Journal Evaluation in Education (JEE)
Vol. 7, No. 3, July 2026, pp. 632~642
ISSN: 2716-4160, DOI: 10.37251/jee.v7i3.3456

Review Article

Automated Scoring for Four-Tier Diagnostic Tests in Physics: A Systematic Review and Needs Analysis

Kania Fatikasari[1,*], Asrial[1], Tugiyo Aminoto[1], Tanti[2]
[1] Master' Program in Science Education, Universitas Jambi, Jambi, Indonesia
[2] Faculty of Tarbiyah and Teacher Training, Universitas Islam Negeri Sulthan Thaha Saifuddin Jambi, Jambi, Indonesia

Article Info

Article history:

Received May 25, 2026
Revised Jun 05, 2026
Accepted Jun 30, 2026
OnlineFirst Jul 15, 2026

Keywords:

Automated Scoring
Four-Tier Diagnostic Test
Misconception Diagnosis
Physics Education
Systematic Literature Review

ABSTRACT

Purpose of the study: This study aimed to establish an empirical foundation for developing an automated scoring system for four-tier diagnostic tests in physics education by mapping current research on tiered diagnostic instruments and identifying the practical scoring challenges teachers face in classroom misconception assessment.

Methodology: A two-phase sequential design was used. A systematic literature review followed the PRISMA protocol across Scopus, ScienceDirect, and Google Scholar (compiled in Mendeley), with record screening managed in Rayyan and narrative synthesis of the selected studies, complemented by semi-structured interviews with four junior high school science teachers, analyzed thematically.

Main Findings: Of 87 four-tier diagnostic studies, only 12 (13.8%) incorporated digital or automated elements, and most digital systems functioned merely as test-delivery platforms while misconception analysis remained manual. Interviews confirmed four recurring barriers: a heavy manual scoring burden, delayed feedback, difficulty distinguishing genuine understanding from guessing, and the need for a simple and accessible automated system. Teachers consistently requested a simple system generating immediate diagnostic results and misconception classifications.

Novelty/Originality of this study: Unlike prior reviews that emphasize instrument construction, this study integrates systematic review evidence with concrete teacher needs to justify an accessible, low-tech automated scoring system built on Google Forms and Google Sheets, reframing the central barrier of four-tier assessment as scoring complexity rather than instrument design.

Corresponding Author:
Kania Fatikasari,
Master' Program in Science Education, Universitas Jambi, A. Manaf Street, Telanaipura, Jambi, 36122, Indonesia
Email: kania.fatikasari.21@gmail.com

1. INTRODUCTION

Misconceptions in physics education represent a systemic challenge because they frequently contradict well-established scientific theory yet prove highly resistant to change in students' thinking [1]. This problem is particularly pronounced in abstract concepts such as work and energy, where students tend to construct ideas grounded in everyday experience that are scientifically inaccurate [1]-[3]. Research conducted at senior high schools in Jambi, for instance, found that misconceptions in work and energy sub-concepts reached as high as 80% in certain categories, with persistent errors in understanding potential energy, the relationship between kinetic and mechanical energy, and the concept of positive and negative work [1]. These findings from the senior high school level are relevant to the present study because students entering junior high school often carry the same experiential foundations that give rise to such errors; addressing misconceptions at an earlier stage may therefore prevent their entrenchment in subsequent learning. A fundamental limitation of conventional assessment is its inability to detect students who produce correct answers on the basis of flawed reasoning, false positives, or rote memorization [2], [4]. Just as a physician requires an accurate diagnosis before prescribing treatment, educators need valid instruments because misdiagnosis of students' conceptual understanding can be argued to be not merely ineffective but potentially harmful to students' cognitive development [2].

To address the limitations of ordinary multiple-choice tests, multi-tier diagnostic instruments have been progressively developed. This evolution began with the two-tier format, which pairs an answer with a corresponding reason [5], and continued with the three-tier format, which adds a confidence level item [6]. The four-tier diagnostic test, however, is regarded as the superior standard because it independently assesses students' confidence in their content answer and their confidence in their reasoning [7], [2]. This format enables researchers to rigorously distinguish genuine misconceptions from guessing and lack of knowledge, yielding a multidimensional portrait of conceptual understanding [2], [8], [9]. Its diagnostic utility has been demonstrated and validated internationally, including studies of Newton's laws [10], [11], density [12], liquid pressure [13], and electric charge [14], as well as across diverse topics in Indonesian research such as static electricity [15], work and energy [16], [3], and fluid mechanics [17], [18].

Despite its theoretical advantages, the use of four-tier tests in practice is impeded by the substantial burden of manual analysis. This complexity originates from what may be termed a two-layered scoring logic: teachers must not only categorize students into understanding profiles using an intricate decision table [2], [19], but also classify the specific type of misconception from each student's response pattern [1]. Such manual analysis is widely regarded as inefficient and prohibitively time-consuming, making it unrealistic for teachers managing large class sizes and limited resources [4], [1]. A systematic review by Unsal and Aydogdu [20] examining technological tools in misconception studies found that while technology adoption has grown significantly, tools oriented toward misconception detection, as opposed to remediation, remain underrepresented, with most digital interventions focusing on animations and simulations rather than automated diagnostic scoring systems.

Automated scoring holds considerable promise as a bridge to delivering instant diagnostic feedback. Recent work has begun to formalize the scoring of four-tier responses; for example, systematic scoring frameworks based on partial credit and modern test theory have been proposed to make four-tier categorization replicable and analyzable [21]-[23]. The broader integration of artificial intelligence into assessment for learning has likewise been advanced as a means of generating timely, individualized feedback [24], and intelligent teacher-assistant prototypes have demonstrated rapid automated diagnosis of misconceptions [25]. Evidence from adjacent fields further suggests that automated scoring can achieve substantial agreement with human scoring [26]. Yet the systematic integration of such technology to facilitate conceptual diagnosis in everyday physics teaching remains scarcely explored [8], and without systems that reduce the evaluative burden, even sophisticated diagnostic instruments will struggle to achieve classroom adoption [4].

Instrument effectiveness is also highly contingent on alignment with teachers' technical capabilities and available resources [2]. Most educators are already familiar with Google's digital ecosystem and hold positive perceptions of Google Forms as an accessible and efficient assessment medium [27]. Consistent with this, several diagnostic instruments have been delivered through Google Forms and mobile or web platforms [28]-[31], indicating that the integration of Google Forms with Google Sheets is a realistic and accessible solution because it requires no specialized programming knowledge [27].

Several prior reviews have mapped this field. Çelikkanlı and Kızılcık [32] reviewed four-tier diagnostic studies in physics, Dirman et al. [33] compared four-tier and five-tier formats, Resbiantoro et al. [34] reviewed the diagnosis, causes, and remediation of physics misconceptions, and Budiyono et al. [35] mapped the evolution of tiered instruments in Indonesian research; related syntheses have also addressed five-tier formats [36] and diagnostic tests more broadly [37]. These reviews, however, concentrate predominantly on instrument development, formats, and reported misconceptions, and pay limited attention to how diagnostic results are scored and to whether existing tools meet the practical needs of classroom teachers. The present study addresses this gap by combining a systematic review of automation in four-tier assessment with a primary needs analysis of teachers, thereby connecting the research literature to concrete classroom requirements.

Against this backdrop, this study pursues two research questions: (RQ1) What are the trends in tiered diagnostic instruments in physics education between 2015 and 2025, and to what extent have four-tier studies integrated automated or digital scoring? (RQ2) What challenges do science teachers face in implementing and scoring diagnostic assessments, and what features do they require in an automated scoring system? By mapping the existing literature and surfacing the concrete barriers encountered in schools, this study seeks to close the functional gap between the complexity of four-tier instrument design and the practical need of teachers for an efficient scoring system built on Google Forms and Google Sheets.

2. RESEARCH METHOD

This study used a two-phase sequential design consisting of a Systematic Literature Review (SLR) and a needs analysis. The SLR was conducted to identify trends in four-tier diagnostic instruments and automated scoring in physics education (RQ1). The needs analysis was carried out through semi-structured interviews with science teachers to identify practical challenges in classroom assessment (RQ2). Findings from both phases were used as the basis for designing an automated scoring system for four-tier diagnostic tests.

The SLR followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol to guide the identification, screening, and selection of relevant studies. Literature searches were conducted in three databases—Scopus, ScienceDirect, and Google Scholar—and all retrieved records were compiled and managed in Mendeley. Three search strings were applied in each database: (1) “four-tier diagnostic test” AND misconception AND physics; (2) (“two-tier” OR “three-tier” OR “four-tier”) AND misconception AND physics; and (3) “four-tier diagnostic test” AND (“digital system” OR “web-based” OR “automated scoring”). Articles were screened using the inclusion and exclusion criteria shown in Table 1.

Table 1. Inclusion and Exclusion Criteria

Criteria	Inclusion	Exclusion
Publication year	2015–2025	Before 2015
Language	English, Indonesian	Other languages
Document type	Peer-reviewed journal articles and conference proceedings	Theses, books, grey literature
Topic focus	Diagnostic tests, misconception identification, automated scoring in education	Assessment topics unrelated to diagnostic instruments
Subject area	Physics or science education	Non-science disciplines

A total of 882 records were identified across the three databases and compiled in Mendeley. After 176 duplicate records were removed, 706 records were screened on title and abstract, of which 514 were excluded, leaving 192 reports sought for retrieval and descriptively mapped by instrument type. Of these, 13 could not be obtained, leaving 179 reports assessed for full-text eligibility. Following full-text assessment, 92 reports were excluded for not employing four-tier diagnostic instruments, yielding 87 four-tier diagnostic studies. From these, 12 studies incorporating digital or automated assessment features (9 primary and 3 supporting) were included in the qualitative synthesis. The complete selection process is presented in Figure 1.

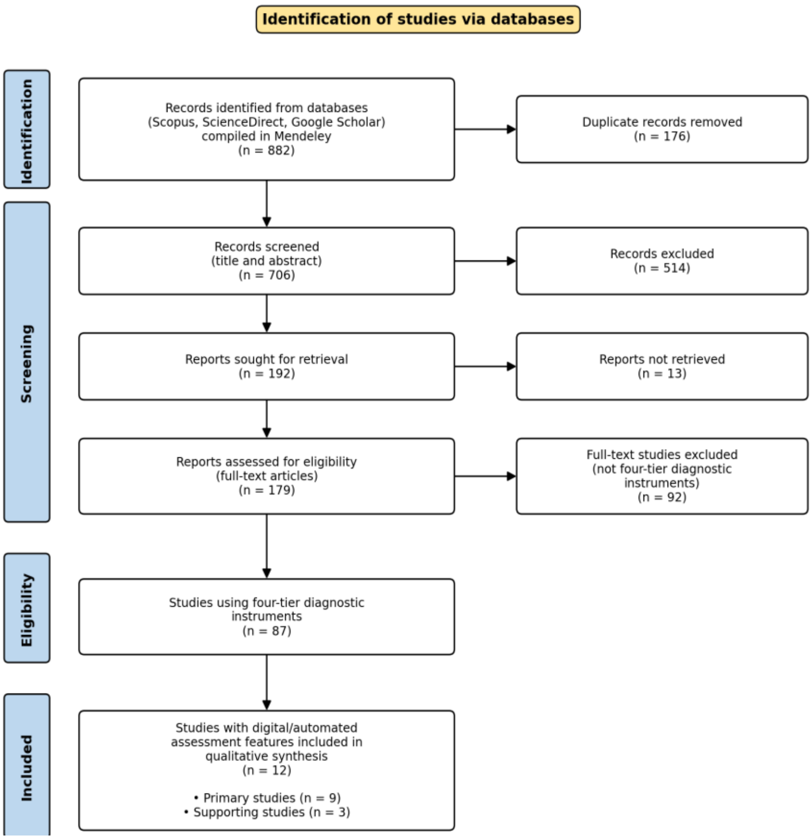

Figure 1. PRISMA flow diagram of the study selection process

Data extracted from each article included: (1) author and publication year, (2) diagnostic instrument type, (3) automated or digital scoring features, (4) subject matter and educational level, and (5) scoring method. Article screening and organization were conducted using Rayyan. The extracted data were analyzed using narrative synthesis to identify trends in instrument types, scoring approaches, and the use of automated systems in four-tier diagnostic assessment.

The needs analysis involved four junior high school science teachers from SMPN 6 Muaro Jambi selected through purposive sampling. All participants had experience using digital assessment platforms in science learning. Participant profiles are presented in Table 2.

Table 2. Participant Profile

Code	Teaching Experience	Current Classes	Digital Assessment Experience
T1	±18 years	Grade 9, 6 classes	Google Form, Quizizz
T2	±7 years	Grade 7–8, 5 classes	Google Form, Quizizz
T3	±38 years	Grade 7, 6 classes	Google Form
T4	±20 years	Grade 8–9, 6 classes	Google Form

Data were collected through semi-structured interviews guided by a protocol addressing four thematic areas: (1) current practices and time demands of assessment correction, (2) use of digital platforms for assessment, (3) experiences identifying misconceptions, and (4) expectations for automated systems. Interviews were audio-recorded and transcribed verbatim. Thematic analysis following Braun and Clarke [38] was used to organize and interpret the data: initial codes were generated inductively from the transcripts and then grouped into broader themes. Feature requirements reported in Table 7 were derived through a second-pass content analysis of participant responses specifically addressing desired system functionalities, cross-referenced across informants for consistency.

3. RESULTS AND DISCUSSION

3.1. Systematic Literature Review Results

The systematic search identified 192 records that passed title and abstract screening and were retained for descriptive mapping of tiered diagnostic instruments published between 2015 and 2025. The volume of research grew markedly over this period, particularly from 2021 onward, indicating intensifying interest in tiered diagnostic assessment. Four-tier diagnostic tests represented the largest proportion of the mapped literature, accounting for 45.3% of the studies. The distribution of instrument types is presented in Table 3 and visualized in Figure 2.

Table 3. Distribution of Instrument Types

Instrument Type	n	Percentage
Four-tier	87	45.3%
Three-tier	42	21.9%
Two-tier	28	14.6%
Five-tier	20	10.4%
Six-tier	10	5.2%
Multi-tier	5	2.6%
Total	192	100%

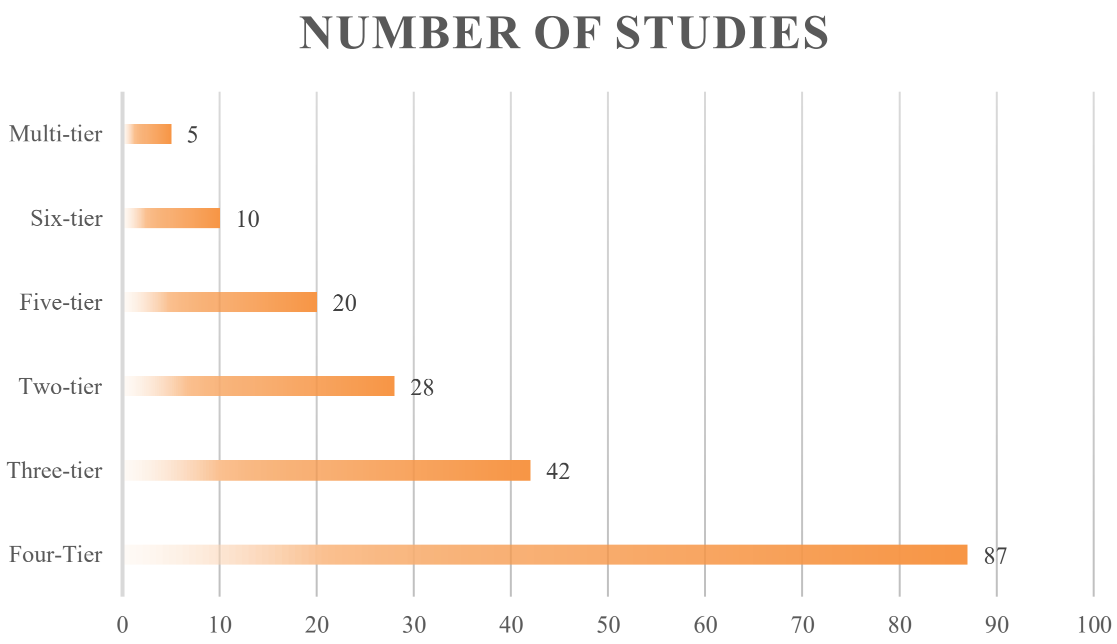

Figure 2. Number of Studies Using Different Tiered Diagnostic Instrument Types

The dominance of four-tier instruments indicates increasing interest in comprehensive diagnostic approaches that capture not only answer accuracy but also reasoning and confidence levels, consistent with the broader development of tiered assessment from two- and three-tier formats toward more detailed misconception identification [5]-[7]. Across the corpus, four-tier tests were applied to a broad range of physics topics, including Newton's laws and kinematics [8], [39], [40], static and dynamic fluids [41]-[44], heat and temperature [45]-[47], electricity and magnetism [48], work and energy [49], waves [50], and thermodynamics [51]. The five- and six-tier formats appeared far less frequently; despite their theoretical comprehensiveness, they remained rarely applied owing to their complexity and the high cognitive load they place on respondents [35].

Only a small proportion of the four-tier studies incorporated automated or digital scoring features. Within the subset of four-tier studies, only 12 of 87 articles (13.8%) included automated or digital elements, indicating that although four-tier instruments are increasingly developed, their analytical implementation remains largely manual; similar patterns have been reported in recent reviews of technology integration in misconception analysis [20]. The 12 studies meeting the inclusion criterion are summarized in Table 4.

Table 4. Classification of Included Four-Tier Diagnostic Studies with Digital/Automated Features

Year	Author(s)	Platform	Automation Status	Note
2018	Pujayanto et al. [52]	Web-based system	Semi-automated	Claims automation but scoring mechanism not transparently described
2021	Septiyani & Nanto [53]	Website	Semi-automated	Web-based delivery; not described as full auto-scoring
2022	Istiyono et al. [54]	Web-based system	Automated (strong)	8-category misconception classification with clear algorithmic mapping
2023	Astuti et al. [55]	Android app	Semi-automated	Misconception profiles generated automatically; scoring algorithm not detailed
2021	Surmaini et al. [56]	Online test	Non-automated	CRI calculated manually; no automated system
2022	Huda et al. [57]	Digital test	Non-automated	Manual analysis; no auto-classification system
2024	Rusilowati et al. [58]	Web-based e-diagnostic	Automated (strong)	Automatic misconception classification + integrated database
2025	Yuliana et al. [59]	Web-based system	Automated (strong)	Web system with automatic misconception classification
2021	Ekacitra et al. [60]	Website	Semi-automated	Automatic output of categories & graphs; rule-based (no AI/inference engine)
2024	Laili et al. [61]	Web/mobile app	Semi-automated	Focused on app performance, not misconception diagnosis system
2024	Hasibuan et al. [62]	Moodle (CBT)	Non-automated	CRI-based analysis and manual interpretation
2021	Saputri et al. [63]	Web-based system	Semi-automated	Automatic classification & graphs; rule-based (no inferential model)

These findings indicate that digitalization does not necessarily produce automated diagnostic analysis. In many studies, web-based systems, Android applications, and Moodle platforms functioned primarily as media for test administration rather than complete diagnostic engines, with the analytical burden of identifying misconceptions still depending on teacher interpretation [53], [55], [57]. Only a few studies, notably Istiyono et al. [54], Rusilowati et al. [58], and Yuliana et al. [59], demonstrated stronger automation through explicit misconception classification. Notably, where strong automation was achieved, it rested on an explicit scoring model, such as the eight-category partial-credit approach formalized for four-tier scoring [21], [23]. A recurring concern, however, is the limited transparency of scoring algorithms: several studies reported automatic classification results without explaining how the categorization was performed, which affects replicability and reduces practical usability. This positions the present review against prior syntheses [32], [35], which emphasized instrument formats and reported misconceptions but did not foreground the scoring mechanism itself as the principal barrier to classroom adoption.

3.3 Needs Analysis Results

Semi-structured interviews with four junior high school science teachers revealed four major issues: scoring burden, limited diagnostic feedback from digital platforms, difficulty identifying misconceptions, and the need for an accessible automated system. The thematic analysis results are summarized in Table 5.

Table 5. Summary of Thematic Analysis Findings from Teacher Interviews

Theme	Theme Label	Key Finding	Evidence (Informants)
1	Manual scoring burden	All four teachers reported that correcting and analyzing test results is time-consuming, limiting the timeliness of feedback.	T1: 1–2 hours/class; T2: up to 1 day/class; T3: limits essay items to ≤5; T4: 1–2 days/class

Theme	Theme Label	Key Finding	Evidence (Informants)
2	Digital delivery without diagnostic output	Google Form is widely used for delivery, but all teachers reported it provides only scores, not diagnostic information.	T1–T4: all use Google Form; none obtained misconception diagnosis from it
3	Hidden misconceptions	Teachers struggle to identify students with correct answers based on wrong reasoning, or students who are silent in class.	T1: cannot tell understanding from guessing; T2, T3: rote memorizers appear adequate
4	Demand for accessible, non-technical system	All four teachers expressed interest but emphasized that usability and simplicity are non-negotiable for adoption.	T1: 'jangan ribet'; T2: 'tidak rumit'; T3: 'kalau terlalu teknis jadi beban'; T4: must work with weak internet

All informants reported that analyzing student answers required substantial time, particularly for reasoning-based items, which delayed feedback and led several teachers to reduce the number of essay questions. These findings support the SLR result that most four-tier studies still depend on manual analysis, and echo challenges reported in four-tier studies on electricity, heat, and mechanics where misconception identification relied heavily on teacher interpretation [15], [45], [57], reinforcing that the complexity of four-tier analysis is a major barrier to classroom implementation [2]. Although all teachers used Google Forms, the platform served only for score collection rather than misconception analysis, consistent with evidence that many digital platforms prioritize administration efficiency over automated conceptual diagnosis [27], [20], [53]. Teachers also struggled to distinguish genuine understanding from guessing or memorization, mirroring documented misconceptions in work and energy [3], [16], momentum and impulse [64], [65], static fluids [66], [67], and Newtonian mechanics [39], [8], precisely the discriminations four-tier tests are designed to make [10], [68]. Teachers' willingness to adopt four-tier diagnosis when supported by appropriate tools has likewise been reported elsewhere [69].

All informants expressed interest in an automated diagnostic system, provided it remained simple and practical. Teachers emphasized automatic result generation, misconception identification, visual summaries, and compatibility with limited internet access. The requested features are summarized in Table 6.

Table 6. Feature Requirements Identified from Teacher Interviews

Feature Category	Specific Request	Mentioned by
Immediate result display	System displays results automatically after test completion, without manual input	T1, T2, T3, T4
Per-student diagnostic recap	Identifies which students have not yet understood and on which specific items	T1, T2, T3, T4
Misconception type identification	Shows the type of error or misconception, not just the score	T1, T2
Confidence level visibility	Shows whether a student's answer reflects genuine understanding or uncertainty	T1
Simplified visual output	Graphs or summaries that are easy to read without additional calculation	T1, T2, T3, T4
Tab-lock / anti-cheating feature	Prevents students from opening other tabs during the test	T2, T3, T4
Low-connectivity functionality	System should be usable even with an unstable internet connection	T4
Follow-up action guidance	Suggests which material to re-teach and how to address specific misconceptions	T4

Read together, the two phases converge on a single conclusion: the principal obstacle to four-tier diagnostic assessment in classrooms is not instrument construction but the complexity of scoring and interpretation. The SLR shows that automated misconception analysis is rare and, where present, often opaque, while the needs analysis shows that teachers already deliver tests digitally yet obtain only scores, not diagnoses. The features teachers prioritized immediate results, per-student and per-item recaps, misconception typing, and simple visual output to map directly onto an accessible scoring pipeline that pairs a Google Form (one item per tier) with a Google Sheets engine that converts responses to a binary answer–confidence pattern and assigns a conceptual category, an approach grounded in established four-tier decision tables [19] and formal scoring models [23]. Such a design would translate the diagnostic logic that currently burdens teachers into an automatic, transparent, and low-technical workflow, consistent with formative-assessment principles that emphasize timely, actionable feedback [70], and with the broader movement toward AI-supported, feedback-oriented assessment [24].

This study has several limitations. The needs analysis involved four teachers from a single school, so its findings are illustrative rather than generalizable and should be read as a preliminary needs assessment; future work should broaden the sample across schools and regions and triangulate interviews with classroom observation. The SLR included Google Scholar, which is less systematically indexed than Scopus or ScienceDirect, and a formal quality appraisal of included studies was not performed; subsequent reviews could add risk-of-bias appraisal and inter-rater reliability for screening. Finally, this study establishes only the empirical rationale for an automated scoring system; the design, development, and classroom validation of the Google Forms–Google Sheets system, including features that cannot be fully realized within Google Forms such as robust anti-cheating controls, remain to be carried out in a subsequent development study.

4. CONCLUSION

This study establishes an empirical foundation for developing an automated scoring system for four-tier diagnostic tests in physics education. The systematic literature review showed that although four-tier diagnostic instruments have become the most widely used tiered assessment format, only 13.8% of the 87 four-tier studies reviewed incorporated digital or automated elements, and stronger automation depended on explicit, transparent scoring models. The needs analysis confirmed that teachers experience this limitation directly, facing difficulties in analyzing student reasoning, identifying misconceptions, and providing timely feedback, while existing digital platforms are generally used only for test administration. Together these results indicate that the central challenge in implementing four-tier diagnostic assessment is the complexity of scoring and misconception analysis rather than instrument construction. The study therefore contributes evidence-based justification for developing a practical and accessible automated diagnostic system that integrates misconception identification, confidence analysis, and teacher-oriented feedback, suitable for real classroom conditions.

ACKNOWLEDGEMENTS

The authors thank the science teachers of Junior High School 6 Muaro Jambi for their participation in the interviews. This research received no external funding.

AUTHOR CONTRIBUTIONS

Conceptualization, K.F. and A.; Methodology, K.F.; Formal Analysis, K.F.; Investigation, K.F.; Data Curation, K.F.; Writing – Original Draft Preparation, K.F.; Writing – Review & Editing, A. and T.A.; Supervision, A., T.A., and T. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The author(s) declare no conflict of interest.

INFORMED CONSENT STATEMENT

Informed consent was obtained from all participants involved in the study. Prior to the interviews, each teacher was provided with an explanation of the study's objectives, procedures, and intended use of the data, and voluntarily agreed to participate. All images in this article have been authorized and are the property of the authors.

USE OF ARTIFICIAL INTELLIGENCE (AI)-ASSISTED TECHNOLOGY

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

REFERENCES

[1] M. Maison, N. Lestari, and A. Widaningtyas, "Identifikasi miskonsepsi siswa pada materi usaha dan energi [Identifying student misconceptions on the topic of work and energy," *Jurnal Penelitian Pendidikan IPA*, vol. 6, no. 1, pp. 32–39, 2020, doi: 10.29303/jppipa.v6i1.314.

[2] D. K. Gurel, A. Eryilmaz, and L. C. McDermott, "A review and comparison of diagnostic instruments to identify students' misconceptions in science," *Eurasia J. Math., Sci. Technol. Educ.*, vol. 11, no. 5, pp. 989–1008, 2015, doi: 10.12973/eurasia.2015.1369a.

[3] S. K. Aripiani, A. Samsudin, D. Saepuzaman, I. Kaniawati, and E. Suhendi, “Diagnostic instruments of four-tier test work and energy (FORTUNE) to identify the level of students' conceptions,” *Tadris: Jurnal Keguruan dan Ilmu Tarbiyah*, vol. 8, no. 1, pp. 19–32, 2023, doi: 10.24042/tadris.v8i1.13524.

[4] D. S. Jubaedah, A. Halim, and Murniati, “Development of two-tier multiple choice diagnostic instrument to identify misconceptions of junior high school students on Newton’s law,” *J. Phys.: Conf. Ser*., vol. 895, 2017, Art. no. 012054.

[5] D. F. Treagust, “Evaluating students' misconceptions by means of diagnostic multiple choice items,” *Res. Sci. Educ*., vol. 16, no. 1, pp. 199–207, 1986, doi: 10.1007/BF02356835.

[6] H. Peşman and A. Eryılmaz, “Development of a three-tier test to assess misconceptions about simple electric circuits,” *J. Educ. Res*., vol. 103, no. 3, pp. 208–222, 2010, doi: 10.1080/00220670903383002.

[7] I. Caleon and R. Subramaniam, “Development and application of a three-tier diagnostic test to assess secondary students’ understanding of waves,” *Int. J. Sci. Educ*., vol. 32, no. 7, pp. 939–961, 2010, doi: 10.1080/09500690902890130.

[8] A. Wandi, N. Suprapto, and H. Mubarok, “Four-tier diagnostic test to identify students' misconceptions on Newton’s law of gravitation,” *Jurnal Penelitian Pendidikan IPA*, vol. 9, no. 1, pp. 118–127, 2023.

[9] J. Nainggolan, B. Silaban, D. Sinaga, and F. Zendrato, “Analysis of physics misconceptions of students in mechanic materials using the tier multiple choice diagnostic test,” *Al-Ishlah: Jurnal Pendidikan*, vol. 15, no. 3, pp. 3023–3034, 2023.

[10] N. J. Fratiwi, I. Kaniawati, E. Suhendi, I. Suyana, and A. Samsudin, “The transformation of two-tier test into four-tier test on Newton's laws concepts,” *AIP Conf. Proc*., vol. 1848, 2017, doi: 10.1063/1.4983967.

[11] I. Kaniawati, N. J. Fratiwi, A. Danawan, I. Suyana, A. Samsudin, and E. Suhendi, “Analyzing students’ misconceptions about Newton's laws through Four-Tier Newtonian Test (FTNT),” *J. Turk. Sci. Educ*., vol. 16, no. 1, pp. 110–122, 2019, doi: 10.12973/tused.10269a.

[12] S. A. Kiray and S. Simsek, "Determination and evaluation of the science teacher candidates' misconceptions about density by using four-tier diagnostic test," Int. J. Sci. Math. Educ., vol. 19, no. 5, pp. 935–955, 2021, doi: 10.1007/s10763-020-10087-5.

[13] T. Taban and S. A. Kiray, “Determination of science teacher candidates' misconceptions on liquid pressure with four-tier diagnostic test,” *Int. J. Sci. Math. Educ*., vol. 20, no. 8, pp. 1791–1811, 2022, doi: 10.1007/s10763-021-10224-8.

[14] N. Onder-Celikkanli and M. Tan, “Determining Turkish high school students' misconceptions about electric charge imbalance by using a four-tier misconception test,” *Phys. Educ*., vol. 57, no. 5, 2022, Art. no. 055014, doi: 10.1088/1361-6552/ac68c1.

[15] R. Aisahsari and F. U. Ermawati, “Evaluating student's misconceptions and the causes on direct current concepts by means of four-tier multiple choice test,” *J. Phys.: Conf. Ser*., vol. 1491, 2020, Art. no. 012009, doi: 10.1088/1742-6596/1491/1/012009.

[16] A. Samsudin, N. J. Fratiwi, A. A. Amin, T. R. Wirakusumah, I. Kaniawati, and Y. Coştu, “Reconstructing students’ misconceptions on work and energy through the PDEODE*E tasks with think-pair-share,” *J. Turk. Sci. Educ*., vol. 18, no. 1, pp. 118–144, 2021, doi: 10.36681/tused.2021.56.

[17] F. Kafiyani, A. Samsudin, and D. Saepuzaman, “Development of four-tier diagnostic test (FTDT) to identify student's mental models on static fluid,” *J. Phys.: Conf. Ser*., vol. 1280, no. 5, 2019, Art. no. 052030, doi: 10.1088/1742-6596/1280/5/052030.

[18] Hunaidah, Erniwati, and M. A. Mahdiannur, “Four-tier diagnostic test to assess students' misconceptions about fluids,” *Jurnal Penelitian Pendidikan IPA*, vol. 8, no. 2, pp. 612–620, 2022.

[19] D. Kaltakçi, “Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics,” Ph.D. dissertation, Grad. Sch. Natural and Appl. Sci., Middle East Tech. Univ., Ankara, Turkey, 2012.

[20] A. A. Unsal and C. Aydogdu, “Technological tools used in misconceptions studies in physics education: A systematic review,” *J. Educ. Sci., Environ. Health (JESEH)*, vol. 11, no. 3, pp. 210–220, 2025, doi: 10.55549/jeseh.814.

[21] E. Istiyono, “Diagnostic tests as an important pillar in today’s physics learning: Four-tier diagnostic test, a comprehensive diagnostic test solution,” *J. Phys.: Conf. Ser*., vol. 2392, 2022, Art. no. 012001, doi: 10.1088/1742-6596/2392/1/012001.

[22] E. Istiyono, W. S. B. Dwandaru, K. Fenditasari, M. R. S. N. A. Shanti, and D. Saepuzaman, “The development of a four-tier diagnostic test based on modern test theory in physics education,” *Eur. J. Educ. Res*., vol. 12, no. 1, pp. 371–385, 2023, doi: 10.12973/eu-jer.12.1.371.

[23] E. Istiyono, K. Fenditasari, M. R. S. Shanti, D. Saepuzaman, and W. S. B. Dwandaru, “An eight-category partial credit model as very appropriate for four-tier diagnostic test scoring in physics learning,” *AIP Conf. Proc*., vol. 2622, 2024, Art. no. 040001, doi: 10.1063/5.0133862.

[24] B. Memarian and T. Doleck, “A review of assessment for learning with artificial intelligence,” *Comput. Hum. Behav.: Artif. Humans*, vol. 2, no. 1, 2024, Art. no. 100040, doi: 10.1016/j.chbah.2023.100040.

[25] S. Sutrisno, Maison, J. Marzal, W. Kurniawan, and A. Saputra, “Rapid artificial diagnostic test in intelligent teacher assistant system (ITAS) to identify misconceptions on kinematics,” in *Proc. Int. Conf. Math., Sci., and Education*, 2023, pp. 316–325.

[26] A. Wahlen, C. Kuhn, O. Zlatkin-Troitschanskaia, C. Gold, T. Zesch, and A. Horbach, “Automated scoring of teachers' pedagogical content knowledge: A comparison between human and machine scoring,” *Front. Educ*., vol. 5, 2020, Art. no. 149, doi: 10.3389/feduc.2020.00149.

[27] M. N. Qomar, and E. S. Utaminingsih, “Development of evaluation instruments based on google forms for grade iv elementary school on the force topic,” *Research and Development Journal of Education*, vol. 11, no. 2, pp. 1313-1324, 2025, doi: 10.30998/rdje.v11i2.26205.

[28] A. Royani and W. Setyarsih, "Development of Google Form-based five-tier e-diagnostic test to identify conception levels and track students' misconceptions on thermodynamics materials," *Prisma Sains*, vol. 10, no. 3, 2022, doi: 10.33394/j-ps.v10i3.5209.

[29] I. T. Pratiwi, V. Viyanti, and D. Permadi, "The development of the two-tier diagnostic test instrument with Google Form to measure student misconceptions on energy and energy forms," *Impulse: Journal of Research and Innovation in Physics Education*, vol. 4, no. 1, pp. 22-30, 2024, doi: 10.14421/impulse.2024.41-03.

[30] P. Rahayu and D. R. Haristy, "Developing an online mobile-based diagnostic instrument to identify science misconceptions for junior high school students," in *Proc. 6th Int. Seminar on Science Education (ISSE),* 2021, pp. 1–9.

[31] W. Kurniawan, M. A. Amin, R. O. Sandra, and M. Iqbal, "The development of web diagnostic test as a misconception tool in work and energy," *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, vol. 9, no. 1, 2023, doi: 10.21009/1.09109.

[32] N. Ö. Çelikkanlı and H. Ş. Kızılcık, "A review of studies about four-tier diagnostic tests in physics education," *J. Turk. Sci. Educ*., vol. 19, no. 4, pp. 1291–1311, 2022, doi: 10.36681/tused.2022.175.

[33] H. M. Dirman, F. Mufit, and Festiyed, "Review and comparison of four-tier multiple choice and five-tier multiple choice diagnostic tests to identify mastery of physics concepts," *Jurnal Penelitian Pendidikan IPA*, vol. 8, no. 1, pp. 1–12, 2022, doi: 10.29303/jppipa.v8i1.838.

[34] G. Resbiantoro, R. Setiani, and Dwikoranto, "A review of misconception in physics: The diagnosis, causes, and remediation," *J. Turk. Sci. Educ*., vol. 19, no. 2, pp. 403–427, 2022, doi: 10.36681/tused.2022.128.

[35] A. Budiyono, S. A. Mulyadi, A. Wildani, N. A. S. Pertiwi, and F. Alatas, "Mapping the evolution of tiered diagnostic instruments for identifying students' misconceptions in physics: A systematic review of Indonesian research (2015–2025)," *Jurnal Pendidikan Fisika*, vol. 13, no. 2, pp. 225–242, 2025, doi: 10.26618/jpf.v13i2.17879.

[36] E. Triani and I. Fahlani, "Five-tier format misconception diagnostic instrument: Systematic literature review," *EduFisika: Jurnal Pendidikan Fisika*, vol. 9, no. 2, 2024, doi: 10.59052/edufisika.v9i3.38725.

[37] D. Lufita, H. Kuswanto, and D. Rosana, "Identification of misconception using diagnostic tests: Systematic literature review," *Jurnal Penelitian Pendidikan IPA*, vol. 11, no. 1, 2025, doi: 10.29303/jppipa.v11i.9986.

[38] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qual. Res. Psychol*., vol. 3, no. 2, pp. 77–101, 2006, doi: 10.1191/1478088706qp063oa.

[39] S. A. Kiray, "Development and evaluation of a four-tier diagnostic test to assess preservice science teachers' misconceptions about Newton's laws," *J. Sci. Educ. Technol*., vol. 29, pp. 575–599, 2020, doi: 10.1007/s10956-020-09831-9.

[40] N. Hermita, A. Suhandi, E. Syaodih, A. Samsudin, T. Iswari, and R. Suhandi, "Constructing and implementing a four-tier test about static electricity to diagnose pre-service elementary school teachers' misconceptions," *J. Phys.: Conf. Ser*., vol. 895, 2017, Art. no. 012167, doi: 10.1088/1742-6596/895/1/012167.

[41] N. Nurdini, A. Suhandi, T. Ramalis, and A. Samsudin, "Developing Multitier Instrument of Fluids Concepts (MIFO) to measure student's conception: A Rasch analysis approach," *J. Adv. Res. Dyn. Control Syst*., vol. 12, no. 7, pp. 1–10, 2020.

[42] R. Diani, J. Alfin, Y. M. Anggraeni, M. Mustari, and D. Fujiani, "Four-tier diagnostic test with certainty of response index on the concepts of fluid," *J. Phys.: Conf. Ser*., vol. 1155, 2019, Art. no. 012078, doi: 10.1088/1742-6596/1155/1/012078.

[43] O. Saputra, I. M. Hermanto, A. I. Safitri, H. T. Putra, A. Nadzirin, and D. Lusiyanti, "Development of 4-Tier Diagnostic Fluid Static Test (4T-DFST) to identify profile students' conception," *J. Phys.: Conf. Ser*., vol. 2900, 2024, Art. no. 012003, doi: 10.1088/1742-6596/2900/1/012003.

[44] N. Bessas, E. Tzanaki, D. Vavougios, and V. P. Plagianakos, "Diagnosing students' misconception in hydrostatic pressure through a 4-tier test," *Heliyon*, vol. 10, no. 23, 2024, doi: 10.1016/j.heliyon.2024.e40425.

[45] A. Devialita, A. Suhandi, A. Samsudin, and D. Rusdiana, "Analysis conception of secondary student's using four tier test on heat and temperature," *J. Phys.: Conf. Ser*., vol. 2098, 2021, Art. no. 012002, doi: 10.1088/1742-6596/2098/1/012002.

[46] M. I. Sukarelawan, S. Sriyanto, S. Sumarna, J. Juliasih, and S. Suparno, "Four-tier heat and temperature diagnostic test (4T-HTDT) to identify student misconceptions," *Jurnal Ilmu Pendidikan Fisika dan Riset*, vol. 4, no. 2, 2021, doi: 10.30599/jipfri.v5i1.856.

[47] K. Fenditasari, J. Jumadi, and E. Istiyono, "Identification of misconceptions on heat and temperature among physics education students using four-tier diagnostic test," *J. Phys.: Conf. Ser*., vol. 1470, 2020, Art. no. 012055, doi: 10.1088/1742-6596/1470/1/012055.

[48] R. Rahmawati, W. Widiasih, D. H. Marisda, and R. Riskawati, "Using four-tier test to identify prospective elementary teacher students' misconception on electricity topic," *Jurnal Penelitian Pendidikan IPA*, vol. 9, no. 2, pp. 876–885, 2023, 10.29303/jppipa.v9i10.3272.

[49] S. Anggrayni and F. U. Ermawati, "The validity of four-tier's misconception diagnostic test for work and energy concepts," *J. Phys.: Conf. Ser*., vol. 1171, 2019, doi: 10.1088/1742-6596/1171/1/012037.

[50] A. M. R. Tumanggor, Supahar, H. Kuswanto, and E. S. Ringo, "Using four-tier diagnostic test instruments to detect physics teacher candidates' misconceptions: Case of mechanical wave concepts," *J. Phys.: Conf. Ser*., vol. 1440, 2020, Art. no. 012059, doi: 10.1088/1742-6596/1440/1/012059.

[51] Y. B. Bhakti, I. A. D. Astuti, and R. Prasetya, "Four-tier thermodynamics diagnostic test (4T-TDT) to identify student misconception," *KnE Soc. Sci*., vol. 7, no. 14, pp. 106-116, 2022, doi: 10.18502/kss.v7i14.11958.

[52] P. Pujayanto, R. Budiharti, E. Adhitama, N. R. A. Nuraini, and H. V. Putri, "The development of a web-based assessment system to identify students' misconception automatically on linear kinematics with a four-tier instrument test," *Phys. Educ*., vol. 53, no. 4, 2018, doi: 10.1088/1361-6552/aac695.

[53] E. Septiyani and D. Nanto, "Four-tier diagnostic test assisted website for identifying misconceptions of heat and temperature," *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, vol. 7, no. 1, 2021, doi: 10.21009/1.07106.

[54] E. Istiyono, W. S. B. Dwandaru, and R. Faizah, "A four-tier web-based assessment with eight categories diagnostic," *Int. J. Eval. Res. Educ. (IJERE)*, vol. 11, no. 2, pp. 735–744, 2022, doi: 10.11591/ijere.v11i2.21938.

[55] I. A. D. Astuti, Y. B. Bhakti, R. Prasetya, and Z. Zulherman, "Android-based 4-tier physics test app to identify student misconception profiles," *Int. J. Eval. Res. Educ. (IJERE)*, vol. 12, no. 3, pp. 1356–1363, 2023, doi: 10.11591/ijere.v12i3.25536.

[56] Surmaini, I. Syafe'i, and R. Diani, "An analysis of students' physics misconceptions in online learning using the four-tier diagnostic test with certainty of response index (CRI)," *J. Phys.: Conf. Ser*., vol. 1796, 2021, doi: 10.1088/1742-6596/1796/1/012099.

[57] C. Huda, A. L. Ma'ani, and U. Kaltsum, "Analysis of student misconceptions using digital four-tier diagnostics test on Newton's law," *Phys. Educ. Res. J*., vol. 4, no. 1, pp. 17–22, 2022, doi: 10.21580/perj.2022.4.1.8631.

[58] A. Rusilowati, T. Darsono, F. O. Sekarningtias, and Q. Fariyani, "Development of a four-tier e-diagnostic test on the topic of momentum to measure and reduce student misconception," *Jurnal Pendidikan Fisika Indonesia*, vol. 20, no. 2, pp. 173–187, 2024, doi: 10.15294/jpfi.v20i2.1297.

[59] E. Septiyani, and D. Nanto, "Four-tier diagnostic test assisted website for identifies misconceptions heat and temperature," *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, vol. 7, no. 1, pp. 35-42, 2021, doi: 10.21009/1.07104.

[60] T. N. Ekacitra, M. Maison, and W. Kurniawan, "Pengembangan instrumen four-tier test berbasis website pada materi rangkaian listrik di perguruan tinggi," *Jurnal Riset dan Konseptual*, vol. 5, no. 2, pp. 314–321, 2021, doi: 10.28926/riset_konseptual.v5i2.349.

[61] M. Wu, P. Tian, D. Sun, D. Feng, and M, Luo, "Evaluating students' conceptual understanding of isomers based on a four-tier diagnostic tool in upper secondary schools," *International Journal of Science and Mathematics Education*, vol. 23, no. 4, pp. 907-947, 2025, doi: 10.1007/s10763-024-10494-y.

[62] F. A. Hasibuan, A. M. Harahap, and T. Hasibuan, "Analisis penguasaan konsep fisika mahasiswa melalui four tier diagnostik test berbasis Moodle [Analysis of students' mastery of physics concepts through a Moodle-based four tier diagnostic test]," *MIND: Jurnal Ilmu Pendidikan dan Kebudayaan*, vol. 4, no. 2, pp. 88–97, 2024, doi: 10.55266/jurnalmind.v4i2.424.

[63] L. Saputri, M. Maison, and W. Kurniawan, "Pengembangan four-tier diagnostic test berbasis website untuk mengidentifikasi miskonsepsi pada materi suhu dan kalor [Development of a website-based four-tier diagnostic test to identify misconceptions on temperature and heat material]," *Jurnal Ilmiah Teknologi Informasi Asia (JITIKA)*, vol. 15, no. 1, pp. 61–70, 2021, doi: 10.32815/jitika.v15i1.563.

[64] S. A. Amalia, E. Suhendi, I. Kaniawati, A. Samsudin, and N. J. Fratiwi, "Diagnosis of student's misconception on momentum and impulse through inquiry learning with computer simulation (ILCS)," *J. Phys.: Conf. Ser*., vol. 1204, 2019, doi: 10.1088/1742-6596/1204/1/012073.

[65] A. Setiani, D. Firmansyah, R. Nurliani, and A. Samsudin, "Analysis of momentum and impulse misconceptions of college students using four-tier diagnostic tests," *Int. J. Educ. Teach. Zone*, vol. 2, no. 2, pp. 146–155, 2023, doi: 10.57092/ijetz.v2i3.146.

[66] M. G. Purwanto, R. Nurliani, I. Kaniawati, and A. Samsudin, "Promoting the hydrostatic conceptual change test (HCCT) with four-tier diagnostic test item," *J. Phys.: Conf. Ser*., vol. 1013, 2018, doi: 10.1088/1742-6596/1013/1/012035.

[67] S. Sudirman, N. J. Ningsih, Saparini, M. Ariska, N. Andriani, and A. Pasaribu, "The identification of misconceptions of physics education students at Universitas Sriwijaya using four-tier diagnostic test on static fluid material," *J. Math. Sci. Comput. Educ*., vol. 3, no. 2, 2023, doi: 10.20527/jmscedu.v3i1.8664.

[68] M. Eliya, W. Purwianingsih, A. Supriatna, A. Samsudin, A. H. Aminudin, and M. Mardiana, "Electromotive forces on multitier instruments (EFoMI): Development to measure student misconception with Rasch," *Journal of Environment and Sustainability Education*, vol. 2, no. 1, pp. 40-52, 2024, doi: 10.62672/joease.v2i1.32.

[69] I. Khasana, Viyanti, and U. Rosidin, "Teachers' perception of four-tier diagnostic test instrument equipped with self-diagnosis sheet to identify students' misconception on linear motion material," *Jurnal Penelitian Pendidikan IPA*, vol. 10, no. 1, pp. 234–245, 2024, doi: 10.29303/jppipa.v10i9.7720.

[70] P. Black and D. Wiliam, "Assessment and classroom learning," *Assess. Educ.: Princ., Policy Pract*., vol. 5, no. 1, pp. 7–74, 1998, doi: 10.1080/0969595980050102.