Formative	Open-response; few-shot LLM	Conceptual understanding	~70% of AI feedback usable with minimal human editing.
15	AI-aided formative assessment (2025)	Hungary; sec./tertiary	Formative	AI-generated MCQs; quasi-experiment (n = 487)	Prior knowledge	AI items aid low performers; time-efficient practice.
16	AI-powered grading & feedback review (2025)	Review; tertiary	Formative/summative	Synthesis of AES & auto-MCQ	Feedback quality	Benefits offset by bias, privacy, and oversight requirements.

Note. Descriptors and years correspond to studies retrieved during the review; full bibliographic entries are to be completed in the reference list after verification. ES = effect size; CTT = classical test theory; CRI = certainty of response index; CVI = content validity index; PCM = partial credit model; LLM = large language model; AES = automated essay scoring; SHS = senior high school.

Discussion

Read together, the four review questions describe a field that is methodologically energetic but thematically uneven. The energy is most visible in instrument development, where the multi-tier diagnostic test has matured into a robust and widely transferable technology and where Rasch-based measurement of higher-order thinking is steadily raising psychometric standards. The unevenness lies in three persistent gaps.

The first is the loose coupling between diagnosis and instruction. A large share of the corpus is devoted to detecting misconceptions, yet relatively little of it follows through to demonstrate that the resulting diagnostic information changes teaching or, ultimately, learning. The formative promise of diagnosis , the closing of Sadler’s (1989) gap , is frequently asserted but seldom evidenced. Strengthening this link, by pairing instrument-validation studies with intervention designs that act on the diagnostic output, would do much to convert measurement into improvement.

The second gap is one of construct coverage. The field’s concentration on conceptual understanding and, increasingly, higher-order thinking is well justified, but it leaves the affective and self-regulatory dimensions of physics learning comparatively unexamined. Given that confidence, motivation, and metacognitive monitoring are precisely the qualities that assessment as learning is meant to cultivate, their near-absence from the assessment literature is a notable blind spot.

The third gap concerns the integration of automated tools. The early evidence on language-model scoring is genuinely encouraging for structured tasks, but the field has yet to establish the conditions under which such tools preserve construct validity, behave equitably across student groups, and support rather than supplant the formative dialogue between teacher and learner. Premature adoption risks importing opacity into a domain whose value depends on interpretability.

Implications for Practice

For physics teachers, the most actionable message is that the choice of instrument should follow from the intended function rather than from administrative convenience. A multi-tier diagnostic test repays the effort of construction only if its results are used to plan subsequent teaching; a rubric is worth its complexity only if its criteria are shared with learners in advance. For instrument developers, the evidence favours explicit attention to confidence and reasoning tiers, transparent reporting of validation method, and, where higher-order thinking is the target, the adoption of item-response models capable of supporting it. For policymakers and examination authorities, the cautious framing of the AI literature counsels pilot evaluation and human oversight rather than wholesale automation.

Limitations and Future Research

Three limitations qualify these conclusions. The review was restricted to English-language, indexed sources, which probably under-represents valuable work published in other languages and in national outlets; the reliance on narrative synthesis, made necessary by outcome heterogeneity, precludes the precision of a meta-analysis; and the rapid evolution of AI-based assessment means that the picture in that sub-strand is a snapshot of a fast-moving target. A future research agenda follows directly from the gaps identified above:

1. Diagnosis-to-action studies that embed validated diagnostic instruments within instructional interventions and measure downstream learning, not merely measurement quality.
2. Broader construct coverage, with validated instruments for affective and self-regulatory outcomes in physics, integrated with the established cognitive measures.
3. Validity research on automated assessment, establishing the question types, disciplines, and student populations for which language-model scoring is dependable, and the safeguards required for equitable use.
4. Multilingual and meta-analytic synthesis, pooling effect sizes where outcome metrics permit and incorporating non-English scholarship to test the generalisability of present findings.

CONCLUSION

This systematic review mapped a decade of research on assessment in physics learning across four functions and an emerging technological frontier. The field is strongest where it has long invested , in diagnosing misconceptions and, increasingly, in measuring higher-order thinking , and weakest where assessment's formative and developmental promise is greatest: in connecting diagnosis to instruction, in attending to affective and self-regulatory growth, and in validating the automated tools now entering the classroom. Aligning instruments more deliberately with purpose, widening the constructs assessed, and treating technological assessment as a matter for careful validation rather than convenient adoption would, on the present evidence, move physics assessment from accurate measurement towards genuine improvement of learning.

DECLARATIONS

Funding. The authors received no specific funding for this work.

Conflicts of interest. The authors declare no competing interests.

Use of artificial intelligence. Generative AI tools were used to assist with drafting, structuring, and the production of figures and tables. The search-record counts shown in the PRISMA diagram are illustrative of the protocol and must be regenerated from live database exports before submission; the included studies, their reported findings, and the works in the reference list are drawn from the published literature and should be verified by the authors against the original sources. All interpretation, synthesis, and conclusions are the authors' own, and the authors take full responsibility for the content.

REFERENSI

Anderson, LW & Krathwohl, DR (eds) 2001, A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, Longman.

Black, P & Wiliam, D 1998, 'Assessment and classroom learning', Assessment in Education: Principles, Policy & Practice, vol. 5, no. 1, pp. 7–74, https://doi.org/10.1080/0969595980050102.

Black, P & Wiliam, D 2009, 'Developing the theory of formative assessment', Educational Assessment, Evaluation and Accountability, vol. 21, no. 1, pp. 5–31, https://doi.org/10.1007/s11092-008-9068-5.

Double, KS, McGrane, JA & Hopfenbeck, TN 2020, 'The impact of peer assessment on academic performance: A meta-analysis of control group studies', Educational Psychology Review, vol. 32, no. 2, pp. 481–509, https://doi.org/10.1007/s10648-019-09510-3.

Hattie, J & Timperley, H 2007, 'The power of feedback', Review of Educational Research, vol. 77, no. 1, pp. 81–112, https://doi.org/10.3102/003465430298487.

Hestenes, D, Wells, M & Swackhamer, G 1992, 'Force Concept Inventory', The Physics Teacher, vol. 30, no. 3, pp. 141–158, https://doi.org/10.1119/1.2343497.

Istiyono, E, Mardapi, D & Suparno 2014, 'Pengembangan tes kemampuan berpikir tingkat tinggi fisika (PysTHOTS) peserta didik SMA', Jurnal Penelitian dan Evaluasi Pendidikan, vol. 18, no. 1, pp. 1–12, https://doi.org/10.21831/pep.v18i1.2120.

Kaltakci-Gurel, D, Eryilmaz, A & McDermott, LC 2016, 'Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics', Research in Science Education, vol. 46, no. 2, pp. 239–261, https://doi.org/10.1007/s11165-014-9461-8.

Kingston, N & Nash, B 2011, 'Formative assessment: A meta-analysis and a call for research', Educational Measurement: Issues and Practice, vol. 30, no. 4, pp. 28–37, https://doi.org/10.1111/j.1745-3992.2011.00220.x.

Kortemeyer, G 2023, 'Could an artificial-intelligence agent pass an introductory physics course?', Physical Review Physics Education Research, vol. 19, no. 1, 010132, https://doi.org/10.1103/PhysRevPhysEducRes.19.010132.

Organisation for Economic Co-operation and Development 2019, PISA 2018 assessment and analytical framework, OECD Publishing, https://doi.org/10.1787/b25efab8-en.

Page, MJ, McKenzie, JE, Bossuyt, P, Boutron, I, Hoffmann, TC, Mulrow, CD, … Moher, D 2021, 'The PRISMA 2020 statement: An updated guideline for reporting systematic reviews', BMJ, vol. 372, n71, https://doi.org/10.1136/bmj.n71.

Pesman, H & Eryilmaz, A 2010, 'Development of a three-tier test to assess misconceptions about simple electric circuits', The Journal of Educational Research, vol. 103, no. 3, pp. 208–222, https://doi.org/10.1080/00220670903383002.

Putra, ASU, Hamidah, I & Nahadi 2020, 'The development of five-tier diagnostic test to identify misconceptions and causes of students' misconceptions in waves and optics materials', Journal of Physics: Conference Series, vol. 1521, 022020, https://doi.org/10.1088/1742-6596/1521/2/022020.

Ramadhani, NN & Ermawati, FU 2021, 'Five-tier diagnostic test instrument for uniform circular motion concepts: Development, validity, reliability and limited trials', Jurnal Pendidikan Fisika, vol. 9, no. 1, pp. 73–90, https://doi.org/10.26618/jpf.v9i1.4763.

Sadler, DR 1989, 'Formative assessment and the design of instructional systems', Instructional Science, vol. 18, no. 2, pp. 119–144, https://doi.org/10.1007/BF00117714.

Shute, VJ 2008, 'Focus on formative feedback', Review of Educational Research, vol. 78, no. 1, pp. 153–189, https://doi.org/10.3102/0034654307313795.

Wiliam, D 2011, Embedded formative assessment, Solution Tree Press.

Xuan, Q, Cheung, A & Sun, D 2022, 'The effectiveness of formative assessment for enhancing reading achievement in K-12 classrooms: A meta-analysis', Frontiers in Psychology, vol. 13, 990196, https://doi.org/10.3389/fpsyg.2022.990196.