International Journal of Instruction
e-ISSN: 1308-1470 ● www.e-iji.net

April 2026 ● Vol.19, No.2
p-ISSN: 1694-609X
pp. 703-724

Article submission code: 20250817144614

Received: 17/08/2025
Revision: 27/11/2025

Accepted: 05/12/2025
OnlineFirst: 07/01/2026

Artificial Intelligence in Assessment: A Bibliometric Review of Research Development, Thematic Patterns and Research Clusters

Ashmimi Maisara Asha'ari
Faculty of Educational Sciences and Technology, Universiti Teknologi Malaysia, Malaysia, *ashmimi@graduate.utm.my*

Nor Farahwahidah Abdul Rahman
Corresponding author, Faculty of Educational Sciences and Technology, Universiti Teknologi Malaysia, Malaysia, *nfwahidah@utm.my*

Anis Diyana Halim
Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Malaysia, *anis.diyana@fsmt.upsi.edu.my*

Ketang Wiyono
Faculty of Teacher Training and Education, Universitas Sriwijaya, Indonesia, *ketang_wiyono@fkip.unsri.ac.id*

This study presents a bibliometric analysis of 68 Scopus-indexed publications from 2011 to July 2025 on artificial intelligence (AI) in assessment. The dataset recorded 1,581 total citations, with an h-index of 22 and g-index of 38, indicating steady growth that intensified after 2021 as AI adoption expanded in post-pandemic education. The United States, Germany, and the United Kingdom emerged as the most productive contributors, supported by high-impact institutions such as Michigan State University, Purdue University, and the IPN Leibniz Institute for Science and Mathematics Education. Influential authors, including Haudek and Zhai (2024), Kubsch et al. (2023), and Wulff (2023), have advanced the field through interdisciplinary research on automated feedback, machine learning, and data-driven evaluation. Keyword co-occurrence analysis revealed three dominant thematic clusters: technology-enhanced learning and automated feedback, machine learning with ethical considerations, and AI-supported formative assessment in science and engineering education. While the findings highlight AI's transformative potential for adaptive, feedback-oriented learning, challenges persist in ensuring accessibility, curricular integration, teacher readiness, and ethical governance. Overall, this study provides a comprehensive overview of global trends, leading contributors, and emerging themes that inform future research and policy directions in AI-driven educational assessment.

Keywords: artificial intelligence, education, bibliometric analysis, assessment, research trends

Citation: Asha'ari, A. M., Abdul Rahman, N. F., Halim, A. D., & Wiyono, K.. (2026). Artificial intelligence in assessment: A bibliometric review of research development, thematic patterns and research clusters. *International Journal of Instruction*, *19*(2), 703-724. https://doi.org/10.29333/iji.2026.19238a

INTRODUCTION

The integration of Artificial Intelligence (AI) into teaching, learning, and assessment has gained significant momentum, especially following the post-COVID shift toward online and hybrid instruction. Assessment is now expected to be innovative, adaptable, and supported by data-driven instruction (Almasri, 2024). Despite growing interest in AI in teaching and learning, the demand for AI-supported assessment that enables customized learning has been rising. Much of the literature has concentrated on real-time feedback, adaptive learning paths, and personalized evaluation and engagement (Messer et al., 2024). Consequently, large language models and adaptive learning algorithms are increasingly being utilized to provide personalization (Mok et al., 2025; Yamtinah et al., 2025).

As AI continues to shape pedagogical strategies, its applications in assessment have expanded beyond traditional testing (Kortemeyer, 2023). This development has generated substantial literature on assessment as learning, emphasizing students' epistemic agency and self-regulation. Given this rapid growth and evolving landscape, a bibliometric analysis is essential to systematically map research developments at the intersection of AI and assessment. Such an analysis enables the identification of influential works, leading contributors, thematic trends, and methodological advancements. By examining a comprehensive dataset of 68 publications published between 2011 and 2025, this study presents a structured overview of how AI has been applied in assessment. The insights generated are expected to support future empirical inquiries and guide educators, policymakers, and developers in leveraging AI more effectively within educational assessments (Belland, Walker, Kim, et al., 2017).

LITERATURE REVIEW

Assessment in education has increasingly shifted toward AI-driven approaches, given their potential to provide timely, individualized, and scalable feedback that is often unattainable in traditional settings (Mok et al., 2025; Wei et al., 2025). The integration of AI into assessment practices addresses long-standing challenges such as delayed feedback, limited personalization, and the inefficiency of processing large volumes of student data (Chan et al., 2025; Yamtinah et al., 2025). Several researchers have systematically examined the state of knowledge on AI in education, highlighting its applications in automated grading (Mok et al., 2025), intelligent problem-solving assessment (Wei et al., 2025), fine-tuned feedback generation in STEM contexts (Yamtinah et al., 2025), and automated item creation to expand assessment coverage (Chan et al., 2025). Collectively, these studies emphasize that AI-driven systems not only enhance efficiency in scoring but also provide deeper diagnostic insights into students' learning processes, thereby strengthening formative assessment practices and enabling more targeted instructional interventions.

Previous studies investigations have underscored the role of AI in facilitating real-time feedback and supporting students' independent learning processes within science laboratory environments. Coban et al., (2025), for example, demonstrated that integrating ChatGPT-based formative feedback within an augmented reality (AR) experiment significantly improved students' conceptual understanding ($t = 2.721$, $p =$

0.014, d = 1.189), with higher D-scores observed for Group A (with feedback) compared to Group B (without feedback). Building on this, recent studies show diverse applications of AI in education, from improving real-time feedback in laboratory settings (Coban et al., 2025) to enhancing student engagement and cognitive performance in virtual and augmented environments (Asmussen et al., 2025; Bewersdorff et al., 2025). Within physics and chemistry education, AI has been employed to assess open-ended responses with inter-rater reliability comparable to human evaluators, as GPT-4 achieved intraclass correlation coefficients above 0.6 in seven of ten scoring groups (Xu et al., 2025), to generate assessment items aligned with established cognitive taxonomies (Chan et al., 2025; Habiballa et al., 2025), and to personalize problem-solving support through intelligent tutoring systems (Asmussen et al., 2025). Beyond improvements in assessment accuracy and efficiency, research has also highlighted AI's capacity to scaffold learning, foster self-regulation, and promote student reflection (Chun et al., 2025; Terrazas-Arellanes et al., 2025). Nevertheless, persistent ethical considerations, including transparency, accountability, and the necessity of balanced human oversight—remain critical in ensuring that AI serves pedagogical purposes rather than merely technical objectives (Khodadad, 2025; Tang & Cooper, 2025). Taken together, the literature suggests that AI is progressively reshaping the design, delivery, and interpretation of educational assessments. As technology continues to advance, further scholarly inquiry is required to address its limitations and to refine its integration into routine teaching practices to optimize learning outcomes.

Previous Studies on Bibliometric Analysis

Previous bibliometric studies on AI in education have provided valuable insights into the development of the field. For instance, Delen et al. (2024) examined 4,673 publications indexed in the Web of Science between 1975 and 2023, mapping highly cited topics, prolific authors and institutions, as well as leading publication sources in educational AI research. Similarly, Rodríguez Flores et al. (2025) analyzed 101 publications indexed in both Web of Science and Scopus up to 2023, identifying influential works, core journals, and contributing countries, thereby offering an overview of the thematic development of AI within broader scientific research. Reyes Flores and Mejía Rivera (2024) concentrated on AI applications in qualitative research, drawing on 111 Scopus-indexed publications (2006–2024) to highlight core keywords, publication classifications, and thematic orientations in non-quantitative educational contexts. At a larger interdisciplinary scale, Carchiolo and Malgeri (2024) mapped more than two million Scopus-indexed documents (1995–2023) to trace the distribution and evolution of AI research across multiple scientific domains, with attention to keyword trends, subject classifications, geographical contributions, and citation patterns.

The intellectual progression and knowledge framework derived from previous bibliometrics reviews (Delen et al., 2024; Rodríguez Flores et al., 2025; Reyes Flores & Mejía Rivera, 2024; Carchiolo & Malgeri, 2024) collectively provide only a macro-level view of publication trends and thematic clusters in AI research. However, from these studies, the dominant thematic clusters tend to emphasize technological affordances rather than the learning mechanisms that foster learner autonomy without directly addressing the intersection of AI and assessment. Accordingly, the purpose of

this study is to analyze the landscape of research on AI in assessment through both bibliometric and critical perspectives. The study aims to offer insights into the evolving focus areas and conceptual foundations of AI-related scholarship in assessment. Specifically, it seeks to map the current state of AI-driven assessment, highlight emerging developments, and identify landmark works and publication sources that have shaped the field. Employing quantitative topic modeling and network analysis techniques, this study seeks to:

RQ 1: What is the current landscape of research on the use of AI in assessment?

RQ 2: What are the emerging trends and developments in AI-driven assessment?

RQ 3: Which journals and publication sources have played a central role in advancing research on AI in assessment?

RQ 4: What landmark papers have shaped the discourse and direction of AI in assessment?

RQ 5: Which key players authors, institutions, and countries are driving advancements in AI in assessment?

RQ 6: What core research themes and areas of focus define the progression and evolution of AI in assessment?

METHOD

This study employed data extracted from the Scopus database as of July 2025. The dataset encompassed a range of variables, including document types and sources, languages, subject classifications, publication trends, average authorship per paper, institutional affiliations, country-level publication output, and frequently occurring keywords.

Search Strategy & Selection Procedure

This bibliometric review adopted the modified PRISMA guidelines (Moher et al., 2009) to ensure a systematic and transparent approach. The Scopus database was searched on 4 August 2025 using a Boolean combination of two keyword groups:

Search string

TITLE-ABS-KEY(("artificial intelligence" OR "AI" OR "machine learning" OR "natural language processing" OR "large language model" OR "automated assessment" OR "intelligent tutoring system") AND ("assessment" OR "student assessment" OR "formative assessment" OR "automated grading" OR "computer-based assessment" OR "learning analytics"))

The Scopus search, limited to articles, reviews, and conference papers published in English between 2011 and 2025, initially yielded 495 documents. After removing 133 non-educational records (e.g., medical or engineering automation), 362 remained for screening. Of these, 213 were excluded for not being research articles or unrelated to the topic, leaving 149 reports for further review. From these, 79 were inaccessible and 2 were removed for not being in English or in their final publication stage, resulting in a

final dataset of 68 relevant studies on artificial intelligence in assessment. Metadata containing citation information, abstracts, author affiliations, and keywords were exported in CSV format, cleaned using OpenRefine, and analyzed in VOSviewer and biblioMagika® for bibliometric mapping and co-occurrence analysis.

Data Cleaning

This study employed OpenRefine and biblioMagika® (Ahmi, 2024) to resolve inconsistencies in bibliographic data, particularly those related to author names, institutional affiliations, and keywords. Metadata were exported from the Scopus database in CSV format, after which key columns such as author, keyword, and affiliation were systematically reviewed and refined using clustering and filtering functions. The records were then categorized by publication year, source title, author, institution, and country. The harmonization process also included identifying and completing missing data, thereby improving the overall reliability of the dataset. Following the removal of duplicates and the screening of abstracts, the final dataset consisted of 68 publications.

Data Analysis

This study's data analysis was designed to align with the research questions by examining the existing body of literature on artificial intelligence in assessment.

RQ 1: What is the current landscape of research on the use of AI in assessment?

To examine the current research landscape, this study analyzed bibliometric indicators including total publications, cited items, citations, h-index, g-index, m-index, and h-core citations, alongside publication year, document type, authorship patterns, and collaboration networks.

RQ 2: What are the emerging trends and developments in AI-driven assessment?

To trace the evolution of the field, publication and citation trends from 2011 to 2025 were examined chronologically to identify growth phases, landmark years, and shifts in research activity over time.

RQ 3: Which journals and publication sources have played a central role in advancing research on AI in assessment?

To identify leading journals, analysis was conducted based on total publications, citation performance, and quartile ranking, with subject classifications examined to assess disciplinary scope and influence.

RQ 4: What landmark papers have shaped the discourse and direction of AI in assessment?

To determine landmark papers, publications with high citation counts and strong citation-per-year performance were reviewed, with particular attention to their thematic relevance and methodological contributions.

RQ 5: Which key players authors, institutions, and countries are driving advancements in AI in assessment?

To profile key contributors, prolific authors, institutions, and countries were analyzed through authorship and institutional data, and collaboration networks were subsequently mapped at both national and international levels.

RQ 6: What core research themes and areas of focus define the progression and evolution of AI in assessment?

To uncover core research themes, co-occurrence keyword analysis, thematic mapping, and factorial analysis were applied to identify dominant themes such as technology-enhanced learning, machine learning and bias, generative AI in feedback, and formative assessment in STEM, as well as to explore their interconnections and evolution over time.

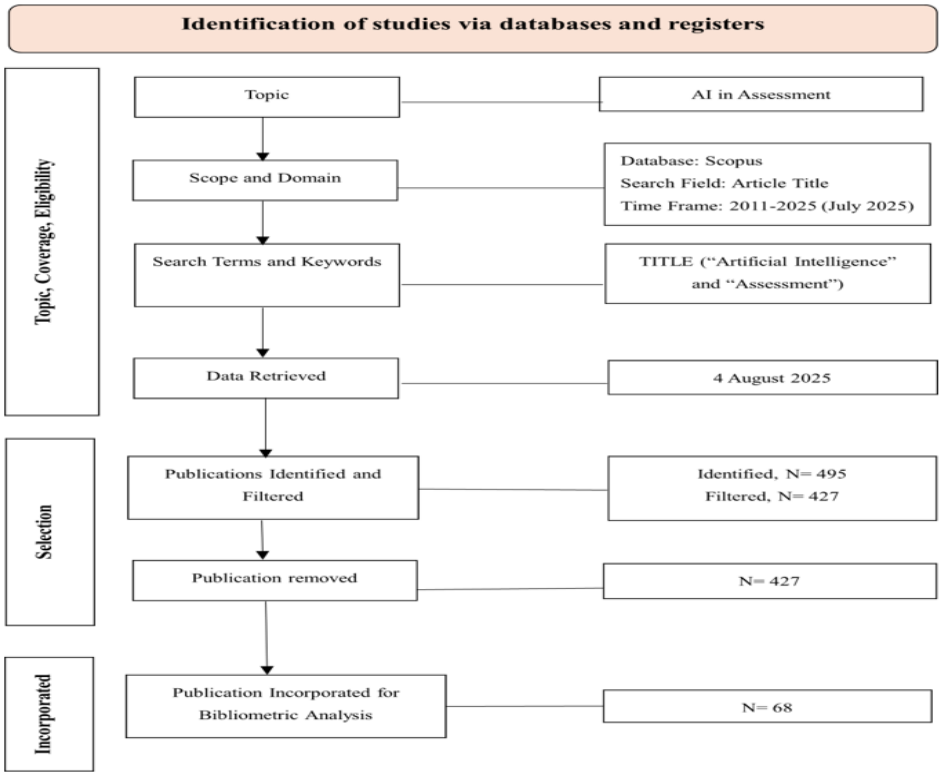

Figure 1
Flow diagram of the search strategy

Tools

This study employed a range of tools for dataset organization and analysis. Microsoft Excel was initially used to organize and preprocess the dataset, while biblioMagika® facilitated the harmonization, cleaning, and standardization of author affiliations, country names, and keywords. OpenRefine was applied to resolve inconsistencies in author keywords. Once the dataset was cleaned and standardized, VOSviewer was utilized to generate co-occurrence network visualizations, keyword clusters, and thematic maps. Mendeley served as the reference management software, supporting the systematic organization and citation of sources. The integration of these tools ensured a rigorous and comprehensive examination of the research landscape in this emerging field.

FINDINGS

This section describes the findings of the study based on the research questions and the analyses we carried out, as described in the previous sections.

What is the current landscape of research on the use of AI in assessment?

The landscape analysis encompassed trends in publication output, citation performance, author collaboration, and core bibliometric indices to evaluate the progression and scholarly impact of research in this domain. Drawing on data from 2011 to 2025, Table 1 presents a summary of the key findings. Over this period, 68 documents were published by 272 contributing authors. Of these, 62 publications received citations, indicating substantial academic visibility. In total, the dataset accumulated 1,581 citations, reflecting a steady increase in scholarly attention to this field.

The dataset reveals an average citation count of 23.25 per paper, with a slightly higher mean of 25.50 among cited publications, underscoring the notable influence of referenced works within the field. On average, each author received 5.81 citations, and publications involved approximately four co-authors per paper, reflecting active and sustained patterns of research collaboration. Notably, the h-core accounted for 1,516 citations, suggesting that a concentrated body of highly cited publications substantially contributes to the visibility and advancement of this domain. Index-based metrics provide further insight: an h-index of 22 and a g-index of 38 demonstrate the consistent presence of impactful research, while an m-index of 1.467 indicates sustained academic productivity across a 15-year period. The annual citation rate of 112.93 further illustrates the accelerating scholarly momentum surrounding AI in educational assessment. Collectively, these metrics depict a maturing and increasingly influential research landscape, establishing a robust foundation for future innovation, theoretical refinement, and empirical exploration.

Table 1
Citation metric

Main Information	Data
Publication Years	2011 - 2025
Total Publications	68
Citable Year	15
Number of Contributing Authors	272
Number of Cited Papers	62
Total Citations	1,581
Citation per Paper	23.25
Citation per Cited Paper	25.50
Citation per Year	112.93
Citation per Author	5.81
Author per Paper	4.00
Citation sum within h-Core	1,516
h-index	22
g-index	38
m-index	1.467

The results, summarized in Table 2 and illustrated in Figure 2, reveal a distinct upward trajectory in scholarly activity over the past decade. While publication output in the early years remained limited and relatively stable, a marked surge occurred from 2022 onward, culminating in a peak in 2024 with 17 publications and a total of 306 citations. This sharp growth reflects an intensification of academic interest and indicates the consolidation of AI in assessment as a prominent strand within science education research.

In the formative years, research productivity was modest, with only two publications recorded in 2011, 2017, and 2018. Despite the limited volume, several of these early contributions exerted a disproportionate scholarly influence. Notably, in 2017, just two publications collectively attracted 276 citations, averaging 138 citations per paper, while the two studies published in 2011 received a combined total of 120 citations, averaging 60 citations per document. These findings suggest that foundational works produced in the initial phase of the field served as intellectual anchors, shaping subsequent inquiry and setting the trajectory for later expansion.

From 2021 onwards, the research landscape demonstrated a discernible shift, marked by a steady increase in both publication output and citation activity. In 2022, the field recorded seven publications accompanied by 212 citations, followed by a notable expansion in 2023 with 12 publications and 270 citations. This upward trajectory continued into 2025, during which the number of contributing authors reached 75, reflecting the field's increasingly collaborative and interdisciplinary orientation. The growing pool of contributors highlights not only expanding global interest but also the diversification of scholarly perspectives engaging with AI-driven assessment in science education.

Impact indicators further underscore the consolidation and rising relevance of this research domain. As shown in Table 3, the h-index attained a value of 22, while the g-index rose to 38, together evidencing a substantial corpus of influential and frequently

cited scholarship. The m-index, which reached 3.500 in 2024, points to a surge of high-impact publications in recent years, signaling the maturation of the field. Although citation averages fluctuated across individual years, such variability is characteristic of a discipline progressing from an exploratory phase toward more systematic integration and application. Taken together, these findings affirm the emergence of AI in assessment as an increasingly significant research trajectory, with clear implications for shaping pedagogical practices and advancing the broader science education discourse.

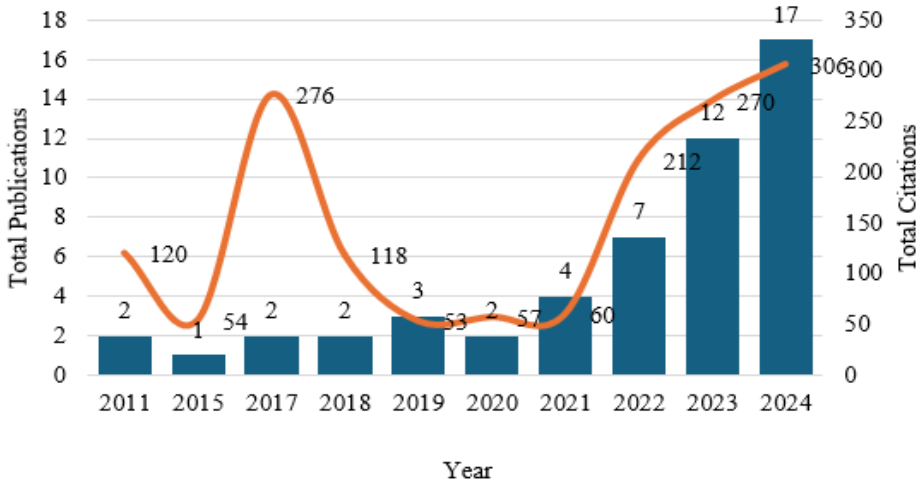

Figure 2
Total publications and citations by year (Excluding the year 2025 as data is only available up to July 2025)

Table 2
Publication by year

Year	TP	NCA	NCP	TC	C/P	C/CP	h	g	m
2011	2	13	2	120	60.00	60.00	2	2	0.133
2015	1	5	1	54	54.00	54.00	1	1	0.091
2017	2	7	2	276	138.00	138.00	2	2	0.222
2018	2	5	2	118	59.00	59.00	2	2	0.250
2019	3	9	3	53	17.67	17.67	2	3	0.286
2020	2	8	2	57	28.50	28.50	2	2	0.333
2021	4	18	4	60	15.00	15.00	4	4	0.800
2022	7	33	7	212	30.29	30.29	7	7	1.750
2023	12	46	12	270	22.50	22.50	8	12	2.667
2024	17	53	15	306	18.00	20.40	7	17	3.500
Total	68	272	62	1581	23.25	25.50	22	38	1.467

Notes: TP = total number of publications; NCA = number of contributing authors; NCP = number of cited publications; TC = total citations; C/P = average citations per publication; C/CP = average citations per cited publication; h = h-index; g = g-index; m = m-index.
* Publication data for the year 2025 is only up until July 2025

Which journals and publication sources have played a central role in advancing research on AI in assessment?

Table 3 presents the most active source titles contributing to the field of artificial intelligence in the assessment focusing on outlets that have published at least two documents. Among them, Frontiers in Education emerge as the leading contributor with a total of six publications and 85 citations, resulting in an average of 14.17 citations per paper. It also shows strong index performance with an h-index of 5, g-index of 6, and an m-index of 1.250, highlighting its consistent influence in this area of research. Another key journal, Computers and Education: Artificial Intelligence, published five documents with a total of 68 citations, achieving a C/P of 13.60 and matching h- and g-index scores of 3 and 5, respectively. This reflects a steady citation performance and active engagement with AI topics in education. The Journal of Research in Science Teaching also demonstrated substantial scholarly impact, with 139 citations across four papers and an impressive citation rate of 34.75 per paper. Its h- and g-index scores of 4 support its position as a high-impact journal in the domain.

High citation averages were particularly evident in IEEE Transactions on Learning Technologies and Physical Review Physics Education Research, each contributing three publications that collectively garnered 151 citations, yielding a citations-per-paper (C/P) ratio of 50.33. This pattern highlights their significant scholarly influence despite a comparatively limited number of outputs. Similarly, ACM Transactions on Computing Education accumulated 79 citations across two papers, corresponding to a C/P of 39.50 and an m-index of 1.000, further illustrating sustained impact relative to its scale of contribution. Other actively contributing journals include Education Sciences and CBE Life Sciences Education, both of which, although generating modest citation counts, provide foundational work that supports the growth of the field. Notably, Review of Educational Research emerges as an outlier, with only two documents achieving an exceptional 276 citations and an average C/P of 138.00, positioning it as the most influential journal on a per-paper basis. In addition, multidisciplinary outlets such as Sustainability (Switzerland) and SAGE Open demonstrate balanced performance across the h- index, g- index, and m-index measures, suggesting steady and diversified scholarly engagement.

Table 3
Most active source titles that published two (2) or more documents

Source Title	TP	NCA	NCP	TC	C/P	C/CP	h	g	m
Frontiers in Education	6	28	5	85	14.17	17.00	5	6	1.250
Computers and Education: Artificial Intelligence	5	22	5	68	13.60	13.60	3	5	0.600
Journal of Research in Science Teaching	4	18	4	139	34.75	34.75	4	4	1.000
Education Sciences	4	13	3	9	2.25	3.00	2	3	1.000
IEEE Transactions on Learning Technologies	3	10	3	151	50.33	50.33	3	3	0.375
Physical Review Physics Education Research	3	10	3	151	50.33	50.33	3	3	0.600
CBE Life Sciences Education	2	13	2	120	60.00	60.00	2	2	0.133
ACM Transactions on Computing Education	2	5	2	79	39.50	39.50	2	2	1.000
Applied Sciences (Switzerland)	2	13	2	4	2.00	2.00	1	2	1.000
Journal of Science Education and Technology	2	10	2	41	20.50	20.50	2	2	0.500
Eurasia Journal of Mathematics, Science and Technology Education	2	8	2	16	8.00	8.00	2	2	0.250
Review of Educational Research	2	7	2	276	138.00	138.00	2	2	0.222
SAGE Open	2	6	2	24	12.00	12.00	2	2	0.286
Physics Education	2	8	0	0	0.00	0.00	0	0	0.000
Sustainability (Switzerland)	2	7	2	66	33.00	33.00	2	2	0.286

Note: TP=total number of publications; NCA=number of contributing authors; NCP=number of cited publications; TC=total citations; C/P=average citations per publication; C/CP=average citations per cited publication; h = h-index; g = g-index; m = m-index

What landmark papers have shaped the discourse and direction of AI in assessment?

Table 4 presents the five most frequently cited documents in the field of AI in science assessment, highlighting studies that have significantly shaped scholarly discourse. Leading this list is the meta-analysis by Belland, Walker, and Kim (2017), which synthesizes findings from empirical research on computer-based scaffolding in STEM education. This work has received a total of 237 citations, averaging 26.33 citations per year, underscoring its foundational influence in advancing instructional support systems with AI technologies. Closely following is the study by Kortemeyer (2023), which examines the feasibility of artificial intelligence agents in managing introductory physics coursework. This article has accumulated 107 citations, with an impressive average of 35.67 citations per year, reflecting its relevance in evaluating AI's cognitive and pedagogical capabilities in structured science instruction.

Benotti et al. (2018) secured the third position with their investigation of automatic formative assessment tools for computer science education, receiving 104 citations with an average of 13.00 citations per year. The study makes a significant contribution to the understanding of AI-driven adaptive assessment systems, particularly in relation to automation and feedback personalization. Almasri (2024) provided a systematic review of empirical research on AI in science teaching and learning, which has accumulated 79 citations. Notably, this work demonstrates the highest citation-per-year average among

the five (39.50), indicating strong early recognition and scholarly relevance. The review critically examines prevailing trends and research gaps, offering strategic directions for future studies. Completing the list is the work by Messer et al. (2024), which evaluates automated grading and feedback tools in programming education. With 75 citations and an average of 37.50 citations per year, this article exemplifies the accelerating interest in AI-enhanced assessment frameworks across STEM domains. Collectively, these studies illustrate the intellectual impact and expanding integration of AI in science education assessment, establishing pivotal reference points for ongoing and future research.

Table 4
Top five (5) highly cited articles

No.	Author(s)	Title	TC	C/Y
1	Belland, Walker, Kim, et al. (2017)	Synthesizing Results from Empirical Research on Computer-Based Scaffolding in STEM Education: A Meta-Analysis	237	26.33
2	Kortemeyer (2023)	Could an artificial-intelligence agent pass an introductory physics course	107	35.67
3	Benotti et al. (2018)	A Tool for Introducing Computer Science with Automatic Formative Assessment	104	13.00
4	Almasri (2024)	Exploring the Impact of Artificial Intelligence in Teaching and Learning of Science: A Systematic Review of Empirical Research	79	39.50
5	Messer et al. (2024)	Automated Grading and Feedback Tools for Programming Education: A Systematic Review	75	37.50

Which key players, authors, institutions, and countries are driving advancements in AI in assessment?

Table 5 highlights the most prolific authors in the domain of artificial intelligence in science education assessment, each of whom has published more than four documents. The list is largely composed of scholars from North America, Europe, and Asia. Leading the list is Zhai (2024) from the University of Georgia, United States, who has authored eight papers and received 205 citations, yielding an average of 25.63 citations per publication. Zhai's (2024) scholarly impact is further reflected in an h-index of 6, a g-index of 8, and an m-index of 1.500. Another prominent contributor is Kevin C. Haudek (2024) from Michigan State University, United States, with eight publications and 144 citations, averaging 18.00 citations per paper. He shares similar h- and g-index values with Zhai (2024) but has a slightly lower m-index of 0.400. Other U.S.-based researchers, including Ross H. Nehm (2023) of Stony Brook University, Leonora Kaldaras (2022) of the University of Colorado, and James Lehman (2018) of Purdue University, also appear prominently, each contributing four papers with consistent citation metrics and h-indices of 4.

Germany is represented by Wulff (2023) of the Heidelberg University of Education, who authored five papers with a total of 89 citations, as well as by Sascha Bernholt and Michael Striewe (2022), both of whom contributed four papers with 54 citations each. Marcus Kubsch of the IPN – Leibniz Institute for Science and Mathematics Education stands out with 108 citations from four publications, averaging 27 citations per paper and demonstrating a strong research influence. Asia's contribution is led by Yizhou

Qian (2018) from Jiangnan University, China, with four papers and 72 citations. Meanwhile, Kobi Gal (2024) of the University of Edinburgh, United Kingdom, represents Europe with four publications and 26 citations. Collectively, these authors illustrate a diverse and collaborative global research effort on the use of AI in science education assessment, emphasizing both the interdisciplinary nature and international scope of this expanding field.

Table 5
Most productive authors that published more than four (4) documents

Full Name	Current Affiliation	Country	TP	NCP	TC	C/P	C/CP	h	g	m
Zhai, Xiaoming	University of Georgia	United States, United States	8	8	205	25.63	25.63	6	8	1.500
Haudek, Kevin C.	Michigan State University	United States, United States	8	8	144	18.00	18.00	6	8	0.400
Wulff, Peter	Heidelberg University of Education	Germany, Germany	5	5	89	17.80	17.80	5	5	1.250
Nehm, Ross H.	Stony Brook University	United States, United States	4	4	138	34.50	34.50	4	4	0.267
Qian, Yizhou	Jiangnan University	China, China	4	4	72	18.00	18.00	4	4	0.500
Bernholt, Sascha	Leibniz Institute for Science and Mathematics Education	Germany, Germany	4	4	54	13.50	13.50	2	4	0.500
Striewe, Michael	University of Duisburg-Essen	Germany, Germany	4	4	54	13.50	13.50	4	4	1.000
Kaldaras, Leonora	University of Colorado	United States, United States	4	4	56	14.00	14.00	4	4	1.000
Lehman, James	Purdue University	United States, United States	4	4	72	18.00	18.00	4	4	0.500
Gal, Kobi	University of Edinburgh	United Kingdom, United Kingdom	4	4	26	6.50	6.50	3	4	0.600
Kubsch, Marcus	IPN – Leibniz Institute for Science and Mathematics Education	Germany, Germany	4	4	108	27.00	27.00	4	4	1.000

Notes: TP = total number of publications; NCP = number of cited publications; TC = total citations; C/P = average citations per publication; C/CP = average citations per cited publication; h = h-index; g = g-index; m = m-index

Table 6 presents the leading institutions engaged in research on artificial intelligence in assessment, focusing on those with at least two publications. Michigan State University and Purdue University, both located in the United States, top the list with five and four publications, respectively. Michigan State University has recorded 170 total citations, averaging 34.00 citations per publication. Its h-index of 5 and g-index of 8 underscore its strong scholarly influence in this field. Purdue University follows with 130 citations and an average of 32.50 citations per publication, supported by an h-index of 4 and a g-index of 6, reflecting a consistent academic contribution.

The IPN – Leibniz Institute for Science and Mathematics Education in Germany also demonstrates notable performance, contributing three publications and 75 citations, yielding an average of 25.00 citations per publication with an h-index of 3 and a g-index of 4. In South Korea, Kyungbok University produced three publications with nine citations, while the Technical University of Munich in Germany achieved 24 citations from two papers, corresponding to an average of 12.00 citations per publication. Both institutions maintain an h-index of 2 and a g-index of 3, indicating steady visibility in this domain.

Other contributing institutions include De La Salle University in the Philippines, the University of Gävle in Sweden, Ludwig-Maximilians-Universität München (LMU) in Germany, Raffles Institution in Singapore, and DePaul University in the United States. Among them, the University of Gävle demonstrated comparatively higher citation strength, with 28 total citations averaging 14.00 citations per publication, along with an h-index of 2 and g-index of 3. Collectively, these institutions exhibit active participation in advancing the scholarly conversation on artificial intelligence in assessment, highlighting the global and interdisciplinary reach of this emerging research area.

Table 6
Most productive institutions with a minimum of two (2) publications

Institution Name	Country	TP	NCP	TC	C/P	C/CP	h	g	m
Michigan State University, United States	United States	5	5	170	34	34	5	8	0.33
Purdue University, United States	United States	4	4	130	32.5	32.5	4	6	0.27
IPN – Leibniz Institute for Science and Mathematics Education, Germany	Germany	3	3	75	25	25	3	4	0.6
Kyungbok University, South Korea	South Korea	3	3	9	3	3	2	3	0.4
Technical University of Munich, Germany	Germany	2	2	24	12	12	2	3	0.4
De La Salle University, Philippines	Philippines	2	2	21	10.5	10.5	2	3	0.4
University of Gävle, Sweden	Swedec	2	2	28	14	14	2	3	0.4
Ludwig-Maximilians-Universität München (LMU), Germany	Germany	2	2	2	1	1	1	1	0.2
Raffles Institution, Singapore	Singapore	2	2	4	2	2	1	1	0.2
DePaul University, United States	United States	2	2	16	8	8	2	2	0.2

Notes: TP = total number of publications; NCP = number of cited publications; TC = total citations; C/P = average citations per publication; C/CP = average citations per cited publication; h = h-index; g = g-index; m = m-index

Based on the data from Figure 3 and Table 7, global contributions to research on artificial intelligence in assessment are led by countries across North America, Europe, and Asia. The United States stands out as the leading contributor, producing 26 publications and accumulating 900 citations, with an average of 34.62 citations per paper. Its h-index (14) and g-index (26) demonstrate sustained scholarly leadership and broad influence within this domain.

Germany follows as the second-largest contributor with 12 publications and 201 citations, averaging 16.75 citations per paper. Its h-index (8), g-index (12), and m-index

(1.6) indicate consistent academic productivity and a growing research footprint in this field.

The United Kingdom, China, and Spain each record four publications. The United Kingdom achieved 157 citations (39.25 citations per paper) and holds an h-index of 3 and g-index of 4, reflecting visible engagement within the global AI-in-assessment landscape. China contributes 98 citations (24.50 citations per paper), while Spain records 88 citations (22.00 citations per paper). Both countries maintain moderate h- and g-indices ranging from 2 to 4, indicating expanding but still developing research influence.

In Asia, South Korea contributes three publications with seven citations, while in Europe, Sweden also produces three publications and 28 citations, averaging 9.33 citations per paper. Both countries exhibit emerging participation supported by h-indices of 2, suggesting active but early-stage engagement in AI-based assessment research. Overall, these findings reveal that research in AI-driven assessment is dominated by Western and East Asian nations, with the United States and Germany at the forefront. Increasing involvement from the United Kingdom, China, Spain, South Korea, and Sweden further highlights the growing international collaboration and expanding global attention devoted to this emerging research area.

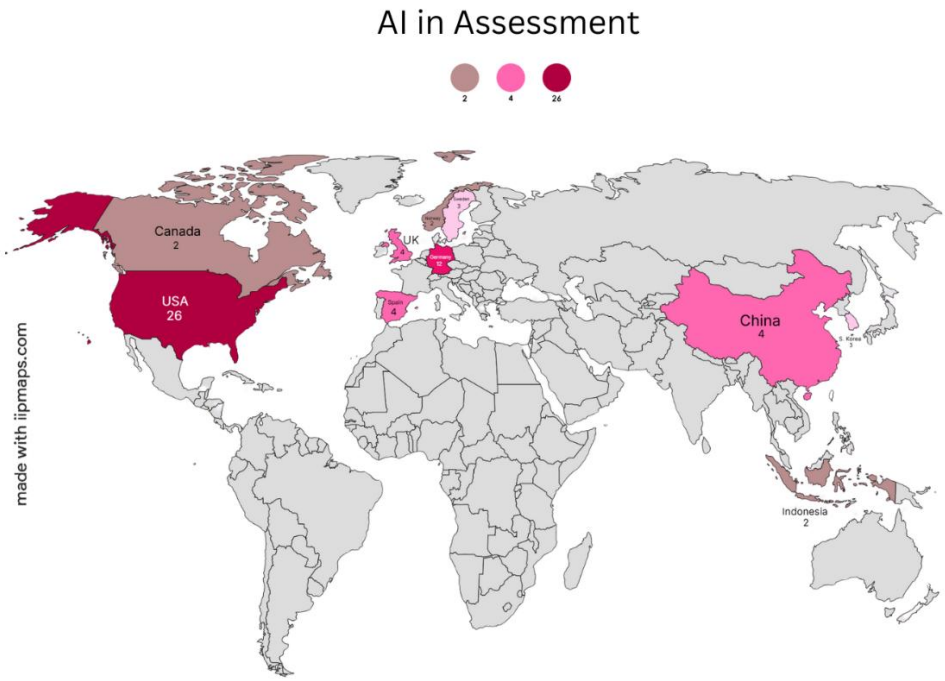

Figure 3
Visualisation of global distribution of AI in Assessment

Table 7
Countries that contributed with at least three (3) or more publications

Country	TP	NCP	TC	C/P	C/CP	h	g	m
United States	26	23	900	34.62	39.13	14	26	0.9
Germany	12	12	201	16.75	16.75	8	12	1.6
United Kingdom	4	3	157	39.25	52.33	3	4	1.5
China	4	3	98	24.50	32.67	2	4	0.3
Spain	4	4	88	22.00	22.00	3	4	1.5
South Korea	3	3	7	2.33	2.33	2	2	1
Sweden	3	2	28	9.33	14.00	2	3	0.4

Notes: TP = total number of publications; NCP = number of cited publications; TC = total citations; C/P = average citations per publication; C/CP = average citations per cited publication; h = h-index; g = g-index; m = m-index

What core research themes and areas of focus define the progression and evolution of AI in assessment?

The co-occurrence network illustrated in Figure 4 provides an in-depth overview of prominent research themes and their interrelations within the field of artificial intelligence in assessment. The maps demonstrate the interconnectedness of education, computer science, and technological innovation. Three clusters are highlighted: the Red Cluster, which focuses on technology-enhanced learning and automated feedback, with related terms such as computer science education, automation, higher education, and automated grading; the Yellow Cluster, which centers on machine learning and algorithmic transparency, including keywords such as machine learning, learning algorithms, bias, argumentation, and assessment; and the Blue Cluster, which emphasizes AI applications in science and engineering education, with associated terms such as science education, engineering students, learning, and education. The network was constructed based on keywords with a minimum of three occurrences, and a summary of findings is presented in Table 8. The most frequent terms—science education, assessment, and students—emerge as dominant nodes, underscoring their foundational role in advancing AI-driven assessment research and establishing distinct areas of focus.

One significant cluster is centered on technology-enhanced learning (Red) and includes keywords such as *automated feedback, automation,* and *computer science education*. This cluster highlights the integration of AI technologies in improving assessment methods, particularly through personalized and automated systems. It signals a shift away from traditional assessment practices toward more adaptive and efficient digital approaches, especially within higher education contexts (Asmussen et al., 2025; Khodadad, 2025; Messer et al., 2024). Another thematic cluster revolves around *machine learning, learning algorithms, bias,* and *argumentation*, reflecting research emphasis on algorithmic transparency, ethical considerations, and fairness in AI-driven assessment systems (Chan et al., 2025; Haudek and Zhai, 2024; Zhai et al., 2022). Closely related to this is a cluster focused on *generative artificial intelligence, ChatGPT,* and *large language models (LLM)*, pointing to the growing scholarly interest in employing generative tools to support tasks such as feedback generation and

automated response evaluation (Bewersdorff et al., 2025; Coban et al., 2025; Mok et al., 2025).

The Blue Cluster features keywords such as AI, engineering students, and formative assessment, indicating a focus on employing AI to design formative assessments that guide learning processes rather than merely evaluate outcomes. This reflects a broader pedagogical shift toward feedback-driven learning environments (Almasri, 2024a; Khodadad, 2025; Xu et al., 2025).

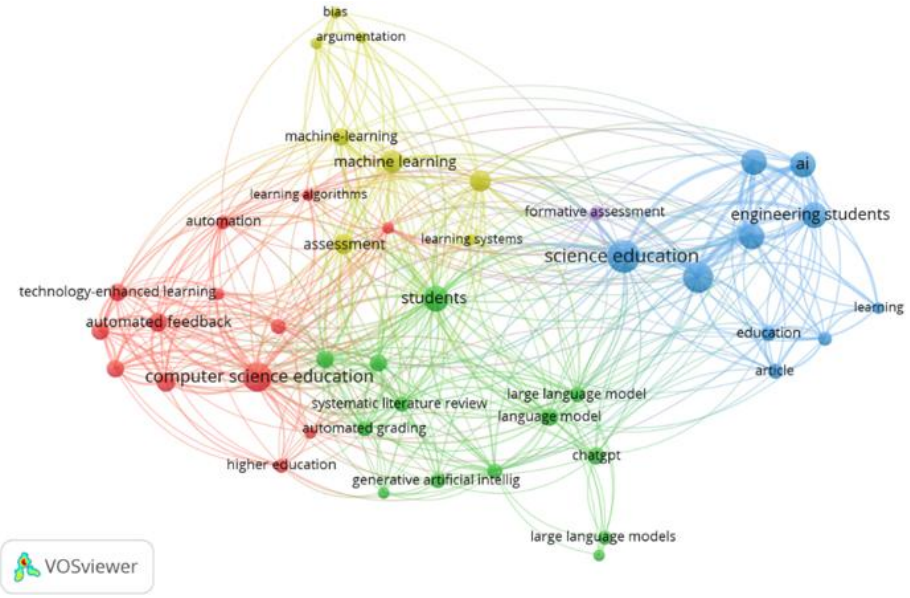

Figure 4
Co-occurrence network of the author's keywords with at least 3 Occurrences

Table 8
Top 5 Co-occurrence network of the author's keywords

Research topics	Occurrences	Total link strength
Science education	22	132
Artificial intelligence	18	105
Computer science education	17	104
Science learning	14	89
Students	13	109

DISCUSSION

Drawing on the synthesis of 68 publications between 2011 and 2025, the analysis reveals a steady upward trajectory in research on AI in assessment, with a marked rise in publication output peaking in 2024 (17 papers, 306 citations) and total citations reaching 1,581. This growth parallels the post-2021 technological advancements and the increasing adoption of AI tools in educational practice, signaling the transition of AI in assessment from emerging innovation to established research focus. Indicators such as a

cumulative citation count of 1,581, an h-index of 22, and a g-index of 38 reflect substantial scholarly impact.

The reviewed literature highlights a shift toward adaptive, feedback-oriented assessment systems that leverage AI to personalize learning support. Recent studies in the dataset, particularly those published since 2023, have examined generative AI, large language models (LLMs), and automated feedback tools as mechanisms for delivering timely, individualized responses to students. This development marks a move away from static, summative approaches toward formative, responsive models capable of adapting to learner needs in real time. For example, Messer et al. (2024) documented AI-powered feedback loops in several high-impact journals specializing in science education, educational technology, and STEM pedagogy. The thematic patterns derived from the co-occurrence network of authors' keywords reveal science education, artificial intelligence, and computer science education as core foci. Their sustained prominence has helped establish benchmarks for rigor, replicability, and pedagogical relevance in AI-related assessment research.

Compared with earlier bibliometric reviews on AI in education by Delen et al. (2024), Rodríguez Flores et al. (2025), and Carchiolo and Malgeri (2024), which primarily examined technological affordances and general pedagogical applications, this study extends the scope to assessment-specific mechanisms such as adaptive feedback, automated grading, and learning analytics. The concentration of studies in journals like Frontiers in Education (6 publications, 85 citations) and Computers and Education: Artificial Intelligence (5 publications, 68 citations) indicates that AI-based assessment has become a central theme within mainstream educational technology research. Landmark works such as Belland et al. (2017) on scaffolding in STEM, Benotti et al. (2018) on automated formative assessment, and Messer et al. (2024) on feedback automation represent key intellectual anchors shaping this domain.

The methodological and conceptual approaches in AI include automation of formative feedback (Haudek et al., 2011), application of machine learning (Zhai et al., 2022), and integration of AI into science curricula (Almasri, 2024) have been referenced across multiple reviewsThe dataset confirms that the United States (26 publications, 900 citations), Germany (12 publications, 201 citations), and United Kingdom (4 publications, 157 citations) dominate the field in both research output and citation impact. Michigan State University, Purdue University, and the IPN Leibniz Institute for Science and Mathematics Education emerge as central institutional hubs. Key contributors such as Haudek and Zhai (2024), Kubsch et al. (2023), and Wulff (2023) have advanced AI-supported assessment through studies on automated feedback systems, machine learning models, and data-driven frameworks that enhance accuracy, personalization, and scalability in educational evaluation.

The prominence of clusters on technology-enhanced learning, machine learning, and AI-supported formative assessment indicates a collective shift toward adaptive, feedback-driven models that prioritize learner engagement and personalization. This aligns with findings from several studies in the dataset, which report improvements in student motivation (Gordillo, 2019) and deeper conceptual understanding (Haudek et

al., 2011). Consequently, timely and tailored feedback is positioned as an epistemic agent in students' learning.

Accessibility remains a critical issue, particularly in contexts with limited technological infrastructure, echoing concerns highlighted by Asunda et al. (2023). There is a pressing need for longitudinal and cross-context studies to evaluate the sustained impact of AI-driven assessment on students' learning outcomes. Additionally, future research should explore how AI tools, such as large language models and machine learning, can be integrated into existing curricula without displacing essential teacher roles, ensuring that feedback remains pedagogically meaningful. Ethical considerations, including bias, transparency, and data privacy must also be central to forthcoming investigations. By addressing these gaps, future studies can advance inclusive, context-sensitive, and pedagogically robust frameworks that maximize the benefits of AI in assessment while safeguarding equity and quality in education.

CONCLUSION

In conclusion, this bibliometric analysis provides a structured and detailed overview of the scholarly development of artificial intelligence in the assessment of science education. Drawing on 68 publications over 15 citable years, the study maps the rapid growth of the field, identifies influential contributors, and highlights emerging research themes and priorities. The United States, Germany, and United Kingdom have led global efforts, supported by prominent institutions and researchers. The thematic analysis indicates that AI is widely applied in science education assessment, with central themes including technology-enhanced learning, automated feedback, and formative assessment. The emergence of generative AI tools further expands the possibilities for responsive and adaptive assessment practices. As the field continues to evolve, future research should explore scalable and equitable solutions, ensuring that technological innovations benefit diverse educational contexts. Overall, this study offers valuable insights for educators, researchers, and decision-makers seeking to advance assessment practices through the responsible and effective use of artificial intelligence.

ACKNOWLEDGEMENT

This work was supported/funded by the Universiti Teknologi Malaysia under UTM Fundamental Research Grant Scheme (Q.J130000.3853.23H96).

REFERENCES

Ahmi, A. (2024). *Bibliometric Analysis using biblioMagika® - Second Edition.* Academic Research Society of Malaysia.

Asmussen, G., Rodemer, M., & Bernholt, S. (2025). Steppingstones to success: A qualitative investigation of the effectiveness of adaptive stepped supporting tools for problem-solving in organic chemistry to design an intelligent tutoring system. *International Journal of Science Education*, *47*(10), 1252–1274. https://doi.org/10.1080/09500693.2024.2361933

Asunda, P., Faezipour, M., Tolemy, J., & Do Engel, M. T. (2023). Embracing Computational Thinking as an Impetus for Artificial Intelligence in Integrated STEM

Disciplines through Engineering and Technology Education. *Journal of Technology Education*, *34*(2), 43–63. https://doi.org/10.21061/jte.v34i2.a.3

Belland, B. R., Walker, A. E., & Kim, N. J. (2017). A Bayesian Network Meta-Analysis to Synthesize the Influence of Contexts of Scaffolding Use on Cognitive Outcomes in STEM Education. *Review of Educational Research*, *87*(6), 1042–1081. https://doi.org/10.3102/0034654317723009

Benotti, L., Martínez, M. C., & Schapachnik, F. (2018). A Tool for Introducing Computer Science with Automatic Formative Assessment. *IEEE Transactions on Learning Technologies*, *11*(2), 179–192. https://doi.org/10.1109/TLT.2017.2682084

Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., Kasneci, G., Zhai, X., & Nerdel, C. (2025). Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *Learning and Individual Differences*, *118*. https://doi.org/10.1016/j.lindif.2024.102601

Carchiolo, V., & Malgeri, M. (2024). Navigating the AI Timeline: From 1995 to Today. *Proceedings of the 13th International Conference on Data Science, Technology and Applications*, 577–584. https://doi.org/10.5220/0012856700003756

Chan, K. W., Ali, F., Park, J., Sham, K. S. B., Tan, E. Y. T., Chong, F. W. C., Qian, K., & Sze, G. K. (2025). Automatic item generation in various STEM subjects using large language model prompting. *Computers and Education: Artificial Intelligence*, *8*. https://doi.org/10.1016/j.caeai.2024.100344

Chun, J., Kim, J., Kim, H., Lee, G., Cho, S., Kim, C., Chung, Y., & Heo, S. (2025). A Comparative Analysis of On-Device AI-Driven, Self-Regulated Learning and Traditional Pedagogy in University Health Sciences Education. *Applied Sciences (Switzerland)*, *15*(4). https://doi.org/10.3390/app15041815

Coban, A., Dzsotjan, D., Küchemann, S., Durst, J., Kuhn, J., & Hoyer, C. (2025). AI support meets AR visualization for Alice and Bob: personalized learning based on individual ChatGPT feedback in an AR quantum cryptography experiment for physics lab courses. *EPJ Quantum Technology*, *12*(1). https://doi.org/10.1140/epjqt/s40507-025-00310-z

Delen, I., Sen, N., Ozudogru, F., & Biasutti, M. (2024). Understanding the Growth of Artificial Intelligence in Educational Research through Bibliometric Analysis. In *Sustainability (Switzerland)* (Vol. 16, Issue 16). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/su16166724

Habiballa, H., Kotyrba, M., Volna, E., Bradac, V., & Dusek, M. (2025). Artificial Intelligence (ChatGPT) and Bloom's Taxonomy in Theoretical Computer Science Education. *Applied Sciences (Switzerland)*, *15*(2). https://doi.org/10.3390/app15020581

Haudek, K. C., & Zhai, X. (2024). Examining the Effect of Assessment Construct Characteristics on Machine Learning Scoring of Scientific Argumentation.

International Journal of Artificial Intelligence in Education, *34*(4), 1482–1509. https://doi.org/10.1007/s40593-023-00385-8

Khodadad, D. (2025). ChatGPT in engineering education: a breakthrough or a challenge? *Physics Education*, *60*(4). https://doi.org/10.1088/1361-6552/add073

Kortemeyer, G. (2023a). Could an artificial-intelligence agent pass an introductory physics course. *Physical Review Physics Education Research*, *19*(1). https://doi.org/10.1103/PhysRevPhysEducRes.19.010132

Kubsch, M., Krist, C., & Rosenberg, J. M. (2023). Distributing epistemic functions and tasks—A framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching*, *60*(2), 423–447. https://doi.org/10.1002/tea.21803

Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2024a). Automated Grading and Feedback Tools for Programming Education: A Systematic Review. *ACM Transactions on Computing Education*, *24*(1). https://doi.org/10.1145/3636515

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. In *Open Medicine* (Vol. 3, Issue 2).

Mok, R., Akhtar, F., Clare, L., Li, C., Ida, J., Ross, L., & Campanelli, M. (2025). Using large language models for grading in education: an applied test for physics. *Physics Education*, *60*(3). https://doi.org/10.1088/1361-6552/adb92b

Reyes Flores, L. G., & Mejía Rivera, K. A. (2024). Inteligencia Artificial En La Investigación Cualitativa: Análisis Bibliométrico De La Producción Científica Indizada En Scopus. *New Trends in Qualitative Research*, *20*(4), e1116. https://doi.org/10.36367/ntqr.20.4.2024.e1116

Rodríguez Flores, E. A., Garcés Giraldo, L. F., Valencia, J., & Valencia-Arias, A. (2025). Tendencias investigativas en el uso de técnicas de inteligencia artificial en la investigación científica. *Revista Venezolana de Gerencia*, *30*(109), 351–380. https://doi.org/10.52080/rvgluz.30.109.23

Tang, K.-S., & Cooper, G. (2025). The Role of Materiality in an Era of Generative Artificial Intelligence. *Science and Education*, *34*(2), 731–746. https://doi.org/10.1007/s11191-024-00508-0

Terrazas-Arellanes, F. E., Strycker, L., Alvez, G., Miller, B., & Vargas, K. (2025). Promoting Agency Among Upper Elementary School Teachers and Students with an Artificial Intelligence Machine Learning System to Score Performance-Based Science Assessments. *Education Sciences*, *15*(1). https://doi.org/10.3390/educsci15010054

Wei, Y., Zhang, R., Zhang, J., Qi, D., & Cui, W. (2025). Research on Intelligent Grading of Physics Problems Based on Large Language Models. *Education Sciences*, *15*(2). https://doi.org/10.3390/educsci15020116

Wulff, P. (2023). Network analysis of terms in the natural sciences insights from Wikipedia through natural language processing and network analysis. *Education and Information Technologies*, *28*(11), 14325–14346. https://doi.org/10.1007/s10639-022-11531-5

Xu, Y., Liu, L., Xiong, J., & Zhu, G. (2025). Graders Of the Future: Comparing the Consistency and Accuracy of Gpt4 and Pre-Service Teachers in Physics Essay Question Assessments. *Journal of Baltic Science Education*, *24*(1), 187–207. https://doi.org/10.33225/jbse/25.24.187

Yamtinah, S., Wiyarsi, A., Widarti, H. R., Shidiq, A. S., & Ramadhani, D. G. (2025). Fine-tuning AI models for enhanced consistency and precision in chemistry educational assessments. *Computers and Education: Artificial Intelligence*, *8*. https://doi.org/10.1016/j.caeai.2025.100399

Zhai, X. (2024). *AI and Machine Learning for Next Generation Science Assessments*. http://arxiv.org/abs/2405.06660

Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching*, *59*(10), 1765–1794. https://doi.org/10.1002/tea.21773