JURNAL EDUSCIENCE (JES)
p-ISSN : 2303 - 355X | e-ISSN : 2685 - 2217
PUBLISHED BY : LPPM of UNIVERSITAS LABUHANBATU

# Development of Technology-Based Assessment Instruments in Science Learning: A Narrative Review (2015-2025)

Yul Ifda Tanjung[1], Tiur Malasari Siregar[2], Suci Frisnoiry[3], Elfitra[4], Taufiq Ramadhan[5], Abil Mansyur[6], Rizky Agassy Sihombing[7]

[1] Department of Science Education, Universitas Negeri Medan, North Sumatra, Indonesia
[2, 3, 4]Department of Mathematics Education, Universitas Negeri Medan, North Sumatra, Indonesia
[5]Department of Business Law, Universitas Negeri Medan, North Sumatra, Indonesia
[6]Department of Mathematics, Universitas Negeri Medan, North Sumatra, Indonesia
[7]Graduate Institute of Science Education, National Taiwan Normal University, Taipei City 106, Taiwan
*Email: yuly@unimed.ac.id

**ARTICLE INFO**

*Keywords:*
Assessment
Technology
Narrative Review
Science Learning

**ABSTRACT**

**Purpose -** This study critically synthesizes research on the development of technology-based assessment instruments in secondary science education (junior and senior high school) through a structured Narrative Literature Review (NLR). This review therefore identifies instrument types, implementation contexts, research dimensions examined, and learning impacts, while highlighting gaps in equity, teacher readiness, and longitudinal outcomes.

**Methodology -** Articles were systematically retrieved from Google Scholar and Mendeley within the 2015–2025 publication range using predefined keywords related to technology-based assessment and science learning. Inclusion criteria required peer-reviewed empirical studies focusing on secondary-level science and reporting instrument development or evaluation. Fourteen eligible articles were analyzed using structured content analysis, with emphasis on instrument characteristics, evidence of validity and reliability, implementation contexts, methodological approaches, and reported limitations.

**Findings -** The synthesis indicates that research predominantly centers on Computer-Based Testing (CBT), Computerized Adaptive Testing (CAT), gamified assessments, two-tier diagnostic tests, and LMS-based e-portfolios. Overall, these instruments demonstrate satisfactory validity and reliability and effectively measure higher-order thinking skills and conceptual understanding. They also enhance student motivation and accelerate feedback processes. However, most evidence is based on short-term, small-scale implementations, with limited attention to equity issues, teacher readiness, and sustained learning effects.

**Contribution -** This review integrates psychometric, pedagogical, and technological perspectives into a coherent analytical framework for technology-based assessment in science education. It clarifies existing research gaps and underscores the need for broader contextual validation, stronger methodological rigor, and the integration of learning analytics to ensure sustainable and evidence-based digital assessment practices.

Received 27 October 2025; Received in revised form 15 November 2025; Accepted 21 February 2026
Jurnal Eduscience (JES) Volume 13 No. 1 (2026)
Available online 28 February 2026

## INTRODUCTION

Assessment constitutes a central component in evaluating the effectiveness and quality of learning, as it provides evidence of students' conceptual mastery, reasoning skills, and attainment of instructional objectives. Within science education—including mathematics, physics, chemistry, and biology—assessment is expected to capture not only factual knowledge but also higher-order thinking, scientific reasoning, and problem-solving abilities (Istiyono et al., 2020; Morris et al., 2021). Effective learning is reflected through meaningful teacher–student interactions and active student participation, which can be examined via structured evaluation processes (Akram, 2019; Han, 2025). Consequently, assessment functions not merely as a measurement tool but as a mechanism for enhancing learning.

Traditionally, classroom assessment has relied heavily on written instruments such as quizzes, worksheets, multiple-choice tests, and essay questions (Hani et al., 2020; Tanjung et al., 2020). While these methods are widely used, they often emphasize lower-level cognitive skills and provide limited feedback for improving learning processes. Assessment is no longer solely focused on learning outcomes but also on the learning process and reflection. Student involvement in this assessment will encourage internal feedback, enabling students to take responsibility for improving their learning (Nicol, 2021; Brown, 2019; Tanjung, Wulandari & Ramadhani, 2021). This goal can be achieved by using information and communication technology to facilitate the assessment process.

The rapid advancement of digital technology has significantly influenced educational practices, including assessment. The integration of computers, mobile devices, and online platforms has facilitated the emergence of technology-based assessment environments (Prendes-Espinosa et al., 2020). Research indicates that digital assessment can enhance efficiency, provide immediate feedback, automate scoring, and broaden the scope of measurable competencies (Alruwais et al., 2018; Elmahdi et al., 2018; Rostaminezhad, 2019). Moreover, technology enables the design of interactive simulations, multimedia tasks, adaptive testing, and log-data analysis to capture learning processes (Shute et al., 2016; O'Leary et al., 2018; Molnár, 2021). In science education, these innovations allow for more comprehensive measurement of inquiry skills, conceptual understanding, and higher-order thinking (Zlatkin-Troitschanskaia et al., 2019; Fitriani et al., 2024).

Despite these benefits, the implementation of technology-based assessment remains limited. Many digital assessments simply replicate conventional multiple-choice items in online formats without leveraging interactive features (O'Leary et al., 2018). Technical challenges— such as system compatibility, reliability issues, and network disruptions—may threaten validity and reliability (Jimada-Ojuolape & The, 2020). Moreover, educators often face challenges related to limited understanding of digital assessment principles, including construct alignment, validity assurance, and reliability in technology-based environments (Nugraha et al., 2021). Low digital competence and difficulties in designing multimedia or simulation-based items further constrain effective implementation (Anestasya, Koto & Setiawan, 2025; Fitriani et al., 2024).

These implementation challenges indicate a mismatch between the potential of technology-based assessment and its practical realization in science classrooms. Although previous studies have examined digital assessment tools, most focus on specific disciplines, individual platforms (e.g., quiz-based systems), or isolated competencies. There remains a lack of a comprehensive synthesis that systematically classifies technology-based assessment instruments across science domains, maps the competencies measured, and analyzes their reported impact on learning outcomes. This fragmentation creates a significant research gap in understanding how technology-based assessment functions holistically within science education (O'Leary et al., 2018; Molnár, 2021; Soamole, Amiroh, & Salim, 2023).

Accordingly, this study seeks to address the identified gap by conducting a structured review of empirical research on technology-based assessment in science learning. The novelty of this work lies in: 1) Providing a cross-disciplinary synthesis encompassing mathematics, physics, chemistry, and biology; 2) Developing a systematic classification of technology-based assessment instruments (e.g., simulation-based, game-based, adaptive, multimedia-integrated); 3) Mapping the competencies and aspects measured; and 4) Identifying patterns of impact on science learning outcomes.

By synthesizing findings across disciplines and instrument types, this study aims to establish a more transparent conceptual framework that supports educators, researchers, and policymakers in designing valid, reliable, and meaningful technology-based assessment systems. This study critically synthesizes research on the development of technology-based assessment instruments in secondary science education (junior and senior high school). Although many studies report the effectiveness of digital assessments, integrative analyses addressing theoretical grounding, psychometric rigor, contextual implementation, and long-term sustainability remain limited. This review therefore identifies instrument types, implementation contexts, research dimensions examined, and learning impacts, while highlighting gaps in equity, teacher readiness, and longitudinal outcomes.

This review covers empirical studies published between 2015 and 2025 to capture both foundational and recent developments in digital assessment research. The analysis focuses on technology-based assessment instruments applied in mathematics, physics, chemistry, and biology. Studies addressing general e-learning without explicit assessment components are excluded. The review emphasizes instrument characteristics, measured constructs, implementation strategies, and reported impacts, rather than technical aspects of software engineering.

## METHODOLOGY

This study applies a Narrative Literature Review (NLR) approach. NLR is a literature review method that describes and scientifically examines a particular topic or theme based on the researcher's theoretical and contextual perspective (Pautasso, 2019; Byrne, 2016). Narrative reviews take a less formal approach than systematic reviews because they do not require the presentation of the stricter aspects characteristic of systematic reviews, such as reporting methodology, search terms, databases used, and inclusion and exclusion criteria (Jahan et al., 2016).

A common criticism of narrative reviews is that they do not always follow the rules of literature search (Byrne, 2016; Furunes, 2019). Literature searches must be conducted comprehensively and in accordance with predetermined eligibility criteria (Bramer et al., 2018) in order to increase confidence that the findings and conclusions of the review are reliable and unbiased. With this in mind, the narrative review in this study used article search and data processing stages adopted from systematic review steps to reduce bias.

The research sample consisted of peer-reviewed empirical articles examining the development of technology-based assessment instruments in science education, published between 2015 and 2025. The articles were retrieved from Google Scholar and Mendeley databases. To ensure alignment with the research objectives, a structured Boolean search strategy was applied using predefined keywords related to technology-based assessment, instrument development, and science learning contexts (physics, chemistry, biology, and general science). The search strategy was designed to ensure that retrieved studies explicitly examined the development, validation, or implementation of digital or technology-based assessment instruments in science education. The Boolean search structure used in the article retrieval process is presented in Table 1.

The identification stage retrieved 68,900 records from Mendeley and Google Scholar using predefined keywords related to the development of technology-based assessment instruments in science learning (chemistry, biology, physics, and general science). During screening, articles were limited to journal publications published between 2015 and 2025. Title and abstract screening yielded 1,085 potentially relevant articles, while 67,815 records were excluded for duplication or irrelevance. Full-text eligibility assessment applied explicit inclusion criteria: (1) empirical research; (2) focus on the development, validation, or implementation of technology-based assessment instruments; (3) conducted in science subjects; (4) secondary education level; and (5) published in English. Exclusion criteria included conceptual or review papers, studies outside the field of science education, non-instrument-focused research, and inaccessible full texts. After this stage, 312 articles met the scope, and 773 were excluded. Finally, 14 articles were included for analysis. Data were analyzed using structured content analysis, with each study coded according to the development model, the type of digital assessment, the validation method, and the reported outcomes.

**Table 1.** Boolean search structure used in the literature retrieval process

| Search Component | Keywords / Terms |
|---|---|
| Technology-Based Assessment | “digital assessment” OR “technology-based assessment” OR “computer-based test” OR CBT OR “computerized adaptive test” OR CAT OR “online assessment” OR “e-assessment” |
| Instrument Development | “instrument development” OR “test development” OR validation OR reliability OR “psychometric analysis” |
| Science Context | “science education” OR physics OR chemistry OR biology |
| Educational Level | “secondary school” OR “high school” OR “junior high school” |

The findings were synthesized narratively to identify dominant trends, methodological patterns, and research gaps. At the same time, this systematic approach provides a comprehensive overview of the field; however, it has limitations, including potential bias in database selection, exclusion of inaccessible full texts, and subjective interpretation during narrative synthesis. The selection process is illustrated in Figure 1.

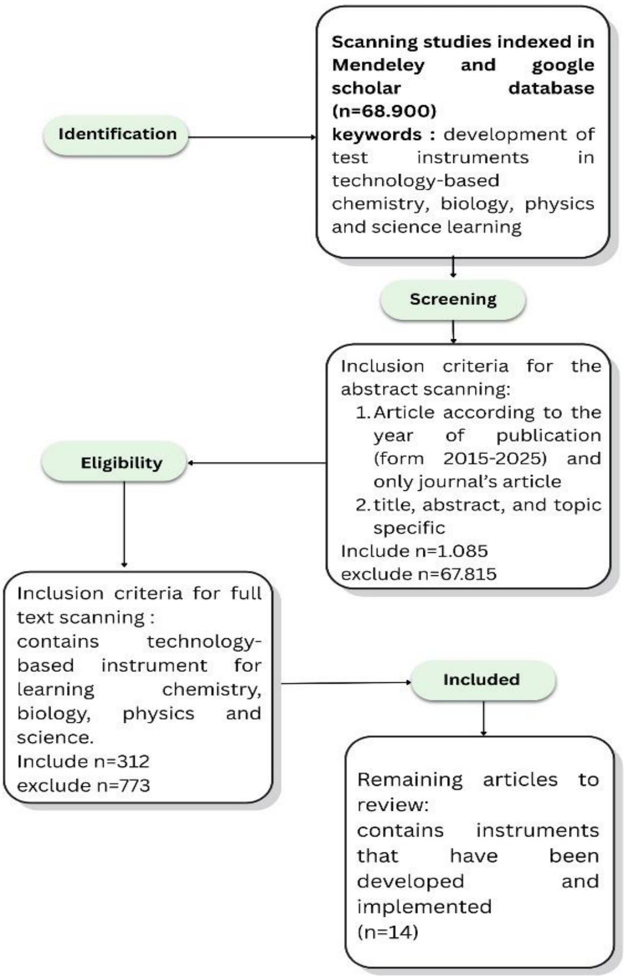
Identification
Scanning studies indexed in Mendeley and google scholar database (n=68.900)
keywords : development of test instruments in technology-based chemistry, biology, physics and science learning
Screening
Inclusion criteria for the abstract scanning:
1. Article according to the year of publication (form 2015-2025) and only journal’s article
2. title, abstract, and topic specific
Include n=1.085
exclude n=67.815
Eligibility
Inclusion criteria for full text scanning :
contains technology-based instrument for learning chemistry, biology, physics and science.
Include n=312
exclude n=773
Included
Remaining articles to review:
contains instruments that have been developed and implemented
(n=14)

**Figure 1.** Article Selection Procedure for Narrative Literature Review

## FINDINGS

### Overall Thematic Synthesis

The initial mapping of the reviewed studies is summarized in Table 1, which presents the distribution of research across science disciplines, types of technology-based assessment instruments developed, research models employed, and the primary learning outcomes measured. This table serves as the foundational overview for the subsequent thematic and analytical discussion.

**Table 1.** Descriptive Synthesis of Technology-Based Assessment Studies (2015–2025)

| Author | Field | Instrument Type | Main Focus | Key Findings |
|---|---|---|---|---|
| Sari et al. (2018) | Physics | CBT (HOTS MCQ) | Analytical & evaluative skills | Valid and reliable for HOTS measurement |
| Istiyono et al. (2020) | Physics | CAT (IRT-based) | Adaptive HOTS measurement | High precision and efficiency |
| Supriyati et al. (2021) | Physics | CBT with equating | Score equivalence | Improved comparability across test forms |
| Haka et al. (2019) | Science | Two-tier diagnostic test | Misconception detection | Highly feasible and attractive instrument |
| Nurdiyanti, Sukarmin & Budiharti (2022) | Physics | Moodle-based CBT | Conceptual understanding | Practical with automatic feedback |
| Firmansyah, Chandra & Aripin (2019) | Biology | LMS-based e-portfolio | Reflective assessment | Improved student learning outcomes |
| Lestari, Sjaifuddin & Resti (2022) | Science | Quizizz-based HOTS | Higher-order thinking | Increased motivation and scores |
| Syahidah et al. (2024) | Chemistry | Blooket gamified test | Engagement & mastery | Valid, reliable, and engaging |
| Asriana, Auliah & Hardin (2023) | Chemistry | Telegram BOT assessment | HOTS evaluation | Valid and practical for mobile-based testing |
| Soto Rodriguez, Vilas & Redondo (2021) | Science | Computer-Based Assessment | Score comparability | Strong correlation with paper-based tests |
| Faradisa et al. (2024) | Science | CBT (4TMC) | Critical thinking | Suitable for Grade 8 evaluation |
| Ihsan et al. (2020) | Chemistry | Computer-based concept test | Concept comprehension | Valid and practical instrument |
| Setiyorini & Lestari (2023) | Chemistry | CBT large-scale trial | Validity & feasibility | Acceptable validation results |
| Dari, Numertayasa & Pradnyana (2025) | Science | Wordwall game-based quiz | Interactive evaluation | Enjoyable alternative assessment |

The synthesis shows that technology-based assessment development in secondary science education is concentrated in five main categories: Computer-Based Testing (CBT), Computerized Adaptive Testing (CAT), gamified assessment platforms, two-tier diagnostic tests, and Learning Management System (LMS)-based e-portfolios. CBT-based instruments are most common, followed by gamified platforms and LMS-integrated systems, while CAT and two-tier diagnostic formats are less common. Most studies emphasize procedural development and feasibility testing rather than comparative effectiveness across contexts. Consequently, although validity and reliability indices are frequently reported, cross-study comparisons remain limited due to variation in constructs measured and implementation scales.

### Distribution of Research Based on Year of Publication

Based on scientific publications on the development of technology-based instruments for science learning across physics, chemistry, and biology from 2015 to 2025, 14 publications were synthesized, as shown in Figure 2 below.

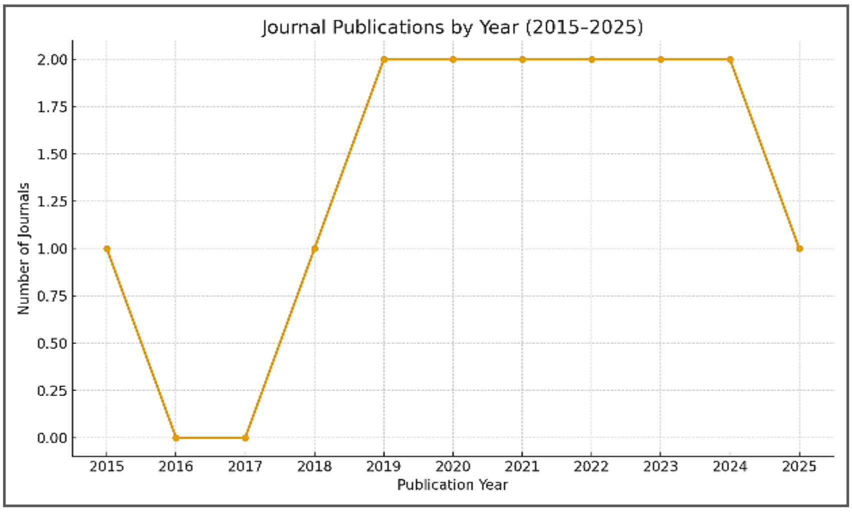

**Figure 2.** Number of Synthesis Journals 2015-2025

Based on the graph of the number of scientific publications synthesized in this article, research and development related to technology-based assessment instruments in science learning have shown fluctuating development since 2015. Initially, publications were still limited and inconsistent, as seen in the emergence of research in early 2015, which continued to increase until 2025 for publications on the development of technology-based instruments in the era of 21st-century education.

### Distribution of Research Based on Research Methods

Over the past 10 years, research on the development of technology-based instruments has generally been conducted using a research and development (R&D) approach with various research models. The graph of the types of research models is presented in Figure 3.

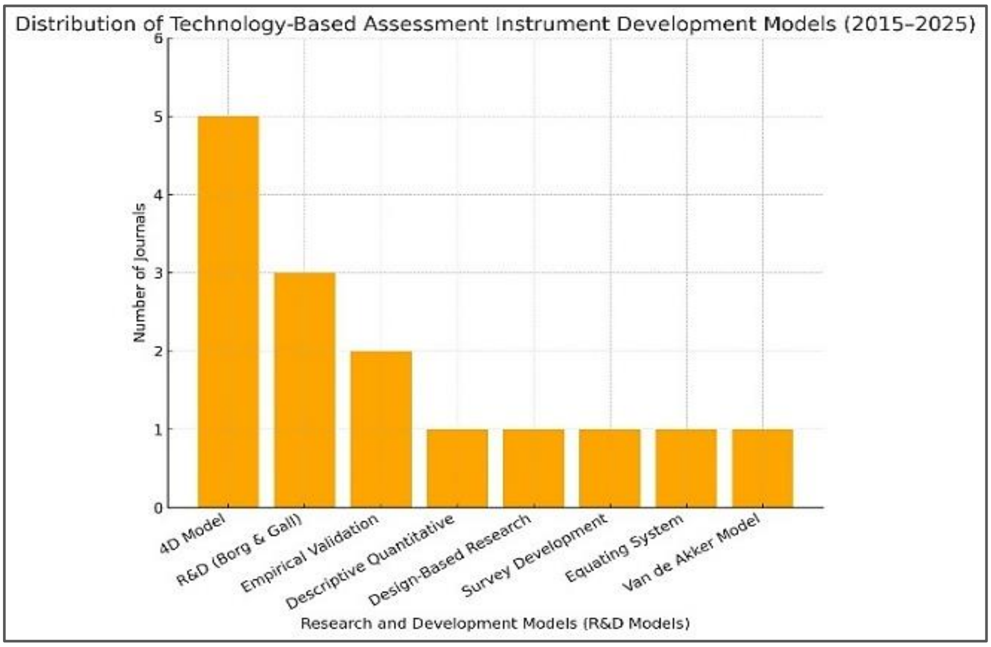

**Figure 3.** Types of Research Development Models

Figure 3 shows that researchers predominantly used the 4D development model. Other models used by researchers included empirical validation, quantitative descriptive, design-based research, survey development, equating systems, and the Van de Akker model. The dominance of the 4D model likely reflects its clear, structured, and practical stages, which make it suitable for educational product development. Its systematic phases facilitate needs analysis, prototype design, and expert validation. However, while the model supports procedural development, it does not inherently emphasize rigorous psychometric testing or large-scale validation. As a result, instruments developed using this model may demonstrate strong content validity but sometimes lack deeper construct validation and broader generalizability.

### Types of Instruments and Their Pedagogical Orientation

The synthesis indicates that Computer-Based Testing (CBT) is the most frequently developed instrument, particularly in physics and chemistry. Gamified assessments using digital quiz platforms represent the second most common category. Other formats include Computerized Adaptive Testing (CAT), LMS-based e-portfolios, and two-tier diagnostic tests. CBT instruments generally emphasize structured item banks and standardized scoring systems, whereas gamified instruments prioritize interactivity and engagement. Adaptive testing and diagnostic instruments appear in fewer studies, indicating limited diversification of assessment models.

### Implementation Context

Most instruments were implemented at the junior and senior secondary school levels using research and development (R&D) designs, predominantly the 4D model. Validation procedures generally include expert judgment and small- to medium-scale classroom trials. Cross-institutional validation, multi-school implementation, and longitudinal testing are rarely reported. Implementation tends to be limited to specific instructional units rather than integrated across full curricula. This pattern indicates that technology-based assessment development is still largely prototype-oriented rather than system-oriented.

### Measured Aspects

Across studies, cognitive aspects dominate measurement focus. Higher-order thinking skills (HOTS) and conceptual understanding are the primary constructs assessed. Psychometric analysis commonly includes validity, reliability, item difficulty, and discrimination indices. Affective dimensions, such as motivation and engagement, are reported primarily in gamified assessments. However, measurement of broader 21st-century competencies—such as collaboration, scientific communication, digital literacy, or problem-solving in authentic contexts—appears limited. Adaptive measurement using IRT is present but not widespread, and simulation- or performance-based digital assessments are underrepresented.

### Impact on Learning

Findings consistently report acceptable levels of validity and reliability across instruments. Reported impacts include increased participation, faster feedback cycles, scoring efficiency, and short-term gains in conceptual mastery. However, the evidence primarily comes from short-term classroom implementation. Few studies evaluate sustained learning trajectories, comparative effectiveness across disciplines, or long-term competency development. The reported impact is therefore focused on immediate instructional outcomes rather than on broader educational transformation.

### Disciplinary Distribution and Research Gaps

The distribution of studies shows a concentration in physics, followed by chemistry, while biology and general science receive comparatively less attention. Physics studies are more likely to employ adaptive systems and equating procedures, whereas chemistry and general science studies tend to emphasize gamified quizzes and CBT formats. Biology-focused research more commonly integrates portfolio-based or LMS-supported assessment. Despite the generally positive validation results, several important gaps remain evident. The development of simulation-based and performance-based digital assessments remains limited, and adaptive testing models are concentrated mainly in physics rather than evenly distributed across science disciplines. Most instruments focus primarily on measuring higher-order thinking skills (HOTS), with little attention to broader competencies such as collaboration, digital literacy, or authentic scientific inquiry. In addition, longitudinal evaluation and large-scale validation studies are rarely conducted, restricting the generalizability of findings. Comparative analysis across physics, chemistry, biology, and general science is also insufficient. Overall, although technology-based assessment instruments in secondary science education have increased in variety and technical sophistication, their development remains heavily centered on cognitive testing formats and short-term validation contexts.

## DISCUSSION

Based on a synthesis of 14 articles, the development of technology-based assessment instruments in physics education shows a very positive trend. Previous researchers have generally used the Research and Development (R&D) method, which allows the development process to be carried out systematically, from the design stage to final validation. The R&D method can provide technology-based assessment instruments that are valid and reliable, thus aligning with the demands of 21st-century learning (Hasyim et al., 2021; Sari et al., 2018).

To date, advances in digital technology have brought about significant changes in learning evaluation practices, particularly in the field of science, which requires a comprehensive assessment of students' conceptual and procedural abilities, as well as higher-order thinking skills (Faradisa et al., 2024; Istiyono et al., 2020; Sari et al., 2018). The implementation of these assessments in authentic classroom contexts shows variations in infrastructure readiness, teacher experience, and student digital literacy, which influence the effectiveness of each technology-based instrument. The implementation of computerized adaptive testing (CAT), computer-based testing (CBT), and various interactive platforms such as Moodle, Quizizz, Blooket, Wordwalss, and Telegram BOT, as well as the use of Android, are innovations that strengthen the effectiveness, efficiency, and reliability of the assessment system and assessment model variations (Ihsan et al., 2020; Supriyati et al., 2021; Syahidah et al., 2024; Dari, Numertayasa & Pradnyana, 2025; Rachma, 2015; Setiyorini & Lestari; Sjaifuddin & Resti, 2022).

In general, the development of technology-based instruments in the field of science is of high urgency, as the nature of scientific materials demands complex visual and conceptual representations. The use of technology enables the presentation of dynamic contexts that are difficult to achieve through conventional tests (Asriana, Auliah & Hardin, 2023; Nurdiyanti, Sukarmin & Budiharti, 2022). In addition, the digitization of assessment provides advantages in automated data processing, instant feedback, and the use of big data analytics (learning analytics) to support evidence-based instructional decision-making (Soto Rodriguez, Vilas & Redondo, 2021; Wijaya et al., 2023). However, the effectiveness of technology-based assessments still depends on the quality of the instrument design, the reliability of the system, and the readiness of teachers and students to operate it (Istiyono et al., 2020; Supriyati et al., 2021). These contextual factors indicate that technology alone does not guarantee successful assessment outcomes without proper implementation strategies and teacher support.

The results of the study, summarized in Table 1, show that technology-based science assessment can be used to measure student learning outcomes, diagnose misconceptions, and map students' critical thinking skills (Haka et al., 2019; Firmansyah, Chandra & Aripin, 2019). This approach aligns with the principles of authentic assessment, which emphasize assessing both the learning process and outcomes with the support of digital technology (Setiyorini & Lestari, Sjaifuddin & Resti, 2022; Soto Rodriguez, Vilas & Redondo, 2021). Therefore, this study not only describes the instruments but also synthesizes cross-study patterns to identify gaps and inform future development of adaptive, equitable, and sustainable assessment tools.

### Types of Instruments and Synthesis Across Studies

Studies show five main categories of technology-based assessment instruments in science education: computer-based testing (CBT), computerized adaptive testing (CAT), gamification-based assessments, two-tier diagnostic tests, and digital authentic assessments such as e-portfolios (Faradisa et al., 2024; Sari et al., 2018; Haka et al., 2019). CBT is the most common, supporting HOTS evaluation and faster, more accurate assessment compared to conventional tests (Supriyati et al., 2021; Tanjung et al., 2020). CAT instruments, using Item Response Theory (IRT), adjust difficulty in real time to student ability, enhancing precision (Istiyono et al., 2020; Revilla et al., 2020).

Gamification elements—Blooket, Quizizz, Telegram BOT—have been shown to increase student motivation and engagement (Syahidah et al., 2024; Lestari, Sjaifuddin & Resti, 2022; Asriana, Auliah & Hardin, 2023). Two-tier tests and e-portfolios provide authentic assessment and conceptual diagnosis, supporting formative and reflective learning (Firmansyah, Chandra & Aripin, 2019; Nurdiyanti, Sukarmin & Budiharti, 2022; Buyarski & Landis, 2014).

Across studies, there is a consistent trend toward integrating multiple assessment types to capture both cognitive and affective learning outcomes. While most instruments successfully measure HOTS and conceptual understanding, few studies systematically address equity, long-term learning effects, or the impact of teacher and student readiness on outcomes, indicating a persistent gap in implementation research. Moreover, the convergence of adaptive testing, gamification, and authentic assessment reflects an emerging paradigm that emphasizes student-centered, participatory, and technology-mediated evaluation. However, challenges remain in standardizing protocols and ensuring fairness across diverse contexts.

Synthesis across studies indicates a shift from assessment of learning to assessment for and as learning, integrating objective testing with authentic, participatory evaluation approaches (Gikandi et al., 2011; Redecker, 2017; Kibble, 2017). This shift suggests that future research should focus not only on instrument development but also on their practical integration within curricula and their effects on student learning trajectories and teacher practices.

### Application of Technology-Based Assessment Instruments

Across studies, there is a clear trend toward combining multiple assessment types to capture both cognitive and affective learning outcomes. Most instruments effectively measure HOTS and conceptual understanding. However, few studies examine issues of equity, long-term learning effects, or the influence of teacher and student readiness, highlighting a gap in implementation research. The integration of adaptive testing, gamification, and authentic assessment reflects an emerging paradigm of student-centered, participatory, and technology-mediated evaluation. However, challenges remain in standardizing protocols and ensuring fairness across contexts. Additionally, the variability in digital platforms, access to technology, and teacher proficiency underscores the need for context-sensitive design and professional development to maximize the effectiveness of these assessments.

Overall, the evidence points to a shift from assessment of learning to assessment for and as learning, blending objective testing with authentic, participatory approaches (Gikandi et al., 2011; Redecker, 2017; Kibble, 2017). Future research should address not only instrument development but also their integration into the curriculum and their impact on student learning trajectories and teacher practices. Moreover, longitudinal studies and cross-context comparisons are needed to evaluate the sustainability, scalability, and equity of technology-based assessments, ensuring they support inclusive and meaningful learning experiences for diverse student populations.

### Aspects of Research

The development of technology-based assessment instruments in science learning generally addresses three interconnected aspects: psychometric, technological, and pedagogical, which together ensure instruments are valid, reliable, and educationally meaningful.

### *Psychometric Aspects*

Most studies prioritize psychometric rigor, analyzing content and construct validity, reliability, discrimination, difficulty level, and distractors. Application feasibility is tested through functionality, usability, and effectiveness trials (Tanjung et al., 2025). Item Response Theory (IRT) is widely used to enhance the precision of computerized adaptive tests (CAT) by more accurately estimating item difficulty and discrimination than classical methods (Istiyono et al., 2020; Van der Linden & Pashley, 2010). Equating procedures, as highlighted by Supriyati et al. (2021), maintain score consistency across test forms, ensuring fairness and comparability (Kolen & Brennan, 2014). Psychometric quality remains the foundation of valid and reliable digital assessments.

### *Technological Aspects*

Technological factors influence system efficiency, accessibility, and user experience. Platforms such as Moodle, Telegram BOT, and Blooket provide practical, interactive, and automated assessment experiences that enhance engagement and responsiveness (Nurdiyanti, Sukarmin & Budiharti, 2022; Asriana, Auliah & Hardin, 2023; Syahidah et al., 2024). The usability, interactivity, and reliability of the technological infrastructure largely determine the success of digital instruments.

### *Pedagogical Aspects*

Pedagogical considerations focus on whether instruments support conceptual understanding, reflection, and continuous skill development. Two-tier diagnostic tests and e-portfolios enable teachers to identify misconceptions, guide interventions, and foster students' self-reflection (Haka et al., 2019; Firmansyah, Chandra & Aripin, 2019). Computer-based assessments also increase engagement without compromising validity (Soto Rodriguez, Vilas & Redondo, 2021). Thus, digital assessments serve a dual purpose: measuring learning outcomes and facilitating learning itself (Gikandi et al., 2011).

Overall, the synthesized results from various studies show that technology-based assessment instruments in science have successfully integrated these three aspects in a balanced manner. This integration reflects the modern assessment paradigm, which is not only oriented toward final results but also supports adaptive, reflective, and continuous teaching and learning processes.

## The Impact of Technology-Based Assessment Instruments on Learning

Overall, studies show that technology-based assessments positively influence student learning outcomes, engagement, and feedback processes. Gamification-based instruments such as Blooket and Quizizz significantly increase participation and performance, with Syahidah et al. (2024) reporting higher achievement in stoichiometry, and Lestari, Sjaifuddin & Resti (2022) noting an 18% increase in average scores compared to conventional tests. Computer-based assessments also maintain strong validity and reliability relative to traditional methods while providing faster automated feedback, as demonstrated by Moodle-based assessments (Soto Rodriguez, Vilas & Redondo, 2021; Nurdiyanti, Sukarmin & Budiharti, 2022). Quick feedback supports motivation, reflection, and deeper learning (Gikandi et al., 2011; Kibble, 2017).

### *Comparison of Research Approaches and Contexts*

Studies reveal differences in methodology, context, and technology. CAT-based instruments and equating procedures (Istiyono et al., 2020; Supriyati et al., 2021) optimize measurement accuracy and efficiency but require large item banks and calibration expertise. Gamification tools—such as Blooket, Quizizz, and Telegram BOTs (Asriana, Auliah & Hardin, 2023; Syahidah et al., 2024)—enhance engagement, particularly in formative assessment contexts. Two-tier diagnostic tests and e-portfolios are oriented toward authentic assessment and conceptual diagnosis; however, inter-rater reliability in e-portfolios remains a challenge that limits objectivity (Buyarski & Landis, 2014; Gikandi et al., 2011).

### *General Patterns and Synthesis of Findings*

From all the studies analyzed, three main patterns can be identified:

1. Integration of assessment with digital learning. Assessment is no longer separate from the learning process but has become an integral part of the online and offline learning ecosystem (Redecker, 2017; Siemens, 2013).
2. Dominance of HOTS-based assessment. Most instruments are designed to assess analytical, evaluative, and creative abilities, which are at the core of modern science learning (Istiyono et al., 2020; Sari et al., 2018).
3. Increased use of adaptive and interactive platforms. Assessment technology is increasingly geared toward adaptive and enjoyable systems, making assessment not just a measuring tool but also an educational learning experience (Dicheva et al., 2019; Caponetto et al., 2021).

Despite demonstrated effectiveness, the success of digital assessments depends on teacher preparedness, student digital literacy, and institutional support. Few studies systematically address equity, long-term learning outcomes, or curriculum integration, leaving a gap for research on sustainable implementation in diverse educational contexts. Future research should explore these factors to optimize the use of technology-based assessments while maintaining fairness and validity.

### Limitations of Previous Studies

A critical analysis of the reviewed literature reveals several limitations that remain common in research on the development of technology-based assessment instruments in science education. First, most studies use relatively small sample sizes that do not fully represent the broader target population. This condition limits the generalization of findings and the external validity of the instruments developed. Second, many studies have not fully reported important technical parameters, such as the results of Item Response Theory (IRT) analyses or equating procedures used to assess test form equivalence. In fact, this aspect is crucial to ensure the stability and fairness of measurement results in various application contexts. Third, there are still few studies examining the long-term impact of using digital instruments on student competency development, particularly in critical thinking, science literacy, and 21st-century skills. Most studies focus only on initial validity and reliability, without evaluating the sustained effects of implementing digital assessments in the classroom.

Fourth, some studies also face technical and ethical constraints in implementation, such as limitations in network infrastructure, human resource readiness, and potential violations of academic integrity during the assessment. These challenges show that the success of digital assessments is not only determined by the quality of the instruments, but also by the readiness of the supporting ecosystem (Chen, 2023; Jurāne-Brēmane; Pals et al., 2023). Fifth, the integration of digital assessment and learning analytics is still rarely implemented optimally. In fact, data analytics can provide deep insights into students' learning processes and help improve the construct validity of instruments. Future research should explore this integration using an artificial intelligence (AI)-based approach that continues to prioritize ethical and data privacy considerations.

Additional methodological challenges include publication bias and variability in peer-review quality across journals. Future research should adopt stricter protocols, such as preregistration, diverse reporting of outcomes, and longitudinal designs, to improve transparency, replicability, and understanding of long-term impacts. Overall, while progress in the development of technology-based assessment is evident, further work is needed to broaden the scope, strengthen methodological rigor, and ensure sustainable, equitable, and context-sensitive application across diverse educational settings.

## CONCLUSION

The development of technology-based assessment instruments in science education has increased significantly over the past decade. Most studies used a Research and Development (R&D) approach with the

4D and Borg & Gall development models. The instruments developed included various forms of digital assessment such as computer-based testing (CBT), computerized adaptive testing (CAT), two-level diagnostic tests, gamification-based assessments, and Learning Management System (LMS)-based e-portfolios.

The results of the study confirm that digital assessment is not merely a tool for evaluating learning but also a medium for learning that encourages active student engagement and strengthens the formative and authentic assessment processes. The use of interactive platforms such as Moodle, Quizizz, Booklet, and Telegram BOT can significantly increase motivation, efficiency, and the quality of learning feedback. The success of technology-based assessment implementation is influenced by infrastructure readiness, teacher and student technological competence, and educational institution policy support. Digital assessment implementation requires a mature integration of the technical, pedagogical, and psychometric aspects to ensure measurement results truly reflect students' abilities holistically. The success of technology-based assessment implementation is influenced by infrastructure readiness, teacher and student technological competence, and educational institution policy support. Digital assessment implementation requires a mature integration of the technical, pedagogical, and psychometric aspects to ensure measurement results truly reflect students' abilities holistically.

This study contributes theoretically by synthesizing current development trends, methodological approaches, and validation practices in technology-based science assessment, thereby providing a comprehensive conceptual map that clarifies the integration of digital technology, assessment design, and psychometric rigor. It also contributes methodologically by identifying dominant development patterns and highlighting the importance of aligning technological innovation with measurement validity and reliability principles. Although many studies report positive impacts, gaps remain in large-scale empirical validation, cross-cultural implementation, long-term effectiveness evaluation, and the integration of advanced analytics for personalized feedback. Future research should therefore focus on experimental and longitudinal designs, broader sample representation, and more vigorous psychometric testing to ensure generalizability and robustness of findings. This study is limited by its reliance on published literature within selected databases and time ranges, which may exclude relevant unpublished or regional studies. Additionally, variations in methodological rigor across the reviewed studies may affect the consistency of the synthesized conclusions.

## REFERENCES

Abduljawad, M., & Ahmad, A. (2023). An analysis of mobile learning (M-Learning) in education. *Multicultural Education*, *9*(2), 145–152.

Akram, M. (2019). Relationship between Students' Perceptions of Teacher Effectiveness and Student Achievement at Secondary School Level. *Bulletin of Education and Research*, *41*(2), 93–108.

Alruwais, N., Wills, G., & Wald, M. (2018). Advantages and challenges of using e-assessment. *International Journal of Information and Education Technology*, *8*(1), 34–37.

Alsaadat, K. (2017). Mobile learning technologies. *International Journal of Electrical and Computer Engineering*, *7*(5), 2833.

Anestasya, F. F., Koto, I., & Setiawan, I. (2025). Implementation of Digital Assessment Based on the Learning Platform" Kahoot!" to Evaluate the Conceptual Understanding of Physics. *Kasuari: Physics Education Journal (KPEJ)*, *8*(1), 204–219.

Asriana, A., Auliah, A., & Hardin. (2023). Pengembangan alat evaluasi HOTS berbasis BOT Telegram pada materi larutan penyangga. *Jurnal Pendidikan Kimia, 15*(2), 122–131.

Bramer, W. M., de Jonge, G. B., Rethlefsen, M. L., Mast, F., & Kleijnen, J. (2018). A systematic approach to searching: An efficient and complete method to develop literature searches. *Journal of the Medical Library Association*, *106*(4), 531–541.

Brown, S. (2019). Using assessment and feedback to empower students and enhance their learning. In *Innovative assessment in higher education* (pp. 50–63). Routledge.

Buyarski, C., & Landis, C. (2014). Using ePortfolios to assess applied learning. *Peer Review, 16*(1), 24–26.

Byrne, J. A. (2016). Improving the peer review of narrative literature reviews. *Research integrity and peer review*, *1*(1), 12.

Caponetto, I., Earp, J., & Ott, M. (2021). Gamification and education: A literature review. *Computers in Human Behavior, 110*, 106385.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2011). Technological issues for computer-based assessment. *Assessment and teaching of 21st century skills*, 143-230.

Chen, D., Jeng, A., Sun, S., & Kaptur, B. (2023). Use of technology-based assessments: A systematic review covering over 30 countries. *Assessment in Education: Principles, Policy & Practice*, *30*(5-6), 396–428.

Dari, N. K. A. U., Numertayasa, I. W., & Pradnyana, P. B. (2025). Pengembangan Instrumen Evaluasi IPA Berbasis Platform Wordwall Untuk Siswa Kelas VI SDN 7 Manukaya Pada Materi Tata. *Dharmas Education Journal (DE_Journal)*, *6*(1), 224-234.

Destiana, D., Suchyadi, Y., & Anjaswuri, F. (2020). Pengembangan Instrumen Penilaian Untuk Meningkatkan Kualitas Pembelajaran Produktif Di Sekolah Dasar. *Jurnal Pendidikan Dan Pengajaran Guru Sekolah Dasar (JPPGuseda)*, *3*(2), 119–123.

Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2019). Gamification in education: A systematic mapping study. *Educational Technology & Society, 22*(3), 75–88.

Elmahdi, I., Al-Hattami, A., & Fawzi, H. (2018). Using Technology for Formative Assessment to Improve Students' Learning. *Turkish Online Journal of Educational Technology-TOJET*, *17*(2), 182–188.

Endang Sri Maruti, Naniek Kusumawati. (2018). Proses Pengembangan Asesmen Alternatif Berupa Penilaian Poduk pada Mata Kuliah Pmebelajaran Bahasa Jawa di SD. *Jurnal pendidikan dasar perkhasa*. Volume 4 Nomor 2.

Faradisa, B. T. Z. V., Kurniasih, S., & Berlian, L. (2024). Pengembangan instrumen tes 4TMC CBT pada materi sistem pernapasan untuk mengukur berpikir kritis siswa SMP kelas VIII. *Jurnal Pendidikan MIPA.* 14(4), 909–918.

Firmansyah, S., Chandra, E., & Aripin, I. (2019). Pengembangan *e-portfolio* sebagai assessment pembelajaran biologi. *Jurnal Pendidikan Biologi Indonesia, 5*(1), 45–53.

Fitriani, F., Pratikto, H., Rahayu, W. P., & Prabowo, A. E. (2024). Pengembangan Instrumen Assessment Pembelajaran Hots Menggunakan iSpring Suite. *Research and Development Journal of Education*, *10*(2), 695-707.

Furunes, T. (2019). Reflections on systematic reviews: Moving golden standards? *Scandinavian Journal of Hospitality and Tourism*, *19*(3), 227–231.

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education, 57*(4), 2333–2351.

Haddaway, N. R., Woodcock, P., Macura, B., & Collins, A. (2015). Making literature reviews more reliable through the application of lessons from systematic reviews. *Conservation biology*, *29*(6), 1596–1605.

Han, X. (2025). Associations between the effectiveness of blended learning, student engagement, student learning outcomes, and student academic motivation in higher education. *Education and Information Technologies*, *30*(8), 10535-10565.

Hani, M. R., Riyoko, E., & Fakhrudin, A. (2020). Efektivitas Strategi Pembelajaran Menyenangkan Berbasis Quantum Teaching And Learning terhadap Hasil Belajar Siswa Kelas V SD Negeri 35 Palembang. *Jurnal Pendidikan Dan Konseling (JPDK)*, *5*(1), 1117–1128.

Harahap, R. D., Bangun, B., & Siregar, S. U. (2025). *The effectiveness of IMLO Biology media in enhancing students ' learning motivation under the Merdeka Curriculum. 8*(2), 184–191. DOI : 10.30821/biolokus.v8i2.4796

Hasyim, M., Swandi, A., Rahmadhanningsih, S., Taqwin, M., Virisi, S. (2021). Pengembangan Instrumen Pembelajaran Fisika dengan Model PenemuanTerbimbing Berbantuan Simulasi Interaktif dan Dampaknya Terhadap Keterampilan Proses Sains Siswa. *Jurnal Riset dan Kajian Pendidikan Fisika*, 8(1), 15-23.

Ihsan, M. S., Hadisaputra, S., Ramdani, A., & Al Idrus, A. (2020). Pengembangan Instrumen Pemahaman Konsep Berbasis Komputer pada Pembelajaran Kimia. *Jurnal Inovasi Pendidikan dan Sains*, 1(1), 26-29.

Istiyono, E. (2020). *Pengembangan Instrumen Penilaian dan Analisis Hasil Belajar Fisika* (Edisi II). UNY Press.

Istiyono, E., Dwandaru, W. S. B., Setiawan, R., & Megawati, I. (2020). Developing Computerized Adaptive Testing to Measure Physics Higher Order Thinking Skills of Senior High School Students and Its Feasibility of Use. *European Journal of Educational Research*, 9(1), 91-101.

Jahan, N., Naveed, S., Zeshan, M., & Tahir, M. A. (2016). How to conduct a systematic review: a narrative literature review. *Cureus*, *8*(11).

Jimada-Ojuolape, B., & Teh, J. (2020). Impact of the integration of information and communication technology on power system reliability: A review. *IEEE Access*, *8*, 24600-24615.

Kibble, J. D. (2017). Best practices in online assessment: Principles, process, and implementation. *Advances in Physiology Education, 41*(1), 110–119.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.

Lestari, S. M., Sjaifuddin, S., & Resti, V. D. A. (2022). Pengembangan instrumen soal lomba cerdas cermat IPA SMP berbasis ICT (information and communication technology) dengan aplikasi quizizz. *PENDIPA Journal of Science Education, 6*(2), 531–540.

Molnár, G. (2021). Challenges and developments in technology-based assessment: possibilities in science education. *Europhysics News*, *52*(2), 16-19.

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, *9*(3), e3292.

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in higher education*, *46*(5), 756–778.

Haka, N. B., Hamid, A., Nurhidayah, N., Kesumawardhani, A. D., Rudhini, M., & Riski, R. A. (2019). Pengembangan instrumen evaluasi two-tier multiple choice terhadap literasi sains berbantuan personal computer. *Biosfer: Jurnal Tadris Biologi*, *10*(2), 201–214.

Nugraha, E. N. L., Salsabila, S., & Ramadhiani, T. S. (2021, March). Implementing an online quiz application in the EFL classroom. In *Proceedings of the International Conference on Education of Suryakancana*.

Nurdiyanti, N., Sukarmin, S., & Budiharti, R. (2022). Pengembangan Media Pembelajaran Fisika Berbasis Moodle Pada Materi Gelombang Bunyi. *Jurnal Materi dan Pembelajaran Fisika*, 12(1), 22-28.

O'Leary, M., Scully, D., Karakolidis, A., & Pitsia, V. (2018). The state-of-the-art in digital technology-based assessment. *European Journal of Education*, *53*(2), 160–175.

O'Connor, A., & Sargeant, J. (2015). Research synthesis in veterinary science: narrative reviews, systematic reviews and meta-analysis. *The Veterinary Journal*, *206*(3), 261–267.

Pals, F. F., Tolboom, J. L., & Suhre, C. J. (2023). Development of a formative assessment instrument to determine students’ need for corrective actions in physics: Identifying students’ functional level of understanding. Thinking Skills and Creativity, 50, 101387.

Pautasso, M. (2019). The structure and conduct of a narrative literature review. *A guide to the scientific career: Virtues, communication, research, and academic writing*, 299–310.

Polanin, J. R., Pigott, T. D., Espelage, D. L., & Grotpeter, J. K. (2019). Best practice guidelines for abstract screening of ample-evidence systematic reviews and meta-analyses. *Research Synthesis Methods*, *10*(3), 330–342.

Prendes-Espinosa, M. P., Gutiérrez-Porlán, I., & García-Tudela, P. A. (2020). Collaborative work in higher education: Tools and strategies to implement the e-assessment. In *Workgroups eassessment: Planning, implementing and analysing frameworks* (pp. 55-84). Singapore: Springer Singapore.

Rachma, N. M. A. (2015). Pengembangan Tes Elektronik (E-test) Berbasis Komputer Pada Materi Bioteknologi Di SMA Negeri 1 Surabaya (Doctoral dissertation, State University of Surabaya).

Redecker, C. (2017). European framework for the digital competence of educators (DigCompEdu). *Publications Office of the European Union*.

Rostaminezhad, M. A. (2019). Students' perceptions of the strengths and limitations of electronic tests, focusing on instant feedback. *Journal of Information Technology Education. Research*, *18*, 59.

Saleem, T. A. (2011). Mobile learning technology: A new step in e-learning. *Journal of theoretical and applied information technology*, *34*(2), 125–137.

Santoso, A., Kartianom, K., & Kassymova, G. K. (2019). Kualitas Butir Bank Soal Statistika (Studi kasus: Instrumen Ujian Akhir Mata Kuliah Statistika Universitas Terbuka). *Jurnal Riset Pendidikan Matematika*, *6*(2), 165–176.

Soamole, S., Amiroh, D., & Salim, A. (2024). Pengembangan e-formative assessment pada materi suhu dan kalor untuk meningkatkan hasil belajar fisika dengan balikan video tik tok. *Jurnal Pendidikan MIPA*. 8(2), 2–5.

Sari, D. R. U., Wahyuni, S., & Bachtiar, R. W. (2018). Pengembangan Instrumen Tes Multiple Choice High Order Thinking Padapembelajaran Fisika Berbasis E-Learning Di Sma. *Jurnal Pembelajaran Fisika,* 7(1), 100-

Setiyorini, S. R., & Lestari, W. (2023). Pengembangan Instrumen Tes pada Materi Kesetimbangan Kimia Kelas XI Berbasis Android. *Jurnal Teknologi Pendidikan*, 1(2), 15-15.

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, *33*(1), 1–19.

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M. W. (2016). Advances in the science of assessment. *Educational Assessment*, *21*(1), 34–59.

Siregar, T. M., Agustina, N., Siahaan, F. B., Sinaga, L. L., Prabawa, I., & Sagala, R. Z. (2024). Analisis Validitas dan Reliabilitas Butir Soal Pilihan Ganda Dalam Mata Pelajaran Matematika Kelas 12 Semester Genap di SMAN 1 Adiankoting. AR-RUMMAN: *Journal of Education and Learning Evaluation*, 1(2), 725–730.

Solihin, A. (2022). Penerapan Supervisi Akademik Kepala Sekolah dalam Peningkatan Kemampuan Guru Menyusun Penilaian Pembelajaran Ahmad. *Meretas : Jurnal Ilmu Pendidikan*, *9*(2), 77–88.

Soto Rodriguez, E. A., Fernández Vilas, A., & Díaz Redondo, R. P. (2021). Impact of computer-based assessments on the Science ranks of secondary students. *Applied Sciences*, *11*(13), 6169.

Suhaila, A., Sani, R., & Zul, A. (2024). Analisis tingkat miskonsepsi siswa MAN Binjai pada materi suhu dan kalor menggunakan three-tier multiple choice diagnostik test. *Jurnal KARST (Jurnal Pendidikan Fisika dan Terapannya)*, *1*(1), 1-8.

Supardi, S. (2015). *Penilaian Autentik Pembelajaran Efektif, Kognitif, dan Psikomotor: Konsep dan Aplikasi*.

Supriyati, Y., Iriyadi, D., & Falani, I.. (2021). The development of an equating application for computer-based tests in the physics HOTS category. *Journal of Technology and Science Education*, 11(1), 117-128.

Syahidah, N., Kusasi, M., Analita, R. N., & Sholahuddin, A. (2024). Pengembangan E-Instrumen Tes Materi Stoikiometri Menggunakan Blooket Game Untuk Mengukur Kemampuan Berpikir Kritis Pada Siswa Sma. *JCAE* (*Journal of Chemistry And Education*), 7(3), 161–174.]

Tanjung, Y. I., Abubakar, Dewi W., dan Rajo, H. L. (2020). Kajian Pengetahuan Konseptual (Teori & Soal). Bandung: Media Sains Indonesia.

Tanjung, Y. I., Wulandari, T., Festiyed, F., Yerimadesi, Y., & Ahda, Y. (2023). Development analysis of creative thinking test instruments on natural science materials. *Jurnal Pendidikan Fisika*. 12(1), 22-27.

Tanjung, Y. I., Wulandari, D., Bakar, A., & Ramadhani, I. (2021, March). The Development of an Online Physics Test System at SMA CT Foundation Medan. In *Journal of Physics: Conference Series* (Vol. 1819, No. 1, p. 012054). IOP Publishing.

Tanjung, Y. I., Azhar, M., Razak, A., Yohandri, Y., Arsih, F., Wulandari, T., ... & Lubis, R. H. (2023). State of The Art Review: Building Computational Thinking on Science Education. *Jurnal Pendidikan Fisika Indonesia*, *19*(1), 65-75.

Yul Ifda Tanjung, Festiyed, Skunda Diliarosta, Asrizal, Fitri Arsih, and Muhammad Aizri Fadillah, "Developing the Physics Learning Management System (PLMS) to Support Blended Learning Models," *International Journal of Information and Education Technology*, vol. 15, no. 1, pp. 18-29, 2025.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3–30). Springer.

Zhao, F., Yu, S., & Wang, L. (2020). Mobile learning and its effectiveness in higher education. *Computers & Education, 159*, 104013.

Zlatkin-Troitschanskaia, O., Kuhn, C., Brückner, S., & Leighton, J. P. (2019). Evaluating a technology-based assessment (TBA) to measure teachers' action-related and reflective skills. *International Journal of Testing*, *19*(2), 148-171