JESEH

Journal of Education in Science, Environment and Health

Volume: 12 Issue: 2 Year: 2026

ISSN: 2149-214X

Journal of Education in Science, Environment and Health
Volume 12, Issue 2, 2026
e-ISSN:2149-214X

EDITORIAL BOARD

Journal of Education in Science, Environment and Health (JESEH)

The Journal of Education in Science, Environment and Health (JESEH) is a peer-reviewed and online free journal. The JESEH is published quarterly in January, April, July and October.The language of the journal is English only. As an open access journal, Journal of Education in Science, Environment and Health (JESEH) does not charge article submission or processing fees. JESEH is a non-profit journal and publication is completely free of charge.

The JESEH welcomes any research papers on education in science, environment and health using techniques from and applications in any technical knowledge domain: original theoretical works, literature reviews, research reports, social issues, psychological issues, curricula, learning environments, book reviews, and review articles. The articles should be original, unpublished, and not in consideration for publication elsewhere at the time of submission to the JESEH.

Abstracting/ Indexing

Journal of Education in Science, Environment and Health (JESEH) is indexed by following abstracting and indexing services: ERIC, Wilson Education Index

Contact Info

Journal of Education in Science, Environment and Health (JESEH)
Email: jesehoffice@gmail.com
Web : www.jeseh.net

JESEH
Journal of Education in Science, Environment and Health
Volume 12, Issue 2, 2026
e-ISSN:2149-214X

CONTENTS

JESEH
Journal of Education in Science, Environment and Health
Volume 12, Issue 2, 2026
https://doi.org/10.55549/jeseh.885

Developing a Four-Tier Concept Test on the Topic of Matter for Middle School Students

Cigdem H. Tamkavas, Cemil Aydogdu

Article Info

Article History

Published:
01 April 2026

Received:
10 August 2025

Accepted:
21 November 2025

Keywords

Misconception,
Four-tier test,
Science education.

Abstract

In this study, a four-tier concept test was developed to determine the conceptual understanding and misconceptions of 8th-grade middle school students regarding the topic of matter. The Matter Concept Test (MCT) was prepared with 16 items, each consisting of four steps, and applied to 175 eighth-grade students studying in Eskişehir. SPSS, Excel, and Factor statistical programs were used to analyze the data. According to the results of the exploratory factor analysis, the test's KMO value was found to be .652, exhibiting a four-factor structure with eigenvalues above one and explaining 42.4% of the total variance. All factor loadings were above .30, supporting the construct validity of the items. In the reliability analysis, the KR-20 coefficient for the scientific knowledge score was 0.816, and the KR-20 coefficient for the misconception score was 0.743. Both values being above .70 indicate that the test is reliable. Furthermore, the false positive average was calculated to be 5.2%, and the false negative average was calculated to be 3.7%. Both ratios being below the 10% threshold specified in the literature support the validity of the test. When the item difficulty and discrimination indices were examined, it was seen that the test consisted of items of medium difficulty and high discrimination. The findings showed that students had clear misconceptions about pure matters, atoms, mixtures and the separation of mixtures. This finding supports that the developed test can identify conceptual errors across multiple dimensions, including content knowledge, reasoning and confidence levels. In conclusion, the developed four-tier concept test was evaluated as a measurement tool that reliably and validly reveals both students' scientific knowledge and their misconceptions.

Introduction

Conceptual misconceptions in education are a fundamental problem in education and teaching, arising from the negative impact of students'/teachers' scientifically incorrect or incomplete knowledge of a subject on the educational process (Smith et al., 1993; Türkdoğan et al., 2015). Although there is erroneous or incomplete information in misconceptions, not every error or incomplete information encountered is a misconception (Eryılmaz, 2002; Önder Çelikkanlı, 2019). To speak of a misconception, it is necessary that a thought/knowledge possessed on a subject does not correspond with scientific knowledge and that, despite this, the false knowledge/thought is defended and accepted as correct (Eryılmaz & Sürmeli, 2002; Önder Çelikkanlı, 2019). Furthermore, misconceptions can hinder students' ability to make sense of newly acquired knowledge and integrate previously acquired knowledge, thereby affecting their scientific thinking skills and complicating the education, teaching, and learning process (Geçgel & Şekerci, 2018; Vosniadou, 2008). Therefore, for practical education and teaching to occur and to prevent disruptions in educational environments, misconceptions must be identified as early as possible and addressed using appropriate methods.

Since misconceptions significantly impact students' learning processes, researchers have developed various tools to identify and address them effectively. In the literature, one of the most common methods used to identify misconceptions is two-tier tests that include multiple-choice questions along with second-tier justification questions (Avcı et al., 2018; Çil et al., 2025; Pan, 2021; Jung, 2020; Küçükkeskin & Kılınç, 2024; Sarı & Bayram, 2018). Two-tier tests not only reveal the correct answer but also how students arrived at it. However, two-tier tests do not always provide sufficient opportunity to analyze students' conceptual knowledge in depth. Three-tier tests have been developed to address these limitations and shortcomings. Three-tier tests reveal the conceptual understanding processes with richer data by revealing whether students are confident in their answers as well as justifying their answers (Akbolat et al., 2023; Akdağ Kılıcı, 2019; Çetinkaya & Taş, 2018; Demirtaş, 2023; Elmas & Pamuk, 2021; Haryono et al., 2021; Haryono & Aini, 2020; Ristanto et al., 2023; Suprapto & Abidah, 2020).

However, because the step of certainty included in these tests does not clearly distinguish whether students' answers or their reasoning reflect conceptual misunderstanding, these tests have certain limitations in the process of diagnosing conceptual misunderstanding. Four-tier tests developed based on this approach have gained increasing attention in recent years because they have the potential to analyze students' responses in terms of scientific accuracy, justification, conceptual consistency, and confidence level (Bessas et al., 2024; Çelik, 2024; Hermita et al., 2017; Kaltakçı Gürel et al., 2017; Kartimi et al., 2021; Jumilah & Wasis, 2023; Putica, 2022).

This diversity supports the multidimensional approach to misconception research, analyzing students' knowledge structures and thinking patterns, not just in terms of measurement and evaluation. Although there are an increasing number of application examples for test types in the literature, tests that allow students' misconceptions to be analyzed in terms of both reasoning consistency and confidence level, as well as scientific accuracy level, are still limited (Çelik, 2024; Ma et al., 2025; Jumilah & Wasis, 2023). This situation highlights the need for new studies to further develop existing measurement tools (Desstya et al., 2025; Kaltakçı Gürel & Eryılmaz, 2015; Mert et al., 2023).

As alternatives to these methods used to identify conceptual misconceptions, other methods include open-ended questions (Alın & İzgi, 2017; Çalgıcı et al., 2020; Gökulu, 2017; Kabasakal & Uygur, 2021; Karaer, 2019; Şener Çoruhlu & Terzioğlu, 2024; Önal & Aksu, 2025), multiple-choice tests (Kardaş et al., 2020; Sancar & Koparan, 2019; Uyanık & Serin, 2016), concept cartoons (Estacio et al., 2024; Siong et al., 2023), questionnaire (Kartal, 2017), learning-oriented letter writing (Uzoğlu & Gürbüz, 2013), concept map (Kordaki & Psomos, 2015; Serttaş & Yenilmez Türkoğlu, 2020), word association tests (Kaya et al., 2019) and diagnostic decision trees (Geçgel & Şekerci, 2018; Karaaslan & Turanlı, 2018). In some studies, semi-structured interviews were conducted in conjunction with diagnostic tests to analyze students' conceptual understanding and misconceptions in depth (Kandemir & Apaydın, 2020). In addition, visual interpretation activities and concept inventories, among qualitative data collection tools, are effectively used to reveal students' mental models (Clement, 1993; Lindell et al., 2007). Some studies also included applications aimed at increasing the awareness level of students or teachers regarding misconceptions (Kandemir & Apaydın, 2020).

This study aimed to develop a valid and reliable four-tier concept test to assess students' conceptual understanding of the concept of matter in the 7th-grade middle school science curriculum in a multidimensional manner. A literature review identified that open-ended questions, interviews, multiple-choice tests, concept cartoons, and tierd tests (two-tier and three-tier) are methods commonly used to identify misconceptions and to reveal information about students' thinking structures. However, these instruments often do not allow for the simultaneous evaluation of multiple variables such as the scientific accuracy of students' answers to questions, the content of the reasons for their answers, their conceptual consistency, and their confidence level. Four-tier tests, developed to overcome this limitation, enable simultaneous analysis of students' cognitive and affective responses.

Problem Statement

A comprehensive literature review revealed that there is no four-tier diagnostic test specific to the topic of "Pure Matters and Mixtures" at the middle school level. In addition, the concept of matter is among the topics where students' prior knowledge gained from their daily life experiences often conflicts with scientific concepts and where misconceptions are frequently observed (Clement, 1993). This situation highlights the need for a novel measurement tool to be developed for this area of study. In this context, this study aims not only to develop an assessment tool but also to contribute to a healthier conceptual understanding of the topic of matter.

It is anticipated that this four-tier test will reveal the underlying thought structures and reasoning behind students' responses to the questions, enabling an assessment of their levels of scientific consistency and confidence. Thus, it will be possible for teachers to identify misconceptions in students and restructure the education-learning processes to eliminate these misconceptions. In addition to these aspects, this study aims not only to provide a functional assessment tool for those who will apply to the developed four-tier test but also to make a comprehensive, original, and innovative contribution to the field of measurement and evaluation. In this context, the sub-problems addressed in the study are presented below:

1. Is the four-tier test developed to reveal middle school students' conceptual understanding of matter a valid measurement tool?
2. Is the four-tier test developed to reveal middle school students' misconceptions about matter a suitable measurement tool in terms of reliability?

3. What are the factor-based findings regarding middle school students' scientific knowledge levels, knowledge gaps, and misconceptions about matter?
4. What are the percentages of scientific knowledge that middle school students possess regarding the matter subject?
5. What are the percentages of knowledge gaps that middle school students possess regarding the matter subject?
6. What are the percentages of middle school students' misconceptions about matters?

Method

Research Design

This study is a test development study aimed at developing a four-tier concept test called the Matter Concept Test (MCT). In the test development process, the survey model was preferred over quantitative research methods to systematically conduct validity, reliability, and item analyses and to identify students' misconceptions in the developed test. The survey model is a research design that aims to present the current situation as it is and to produce statistically relevant results for the phenomenon in question through quantitative data obtained from large sample groups (Freankel et al., 2012; Büyüköztürk et al., 2019).

Data Collection

In the study, non-random sampling was preferred among sampling methods, allowing the application process to be planned according to field conditions. This method was preferred because it minimizes the researcher's limitations in terms of time, access, and implementation processes, making the process more feasible; it also aligns with sampling approaches commonly recommended in test development studies (Frankel et al., 2012; Büyüköztürk et al., 2019). In determining the sample size, it was considered that at least five times the number of items in the measurement tool should be included (Tabachnick & Fidell, 2013). In line with this, the targeted sample size was achieved in the study.

For a pilot study, the first version MCT was applied to 200 eighth-grade students attending public middle schools affiliated with the Ministry of National Education in the province of Eskişehir. Data from students who only answered some tiers of the test or left many questions blank were considered incomplete and excluded from the analysis. Data from 175 students who answered all four tiers of the test were included in the analysis. The ages of the students included in the analysis ranged from 12 to 14 years old. After revisions of the pilot version, the final version of the test was administered to 430 eighth-grade students attending public schools affiliated with the Ministry of National Education in Eskişehir to identify misconceptions. The ages of the students participating in the main study ranged from 12 to 14 years old. Detailed demographic characteristics of both the pilot and main study samples are presented in Table 1.

Table 1. Gender distribution

	Gender	**Number**	**Percentage %**
Pilot study	Female (F)	93	53.1 %
	Male (M)	82	46. 9 %
	Total	175	100 %
Main study	Female (F)	209	48.6 %
	Male (M)	221	51,4 %
	Total	430	100 %

According to the 2018 Science Teaching Program of the Ministry of National Education (MEB), the topic of "matter" is included in the "Pure Matters and Mixtures" unit at the 7th-grade level. In the Turkey Century Education Model Science Curriculum, published in 2024, it is again presented at the 7th-grade level under the heading 'Journey to the Nature of Matter'. However, since the new curriculum was only applied in 5th grade starting in the 2024–2025 education year, the study group consisted of 8th-grade students, considering that they had completed the learning process related to the topic based on the 2018 curriculum.

A comprehensive literature review was conducted before the test development process, and in this context, the distractors in the test were prepared to include misconceptions frequently encountered in the literature (Avcı et al., 2018; Lindell et al., 2007; Özmen & Sever, 2024; Sadler, 1998; Ünal et al., 2010). Following the approach of

Kıray et al. (2015), open-ended questions were prepared to assess students' conceptual understanding of the subject and identify any misconceptions they may have.

Before the pilot study, a preliminary trial process was conducted. The questions were first read to 10 eighth-grade students, who were not part of the main sample, in order to gather feedback on visual appeal, language clarity and comprehensibility. Based on the students! Suggestions, necessary revisions were made. Following this tier, the revised items were administered to another group of 100 eighth-grade students in Eskişehir to examine students' response patterns. The researchers analyzed these responses and common misconceptions were identified. The distractors of the test items were then structured to reflect these misconceptions.

Based on these analyses, frequencies were calculated for the responses obtained, and each item was converted into a three-option multiple-choice format (two false and one correct option) based on these frequencies. Subsequently, the question "Why did you select this option?" was added below these multiple-choice items, asking students to write their reasons for choosing their response. The responses obtained were then subjected to frequency analysis, the most frequently given reasons were converted into options, and the rationale for the correct answer to the question was also included among the options. At this point, the test was structured in four steps, allowing students to select either "I am sure" or "I am not sure" after providing their answers and justifications.

The four-tier draft test was first submitted to three subject matter experts for content and face validity review, and necessary adjustments were made based on their feedback. Following these revisions, the draft test was administered to 200 eight-grade students as a pilot study. After excluding incomplete forms, data from 175 students who completed all four tiers were used to conduct item analyses and finalize the test. After the test development process was completed, the final version of the MCT was administered to larger sample of 430 eighth-grade students during the second semester of the 2024–2025 school year in order to identify students' misconceptions and determine their levels of scientific knowledge, knowledge gaps, and error types. The findings reported in this article are based on this main application.

Data Analysis

The validity and reliability analyses of the developed MCT were performed using statistical software packages, including SPSS, Excel, and Factor. During the data analysis process, different scoring types were considered in accordance with the four-tiertest structure. In this context, separate coding was performed according to categories such as scientific knowledge, conceptual misconceptions, false positives, and false negatives. In the coding, responses given in the first and third tiers were scored as "1" if correct and "0" if false. In the second and fourth tiers, where the confidence level was questioned, the "I am sure" option was coded as '1' and the "I am not sure" option as "0".

Scientific knowledge scores were calculated based on the 1-1-1-1 coding, representing the case where all tiers were answered correctly. In the conceptual misunderstanding scoring, false answers in the first and third tiers and confident answers in the confidence tiers (0-1-0-1) were considered. False positive (correct answer, incorrect reasoning) were coded as 1-1-0-1, while false negative (incorrect answer, correct reasoning) were coded as 0-1-1-1. Other combinations were included in the knowledge deficiency category. This scoring approach has enabled a multidimensional analysis of students' conceptual understanding, taking into account not only their knowledge level but also the justification of their answers and their level of confidence. Four-tier patterns used the MCT and their interpretations are presented in Table 2.

Tablo 2. Four-tier coding patterns used in MCT

First-tier (Content)	Second-tier (Confidence)	Third-tier (Reason)	Fourth-tier (Confidence)	Interpretation
1	1	1	1	Scientific knowledge
1	1	0	1	False positive
0	1	1	1	False negative
0	1	0	1	Misconception

Reliability Analysis of the Test

The developed MCT can be used to reveal both students' scientific knowledge level and their conceptual misconceptions. In this regard, the reliability of the four-tier test was evaluated using two different scoring approaches.

Reliability 1. Reliability of Scientific Knowledge

This is the reliability coefficient calculated by considering the responses where students gave correct answers in the first and third tiers and marked the "I am sure" option in the second and fourth tiers (1-1-1-1).

Reliability 2. Reliability of Misconception

This reliability coefficient is calculated based on the responses where students gave false answers in the first and third tiers and selected the "I am sure" option in the second and fourth tiers (0-1-0-1). In the literature, a reliability coefficient of .70 or above for measurement tools is generally considered acceptable (Freankel et al., 2012; Büyüköztürk et al., 2019). However, it is also noted that this value may be lower in tests designed to identify misconceptions (Kaltakçı, 2012).

Validity Analysis of the Test

The validity of the test was examined through factor analysis, the correlation between correct answers and confidence scores, the probability of positive and negative incorrect responses, and expert opinion.

Validity 1. Factor Analysis

One of the most frequently used methods for gathering evidence regarding construct validity is factor analysis, which aims to reveal the underlying dimensions of the measurement tool by examining the relationships between test items (Pedhazur & Schmelkin, 1991). In this regard, the suitability of the data obtained for the developed MCT for factor analysis was evaluated using the Kaiser–Mayer–Olkin (KMO) coefficient and Bartlett's sphericity test. A KMO value above .60 and a significant Bartlett test indicate that the data are suitable for factor analysis (Alpar, 2022). In this study, exploratory factor analysis (EFA) was conducted on the data obtained from the four-tier test to examine the construct validity of the instrument, and the results of these analyses are reported in the Findings section. Additionally, within the scope of item analysis, the difficulty and discrimination indices of each item were calculated and evaluated according to the relevant criteria.

Validity 2. Correlation Between Correct Answer Scores and Confidence Scores

To examine the relationship between students' correct answers and their confidence levels, three separate correlation coefficients were calculated: "first tier with second tier," "third tier with fourth tier," and "first and third tiers with second and fourth tiers."

Validity 3. False Positive and False Negative Probabilities

To support content validity, probabilities related to knowledge gaps, as well as false positive and false negative situations, were examined, and expert opinions were also utilized. The literature states that the averages of positive and negative incorrect responses in tests designed to measure misconceptions should be below 10% (Hestenes & Halloun, 1995).

Validity 4. Expert Opinion

Expert opinions were utilized in the development of the test items. During the preparation of open-ended questions, the opinions of an academic expert in science, a science teacher, and an academic expert in measurement and evaluation were sought. These experts evaluated the draft test in terms of content and face

validity, and the final version of the test was created by making the necessary adjustments based on their feedback. The structure of the developed test is presented in the sample question in Figure 1.

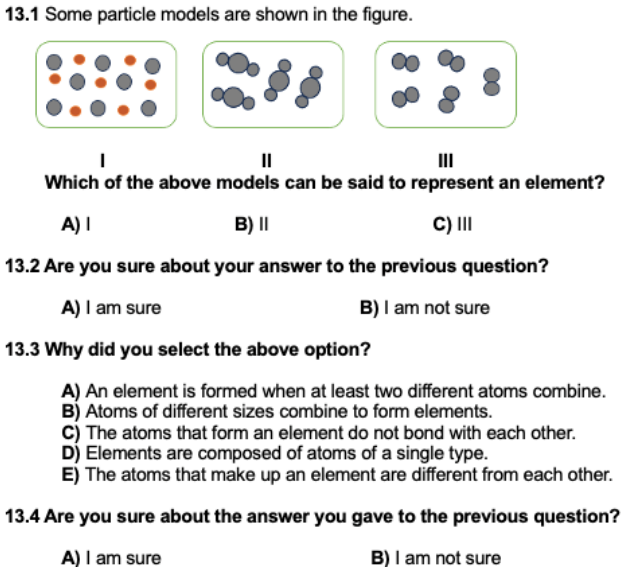


Figure 1. Sample question from the matter concept test

Findings and Discussion

The data were collected during the second semester of the 2024–2025 academic year using the MCT developed by the researcher.

Findings Related to the Reliability of the MCT

As a result of reliability analyses, the KR-20 internal consistency coefficient was calculated for the test's scores of scientific knowledge and misconceptions. The KR-20 value for the scientific knowledge score (calculated based on correct answers; 1 point was awarded when "correct answer" and "I am sure" were marked, and 0 points otherwise) was found to be 0.816. The KR-20 value calculated for the misconception scores (1 point when the false answer and 'I am sure' are marked, 0 points in other cases) was found to be .743. The fact that both coefficients are above .70 indicates that the test is sufficiently reliable in terms of scientific knowledge and misconception dimensions. (Gomez – Rodriguez et al., 2020; Vrotsou et al., 2018). Özmen & Sever (2024) reported a KR-20 value of .79 in their study, where they developed a three-tier test and noted that the contribution of the test's different tiers to measurement affected reliability. Yun et al. (2023) emphasized that internal consistency coefficients above .70 in four-tier tests are a positive indicator of the test's structural integrity and item fit. Hasançelebi et al. (2020) also stated that alpha values above .80 in multidimensional measurement tools indicate reliability. In this context, the obtained KR-20 coefficients indicate that the test can reliably measure students' conceptual understanding, and its four-tier structure supports this process.

Findings Related to the Validity of the MCT

Validity 1. Factor Analysis

According to the EFA results for the misconception test, the KMO value is .652, indicating that the sample size is adequate. The Bartlett's sphericity test (X2 = 179.1, df = 120, p = .000) was statistically significant, indicating that the correlation matrix obtained for EFA is not a unit matrix and is therefore suitable for factor analysis. Figure 2 shows the cluster plot from the EFA results for the Misconception Diagnosis Test.

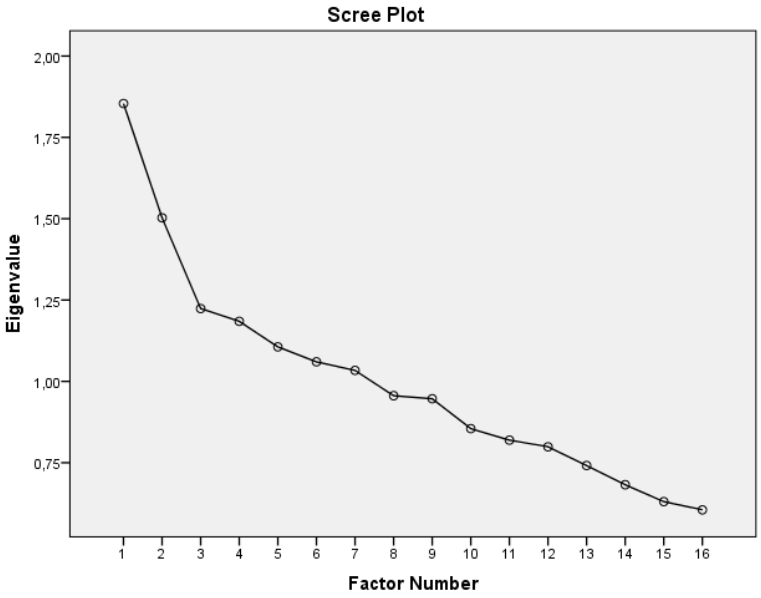


Figure 2. Scree plot

Upon examination of the graph, it can be seen that it begins to flatten from the fifth point onwards, indicating that the test has four factors. The factor loadings and explained variance values are presented in Table 3. The analysis findings show that the factor loadings of all items are at an acceptable level. Furthermore, the fact that the total explained variance ratios are above 40% is considered sufficient evidence of construct validity for tests developed in the field of educational sciences and measurement and evaluation (Tabachnick & Fidell, 2013). However, the high explained variance ratio is interpreted as a situation that increases the power of the measurement tool.

Table 3. MCT factor loadings

Test Questions	**F1**	**F2**	**F3**	**F4**
S1	0,718			
S7	0,642			
S9	0,348			
S14	0,402			
S15	0,51			
S3		0,54		
S8		0,612		
S13		0,432		
S16		0,45		
S2			0,487	
S4			0,432	
S6			0,823	
S11			0,418	
S5				0,834
S10				0,785
S12				0,779
Eigenvalue	2,351	1,793	1,353	1,292
Explained Variance	0,147	0,112	0,085	0,081
Explained Total Variance	0,424			
Reliability	0,737	0,728	0,724	0,708
Overall Scale Reliability	0,743			

As a result of the exploratory factor analysis (EFA) conducted for the MCT, four factors with eigenvalues ranging from 1.292 to 2.351 and eigenvalues above one were obtained. The first factor comprises five items (S1, S7, S9, S14, and S15) and accounts for 14.7% of the variance. The factor loadings for this factor range from 0.348 to 0.718. The second factor comprises 4 items (S3, S8, S13, and S16) and accounts for 11.2% of the variance; factor

loadings range from 0.432 to 0.612. The third factor consists of 4 items (S2, S4, S6, and S11) and explains 8.5% of the variance; factor loadings range from 0.418 to 0.823. The fourth factor comprises 3 items (S5, S10, and S12) and accounts for 8.1% of the variance, with factor loadings ranging from 0.779 to 0.834. The four factors explain a total of 42.4% of the variance. The reliability coefficients based on factors range from 0.708 to 0.737, and the overall reliability coefficient of the test is 0.743. All factor loadings were found to be above 0.30, and as a result, it was determined that the items related to the misconception test exhibit a four-factor structure.

Validity 2. Correlation Between Correct Answer Scores and Confidence Scores

The correlation coefficients calculated for the relationships between correct answers and confidence scores are presented in Table 4. In this context, the values obtained by comparing the first tier with the second tier, the third tier with the fourth tier, and both tiers together are compared.

Table 4. Correlation values of MCT

		Second Tier Scores
First Tier Scores	r	,293**
	p	,000
		Fourth Tier Scores
Third Tier Scores	r	,244**
	p	,000
		Second and Fourth Tier Scores
First and Third Tier Scores	r	,273**
	p	,000

Büyüköztürk et al. (2019) state that a correlation coefficient value of 1.00 indicates a perfect positive relationship between two variables, while a value of -1.00 indicates a perfect negative relationship. A coefficient value of 0.00 indicates that there is no relationship between the variables. According to the generally accepted classification, the absolute value of the coefficient between 0.70 and 1.00 is considered a strong relationship, between 0.30 and 0.70 is considered a moderate relationship, between 0.00 and 0.30 is considered a weak relationship. Pallant (2017) stated that when IrI < .40, the relationship is low. There is a low positive relationship between the first and second tier scores ($r = .293$, $p < .05$). There is a low-level positive relationship between the third and fourth tier scores ($r = .214$, $p < .05$). A low-level positive relationship was found between the first and third tier scores and the second and fourth tier scores ($r = .273$, $p < .05$). All scores increase and decrease in the same direction.

Validity 3. False Positive and False Negative Probabilities

The literature suggests that maintaining the positive and negative error rates below 10% in tests designed to identify misconceptions is crucial for validity (Hestenes & Halloun, 1995). The analyses conducted in this study found that the average false positive rate was 5.2%, while the average false negative rate was 3.7%. The fact that both rates are below 10% is a result that supports the validity of the developed test.

Validity 4. Expert Opinion

The content validity of the test was ensured by consulting with field experts. Consulting expert opinion to support the content validity of measurement tools is a frequently used approach in the literature (Peterson & Treagust, 1989; Zengin & Bozkurt, 2022). Furthermore, it is stated that expert contributions provide structural support to the content validity process by strengthening the consistency of test items with the conceptual framework (Sireci & Faulkner-Bond, 2014).

Some of the 33 items included in the initial development of the MCT were removed from the test based on expert opinions, as they contained content that repeated similar concepts in a way that could harm content validity. Removal of items also shortened the test administration time and made it more economical, and in line with the findings obtained from factor analysis and item analysis. In this context, assessments of the difficulty and discriminative power levels of the items were evaluated based on the criteria presented in Table 5; items with a difficulty (p) below **.20** (difficult), items close to **1.00** (very easy), and items with a discriminative power index (r) below .**20** were not considered.

Table 5. Evaluation criteria based on item difficulty and item discrimination index values

Difficulty Index	Evaluation of the Item	Distinctiveness Index	Evaluation of the Item
0.70 – 1	Very easy	0.19 and smaller	Very weak, must be removed (Weak)
0.5 – 0.69	Easy	0.20 – 0.29	Needs correction and improvement (Moderate)
0.30 – 0.49	Moderately difficult	0.30 – 0.39	Quite good, but still could be improved (Good)
0.29 and below	Difficult	0.40 and larger	Very good item (Very good)

Note: This table was created based on widely accepted principles in the field of measurement and evaluation in education and was adapted from the study by Hasançelebi et al. (2020).

Table 6 presents the difficulty and discriminating power levels of the items in the final version of the MCT, and the values indicate that both are within acceptable ranges. The four-tier structure of the developed test allows for analyzing not only whether students give correct answers but also the underlying reasons for these answers and the students' confidence levels. In this respect, the test has been evaluated as a diagnostic tool that can reveal students' misconceptions in greater depth. Similarly, Bessas et al. (2024) emphasized that the four-tier test structure is effective in identifying misconceptions by revealing the thinking processes behind student responses. Özmen & Sever (2024) stated that three-tier tests reveal students' ability to establish cause-and-effect relationships, allowing for a more precise observation of their orientation toward alternative concepts. The study conducted by Şen et al. (2017) also emphasized that multi-tier tests make significant contributions to distinguishing students' superficial knowledge and identifying false conceptualizations. In this context, the four-tier structure of the developed test provides the opportunity to assess not only students' knowledge levels but also their tendencies toward misconceptions and cognitive confidence processes in a multidimensional manner.

Table 6. Difficulty and distinctiveness levels of questions remaining in the MCT

Question Number	Difficulty Index	Difficulty Level	Discrimination Index	Discrimination Level
1.1	0.42	Moderate	0.48	Very Good
1.3	0.36	Moderate	0.39	Good
2.1	0.38	Moderate	0.31	Good
2.3	0.30	Moderate	0.30	Good
3.1	0.43	Moderate	0.41	Very Good
3.3	0.43	Moderate	0.50	Very Good
4.1	0.46	Moderate	0.34	Good
4.3	0.22	Difficult	0.32	Good
5.1	0.50	Moderate	0.46	Very Good
5.3	0.32	Moderate	0.48	Very Good
6.1	0.39	Moderate	0.45	Very Good
6.3	0.19	Difficult	0.34	Good
7.1	0.38	Moderate	0.41	Very Good
7.3	0.32	Moderate	0.30	Good
8.1	0.51	Moderate	0.43	Very Good
8.3	0.34	Moderate	0.33	Good
9.1	0.41	Moderate	0.52	Very Good
9.3	0.24	Difficult	0.47	Very Good
10.1	0.42	Moderate	0.39	Good
10.3	0.17	Difficult	0.40	Very Good
11.1	0.47	Moderate	0.49	Very Good
11.3	0.39	Moderate	0.38	Good
12.1	0.40	Moderate	0.54	Very Good
12.3	0.12	Difficult	0.35	Good
13.1	0.37	Moderate	0.50	Very Good
13.3	0.30	Moderate	0.48	Very Good
14.1	0.52	Moderate	0.48	Very Good
14.3	0.39	Moderate	0.58	Very Good
15.1	0.42	Moderate	0.33	Good
15.3	0.36	Moderate	0.36	Good
16.1	0.58	Moderate	0.48	Very Good
16.3	0.41	Moderate	0.58	Very Good

Findings Related to Misconceptions

Table 7. Classification of students' responses on the topic of matter in MCT

Content	Factor 1						Factor 2						Factor 3						Factor 4					
Question Number	1	7	9	14	15	Mean	3	8	13	16	Mean	2	4	6	11	Mean	5	10	12	Mean
% First Tier	43,95	35,58	41,63	51,16	41,16	42,70	41,40	45,12	33,95	53,95	43,60	37,67	45,35	35,58	43,72	40,58	48,37	40,93	40,23	43,18
% First Two Tiers	43,95	35,35	41,40	50,93	41,16	42,56	41,40	44,88	32,26	53,95	43,37	37,67	45,35	34,88	43,49	40,35	48,14	40,93	38,84	42,64
% First Three Tiers	19,07	6,98	16,05	36,05	21,40	19,91	26,05	19,77	17,91	28,84	23,14	18,37	3,49	13,02	19,07	13,49	21,63	14,65	8,37	14,88
% First Four Tiers	19,07	6,98	15,58	35,81	21,40	19,77	25,81	19,53	17,91	28,84	23,02	18,37	3,49	13,02	18,84	13,43	21,63	14,65	8,14	14,81
% Scientific Knowledge	19,07	6,98	15,58	35,81	21,40	19,77	25,81	19,53	17,91	28,84	23,02	18,37	3,49	13,02	18,84	13,43	21,63	14,65	8,14	14,81
% False Positive	24,65	28,14	25,12	14,88	19,77	22,51	15,12	25,12	15,35	24,88	20,12	19,30	41,63	21,86	24,19	26,74	26,28	26,05	30,47	27,60
% False Negative	16,28	21,63	6,28	2,56	10,93	11,53	14,42	9,30	12,33	6,28	10,58	11,16	6,05	4,42	14,19	8,95	6,28	5,35	5,12	5,58
% Misconception	39,77	42,33	51,40	45,81	47,44	45,35	43,95	44,88	53,49	39,30	45,41	50,00	48,40	59,30	41,63	49,88	45,35	52,56	53,95	50,62
% Lack of Knowledge 1	0	0	0,47	0	0,23	0,14	0,23	0,23	0	0	0,12	0	0	0	0,23	0,06	0	0	0,23	0,08
% Lack of Knowledge 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
% Lack of Knowledge 3	0	0	0,23	0	0,23	0,09	0	0	0,23	0,06	0	0	0	0	0	0	0	0	0,23	0,08
% Lack of Knowledge 4	0,23	0,23	0,23	0	0	0,14	0,23	0	0	0,23	0,12	0	0,23	0	0,23	0,12	0,23	0,23	0	0,16
% Lack of Knowledge 5	0	0	0	0	0	0	0	0,23	0,23	0	0,12	0	0	0,70	0,23	0,23	0,23	0	0,47	0,23
% Lack of Knowledge 6	0	0,23	0	0	0	0	0	0	0,23	0	0,06	0	0	0	0	0	0	0	0,70	0,23
% Lack of Knowledge 7	0	0	0	0	0	0	0,23	0	0	0	0,06	0	0	0	0	0	0	0	0,23	0,08
% Lack of Knowledge 8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
% Lack of Knowledge 9	0	0,23	0	0	0	0,05	0	0	0	0	0	0	0	0	0	0	0	0	0	0
% Lack of Knowledge 10	0	0	0,47	0	0,47	0,19	0	0,23	0	0	0,06	0,23	0	0,23	0	0,12	0	0,70	0,23	0,31
% Lack of Knowledge 11	0	0	0,23	0	0	0,05	0	0	0	0	0,17	0,23	0	0,23	0,23	0	0	0,47	0,23	0,23
% Lack of Knowledge 12	0	0,23	0	0,47	0	0,14	0	0,47	0,23	0,23	0,23	0,70	0	0,23	0,23	0,29	0	0	0	0

Table 8. Misconceptions in MCT

Misconception Number	Misconceptions	Item–Tier Combinations
KY 1	Water is not a matter.	1.1.a 1.2.a 1.3.a 1.4.a
KY 2	Sugary water is a pure matter.	1.1.a 1.2.a 1.3.b 1.4.a
KY 3	Sugary water has a formula.	1.1.c 1.2.a 1.3.e 1.4.a
KY 4	Pure matters can be separated by physical means.	1.1.c 1. 2.a 1.3 f 1.4.a
KY 5	Ethyl alcohol - water is a heterogeneous mixture.	2.1.a 2.2.a 2.3.b 2.4.a
KY 6	Ethyl alcohol - water separated by evaporation	2.1.a 2.2.a 2.3.e 2.4.a
KY 7	Olive oil has a higher density than water	2.1.b2.2.a2.3.c2.4.a 2.1.b 2.2.a 2.3.d 2.4.a
KY 8	Atoms do not have energy levels.	3.1.b 3.2.a 3.3.a 3.4.a
KY 9	Atoms do not have a nucleus.	3.1.b 3.2.a 3.3.e 3.4.a
KY 10	Some matters do not contain atoms	3.1.c 3.2.a 3.3.c 3.4.a
KY 11	Atoms are not the building blocks of all matters found in nature.	3.1.c 3.2.a 3.3.f 3.4.a
KY 12	Homogeneous solid-liquid mixtures, such as salt water, are separated from each other by filtration.	4.1.b 4.2.a 4.3.a 4.4.a
KY 13	Homogeneous liquid-liquid mixtures are separated from each other by the density difference method.	4.1.b 4.2.a 4.3.a 4.4.a 4.1.c 4.2.a 4.3.c 4.4.a
KY 14	Homogeneous mixtures, such as salt-water and alcohol-water, are separated from each other by density difference.	4.1.b 4.2.a 4.3.b 4.4.a
KY 15	Alcohol and water have different densities, so they are separated using a separating funnel.	4.1.c 4.2.a 4.3.e 4.4.a
KY 16	When the liquid evaporates, the mixture becomes heterogeneous.	5.1.a 5.2.a 5.3.d 5.4.a
KY 17	When the mixture evaporates, the amount of sugar decreases compared to the initial state.	5.1.c 5.2.a 5.3.b 5.4.a
KY 18	The sugar evaporates and disappears compared to the second state.	5.1.c 5.2.a 5.3.e 5.4.a
KY 19	During dissolution, the amount of matter decreases.	6.1.b 6.2.a 6.3.b 6.4.a
KY 20	Sugar dissolves in hot tea.	6.1.b 6.2.a 6.3.d 6.4.a
KY 21	During dissolution, the amount of sugar increases.	6.1.c 6.2.a 6.3.a 6.4.a
KY 22	The atoms of compounds are the same.	7.1.b7.2.a7.3.a7.4.a, 7.1.b7.2.a7.3.e7.4.a, 16.1.a16.2.a16.3.a16.4.a, 16.1.c16.2.a16.3.d16.4.a, 16.1.c16.2.a16.3.e 16.4.a
KY 23	The atoms of elements are different from each other.	7.1.a 7.2.a 7.3.c 7.4.a, 7.1.a 7.2.a 7.3.d 7.4.a, 8.1.b 8.2.a 8.3.a 8.4.a, 13.1.b13.2.a13.3.b13.4.a 13.1.a13.2.a13.3.e13.4.a
KY 24	When at least three different types of atoms come together, an element is formed.	8.1.b 8.2.a 8.3.c 8.4.a
KY25	Compounds are formed when more than one molecule comes together.	8.1.c 8.2.a 8.3.d 8.4.a
KY 26	Steel is a pure matter.	9.1.b 9.2.a 9.3.a 9.4.a
KY 27	Seawater is a heterogeneous mixture.	9.1.a 9.2.a 9.3.b 9.4.a
KY 28	Not all solutions are homogeneous.	9.1.a 9.2.a 9.3.c 9.4.a
KY 29	Steel is a compound.	9.1.b 9.2.a 9.3.d 9.4.a
KY 30	Milk does not mix with anything.	10.1.a10.2.a10.3.a10.4.a
KY 31	Milk is a pure matter.	10.1.a10.2.a10.3.b10.4.a
KY 32	Fog forms a homogeneous mixture because it does not disperse evenly everywhere.	10.1.c10.2.a10.3.c10.4.a
KY 33	Fog is a homogeneous mixture because it has a single color.	10.1.c10.2.a10.3.e10.4.a
KY 34	Increasing the temperature accelerates the melting of the matter.	11.1.b11.2.a11.3.d11.4.a, 11.1.b11.2.a11.3.e11.4.a, 11.1.c11.2.a11.3.c 11.4.a
KY 35	As the disappearance time of matters increases, the melting rate increases.	11.1.c11.2.a11.3.b 11.4.a
KY 36	Salt and coffee mix in water and all turn into the same matter.	12.1.c12.2.a12.3.a 12.4.a

KY 37	In mixtures of water, salt, and coffee, coffee settles to the bottom, forming a homogeneous mixture.	12.1.c12.2.a12.3.c 12.4.a
KY 38	Olive oil and sugar are homogeneous because they do not mix completely in water.	12.1.a12.2.a12.3.b 12.4.a
KY 39	Olive oil is a homogeneous mixture because it forms a separate layer from the sugary water.	12.1.a12.2.a 12.3.f 12.4.a
KY 40	When at least two different atoms combine, an element is formed.	13.1.b13.2.a13.3.a 13.4.a
KY 41	The atoms that make up the element do not bond with each other.	13.1.a13.2.a13.3.c 13.4.a
KY 42	The element is represented by its first letter.	14.1.b14.2.a14.3.a14.4.a, 14.1.b14.2.a14.3.d14.4.a, 14.1.c14.2.a14.3.c14.4.a, 14.1.c14.2.a14.3.e 14.4.a
KY 43	The name of the CO compound is carbon hydrogen.	15.1.b15.2.a15.3.b 15.4.a
KY 44	The name of the CO compound is calcium oxide.	15.1.b15.2.a15.3.c 15.4.a
KY 45	The name of the SO_2 compound is nitrogen oxide.	15.1.c15.2.a15.3.d 15.4.a
KY 46	The name of the SO_2 compound is sodium dioxide.	15.1.c15.2.a15.3.e 15.4.a

The findings of the analysis related to misconceptions are presented in Table 7, Table 8, Table 9, and Figure 3. The findings show that students exhibit various conceptual errors regarding the matter. Table 7 shows the distribution of students' responses on the matter subject across factor groups. The factors represent different dimensions of the test and reveal the general pattern of students' conceptual tendencies regarding the subject of matter. When the obtained average values are examined, it was determined that for items within Factor 1 (1, 7, 9, 14, 15), the average percentage of scientific knowledge among students was 19.77%, while the misconception rate was 45.35%. For items under Factor 2 (3, 8, 13, 16), the average scientific knowledge was 23.02%, and the misconception rate was 45.41%. In Factor 3 (2, 4, 6, 11), the average scientific knowledge of students was 13.43%, and the misconception rate was 49.88%. Finally, in Factor 4 (5, 10, 12), the average scientific knowledge was 14.81%, and the misconception rate was 50.62%. These findings show that students' conceptual understanding of matter is generally weak, with misconceptions being particularly concentrated in the third and fourth factors. Therefore, students have superficial knowledge of some concepts, but this knowledge is underpinned by persistent misinterpretations. Table 8 lists 46 misconceptions in the MCT, grouped into four main categories: "pure matter," "atom," "mixture," and "separation of mixtures."

Table 9 and Figure 1 show the percentage rates of misconceptions in MCT. Upon examining the findings, it was determined that the most common misconceptions among students were related to items KY 22 (*Compounds have the same atoms.*) and KY 42 (*Elements are represented by their first letter.*). This result indicates that students have difficulty establishing a relationship between the structure of the atom and the concept of pure matters. The misconceptions observed at low rates, “sugary water is a pure matter (KY 2)” and “sugary water has a formula (KY 3),” reveal that students evaluate the distinction between matters and mixtures based on superficial characteristics.

Table 9. Percentage distribution of the misconceptions identified in the MCT

	KY 1	**KY 2**	**KY 3**	**KY 4**	**KY 5**	**KY 6**	**KY 7**	**KY 8**	**KY 9**	**KY 10**
N	44	17	24	24	37	45	29	57	11	17
% Mean	10,23	3,95	5,58	5,58	8,60	10,47	6,74	13,26	2,56	3,95
	KY 11	**KY 12**	**KY 13**	**KY 14**	**KY 15**	**KY 16**	**KY 17**	**KY 18**	**KY 19**	**KY 20**
N	35	12	79	8	27	55	30	28	61	50
% Mean	8,14	2,79	18,37	1,86	6,28	12,79	6,98	6,51	14,19	11,63
	KY 21	**KY 22**	**KY 23**	**KY 24**	**KY 25**	**KY 26**	**KY 27**	**KY 28**	**KY 29**	**KY 30**
N.	45	169	152	23	43	60	37	31	25	31
% Mean	10,47	39,30	35,35	5,35	10,00	13,95	8,60	7,21	5,81	7,21
	KY 31	**KY 32**	**KY 33**	**KY 34**	**KY 35**	**KY 36**	**KY 37**	**KY 38**	**KY 39**	**KY 40**
N	56	17	33	72	16	30	44	25	48	48
% Mean	13,02	3,95	7,67	16,74	3,72	6,98	10,23	5,81	11,16	11,16
	KY 41	**KY 42**	**KY 43**	**KY 44**	**KY45**	**KY46**				
N	45	158	38	39	7	11				
% Mean	10,47	36,74	8,84	9,07	1,63	2,56				

Note. N: The number of students who selected the corresponding misconception (total n=430). % Mean: The mean percentage indicating the mean occurrence rate of given misconception

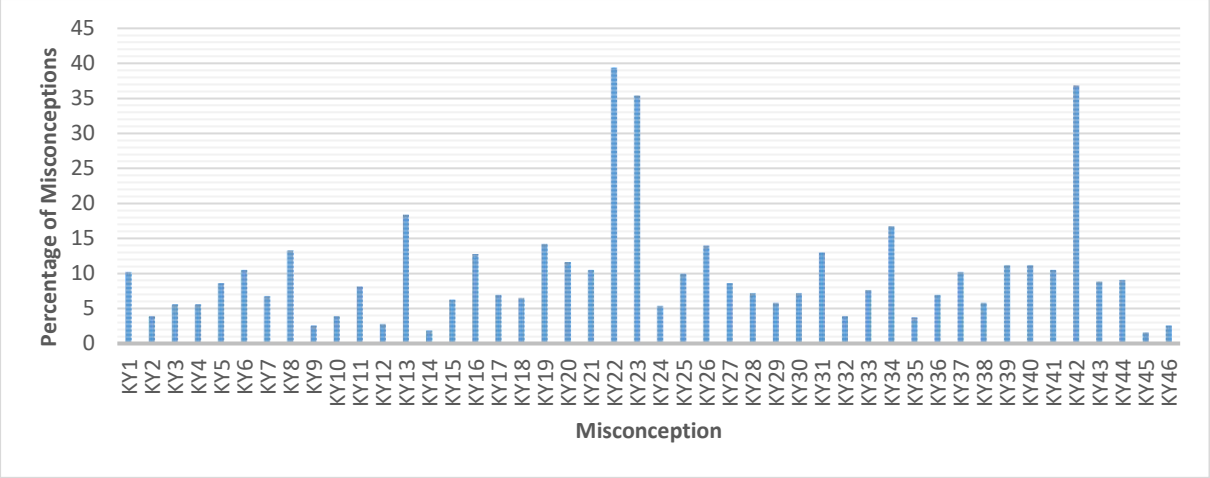


Figure 3. Percentages of misconceptions in MCT

The rates of misconceptions have been determined in the literature, and misconceptions with an average of 10% or higher have been considered significant (Caleon & Subramaniam, 2009; Kaltakçı, 2012; Kıray & Şimşek, 2020; Taban & Kıray, 2022). Considering this criterion, in this study, the misconceptions numbered KY 1, KY 6, KY 8, KY 13, KY 16, KY 19, KY 20, KY 21, KY 22, KY 23, KY 25, KY 26, KY 31, KY 34, KY 37, KY 39, KY 40, KY 41, and KY 42 were found to exceed the 10% threshold. These findings indicate that students continue to experience conceptual difficulties, particularly in pure matters, atoms, mixtures, and the separation of mixtures.

Conclusions and Recommendations

The MCT developed within the scope of this study was designed to assess middle school students' conceptual understanding of the matter subject in a multidimensional manner. The four-tier structure of the test allowed for the evaluation of students' conceptual knowledge not only at the recall level but also in terms of justification and confidence in their responses. As a result of reliability analyses, the KR-20 value for the scientific knowledge score was .816, while the KR-20 value for the misconception score was .743; both values, above .70, supported the test's reliability.

Exploratory factor analysis findings revealed that the test has a four-factor structure. The variance ratios explained by the factors, whose eigenvalues ranged from 1.292 to 2.351 in terms of percentage ranged from 8.1% to 14.7%, and the total explained variance reached was 42.4%. The fact that all factor loadings were above .30 indicates that the test items are consistent with the structures they aim to measure and supports the test's construct validity. In analyses of test validity, the positive false rate was 5.2% and the negative false rate was 3.7%. These rates below 10% indicate that the test is at an acceptable level of conceptual accuracy and validity. Furthermore, correlation analyses revealed low but significant positive relationships among the test tiers. This finding shows that the four-tier structure offers a mutually supportive measurement system and is suitable for evaluating students' conceptual understanding in a multidimensional way.

As a result of item analyses and expert opinions, the initial 33-item version of the test was reduced to 16 items while maintaining content validity. During this process, items showing overlapping conceptual content, inappropriate item difficulty or discrimination indices, weak factor loadings or potential threats to content validity based on expert evaluations were removed from the test. The fact that the difficulty and discriminative levels of the items were within appropriate ranges showed that the test could reliably distinguish students' conceptual levels. This finding indicates that the developed four-tier test is a valid and reliable measurement tool and can make significant contributions to the field of application.

The results of the factor-based analysis show that students' conceptual understanding of the matter subject is weak overall and that misconceptions concentrated in the third and fourth factors indicate learning difficulties in these areas. This situation suggests that students' knowledge levels remain superficial and that they tend to resort to alternative thinking models when trying to understand scientific concepts. The study also identified students' misconceptions regarding matter. The analyses revealed that students experienced conceptual difficulties, particularly in pure matters, atoms, mixtures, and the separation of mixtures. Misconceptions observed at a rate of 10% or higher revealed that students tend to relate the particulate structure of the matter at the macroscopic

level and have difficulty understanding abstract concepts. This result shows that the developed four-tier test can reveal these misconceptions in a multidimensional way.

In conclusion, the developed four-tier test is a valid and reliable measurement tool that can effectively distinguish students' scientific knowledge levels, knowledge gaps, and misconceptions. Although numerous studies in the literature aim to reveal students' misconceptions about matters, the absence of a four-tier diagnostic test developed for middle school students makes this study unique in its field. In this respect, the test is suitable for determining students' conceptual difficulties with matter and for planning instructional interventions to address them. Through this test, the areas where students' misconceptions are concentrated can be identified, and the underlying cognitive processes can be examined in depth. Furthermore, the test can serve as a functional assessment tool to determine the contributions of different teaching methods and instructional materials to the elimination of students' misconceptions.

Ethical Statement

* The research was conducted in accordance with ethical principles, and the necessary permissions were obtained from the Hacettepe University Social and Human Sciences Research Ethics Committee, with decision dated 02/11/2023 and numbered E-51944218-300-00003177317.

Conflict of Interest

* There is no conflict of interest among the authors in the conduct and publication of this study.

Funding

* There is no conflict of interest among the authors, and the study was not financially supported by any institution or organization.

Acknowledgments and Notes

* This study is derived from the first author's doctoral thesis and covers a section of the thesis.

References

Alın, G., & İzgi, Ü. (2017). İlköğretim öğrencilerinin yıldızlar konusuna ilişkin kavram yanılgılarının incelenmesi. *Sosyal Bilimler Dergisi, 4(*10), 202 – 214.

Alpar, R. (2022). *Spor, sağlık ve eğitim bilimlerinden örneklerle uygulamalı istatistik ve geçerlik-güvenirlik: SPSS'de çözümleme adımları ile birlikte* (7. baskı). Detay Yayıncılık.

Akbolat, E., Soyalp, F. & Arsen, S. (2023). Fen bilimleri 7. sınıf güneş sistemi ünitesinde artırılmış gerçeklik kullanımının kavram yanılgıları üzerine etkisi [Bildiri sunumu]. *TRB2 Uluslararası Eğitim Bilimleri Kongresi*, Van.

Akdağ Kılıcı, T. (2019). *Ortaokul 7. sınıf öğrencilerinin atom kavramı hakkındaki kavram yanılgıları* (Yayınlanmamış yüksek lisans tezi). Necmettin Erbakan Üniversitesi.

Avcı, F., Acar Şeşen, B. & Kırbaşlar, F. G. (2018). Maddenin yapısı ve özellikleri ünitesine yönelik iki aşamalı teşhis testinin geliştirilmesi. *Kastamonu Educational Journal*, *26*(4), 1007 – 1019. DOI:10.24106/kefdergi.434239

Bessas, N., Tzanaki, E., Vavougios, D. & Plagianakos, V. P. (2024). Diagnosing students' misconception in hydrostatic pressure through a 4-tier test. *Helion*, *10*(23), 1 – 19. https://doi.org/10.1016/j.heliyon.2024.e40425

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş & Demirel, F. (2019). *Bilimsel araştırma yöntemleri* (26. Baskı). Pegem Akademi Yayıncılık.

Caleon, I., & Subramaniam, R. (2010). Development and application of three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education,* 7(32), 939 – 961. https://doi.org/10.1080/09500690902890130

Clement, J. (1993). Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics. *Journal of Research in Science Teaching, 30*(10), 1241–1257.

Çalgıcı, G., Yıldırım, M., & Duru, M. K. (2020). 5. sınıf öğrencilerinin madde ve hal değişimi konusunda kavram yanılgılarının oyunlaştırma ile giderilmesi. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi (EFMED), 14*(2)*,* 1278 – 6086. Doi: 10.17522/balikesirnef.814908

Çelik, E. (2024). *8. sınıf öğrencilerinin ısı ve sıcaklık konusundaki kavram yanılgılarının belirlenmesi ve kavramsal değişim metinleri ile kavram yanılgılarının iyileştirilmesi* (Yayınlanmamış yüksek lisans tezi). Harran Üniversitesi.

Çetinkaya, M. & Taş, E. (2018). Etkinlik temelli web materyalinin 6. sınıf "vücudumuzda sistemler" ünitesindeki kavram yanılgılarının giderilmesine etkisi. *International e-Journal of Educational Stadies* (IEJES), *2*(4), 92 – 113. https://doi.org/10.31458/iejes.428319

Çil, E., Gökçen, E. & İren, I. (2025). Fen bilimleri öğretmen adaylarının okyanus hakkındaki kavram yanılgılarının belirlenmesi. *Fen, Matematik, Girişimcilik ve Teknoloji Eğitimi Dergisi*, *8*(2), 187 – 213. DOI: 10.36681/ fmgted.24010

Demirtaş, M. (2023). *Ortaokul öğrencilerinin basit elektrik devreleriyle ilgili kavram yanılgılarını ölçmek amacıyla üç aşamalı kavram yanılgısı testi geliştirilmesi* (Yayınlanmamış yüksek lisans tezi). Fırat Üniversitesi.

Desstya, A., Sayekti, I. C., Abduh, M., & Sukartono. (2025). Development of a four-tier diagnostic test for misconception in natural science of primary school pupils. *Journal of Turkish Science Education, 22*(2), 338 – 353.

Elmas, R. & Pamuk, S. (2021). Öğretmen adaylarının kavram yanılgılarının üç aşamalı kavram yanılgısı testi ile belirlenmesi. *Afyon Kocatepe Üniversitesi Sosyal Bilimler Dergisi*, *23*(4), 1386 – 1403. https://doi.org/10.32709/akusosbil.916063

Estacio, R. D., Reyes, E. A. S. & Apusaga, N. C. (2024). Exploring engineering students' misconceptions about motion and forces using concept cartoons, *Mimbar Pendidikan***,** *9*(1), 13 – 32. https://doi.org/10.17509/mimbardik.v9i1

Eryılmaz, A. (2002). Effects of conceptual assignments and conceptual change discussions on students' misconceptions and achievement regarding force and motion. *Journal of Research in Science Teaching*, *39*, 1001–1015. https://doi.org/10.1002/tea.10054

Eryılmaz, A. & Sürmeli, E. (2002). Üç - aşamalı sorularla öğrencilerin ısı ve sıcaklık konularındaki kavram yanılgılarının ölçülmesi. *V. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi.*

Freankel, J. R., Wallen, N. E & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw – Hill.

Geçgel, G. & Şekerci, A. R. (2018). Bazı kimya konularındaki alternatif kavramların tanılayıcı dallanmış ağaç tekniği kullanarak belirlenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *14*(1), 1 – 18. https://doi.org/10.17860/mersinefd.290254

Gomez – Rodriguez, R., Díaz – Pulido, B., Gutierrez – Ortega, C., Sanchez – Sanchez, B. & Torres – Lacomba, M. (2020). Cultural adaptation and psychometric validation of the standardised nordic questionnaire Spanish version in musicians. *International Journal of Environmental Research and Public Health 17*(2), 653. https://doi.org/10.3390/ijerph17020653

Gökulu, A. (2017). 8. sınıf öğrencilerin element, bileşik, karışım kavramlarını anlama düzeyleri ve kavram yanılgılarının incelenmesi. *Kastamonu Eğitim Dergisi*, *25*(2), 1 – 16.

Hasançelebi, B., Terzi, Y. & Küçük, Z. (2020). Madde güçlük indeksi ve madde ayırt edicilik indeksine dayalı çeldirici analizi. *Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *10*(1), 224 – 240.

Haryono, H. E. & Aini, K. N. (2020). Diagnosis misconceptions of junior high school in lamongan on the heat concept using the three – tier test. *Journal of Physics: Conference Series, 1806*(1), 012002.

Haryono, H. E., Aini, K. N., Samsudin, A., & Siahaan, P. (2021). Diagnosis of student misconception in heat material using tier test. *Jurnal Pembelajaran Fısıka*, *9*(2), 155 – 162. http://dx.doi.org/10.23960/jpf.v9.n2.202104

Hermita, N., Suhandi, A., Syaodih, E., Samsudin, A., Isjoni, I., Johan, H., Rosa, F., Setyaningsih, R., Sapriadil, S. & Safitri, D. (2017). Constructing and implementing a four tier test about static electricity to diagnose pre-service elementary school teacher' misconceptions. *Journal of Physics: Conference Series, 895*, 012167. https://doi.org/10.1088/1742-6596/895/1/012167

Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory: a response to huffman and heller. *The Physics Teacher*, *33*, 502 – 506. https://doi.org/10.1119/1.2344278

Jumilah, J., & Wasis, W. (2023). Development of four-tier diagnostic test instrument to introduce misconceptions and identify causes of student misconceptions in the sub-topic of Bernoulli's principle. *Jurnal Penelitian Pendidikan IPA (JPPIPA)*, *9*(7), 5773 – 5781.

Jung, J. (2020). Diagnosing causes of pre-service literature teachers' misconceptions on the narrator and focalizer using a two-tier test. *Educational Sciences*, *10*(4), 104 – 124.

Karaaslan, G., & Turanlı, N. (2018). Karmaşık sayılar konusundaki kavram yanılgıları ve hataları belirlemek için alternatif bir araç: tanılayıcı dallanmış ağaç. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi* (*H.U. Journal of Education), 33*(1), 72 – 89. Doi:10.16986/HUJE.2017030356

Kabasakal, H. Z., & Uygur, S. (2021). Öğretmenlerin özel öğrenme güçlüğüne ilişkin bilgi düzeyleri, görüş ve kavram yanılgılarının incelenmesi. *International Socıal Mentality And Researcher Thinkers Journal*, *7*(51), 2736 – 2754. http://dx.doi.org/10.31576/smryj.1121

Kaltakçı, D. (2012). *Development and application of a four – tier test to asses pre – service physics teachers' misconceptions about geometrical optics* (Yayınlanmamış doktora tezi). Middle East Technical University.

Kaltakçı Gürel, D., Eryılmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *Research in Science & Technological Education, 35*(2), 238 – 260. https://doi.org/10.1080/02635143.2017.1310094

Kandemir, M. A. & Apaydın, Z. (2020). Sınıf öğretmenlerinin fen bilimleri dersinde öğrencilerin sahip olduğu kavram yanılgılarını belirlemelerine ve gidermelerine yönelik bir değerlendirme. *Türkiye Bilimsel Araştırmalar Dergisi (TÜBAD), 5*(2), 183 – 198.

Karaer, H. (2019). Determination of some misconceptions in solution concentrations of the teacher candidates and examination regarding to some comprehension levels. *Ercıyes Journal Of Education (Eje), 3*(2), 87 – 104.

Kardaş, F., Bayrakçeken, S., & Taşdemir, F. N. (2020). Onuncu sınıf öğrencilerinin çözeltiler konusuna yönelik kavram yanılgılarının belirlenmesi. *International Social Sciences Studies Journal, 6*(71), 4405 – 4412. http://dx.doi.org/10.26449/sssj.2713

Kartimi, K., Yunita, Y., Fuadi, F. N., & Addiin, I. (2021). A four-tier diagnostic instrument: An analysis of elementary student misconceptions in science topic. *Jurnal Penelitian Pendidikan IPA (JPPIPA)*, *7*(Special Issue), 61 – 68.

Kaya, B., Ateş, A., & Kılıç, S. (2019). Üniversite öğrencilerinin küresel ısınma konusundaki bilişsel (zihinsel) yapıları ve kavram yanılgılarının belirlenmesi. *International Journal of Social Science, 74*, 29 – 40. http://dx.doi.org/10.9761/JASSS7731

Kıray, S. A., Aktan, F., Kaynar, H., Kılınç, S. & Görkemli, T. (2015). A descriptive study of pre-service science teachers' misconceptions about sinking-floating. *Asia – Pacific Forum on Science Learning and Teaching*, *16*(2).

Kıray, S. A., & Şimşek, S. (2020). Determination and evaluation of the science teacher candidates' misconceptions about density by using four-tier diagnostic test. *International Journal of Science and Mathematics Education, 19*, 935 – 955. https://doi.org/10.1007/s10763-020-10087-5

Kordaki, M., & Psomos, P. (2014). Diagnosis and treatment of students' misconceptions with an intelligent concept mapping tool. *Procedia - Social and Behavioral Sciences, 191*, 838 – 842. https://doi.org/10.1016/j.sbspro.2015.04.478

Küçükkeskin, E. & Kılıç, D. (2024). Hücre bölünmeleri konusunda öğrencilerin kavramsal anlamalarını belirlemeye yönelik iki aşamalı test geliştirilmesi. *Fen Bilimleri Öğretimi Dergisi*, *12*(1), 99 – 121. https://doi.org/10.56423/fbod.1380581

Lindell, R. S., Peak, E., & Foster, T. M. (2007). Are they all created equal? A comparison of different concept ınventory development methodologies. *Physics Education Research Conference Proceedings*.

Ma, H., Yang, H., Li, C., Ma, S., & Li, G. (2025). The effectiveness and sustainability of tier diagnostic technologies for misconception detection in science education: A systematic review. *Sustainability, 17*(7), 3145. https://doi.org/10.3390/su17073145

Mert, E., Apaydın, Z., & Omca Çobanoğlu, E. (2023). Examination of thesis written on misconceptions in primary science education: between 2012 – 2022. *Journal of Computer and Education Research*, *12*(23), 125 – 147.

Millî Eğitim Bakanlığı (MEB) (2018). Fen bilimleri dersi öğretim programı (İlkokul ve ortaokul 3, 4, 5, 6, 7, ve 8. Sınıflar). Ankara. https://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=325

Milli Eğitim Bakanlığı (MEB) (2024). Fen bilimleri dersi öğretim programı (İlkokul ve ortaokul 3, 4, 5, 6, 7, ve 8. Sınıflar). Ankara. https://mufredat.meb.gov.tr/

Kartal, M. (2017). *Fen bilgisi öğretmen adaylarının bazı kimya kavramlarını anlama seviyeleri ve kavram yanılgılarının belirlenmesi* (Yüksek lisans tezi). Necmettin Erbakan Üniversitesi.

Önal, B., & Aksu, Z. (2025). 5. Sınıf öğrencilerinin çevre ve alan konusundaki kavram yanılgılarının giderilmesinde gerçekçi matematik eğitimi yaklaşımının etkililiği. *Social Sciences Studies Journal, 11*(4), 630 – 646. https://doi.org/10.5281/zenodo.15282824

Önder Çelikkanlı, N. (2019). *Elektriklenme konusunda dört aşamalı kavram yanılgısı testi geliştirme* (Yayınlanmamış doktora tezi). Gazi Üniversitesi.

Özmen, F. & Sever, R. (2024). Küresel sorunlar üç aşamalı kavramsal anlama testi geliştirme çalışması. *Turkish Studies – Educational Sciences*, *19*(1), 137 – 164.

Pallant, J. (2017). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (12th ed.). McGraw-Hill Education.

Pan, S. J. – A., (2021). Taiwanese and American graduate students' misconceptions regarding responsible conduct of research: a cross-national comparison using a two-tier test approach. *Science and Engineering Ethics*, *27*(20). https://doi.org/10.1007/s11948-021-00297-7

Pedhazur, E. J., & Pedhazur Schmelkin, L. (1991). *Measurement, design and analysis: An integrated approach*. Psychology Press.

Peterson, R. F. & Treagust, D. F. (1989). Grade – 12 students' misconceptions of covalent bonding and structure. *Journal of Chemical Education*, *66*(6), 459 – 460.

Putica, K. B. (2022). Development and validation of a four-tier test for the assessment of secondary school students' conceptual understanding of amino acids, proteins, and enzymes. *Research in Science Education*, *53*, 651 – 668. https://doi.org/10.1007/s11165-022-10075-5

Ristanto, R. H., Suryanda, A. & Indraswari, L. A. (2023). The development of ecosystem misconception diagnostic test. *International Journal of Evaluation and Research in Education (IJERE), 12*(4), 2246 – 2259.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching, 35*(3), 265 – 296.

Sancar, M., & Koparan, T. (2019). Ortaokul öğrencilerinin çokgenler konusundaki kavram yanılgılarının giderilmesinde kavram karikatürlerinin etkisinin incelenmesi. *Karaelmas Journal of Educational Sciences, 7*, 101 – 122.

Sarı, A. & Bayram, H. (2018). Kavram haritası ve bilgisayar destekli öğretimin 7. sınıf öğrencilerinin madde konusundaki kavram yanılgılarının giderilmesine etkisi. *The Journal of Academic Social Science Studies, 67*, 29 – 47. http://dx.doi.org/10.9761/JASSS7579

Serttaş, S., & Yenilmez Türkoğlu, A. (2020). Diagnosing students' misconceptions of astronomy through concept cartoons. *Participatory Educational Research (PER), 7*(2), 164 – 182. http://dx.doi.org/10.17275/per.20.27.7.2

Siong, L. C., Tyung, O. Y., Phang, F. A. & Pusppanathan, J. (2023). The use of concept cartoons in overcoming the misconception in electricity concepts . *Participatory Educational Research (PER), 10*(1), 310 – 329. http://dx.doi.org/10.17275/per.23.17.10.1

Sireci, S. & Faulkner – Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100 – 107.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115-163. https://doi.org/10.1207/s15327809jls0302_1

Suprapto, N. & Abidah, A. (2020). Usıng onlıne three-tıer dıagnostıc test to assess conceptıons of ıonızatıon energy. *Periodico Tche Quimica*, *17*(36), 196 – 212.

Şen, Ş., Yılmaz, A. & Geban, Ö. (2017). Üç aşamalı elektrokimya kavram testinin geliştirilmesi. *Karaelmes Fen ve Mühendislik Dergisi*, *8*(1), 324 – 330.

Şenel Çoruhlu, T., & Terzioğlu, S. D. (2024). Ortaokul 5. Sınıf fen bilimleri dersinde eğitsel şarkı kullanımının öğrencilerin kavram yanılgılarını gidermedeki etkisi: mantarlar örneği. *Trakya Eğitim Dergisi, 14(2)*, 1146 – 1159.

Taban, T., & Kıray, S. A. (2022). Determination of science teacher candidates' misconceptions on liquid pressure with four-tier diagnostic test. *International Journal of Science and Mathematic, 20*, 1791 – 1811. https://doi.org/10.1007/s10763-021-10224-8

Tabachnick, B. G. & Fidell, L. (2013). *Using Multivariate Statistic* (8th. ed.). Pearson Education

Türkdoğan, A., Güler, G., Bülbül, H. İ., & Danişman, Ş. (2015). Türkiye'de matematik eğitiminde kavram yanılgılarıyla ilgili çalışmalar: Tematik bir inceleme. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 11*(3), 216 – 231. https://doi.org/10.17860/mersinefd.255437

Uyanık, G. & Serin, M. K. (2016). Sınıf öğretmeni adaylarının bazı temel fen konularındaki kavram yanılgılarının belirlenmesi. *Amasya Üniversitesi Eğitim Fakültesi Dergisi, 5*(2), 510 – 538.

Uzoğlu, M., & Gürbüz, F. (2013). Fen ve teknoloji adaylarının ısı ve sıcaklık konusundaki kavram yanılgılarının belirlenmesinde öğrenme amaçlı mektup yazma aktivitesinin kullanılması. *International Journal of Social Science, 6*(4), 501 – 517.

Ünal, S., Coştu, B. & Ayas, A. (2010). Secondary school students' misconceptions of covalent bonding. *Journal of Turkish Science Education*, *7*(2), 3 – 29.

Vosniadou, S. (Ed.). (2008). *International handbook of research on conceptual change*. Routledge.

Vrotsou, K., Machon, M., Rivas - Ruiz,F., Carrasco, E., Contreras - Fernandez, E., Mateo - Abad, M., Güell, C. & Vergara, I. (2018). Psychometric properties of the Tilburg Frailty Indicator in older Spanish people. *Archives of Gerontology And Geriatrics, 78*, 203 – 212. https://doi.org/10.1016/j.archger.2018.05.024

Yun, V. W. S., Ulang, N. M. & Husain, S. H. (2023). Measuring the internal consistency and reliability of the hierarchy of controls in preventing infections diseases on construction sites: the kuder – richardson (KR

– 20) and cronbach's alpha. *Journal of Advanced Researh in Applied Sciences and Engineering Technology, 33*(1), 392 – 405.

Zengin, Y. & Bozkurt, E. (2022). 9. sınıf kuvvet ve hareket konusu ile ilgili kavram yanılgılarını belirlemek için üç aşamalı bir testin geliştirilmesi. *International Journal of Social, Humanities and Administrative Sciences*, *8*(50), 380 – 390. http://dx.doi.org/10.31589/JOSHAS.921

Author(s) Information

Çigdem H. Tamkavas
Turkish Ministry of National Education (TC Milli Egitim Bakanligi), TOBB Bilim ve Sanat Merkezi, Uluönder Mahallesi Malkoç Bey Sokak No 4 /Tepebaşı/Eskişehir/Türkiye
Contact e-mail: c.tamkavas@gmail.com
ORCID iD: https://orcid.org/0000-0002-2310-3876

Cemil Aydogdu
Hacettepe University, Faculty of Education Science Education Department Beytepe/Cankaya/ANKARA
ORCID iD: https://orcid.org/0000-0003-1623-965X

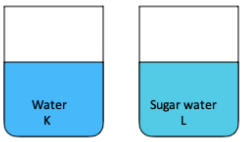
Appendix

THE MATTER CONCEPT TEST

Dear students,
This study aims to **identify students' misconceptions about matter**. Please read each item below carefully and mark the section that best applies to you. Each relevant section is separate and serves a specific purpose. It is vital for the scientific validity of this study that your answers are honest. Thank you for your contribution to this research.
Çiğdem H. TAMKAVAS

Gender: Female ☐ Male ☐ Grade:

1.1 The K container in the figure contains water, as does the L container.



I. Both are homogeneous.
II. The K container contains a compound, while the L container contains a solution.
III. The matter in both containers is represented by a formula.

Which of the previous statements can be made?

A) II and III **B) I and II** **C)** I, II, and III

1.2 Are you sure about your answer to the previous question?
A) I am sure **B)** I am not sure

1.3 Why did you select the above option?

A) There are no matter in water, so water is not a pure matter.
B) Sugary water is a pure matter and is represented by formulas.
C) Water is a compound consisting of hydrogen and oxygen, while sugary water is a homogeneous mixture consisting of sugar and water.
D) The matter in both containers is a pure matter and cannot be represented by formulas.
E) Water and sugary water have formulas.
F) Both water and sugary water can be separated physically.

1.4 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

2.1 In Science class, the teacher wants to separate olive oil-water and ethyl alcohol-water mixtures using the apparatus shown in the diagram

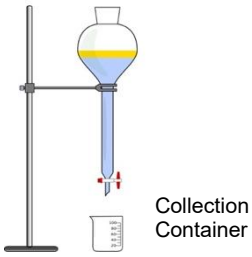


Figure – 1

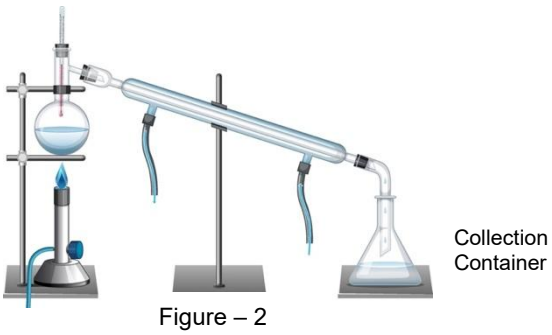


Figure – 2

The teacher performs the following steps using these mixtures.

- They place the olive oil-water mixture into the apparatus shown in Figure 1. Then, they open the spout of the separating funnel and transfer the liquid collected in the collection container to another container.
- He places the ethyl alcohol-water mixture into the apparatus shown in Figure 2 and heats it. After a while, he separates the liquid accumulated in the collection container and transfers it to another container.

Accordingly, which of the following <u>cannot be said</u> about the procedures performed by the teacher<u>?</u>

A) Since the ethanol-water mixture is homogeneous, a fractional distillation apparatus was used.
B) The liquid that accumulates in the collection container in Figure 1 and separates from the mixture is water.
C) In Figure 2, water has collected in the collection container because the boiling point of water is higher than that of ethanol.

2.2 Are you sure about your answer to the previous question?
A) I am sure **B)** I am not sure

2.3 Why did you select the above option?

A) Ethanol begins to boil earlier and its liquid accumulates in the collection container.
B) The ethanol-water mixture in the setup shown in Figure 2 is a heterogeneous mixture.
C) In the setup shown in Figure 1, the first liquid to separate is olive oil.
D) Olive oil has a higher density.
E) The ethanol-water mixture is separated using the evaporation method.

2.4 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

3.1 The teacher asked questions about atoms in class. Some students' answers are given below.
Deniz: Atoms are the building blocks of all matters found in nature.
Su: The nucleus of an atom contains electrons that are mobile and negatively charged.
Güneş: Atoms consist of two parts: the nucleus and the layers (orbits).
Which of the students' answers <u>cannot be said</u> to be correct?
A) Su **B)** Güneş **C)** Deniz

3.2 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

3.3 Why did you select the above option?

A) There are no energy levels in an atom.
B) Electrons are neutral.
C) Some matters do not contain atoms.
D) Electrons are found in the energy levels of the atom.
E) There is no nucleus in an atom.
F) Atoms are not the building blocks of all matters found in nature.

3.4 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

4.1 The table below matches some mixtures with separation methods.

	Mixture	**Separation Method**
I	Salt + Water	Evaporation
II	Alcohol + Water	Density Difference

Which of the pairings given in the table are correct according to the given situation?
A) Only I **B)** Only II **C)** I and II

4.2 Are you sure about your answer to the previous question?
A) I am sure **B)** I am not sure

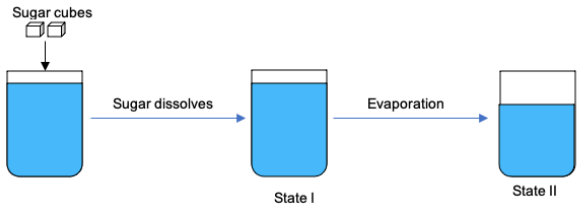
4.3 Why did you select the above option?

A) Homogeneous solid-liquid mixtures, such as salt-water, are separated by filtration, while density-based methods separate homogeneous liquid-liquid mixtures.
B) Homogeneous mixtures such as salt-water and alcohol-water are separated from each other by density difference.
C) Solid-liquid mixtures are separated by evaporation, while homogeneous mixtures formed by the combination of two liquids are separated by the density difference method.
D) All mixtures consisting of solids and liquids are separated from each other using the evaporation method, while liquid-liquid homogeneous mixtures are separated using the difference in boiling points.
E) Salt-water mixtures are separated using the evaporation method, while alcohol and water are separated using a separating funnel because they have different densities.

F) When a salt-water mixture is heated, it evaporates and the salt remains; an alcohol-water mixture is separated using the distillation method.

4.4 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

5.1 Zeynep, who dropped two sugar cubes into a container filled to the brim, ensured the sugar dissolved in the first case. In the second case, she evaporated part of the liquid in the container.



According to the diagram, which of the following can be said about the mixture in State II?

A) It has lost its homogeneity.
B) The amount of solvent has decreased.
C) The amount of sugar has decreased compared to the first situation.

5.2 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

5.3 Why did you select the above option?

A) As water evaporates, sugar turns into a solid state.
B) When the mixture evaporates, the amount of sugar decreases compared to its initial state.
C) The amount of solvent in the container decreases as it evaporates.
D) When the liquid evaporates, the mixture acquires heterogeneous properties.
E) The sugar evaporates and disappears in the second state.

5.4 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

6.1 Ebru is serving tea to her guest, Burcu. During their conversation, Ebru notices that her guest has finished her tea and serves her another hot cup. Meanwhile, she realizes that her own tea has gone cold and that she hasn't added sugar to it. Ebru and Burcu each add one sugar cube to their tea cups and wait for the sugar to dissolve without stirring.

According to this, how does the amount of sugar cubes added to the tea change over time?

A) It does not change **B**) It decreases **C)** Increases

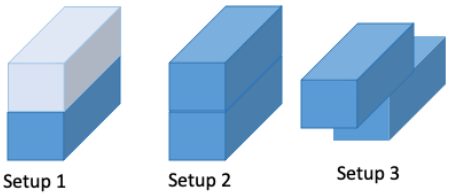
6.2 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

6.3 Why did you select the above option?

A) Temperature accelerates dissolution, increasing the amount of sugar.
B) When sugar dissolves in tea, its amount decreases.
C) Dissolution does not occur in cold water.
D) Sugar added to hot tea dissolves more quickly.
E) Breaking sugar into granules does not change the number of sugar cubes.

6.4 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

7.1 Below are setups created using blue and white Lego pieces that represent molecular models.



Based on this, which of the following statements can be made about the models?

A) Setup – 1 is a molecular structure containing two types of atoms.
B) Setup – 2 is a compound molecule containing two atoms.
C) Setup – 3 is an element molecule containing one type of atom.

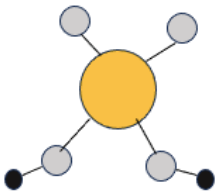
7.2 Are you sure about your answer to the previous question?
A) I'm sure **B)** I'm not sure

7.3 Why did you select the above option?

A) The atoms in the molecules that form compounds are the same.
B) Elements contain only one type of atom in their structure.
C) The atoms of elements are different from each other.
D) The atoms that make up an element have different colors.
E) When two atoms of the same type combine, a compound is formed.

7.4 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

8.1 A student uses Styrofoam spheres and toothpicks of different colors and sizes to create the model shown in the image to illustrate the atomic model of a matter.



Which of the following statements can be made about the prepared model?

A) It is a model of a molecule consisting of 7 atoms.
B) It is a model of an element consisting of 7 atoms.
C) It is a model of a compound consisting of 7 molecules.

8.2 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

8.3 Why did you select the above option?

A) Atoms of different types come together to form elements.
B) Molecules are formed when two or more atoms come together.
C) When at least three different types of atoms come together, an element is formed.
D) Compounds are formed when two or more molecules come together.
E) The atoms that make up a molecule are different from each other.

8.4 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

9.1 Which of the following <u>cannot be said?</u>

A) Seawater is an example of a solution.
B) Steel is an example of a mixture.
C) Heterogeneous mixtures are named as solutions.

9.2 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

9.3 Why did you select the above option?

A) Steel is a pure matter.
B) Seawater is a heterogeneous mixture.
C) Not all solutions are homogeneous.
D) Steel is a compound.
E) Heterogeneous mixtures are called simple mixtures.

9.4 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

10.1 Which of the following can be said to be a homogeneous mixture?

A) Milk **B)** Cola **C)** Fog

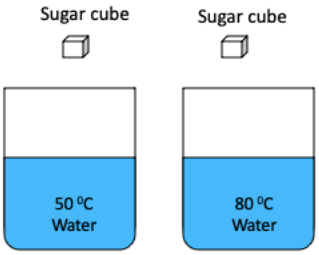
10.2 Are you sure about your answer to the previous question?
A) Yes, I am **B)** I'm not sure

10.3 Why did you select the above option?

A) Milk is a homogeneous mixture because it does not mix with anything.
B) Milk is a homogeneous mixture because it is a pure matter.
C) Fog forms a homogeneous mixture because it does not distribute uniformly everywhere.
D) The matters in cola are evenly distributed throughout the liquid.
E) Fog is a homogeneous mixture because it has a single-colored appearance.

10.4 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

11.1 The science teacher conducted the experiment shown in the diagram below.



The containers shown in the image contain the same amount of water at the specified temperatures. The time it took for all the sugar added to the containers to disappear was measured.
Which of the following can be said to be the hypothesis of this experiment?

A) What is the effect of temperature on the rate of dissolution?
B) How does sugar melt faster depending on the water temperature?
C) Which of the sugar cubes exposed to different temperatures melt faster?

11.2 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

11.3 Why did you select the above option?

A) A higher temperature increases the rate at which matters dissolve.
B) As the time it takes for matters to disappear increases, their melting rate increases.
C) Temperature affects the melting rate of a matter.
D) When the temperature is high, sugar melts faster.
E) The time it takes for sugar to melt depends on the temperature of the water.

11.4 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

12.1 A teacher divides the class into three groups for an activity and gives each group the following materials, asking them to create various mixtures.

GROUP 1
Water Olive oil Sugar

GROUP 2
Water Alcohol Sugar

GROUP 3
Water Coffee Salt

Which of the above groups <u>is not expected</u> to form a heterogeneous mixture?

A) 1 **B) 2** **C)** 3

12.2 Are you sure about your answer to the previous question?
A) I am sure **B)** I am not sure

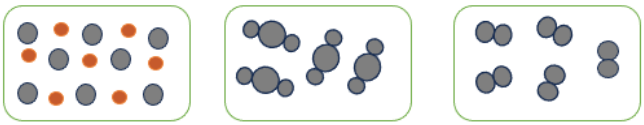
12.3 Why did you select the above option?

A) Salt and coffee mix in water and all turn into the same matter.

B) Olive oil and sugar are homogeneous because they do not mix completely in water.
C) In mixtures of water, salt, and coffee, coffee settles to the bottom, forming a homogeneous mixture.
D) Water, alcohol, and sugar form a homogeneous mixture because they are the same color.
E) The mixture formed by water, alcohol, and sugar is evenly distributed throughout.
F) Olive oil forms a separate layer from sugary water, so it is a homogeneous mixture.

12.4 Are you sure about your answer to the previous question?
A) I am sure **B)** I am not sure

13.1 Some particle models are shown in the figure.



I **II** **III**

Which of the above models can be said to represent an element?

A) I **B)** II **C)** III

13.2 Are you sure about your answer to the previous question?
A) I am sure **B)** I am not sure

13.3 Why did you select the above option?

A) An element is formed when at least two different atoms combine.
B) Atoms of different sizes combine to form elements.
C) The atoms that form an element do not bond with each other.
D) Elements are composed of atoms of a single type.
E) The atoms that make up an element are different from each other.

13.4 Are you sure about the answer you gave to the previous question?
A) I am sure **B)** I am not sure

14.1 Which of the following elements and symbols is correct?

	Element	**Symbol**
A)	Sulfur	S
B)	Nitrogen	A
C)	Magnesium	Ma

14.2 Are you sure about your answer to the previous question?
A) I am sure **B)** I am not sure

14.3 Why did you select the above option?

A) The symbol for nitrogen is the letter A.
B) The symbol for sulfur is represented by the letter S.
C) The symbol for magnesium is represented by the letter Ma.
D) The symbol for sulfur is K, and the symbol for magnesium is Mg.
E) The symbol for nitrogen is Az and the symbol for sulfur is K.

14.4 Are you sure about your answer to the previous question?
A) I am sure **B)** I am not sure

15.1 Formula

CO_2: Calcium oxygen
CO: Carbon monoxide
SO_2: Sulfur dioxide

Which of the compound formulas given above has an <u>incorrect</u> name?
A) SO_2 **B)** CO **C)** CO_2

15.2 Are you sure about your answer to the previous question?
A) Yes **B)** I am not sure

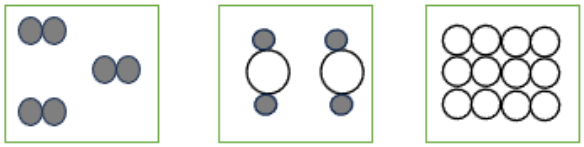
15.3 Why did you select the above option?

A) The name of the CO_2 compound is calcium oxide.
B) The name of the CO compound is carbon hydrogen.
C) The name of the CO compound is calcium oxide.
D) The name of the SO_2 compound is nitrogen oxide.
E) The name of the SO_2 compound is sodium dioxide
F) The name of the CO_2 compound is carbon dioxide.

15.4 Are you sure about your answer to the previous question?

A) I am sure **B)** I am not sure

16.1 The particle models of some matters are given below.



I II III

Which of the following can be said to be a compound?

A) I **B)** II **C)** III

16.2 Are you sure about your answer to the previous question?

A) I am sure **B)** I am not sure

16.3 Why did you select the above option?

A) In Figure 1, atoms of the same type form compounds.
B) In Figure 2, since there are at least 3 atoms together, it is a compound.
C) The atoms of different types in Figure 2 form a compound.
D) The same atoms in Figure 3 form compounds.
E) The atoms in Figure 3 are of the same type, so they form a compound.

16.4 Are you sure about your answer to the previous question?

A) I am sure **B)** I am not sure

JESEH Journal of Education in Science, Environment and Health
Volume 12, Issue 2, 2026 https://doi.org/10.55549/jeseh.871

Primary School Teacher Candidates' Perceptions and Experiences of Real-World Sustainability Problems in Education for Sustainable Development

Nur Utkur-Gulluhan

Article Info

Article History

Published:
01 April 2026

Received:
10 September 2025

Accepted:
22 February 2026

Keywords

SDGs
Sustainability
Real world problems

Abstract

Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs. In modern usage it generally refers to a state in which the environment, economy, and society will continue to exist over a long period of timeThis study explored the perceptions and experiences of primary school candidate teachers trained in Sustainable Development Goals (SDGs) regarding real-world issues. Using a phenomenological approach with focus groups and semi-structured interviews, this study aims to understand the impact of SDG-related education on these candidates. Data analysis involved thematic and content analysis, revealing that teacher candidates saw a direct link between SDGs and global issues, often sharing insights based on personal experiences. They highlighted the importance of integrating the SDGs into the curriculum, suggesting both compulsory courses and projects. The candidates also prioritized specific SDGs when designing activities that were personally significant. This small-scale study aims to inform future research on SDGs and real-world problem-solving in education.

Introduction

There are many problems related to the environment around the world. Every day, we all contribute to consumption, and unfortunately, production-oriented activities are not sufficient to balance this consumption. For this reason, various studies have been conducted on international platforms. The concept of sustainability has a broad content. Sustainability, in accordance with its current meaning and purpose, was first mentioned in the "World Nature Charter" document approved by the World Conservation Union in 1982 (Ruiz-Mallén & Heras, 2020). Sustainability is a type of development that concerns all citizens living worldwide. Here, citizens' respect for ecological life and behavioral changes come into play before politicians take official actions. United Nations Educational, Scientific and Cultural Organization (UNESCO), (1997) has encouraged present generations to take action for the permanent development and preservation of life, including preserving the quality and integrity of the environment and ensuring that future generations are not exposed to the environmental degradation that would endanger their health or existence. With the aim of addressing environmental, economic, and social problems, sustainability is a good example of the interconnectedness of systems, and shows that truly sustainable solutions can only be found by transcending the current limits set by traditional rules (Uiterkamp & Vlek, 2007).

The expansion of the scope of sustainability over time has also led to concrete steps being taken towards the idea of "sustainable development." In this context, combating climate change has become one of the key dimensions of sustainable development, since climate-related risks directly threaten ecosystems, social well-being, and economic stability. At the Paris Climate Summit in 2015, countries supporting the United Nations Framework Convention on Climate Change reached a consensus. The "Global Climate Agreement" was signed. United Nations Framework Convention on Climate Change [UNFCCC], 2015).

In the Sustainable Development Program, which is a continuation of the Millennium Development Goals and was organized in 2015, 193 countries that are members of the United Nations adopted 17 sustainable development goals to eliminate poverty in all its dimensions and promote the well-being of all humanity by 2030, as shown in Figure 1. As seen in Figure 1, priority will be given to efforts to end hunger across the world with the goals of "end poverty and hunger"; raising awareness, and protecting every individual outcomes disease used by harmful chemicals and air, water, and soil pollution until 2030 with the aim of "healthy and quality life"; with the goal of "quality education," the knowledge needed by all students to advance sustainable development through education for sustainable development and sustainable lifestyles, human rights, gender equality, promoting a culture of peace and non-violence, world citizenship and recognition of the contribution of cultural diversity and culture to sustainable development, and the outcomes of skills, creating and developing

educational opportunities sensitive to children, the disabled, and gender equality, and creating safe, non-violent, inclusive, and effective learning environments for all; defending women and girls everywhere from discrimination, aiming to eliminate gender inequality; to reduce inequalities, it is adopted as a principle that respects differences, eliminates languages, religions, races, genders, and ethnicities, and builds a livable world (UNDP, 2016). Overall, Sustainable Development Goals (SDGs) have emerged as universal goals that aim to support inclusive societies, combat inequalities, and recognize the importance of cooperation (Garcia et al. 2017).

Figure 1. United Nations 2030 sustainable development goals (United Nations Development Programme [UNDP], 2015).

Background

The theoretical framework of this study is constructivism. Constructivism is a theory of the nature of knowledge and is based on students' knowledge construction (Bodner, 1986; Brooks & Brooks, 1999; Fosnot, 2007; Hendry, 1996; Hendry, Frommer, & Walker, 1999; Hove & Berv, 2000; Philips & Soltis, 2005; Schunk, 2011; Zimmermann, Peschl, & Nossek-Römmer, 2010). One of the most important factors in the development of constructivism is the research and studies of Piaget, John Dewey and Vygotsky on human development (Schunk, 2011). Constructivism focuses on "cognitive development and deep understanding" rather than on behaviors and skills in teaching. It is based on the process of students being active in constructing knowledge in their minds. Piaget explains this process as cognitive equilibrium, which occurs when individuals develop, seek new knowledge, and encounter unfamiliar situations that challenge their existing understanding. In such cases, individuals strive to restructure their cognitive schemas in order to maintain the continuity of previously formed behaviors (Brooks & Brooks, 1999; Fosnot, 2007). Since constructivism is grounded in the assumption that subjective meanings are formed through individuals' experiences, it is compatible with the scope of this study. In this respect, university students' understanding of a more sustainable system, reconstructing this understanding by generating personal meanings, and integrating it into their daily lives can be explained through a constructivist perspective.

In line with this theoretical background, Education for Sustainable Development (ESD) education has increasingly emphasized the role of learning environments that enable students to actively engage with real-life problems and construct meaning through experience. A persistent focus of academics and practitioners in the field of Sustainable Development Education since the 1980s has been how education systems can prepare students to understand, participate in, and promote more sustainable forms of development. This interest gained momentum after the Brundtland Report (WCED, 1987), and the scope of sustainability expanded to include all levels of education, including higher education and undergraduate programs. When the basis of sustainable developments examined; it is seen that Elkington (1997) put forward the "three pillars" model (in Figure 2) and

suggested that economic development policies, environmental impacts and social consequences should be balanced with equal attention for sustainability (cited in McKenzie, 2004; WECD, 1987).

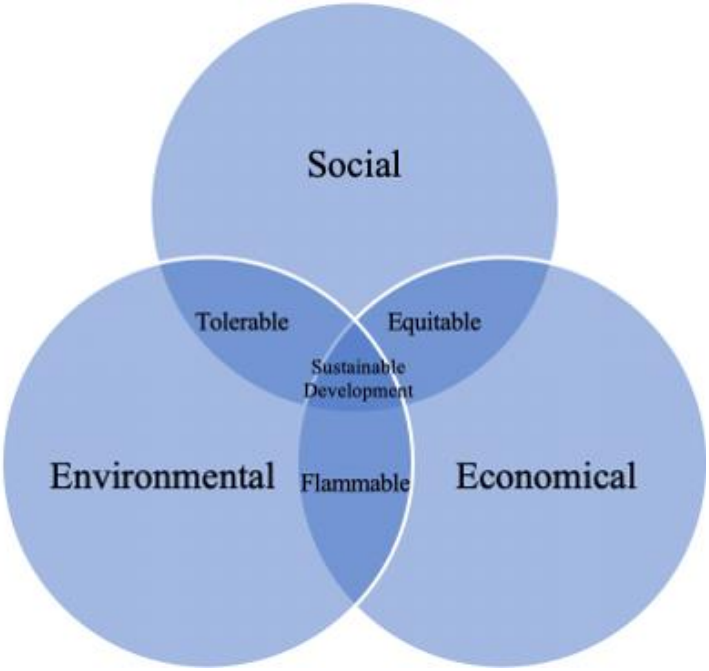


Figure 1. Three basic pillars of sustainability

While efforts have been made at various levels to promote sustainability in higher education, the integration of sustainability across traditional disciplinary boundaries remains insufficient to generate fundamental global change (McKenzie, 2004; WECD, 1987). As illustrated in Figure 2, sustainable development emerges from the balanced and integrated interaction of three interdependent pillars: social equity, environmental protection, and economic viability. However, these dimensions cannot remain merely theoretical constructs. Their realization requires concrete policy frameworks, institutional transformation, and educational initiatives that operate both top-down and bottom-up. In addition to state-led policies, awareness-based ESD initiatives targeting citizens, children, and university students—implemented individually and collectively—is essential to promote more equitable living conditions and to ensure the protection of natural resources (Franz, 2022).

The United Nations declared the years 2005-2014 as the Decade of Education for Sustainable Development. As the decade draws to more than 20 years after the Rio Summit, there is debate about how much progress has been made towards incorporating sustainability into higher education (Aktas, 2015). The important point here is the popularization of sustainability in higher education. This is because university students will be in our future. Various studies should be conducted by considering the opinions of university students on this subject using an integrative approach. University students play an active role in sustainable development because they have the potential to have an important mission as future leaders, decision makers, and shapers of society (Foguet et al., 2018).

Sustainability is a concept that includes the environmental, economic, and social dimensions. The actions that university students can take in these areas can raise awareness of education for a sustainable world. University students can contribute to creating a livable world by offering courses and workshops on sustainability. Rapid global developments have increased the interest of university students in sustainability concepts and education (Lozano at all, 2013). Studies have shown that more than 60% of university students want to learn more about sustainability, and 87% of all students agree that their university requires sustainability awareness. Global events such as Covid (19), climate change, and economic crises have shown that university students are sensitive to sustainable development (UNESCO, 2022).

Young people, who will change the fate of the future world, are also responsible for the sustainable state of the current world (Aleixo, Leal & Azeiteiro, 2021). In other studies in the literature on sustainability (Boca and Chan et al., 2017; Chuvieco et al., 2018; Dagiliūtė, Liobikienė & Minelgaitė, 2018; Karatzoglou Rieckmann, 2012; Saraçlı, 2019; Sibbel, 2009; Velazquez, Munguia & Sanchez, 2005; Wright, 2002) show that the

importance of sustainability has become more important among young people in order not to worry about the future.

As a matter of fact, some studies examining teacher candidates' perceptions and experiences regarding sustainability (Avsec & Savec, 2021; García-González, Jiménez-Fontana & Azcárate, 2021; García-Morís & Martínez-Medina, 2022) are also available in the literature. However, these studies were generally conducted with candidate teachers. This study aimed to examine the perceptions and experiences of candidate primary school teachers who received training on SDGs regarding real-world problems.

Although previous studies have examined teacher candidates' perceptions and experiences regarding sustainability (Avsec & Savec, 2021; García-González, Jiménez-Fontana & Azcárate, 2021; García-Morís & Martínez-Medina, 2022), these studies have generally focused on measuring attitudes or general sustainability awareness. In contrast, the present study investigates the perceptions and lived experiences of primary school teacher candidates who received structured training on the Sustainable Development Goals (SDGs) and engaged with real-world sustainability problems. By focusing on how teacher candidates interpret and internalize sustainability concepts after targeted instruction, this study moves beyond descriptive perception studies and explores the transformative dimension of ESD.Primary school teachers play a foundational role in shaping children's early understanding of social, environmental, and economic issues. Therefore, examining how future classroom teachers construct meaning around sustainability and climate-related challenges is critical, as these interpretations will influence their future pedagogical practices. By adopting a qualitative and experience-based perspective, this study contributes to the literature by providing deeper insight into how sustainability education can inform teacher identity formation and classroom practice.

Teachers in primary schools, where children receive their first education after their families, are very important to them. Together with these teachers, they can shed light on both the life and world in which they live. Therefore, primary school teacher candidates' experiences and perceptions regarding sustainability and climate are important, so that they can convey these to their students when they become teachers. Thus, the opinions of classroom teacher candidates who received this training will be examined, and it is thought that this will contribute to the literature and this important problem. The research problems to be addressed for this purpose are as follows.

The purpose of this study is to explore the perceptions and lived experiences of primary school teacher candidates regarding real-world sustainability problems within the framework of Education for Sustainable Development (ESD), and to examine the curriculum elements and learning activities they propose based on these experiences.

This study seeks to answer the following questions:

1. How do primary school teacher candidates who have received training on the SDGs perceive and experience real-world sustainability problems?
2. Based on these perceptions and experiences, what curriculum elements and learning activities do they propose within the framework of sustainability?

Method

Research Model

This research was conducted using phenomenology, a qualitative paradigm design. This design aims to reveal common practices and describe and explain the meanings created by the participants (Annells, 2006). To understand social reality, the focus is on the human experiences through which social reality is created (Ersoy, 2019). In this design, "the individual's experience and how he experiences it' is determined by generally reducing a universal situation to individual experiences (Moustakas, 1994; *cited in* Ersoy, 2019).

This research focuses on the perceptions and lived experiences of prospective teachers regarding real-world sustainability problems within the framework of ESD. To understand the effects of ESD-related education that teacher candidates received at some point in their lives, a phenomenological design including focus groups and semi-structured interviews was adopted. In other words, it was aimed to obtain the thoughts of the students who received education on Sustainable Development after this experience and to understand what can be done about

this issue in the future through their explanations. The research process within the scope of this study is shown in Figure 3.

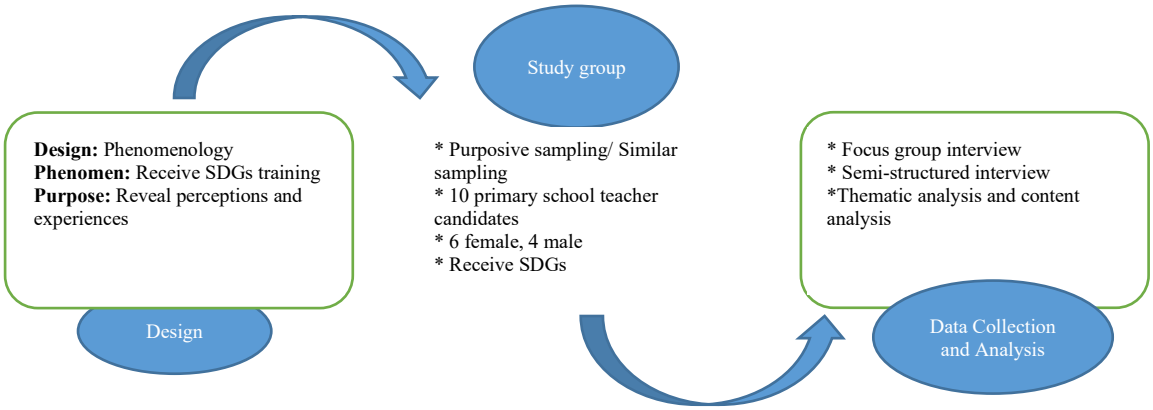


Figure 3. Research design

Study Group

To determine the participants in the study, a similar sampling method, purposive sampling, was adopted (Patton, 1987). Homogeneous purposive sampling, one of the purposive sampling strategies, was employed to determine the participants of the study (Patton, 1987). The aim of this sampling method is to identify a specific subgroup by forming a small and relatively homogeneous sample. Although the number of participants in phenomenological research may vary depending on the phenomenon under investigation, Creswell (2007) indicates that it typically ranges between 5 and 25. In line with these considerations, 10 participants were selected. The homogeneity criterion guiding participant selection was that they were enrolled in the primary school teaching program within the faculty of education and had previously received structured training related to sustainability within the framework of Education for Sustainable Development (ESD). Table 1 presents the general characteristics of the participants.

Table 1. General characteristics of the participants

		(f)
Grade level	3rd grade	5
	4th grade	5
Gender	Woman	6
	Man	4

Five of the participants were studying in the 3rd grade and five in the 4th grade in the department of primary school teaching; six of them were female and four were male. All of them had received sustainability training at some point in their lives. Permission for the study was obtained from the University of Social and Humanities Ethics Committee. In addition, prospective classroom teachers who volunteered to participate in the study were recruited and a consent form was obtained from each.

Data Collection Tools

Focus group interviews were used as data collection tools. According to Patton (1987), data can be effectively collected from groups determined by homogeneous sampling through focus group interviews. First, teacher candidates who wanted to participate in the research explained how the process would work, and ethical principles were observed by having volunteers sign a consent form. Then, a focus group interview was held with the teacher candidates, and the process was examined in detail through semi-structured interviews. The researchers applied these data-collection tools.

The focus group interview lasted 3.5 hours. It took place in 1.5 and 2 hour sessions. After the focus group interviews were transcribed, semi-structured interviews were conducted with each participant separately. Semi-structured interviews took place on a predetermined day for periods ranging from 30 to 50 min. Through semi-structured interviews, the students' opinions, which remained superficial in the focus group discussion,

deepened. In addition, after these interviews, each participant was given 30 min and asked to reflect on their activities.

Semi-structured interview questions were prepared based on the themes that emerged from the focus group discussion. In semi-structured interviews, prospective teachers were asked questions regarding both the general themes obtained from the focus group interviews and the candidate's specific situation.
The questions generally address the following:

- what candidate's perception of sustainability is,
- what they think about SDGs,
- their perceptions of real life problems,
- perceptions and opinions about integrating the SDGs into the education faculty curriculum,
- their views on activities that can be developed to achieve SDGs and solve real-life problems.

After determining the questions, opinions were received from six experts, three of whom were experts in primary school teaching and three experts in the field of sustainability. The final version of the data collection tool was developed based on their opinions.

Data Analysis

The data obtained from the focus group and semi-structured interviews were analyzed using thematic analysis and content analysis (Patton, 1987). The data analysis technique proposed by Moustakas (1994) was taken as the basis for these analyses. In this type of analysis proposed by Moustakas (1994), both types of analysis are used in an integrated manner, intertwined in a way that supports each other.

Accordingly,
• Identifying significant statements,
• Identifying common expressions,
• Thematizing meaning clusters,
• Creating structural and textural descriptions,
• Combining structural and textural descriptions were followed.

The data obtained from the focus group interviews were presented in themes by combining common expressions in order to understand the phenomenon and identify significant statements. Subsequently, as a result of the training they received regarding the SDGs, how the participants structurally formed perceptions of real-life problems and what was experienced texturally were analyzed. The presentation of the findings was also based on the semi-structured interview data, including the focus group interview data.

Validity and Reliability

The data obtained from the focus group discussions and semi-structured interviews were first transcribed and read several times. Words, sentences, and paragraphs in the data were then identified and marked for coding. Important and common expressions were identified. After the researcher coded for this process, two field-expert professors working at a state university were asked to code separately. Encoder comparisons were performed twice to ensure reliability. The two experts' codings were compared first among themselves and then with the researcher's coding, and Miles and Huberman's (1994) Validity and Reliability formula were applied. The details are presented in Table 2.

Table 2. Reliability coefficients

	Reliability value among experts	Final reliability value between researcher and experts
First problem question	0.90	0.92
Second problem question	0.92	0.94

According to this formula, to which coder reliability was applied in the analysis of the first research problem, the final reliability coefficient was 0.92. For the second research problem, this value was found to be 0.94. Therefore, according to Miles and Huberman (1984), if this value is above 0.70, the analysis of data collection tools can be considered reliable. Then, different codings were agreed upon by the experts. Finally, all coding

was divided into categories, and themes were created.While quotes from the teacher candidates' sentences were included, code names matching the initials of their names were given. Thus, their names were kept confidential. Additionally, to ensure credibility, participant confirmation was obtained after the codes, categories, and themes were determined, and teacher candidate quotes were used.

Findings

As a result of the data obtained from the focus group discussions and semi-structured interviews, the themes in Figure 4 reflect the findings.

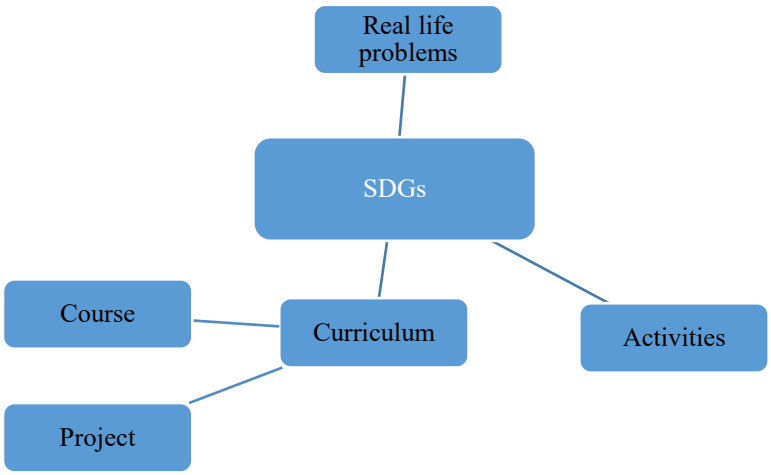


Figure 4. Themes related to SDGs

Real Life Problems Theme

As a result of the analysis, it was seen that there are real-life problems among the themes stated by teacher candidates regarding SDGs. The SDGs generally mentioned by prospective teachers and their sub-themes are listed in Table 3.

Table 3. Real-life problems faced by people related to SDGs

Main themes	Subthemes	(f)
Quality Education	Lack of differentiated education	5
	Problems in girls' education	5
	Problems in early childhood education	4
Climate Action	Climate crisis and drought	5
No Poverty	Rapid population growth	4
	Infectious diseases	3
	Inequality in economic resources	3
Zero Hunger	Mother-child deaths	4
Clean Water and Sanitation	Unconscious use of water resources	4
	Chemical waste pollutes water	3
Gender Equality	Lack of clarification of women's role	4
	Gender inequality	3
Peace, Justice and Strong Institions	Continuation of wars	4
Sustainable cities and communites	Lack of safe playground	3
	High number of regional migrations	2

When the opinions of teacher candidates were examined in general, they expressed their opinions regarding the eight SDGs and related them to real-life problems. “Quality education” is among the SDGs that are mentioned first.

> ***Elliott:*** *...Even though we are in the 21st century, the problem of girls' inability to benefit from education and training and differentiated education in line with the students' wishes and needs still persists in every country. This situation really makes people sad.*
>
> ***Robert:*** *There are problems in pre-school and early childhood education. Education is the basis of everything, but we cannot provide this foundation. If we cannot provide pre-school education, we will experience a fundamental collapse. In addition, individualized and differentiated education is the most important point, for example. We see these all over the world.*

Teacher candidates generally mentioned problems in the education of girls, differentiated and individual education problems, and problems in the education of pre-school age children among the problems related to education. They stated that they experienced these in their daily lives and that every country had a general problem. Regarding the "Climate Action" target, one of the candidate teachers said the following:

> ***Elliott:*** *It is aimed that all states strive to achieve these goals and eliminate many problems that threaten the world. However, in line with the statements made, it is stated that we are still far from achieving these goals. In particular, the world is faced with many problems such as the climate crisis, the impact of which we have felt more in recent years, and the resulting drought, regional migrations, irregular weather events and natural disasters. We already see this in the world we live in. We are getting worse every day...*

With this view, Elliott addressed the global climate crisis with his own experiences and referred to problems in daily life. Another stated goal is "No Poverty". The opinions expressed regarding this include the following.

> ***Elenor:*** *The environment where children live should be healthy in every aspect. There must be a decent environment in many areas such as living, shelter, nutrition, marriage age, gender equality, education, health and psychology. Attention was drawn to the regions of Central and Eastern Europe and Central Asia. There are much more visible unequal and unfair living conditions in these regions, and the consequences of this situation on children are stated. Problems with rapid population growth, infectious diseases and economic resources continue wherever we live.*
>
> ***Hans:*** *It is called no poverty, but there is no end state. Rapid population growth gives rise to infectious diseases and everything is getting worse day by day...*

In her views, Elenor generally mentioned the problems regarding rapid population growth, infectious diseases and economic resources experienced in the world and in the country, in her views under the goal of "No Poverty" from the SDGs. On the other hand, Hans stated that he complained that he lived in a world where poverty did not end. The opinions stated by teacher candidates regarding the "Zero Hunger" goal are as follows:

> ***Robert:*** *Although one of the SDGs prepared by the United Nations is to end hunger, tons of food are wasted every year in the world. While there is enough food for all of us, food is wasted and some people continue to go hungry.*
>
> ***Bailey:*** *I think the fact that more than half of the world is struggling with hunger shows that we cannot meet this goal. I am really sad. And because of this, mother-child deaths are increasing day by day. I also lost a relative this way.*

Robert and Bailey were very emotional as they shared their views and experiences. This shows that they too have experienced similar processes and have internalized this situation quite a bit. They stated that hunger is a serious global problem. Again, Bailey regarding "Clean Water and Sanitation, "*"Although there are many clean energy production methods in the world, we see that energy production methods that will harm the world are used and developed. Instead of helping people who still follow outdated ways to access clean water, efforts are being made to extract water on the Moon and Mars. I think we still have a lot of shortcomings in this regard." he said.* Robert, regarding "Gender quality"; *"In our society, the position of women in the workplace, at home and in all other areas of life is still not clear. For example, my mother was mobbed at work the other day just because she was a woman."* he said. Elliott expressed her views on "Peace, Justice and Strong Institutions" as follows: *"Wars, which have covered the entire history of the world, continue today, and unfortunately, this situation makes it impossible to live in peace."* Again, Robert said the following regarding "Sustainable Cities and Living Spaces": *"When we look at real life, we can see that the buildings are built side by side, there are no*

safe parking areas for children, and there are few large and wooded areas where people can get fresh air and spend time with their families. It is very difficult to say that this system, where people are stuck in boxy houses, creates a sustainable living space. And now I am starting to lose hope, sir."

When the views of Bailey, Elliott, and Robert regarding various SDGs were examined, it was observed that they addressed numerous real-world problems. Overall, their perceptions and experiences reflect a world characterized by negative conditions and ongoing challenges. These conditions appear to evoke feelings of concern, frustration, and, in some cases, hopelessness, particularly when participants refer to issues such as war, gender inequality, and unsustainable urban environments.

Curriculum Theme

As a result of the interviews, data on what teacher candidates would put into the curriculum to solve sustainability and real-life problems if they were the ones who prepared the curriculum applied in schools are given in Table 4.

Table 4. Themes for what could be added to the curriculum

Mean themes	Subthemes	(f)
Project	Should be compared with real life problems	10
	Must be supported with trips	9
	Must be supported with games	8
	Activities supported by drama should be carried out	8
	The concept of visual literacy should be brought to the fore	7
	Seminars and social solidarity events should be added	7
Compulsory Course	Students should be provided with active activities	6
	It should be aimed at gaining skills-attitude-value	5
	A sense of curiosity and creativity should be aroused in students	5
	SDGs should be given under book reading activities in preschool period.	4
	SDGs should be included in a coordinated manner in primary, secondary and high school lessons.	4
	It should be based on problem-based learning	3
	Should be integrated into science courses	3

As seen in Table 4, there are two main themes that should be included in the curriculum regarding the SDGs. The first of these is "Getting a project done under the name of SDGs." The other is, "SDGs should be a compulsory course at every school level." Candidate teacher opinions regarding these issues were included.

Project

When the opinions of teacher candidates are examined in general, it is seen that the opinions that want to add the SDGs to the curriculum as a project are at the forefront. It was observed that all prospective teachers stated that using projects would enable them to compare students with real-life problems. They stated that they could solve these problems by supporting themselves with various trips, games, drama activities, visual activities, seminars, and social solidarity activities. The opinions of the candidate teachers who expressed their opinions on these issues are as follows:

> ***Teresa**: Studies on the importance of sustainability, the cause and consequences of resource insufficiency, and what the state of our environment could be if we use resources regularly, cleanly and reproducibly can be done as a project. I would also add the benefits of practical internships related to sustainability in various organizations to the program. I would aim for students to acquire the ability to collaborate and carry out projects in groups.*

> ***Adam:** If we can instill these issues in teacher candidates, they can instill these goals in the younger generation when they graduate, and thus awareness will be created. You can think of this as a train consisting of locomotives. We may not be able to achieve a result by just taking one course at the faculty of education. Because it is not a compulsory course and remains an elective, many trained teachers may remain unaware of this project. For this purpose, prospective teachers can be informed by giving seminars and awareness can be created by*

printing the magazines and articles published around the world for this project and distributing them to students. By establishing a club, social solidarity can be achieved among students, faculty members and families.

Bailey: *The project I would prepare would be as follows: I would organize a trip to an energy production center (dams, thermal resources, solar panels, windmills, etc.), if there is one in the province we are in. I would like them to examine the items they use all around them. We would chat about how they worked, and I would have users calculate how much electricity these products consume in an hour, a day, and a month. In short, I would try to raise awareness about energy.*

Teresa, Adam, and Bailey included in their comments the project proposals that they wanted to do regarding sustainability. For example, according to Adam's proposal, the teacher would be the head of the train and the students would be its engines. Thus, the transfer of knowledge and experience continues. Bailey also discussed the importance of raising awareness by seeing and experiencing on-site in his project, which he aims to have done through trips to energy production centers. Teacher candidates suggested these projects to solve their perceptions of the problems encountered in their daily lives. What they all have in common is to gather society and students in one place and try to raise awareness of the SDGs through various activities and events.

Compulsory Course

When teacher candidates' opinions were examined, it was observed that they attached importance to the SDGs being reflected as a compulsory course at every school level. Some of their opinions on this are given below:

Elliott: *These goals should be included in primary, secondary and high school courses in a coordinated manner. In art class, you can have a painting or sculpture work on life in water, in music class, you can have a musical instrument made from waste materials, and in life sciences class, you can have a drama work on hunger. Problem-based learning, project-based learning, etc. I would focus on activities where methods could be used effectively. In this process, I would especially organize the curriculum in a way that would enable students to produce solutions to the problems they encounter in real life by adding them to the curriculum.*

Hans: *We can include activities in our lessons based on sustainability-related achievements in science lessons. For example, we can have our students go out to the school garden and pick up the garbage on the ground during a free activity. We can have art exhibitions related to science. A club related to sustainability can be established. Regarding recycling, waste separation activities can be carried out.*

Bailey: *I believe that in order for teacher candidates to handle these issues, they must first experience them and produce solutions themselves. For this reason, first of all, the information learned in the compulsory course; I recommend that it be reinforced with activities outside the classroom. For example; Vineyard, garden and nature trips for environmental awareness; I would like them to visit their homes and neighborhoods about recycling and examine the living things in these environments they live in. Thereupon, I would design and have activities designed for recycling and for those creatures to see what they should pay attention to and what they could do. Also, during these trips, I would ask people living on the streets and to examine the garbage bins closest to their homes. I would ask if food was thrown away. I would ask those people living on the streets what they need most. I would try to point out how close waste and hunger are.*

Bella: *I would include a course titled "SDG" in the curriculum. I would set it up so that one goal was covered each week. I would take care to teach interactive lessons based on questions such as: What are the problems encountered in the world related to the target, what has been done to solve these problems, and what would you do differently? I would divide the students into groups, give each group a goal, and ask the groups to do extensive research on their goals. I would organize trainings to raise public awareness by ensuring cooperation between Education Faculties and necessary places.*

When the opinions of teacher candidates are examined, it is seen that, based on their own experiences and perceptions, they recommend that development goals be taught as a course in the curriculum at almost all levels. Elliott's idea, which suggests that these objectives should be included in each lesson in a coordinated manner, is quite interesting. Curricula are used in a spiral and intertwined manner. The aim was to include SDGs in these programs. Hans mentioned that the compulsory course should be integrated into the science course because it is already among the achievements in the content of this course. On the other hand, Bailey talked about the importance of candidate teachers experiencing and coming up with solutions themselves. She believes that the issues experienced and perceived on-site will have a greater impact. Bella also stated that collaborating with other institutions and organizations within the scope of the course is important for raising public awareness.

Activities Theme

In the study, after the interviews, participants were given additional time and asked to create activities related to real-life problems related to SDG. One of them, Teresa, suggested that an event could be held regarding the "Zero Hunger"'s goal.

> ***Teresa:*** *A group of 10-12 people gathers and seeds of plants, flowers and legumes from different parts of the country are collected. It should be a priority for the seeds to be productive and natural seeds, also known as 'ancestral seeds'. A seed bank should then be obtained with the collected seeds. The seed bank should be detailed by separating all seeds with great care and grouping only those with the same seeds. In the first stage, farming should be started using only a small portion of the seeds. The remaining seeds should be stored in a sheltered place as a precaution against disasters such as war, famine, fire and flood, taking into account the storage conditions. Seed collection centers can be used in certain parts of cities to prevent the seeds of the food produced by the seeds used in the first place from being thrown away. These centers must also comply with the seed storage conditions and be hygienic. In order to encourage people, seeds of foods they do not have can be given as gifts in return for the seeds they bring. Thus, both seed productivity and diversity will increase, especially in agricultural regions.*

Hans expressed her views regarding the goal of "Quality Education" as follows:

> ***Hans:*** *In my opinion, a qualified teacher is at the beginning of a quality education. That's why I would organize events involving children. For example, there would be schools with which the university has an agreement. A science festival can be organized in one of the schools. Teacher candidates can also take part in this festival. Can work with children and learn with them. He/she will both experience the teaching profession and become more qualified with these experiences. In this way, I think we will take another step towards quality education.*

Robert and Bailey suggested that the following activities could be conducted regarding the goal of "Gender Equality".

> ***Robert:*** *People with whom the activity will be held are asked to take note of the situations of gender inequality they encounter in real life, in their family lives, on social media, in the books and news they read, and in the TV series and movies they watch, for a day or a week. Drama work is performed by selecting one or more of these situations. In the interim evaluations made throughout the drama work, thought-provoking questions are asked about what people who are exposed to gender inequality, who do this, and who have witnessed this situation feel and what they should do.*
>
> ***Bailey:*** *I would prepare dramatic situations in which they could put themselves in the place of the opposite sex in social life and ensure that each of the actors played the opposite sex. Thus, I would try to make them empathize with the opposite sex. I would like them to think about how they can impart what they have learned here to children and prepare activities for this. Thus, they can understand gender equality.*

> ***Bailey:*** *You have seen many times that the water in the seas is dirty. Garbage thrown unconsciously can pollute sea waters. For this purpose, the following activity can be done to raise awareness and sensitivity towards the people around: Under the leadership of one person, a small address can be made to the people sitting on the beach. After talking about sustainability, the garbage on the beach is collected together. There is lifelong learning here. Because there is a 7-year-old girl on the beach and a 40-year-old aunt who will collect the garbage. As is known, there are hunting bans for aquatic life. Individuals should be sensitive to these prohibitions. One should not fish during the prohibited time. For this reason, we must explain this to individuals at a young age and raise their awareness. Because hunting bans are a measure taken to protect our seas and fish.*

Elliot stated that an event could be held regarding both of the "Climate Action and Life on Land" targets.

> ***Elliot:*** *We can have an activity based on multiple intelligence types. In this way, we also use individualized teaching. After each intelligence group did its own work, their work was opened by the teacher on "Instagram, YouTube, blogs, etc." It can be published on the platforms with the permission of the parents. Students were asked: "What can be done to raise people's awareness about the climate crisis and the problems occurring in terrestrial life and to solve the problems?" The question is asked. Students are divided into groups according to the multiple intelligence method. Each student is in the group he thinks is dominant.*

When we look at the activities suggested by teacher candidates, it can be seen that they generally design activities that can make students active and aim to solve real-life problems related to these SDGs based on their experiences. It was noted that candidate teachers said that they wanted to use these activities in their own classrooms when they became teachers.

Results and Discussion

This study aimed to examine the perceptions and experiences of candidate teachers who received training on SDGs regarding real-world problems. When the topic of sustainability first emerged in the interviews, prospective teachers started to express their opinions starting from real-life problems. This situation was examined in the first theme of this study. Teacher candidates stated that SDGs have a direct relationship with global problems. They also expressed problems arising from this by using their personal experiences. When the opinions of teacher candidates are examined in general, it is seen that they express their opinions about the goals of "Quality education, climate action, no poverty, zero hunger, clean water and sanitation, peace-justice-strong institutions, sustainable cities and communities." In general, their experiences and perceptions are that they live in a world that is full of negative conditions. This situation makes them feel bad and they express that they are unhappy.

According to a study conducted by Students Organizing for Sustainability [SOS] (2021), it is generally stated that sustainable development requires all universities and colleges to be actively involved and promoted. They also stated that the concept of sustainable development was given little or no place in the course curriculum, and that they were disturbed by this. When asked to describe their feelings about climate change and their future, they said that they were worried about. Emanuel and Adams (2011) state in their study that university students are concerned about sustainability and want to volunteer in sustainability projects. Therefore, in the current study, it is understandable that candidate teachers were generally anxious. However, university candidates want to do something to solve real-life problems, both in this study and in other supporting studies.

As a result of their experiences, candidate teachers stated that the world is under some negative conditions regarding sustainability. Their suggestions focus on the development of activities integrated and parallel to the curriculum. "Curriculum" is considered as the second theme, and the sub-themes are "project and compulsory

course.” They also talked about integrating the SDGs into the curriculum. When teacher candidates' opinions are examined, it is seen that opinions that want to add the SDGs to the curriculum as a project are at the forefront; It was observed that they stated that using projects would allow students to compare real-life problems. They stated that they could solve these problems by supporting them with "various trips, games, drama activities, visual activities, seminars, and social solidarity activities". Teacher candidates suggested these "projects" to solve their perceptions of the problems they encountered in their daily lives. The common point of view is to gather society and students at a single point and try to raise awareness of the SDGs through various activities and events. This may indicate that students are searching for solutions to their experiences in their own lives. In addition, when the opinions of prospective teachers are examined, it is seen that in the second sub-theme, based on their own experiences and perceptions, they recommend that development goals be taught as a "compulsory course" in the curriculum at almost all levels. Curricula are used in a spiral and intertwined manner. The aim was to include SDGs in these programs. They also stated that collaborating with other institutions and organizations within the scope of the course is important for raising public awareness.

In fact, when looking at the literature, the sustainability factors defined in Menon and Suresh's (2021) study support the views of candidate teachers in this study. These are: "adding sustainability courses to the curriculum, adopting student-centered, interactive, participatory learning approaches, supporting interdisciplinary studies, education based on innovative pedagogy, teacher training, leadership, incorporating sustainability perspectives and values into the curriculum, networking skills, change and transformation skills, being able to work with different stakeholders, cooperation between academics, university-industry collaborations, community programs and projects, research on social problems, holistic approaches, interdisciplinary research; "increasing leadership capacity, increasing public awareness, dissemination of information, use of social media. “Therefore, the project and compulsory course suggestions that teacher candidates consider integrated and parallel to the curriculum are supported. In addition, the opinions emerging from the interviews coincide exactly with Menon and Suresh's (2021) efforts to integrate sustainability factors into higher education.

For example, let us draw attention to the following sentence among the opinions of a teacher candidate*: "Achievements such as sustainable living, economical use of resources and recycling are included in the program." We can include activities in our lessons based on these achievements. Another teacher candidate said; “Studies on the importance of sustainability, the cause and consequences of resource insufficiency, and what the state of our environment could be if we use resources regularly, cleanly and reproducibly can be done as a project. “I would also add the benefits of practical internships related to sustainability in various organizations to the program.”* It supports these views; Yıldız et al. (2021) also stated in their study that providing university students with training on environmental problems, developing projects to raise environmental awareness and creating a sustainable natural environment will make positive contributions to the development of individuals' attitudes towards the environment. There are also studies indicating that it would be appropriate to present sustainability-based environmental education both as a separate course and as intertwined with other courses (Alım, 2006; Stokes, Edge, & West, 2001; Tanrıverdi, 2009).

Sustainable environmental education can be associated with students learning the importance of the environment, correcting mistakes regarding environmental problems, and implementing correct practices. It is thought that students receiving effective environmental education for a sustainable future will enable them to have a correct environmental perspective and develop attitudes towards environmental problems by transforming this perspective into behavior and becoming role models for future generations. When the relevant literature is examined, it is possible to come across studies in which students receiving education on environmental sustainability improve their attitudes towards environmental problems. (Gallagher et al., 2000; Larijani & Yeshodhara, 2008; Raut & Pendse, 2013; Tanrıverdi, 2009; Kayalı, 2010; Yıldız, et al. 2021).

In this study, it was observed that teacher candidates prepared activities by prioritizing the SDGs that they thought were important in their questions regarding activity suggestions for solving real-life problems. According to these activities, students at all levels and every person in daily life stated that they should be mobilized, raise awareness in some way, and protect our world. This situation shows that candidate teachers make an effort despite everything and is pleasing. When we look at the activities suggested by the candidate teachers, it is generally seen that they design activities that can make students active, aim to solve real-life problems related to these SDGs based on their experiences, and are based on the principle of near to far. To give an example; “creating seed banks, organizing science festivals where children can be together, more teacher support to regions in difficult situations, identifying and discussing gender inequalities in social media and personal life, empathy and drama studies, collecting garbage on the beach together with all the people and multiple intelligence for climate action.” studies, ” are among them. What is meant by the principle of near to far here is that it is necessary to start by taking advantage of elements close to the child's own life and to reach

those far away and the whole world. Tanrıverdi (2009) states that this issue should be given importance. Again, this supports these results; Evangelinos et al. (2009) showed that university students stated that courses and activities at the academy were important in protecting the environment and nature. However, they stated that they expected support from universities and other organizations to be sufficiently active.

Conclusion

As a result, when the opinions and activities stated by the candidate teachers based on their own experiences are examined, it is seen that they have an accumulation of knowledge towards sustainable development. However, they are also aware of real-life problems experienced in our country and the world. Therefore, it is obvious that they have some suggestions and things that they want to do regarding these issues. In this regard, all the humanities, especially educators and researchers, have great duties.

This study is limited to teacher candidates in the city center of Istanbul, which is the most cosmopolitan and most populated province in Turkey and therefore has all kinds of environmental and sustainability problems. With its crowdedness and constant immigration from different countries, it brings globalization and various environmental problems. In this context, the study results can be generalized to similar situations in similar countries.

Recommendations

Based on the results of this research, the following can be suggested for future studies:

- Teacher candidates focused on certain goals in the real-life problems and activities section. However, there are also SDGs other than those set by the United Nations. For these, events can be prepared and various studies can be conducted.
- In this study, two parts–courses and projects–are mentioned in the process of integrating the SDGs into the curriculum, and brief examples are provided. In future studies, this can be addressed specifically for various courses, and if there is a truly integrated goal, its impact on future students and humanity can be examined.
- In this study, various activities were prepared by considering the SDGs. Future studies may include the application of these activities to the classroom environment. Thus, their impact on students can be examined using qualitative or quantitative research methods.

Scientific Ethics Declaration

* The author declares that the scientific, ethical, and legal responsibility of this article published in the JESEH journal belongs to the author..

*The author acknowledged that all ethical rules were followed in the study.

Conflict of Interest

* The author declares that there is no conflict of interest.

Funding

* This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgements or Notes

* The authors would like to thank the conference scientific committee and referees for their feedback on the article.

References

Aktas, C. B. (2015). Reflections on interdisciplinary sustainability research with undergraduate students. *International Journal of Sustainability in Higher Education, 16*(3), 354-366.

Aleixo A. M., Leal S., & Azeiteiro, U. M. (2021) Higher education students' perceptions of sustainable development in Portugal. *J Cleaner Prod*, 327, 1-15. https://doi.org/10.1016/j.jclepro.2021.129429

Alım, M. (2006). Avrupa birliği üyelik sürecinde Türkiye'de çevre ve ilköğretimde çevre eğitimi. *Kastamonu Eğitim Dergisi, 14*(2), 599- 616.

Annells, M. (2006). Triangulation of qualitative approaches: hermeneutical phenomenology and grounded theory. *Journal of Advanced Nursing, 56*(1), 55-61. https://doi.org/10.1111/j.1365-2648.2006.03979.x

Avsec, S., & Ferk Savec, V. (2021). Pre-service teachers' perceptions of, and experiences with, technology-enhanced transformative learning towards education for sustainable development. *Sustainability*, *13*(18), 10443. doi: https://doi.org/10.3390/su131810443

Boca, G. D., & Saraçlı, S. (2019) Environmental education and student's perception, for sustainability. *Sustainability, 11*, 1553. https://doi.org/10.3390/su11061553

Bodner, G. M. (1986). Constructivism: a theory of knowledge. *Journal Chemical Education, 63* (10), 873-878.

Brooks, J. G., & Brooks, M. G. (1999*). In search of understanding the case for constructivist classrooms*. Virginia: Association For Supervision And Curriculum Development.

Chan, C. K. Y, Fong E. T. Y., & Luk L. Y. Y. (2017). A review of literature on challenges in the development and implementation of generic competencies in higher education curriculum. *Int J Educ Dev*, 57, 1-10. https://doi.org/10.1016/j.ijedudev.2017.08.010

Chuvieco, E., Burgui, M., & Silva, E. (2018) Factors affecting environmental sustainability habits of university students: intercomparison analysis in three countries (Spain, Brazil and UAE). *J Cleaner Prod., 198,* 1372-1380. https://doi.org/10.1016/j.jclepro.2018.07.121

Creswell, J. W. (2007). *Qualitative inquiry & research design: choosing among five approaches (3rd edition).* Thousand Oaks: Sage.

Dagiliūtė, R., Liobikienė, G., & Minelgaitė, A. (2018). Sustainability at universities: students' perceptions from green and non-green universities. *J Cleaner Prod., 181*, 473-482. https://doi.org/10.1016/j.jclepro.2018.01.213

Emanuel, R., & Adams, J. N. (2011). College students' perceptions of campus sustainability. *International Journal of Sustainability in Higher Education*, *12*(1), 79-92.

Ersoy, A. F. (2019). Fenomenoloji. A. Saban ve A. Ersoy (Ed.), *Eğitimde nitel araştırma desenleri* içinde (s. 81-139). Anı.

Evangelinos, K. I., Jones, N., & Panoriou, E. M. (2009). Challenges and opportunities for sustainability in regional universities: a case study in Mytilene, Greece. *Journal of Cleaner Production*, *17*(12), 1154-1161.

Fosnot, C. T. (2007). *Oluşturmacılık: teori, perspektifler ve uygulama.* S. Durmuş (Trans.). Ankara: Nobel Yayın Dağıtım.

Franz, N. (2022). *The technology for a fruitful future? analysing un policies on blockchain for sustainable development* [Unpublished Master's Thesis]. Malmö University.

Hendry, G. D. (1996). Constructivism and educational practice. *Australian Journal of Education, 40* (1), 19-45.

Hendry, G. D., Frommer, M., & Walker, R. A. (1999). Constructivism and problem-based. *Journal of Further and Higher Education, 23*(3), 359-371.

Howe, K. R., & Berv, J. (2000). Constructing constructivism, epistemological and pedagogical. In D. C. Phillips (Ed.), *Constructivism in Education: Opinions and Second Opinions on Controversial Issues*. The University of Chicago Press.

Gallagher, J., Wheeler, C., Mcdonough, M., & Namfa, B. (2000). Sustainable environmental education for a sustainable environment: lessons from thailand for other nations. S. Belkin (Ed.), in *Environmental Challenges.* Springer.

Garcia J., da Silva, S. A., Carvalho, A. S., de Andrade Guerra, J. B. S. O. (2017). Education for sustainable development and its role in the promotion of the sustainable development goals. J. Davim (Ed.), in *Curricula for Sustainability in Higher Education Management and Industrial Engineering (*pp. 1-18). Springer. https://doi.org/10.1007/978-3-319-56505-7

García-González, E., Jiménez-Fontana, R., & Azcárate, P. (2020) . Education for sustainability and the sustainable development goals: pre-service teachers' perceptions and knowledge. *Sustainability, 12*, 7741. doi:10.3390/su12187741.

García-Morís, R., & Martínez-Medina, R. (2022). Trainee teachers' perceptions of socio-environmental problems for curriculum development. *Soc. Sci. 11*(10), 445. doi: https://doi.org/10.3390/socsci11100445

Karatzoglou, B. (2013). An in-depth literature review of the evolving roles and contributions of universities to education for sustainable development. *J Cleaner Prod, 49,* 44-53. https://doi.org/10.1016/j.jclepro.2012.07.043

Kayalı, H. (2010). Sosyal bilgiler, türkçe ve sınıf öğretmenliği öğretmen adaylarının çevre sorunlarına yönelik tutumları. *Marmara Coğrafya Dergisi, 21,* 258-268.

Larijani, M., & Yeshodhara, K. (2008). An empirical study of environmental attitude among higher primary school teachers of India and Iran. *Journal of Human Ecology, 24*(3), 195-200.

Lozano, R., Lukman, R., Lozano, F. J., et al. (2013). Declarations for sustainability in higher education: becoming better leaders, through addressing the university system. *J Cleaner Prod, 48*, 10-19. https://doi.org/10.1016/j.jclepro.2011.10.006

McKenzie, S. (2004). Social sustainability: towards some definitions. *Hawke Research Institute Working Paper Series, 27*, 1-29.

Menon, S., & Suresh, M. (2021). Modelling the enablers of sustainability in higher education institutions. *Journal of Modelling in Management.* https://doi.org/10.1108/JM2-07-2019-0169

Miles, M. B., & Huberman, A. M. (1994). Q*ualitative data analysis: An expanded sourcebook.* Sage.

Moustakas, C. (1994). *Phenomenological research methods.* Sage.

Patton, M. Q. (1987). *How to use qualitative methods in evaluation.* Sage.

Philips, D. C., & Soltis, J. K. (2005). In S. Durmuş (Trans. Ed.), *Perspectives on Learning*. Ankara: Nobel Yayın Dağıtım.

Perez-Foguet, A., Lazzarini, B., & Gine, R. (2018). Promoting sustainable human development in engineering: assessment of online courses within continuing professional development strategies. *J Cleaner Prod*, *172,* 4286-4302. https://doi.org/10.1016/j.jclepro.2017.06.244

Raut, N., & Pendse, M. (2023). A study of conscious consumerism of sustainable products among the university students. In *Transformation for Sustainable Business and Management Practices: Exploring the Spectrum of Industry 5.0* (pp. 105-117). Emerald.

Rieckmann, M. (2012). Future-oriented higher education: which key competencies should be fostered through university teaching and learning? *Futures*, 44, 127-135. https://doi.org/10.1016/j.futures.2011.09.005

Ruiz-Mallén, I., & Heras, M. (2020). What sustainability? Higher education institutions' pathways to reach the agenda 2030 goals. *Sustainability*, *12*(4), 1290. https://doi.org/10.3390/su12041290

Schunk, D. H. (2011). Yapılandırmacı teori. In M. Y. Demir (Trans.), *learning theories an educational perspective: öğrenme teorileri eğitimsel bir bakışla* (pp. 234-277). Nobel Yayınları.

Sibbel, A. (2009) Pathways towards sustainability through higher education. *Int J Sustainability Higher Educ., 10*, 68-82. https://doi.org/10.1108/14676370910925262

Stokes, E., Edge, A., & West, A. (2001). *Environmental education in the educational systems of the European Union.* Centre for educational research london school of economics and political science. Commissioned by the Environment Directorate-General of the European Commission.

Students Organising for Sustainability [SOS]. (2021). *Students organising for sustainability international summit.* https://sos.earth/wp-content/uploads/2021/02/SOS-International-Sustainability-in-Education-International-Survey-Report_FINAL.pdf

Tanrıverdi, B. (2009). Sürdürülebilir çevre eğitimi açısından ilköğretim programlarının değerlendirilmesi. *Eğitim ve Bilim, 34*(151), 89-103.

Uiterkamp, A. J. M. S., & Vlek, C. (2007), Practice and outcomes of multidisciplinary research for environmental sustainability, *Journal of Social Issues, 63*(1), 175-197.

UNDP, (2015). *Sustainable Development Goals.* http://www.tr.undp.org

UNFCCC. (2015). *The Paris Agreement*. United Nations Framework Convention on Climate Change. Available at: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement (Accessed: 18 February 2026).

United Nations Educational Scientific and Cultural Organization (UNESCO), (1997). *Record of the general conference*, Paris. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000114588.

United Nations Educational Scientific and Cultural Organization (UNESCO), (2022). The concept of sustainability and its contribution towards quality transformative education: thematic paper, Paris. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000381528.

Velazquez, L., Munguia, N., & Sanchez, M. (2005) Deterring sustainability in higher education institutions: An appraisal of the factors which influence sustainability in higher education institutions. *Int J Sustainability Higher Educ.,* 6, 383-391. https://doi.org/10.1108/1467637051062386

World Commission on Environment and Development (WECD), (1987). *Our Common Future*. Oxford University Press. https://sustainabledevelopment.un.org/content/documents/5987our-common-future.pdf.

Wright, T. S. A. (2002). Definitions and frameworks for environmental sustainability in higher education. *High Educ Policy,* 15, 105-120. https://doi.org/10.1016/S0952-8733(02)00002-8

Yıldız, K., Güzel-Gürbüz, P., Esentaş, M., Beşikçi, T., & Balıkçı, İ. (2021). Üniversite öğrencilerinin sürdürülebilir çevre eğitimi ve çevre sorunlarına yönelik tutumları arasındaki ilişkinin incelenmesi. *International Journal of Social Science Research, 10*(1), 35-49.

Zimmerman, E., Peschl, F. M., & Nassek-Röhmer, B. (2010). Constructivist curriculum design for the ınterdisciplinary study programme. *Education Science, 5*(3), 144-157.

Author(s) Information

Nur Utkur-Gulluhan
Istanbul University-Cerrahpaşa,
Faculty of Hasan Ali Yucel Education, Department of
Primary Education, İstanbul/Türkiye
Contact e-mail: nur.utkur@iuc.edu.tr
ORCID iD: https://orcid.org/0000-0003-2062-5430

JESEH
Journal of Education in Science, Environment and Health
Volume 12, Issue 2, 2026
https://doi.org/10.55549/jeseh.894

Argumentation Research in Science Education: Global Publication Trends, Intellectual Structure, and Thematic Transformation (2001–2025)

Esra Ergunt, Serkan Yilmaz

Article Info

Article History

Published:
01 April 2026

Received:
02 January 2026

Accepted:
15 February 2026

Keywords

Science education,
Argumentation,
Bibliometric analysis,
Scientific mapping,
Thematic transformation

Abstract

Argumentation has become a cornerstone of science education research, essential for fostering evidence-based thinking, scientific reasoning, and scientific literacy. This study employs bibliometric methods to examine global publication trends, intellectual structures, and thematic transformations in argumentation-focused research between 2001 and 2025. A final dataset of 474 peer-reviewed articles, retrieved from the Web of Science Core Collection after applying stringent inclusion and exclusion criteria, was analyzed using the R-based Bibliometrix package. To ensure a holistic interpretation, author keywords were standardized and categorized into overarching themes, allowing the field's conceptual landscape to be mapped in a more integrated manner rather than through fragmented indicators. The results indicate that argumentation studies emerged between 2001 and 2010, experienced rapid growth from 2011 to 2020, and approached a maturation phase by 2025. Early research focused mainly on cognitive argument structures, while later work increasingly engaged with socioscientific issues, epistemic practices, and classroom discourse. During this process, classical models were recontextualized within contemporary pedagogical settings and underwent terminological transformation. By illustrating the transformation of the science education argumentation literature from a pedagogical tool to a foundational epistemic framework, this study offers an empirically grounded perspective that may inform future research directions in the field.

Introduction

Science education aims not only to teach individuals scientific concepts, but also to enable them to question, evaluate, and justify this knowledge based on evidence. In this context, argumentation stands out as a fundamental reasoning process that enables students to support their claims with data and to substantiate these claims through reasoning (Toulmin, 1958). Argumentation-based learning environments support the development of scientific reasoning skills by allowing students to interact with different perspectives (Osborne et al., 2004). Scientific knowledge is recognized as a dynamic process that is socially constructed through classroom discussions and reasoned discourse (Driver et al., 2000). Accordingly, argumentation is positioned not only as a form of expression but also as a central epistemic practice that enables an understanding of the nature of science.

Argumentation-based approaches in science education are considered effective instructional approaches that not only deepen conceptual learning but also support the development of critical thinking and scientific literacy skills (Osborne, 2010). In recent years, there has been a marked increase in the number of studies on this topic, indicating that argumentation is considered not only a teaching strategy but also a comprehensive approach to learning and thinking. This upward trend can be attributed not only to global research dynamics but also to the explicit inclusion of argumentation in science education curricula in some countries. Indeed, the emphasis on scientific reasoning and argumentation in science curricula updated in Turkey since the mid-2010s has shown a parallel development with the increase in academic production in this field. However, individual studies may be limited in revealing the general publication trends, dominant thematic orientations, and intellectual structure of the field. This lack of a holistic perspective in the literature, coupled with the increasing volume of publications, further underscores the need for bibliometric approaches capable of systematically examining its structural characteristics and longitudinal dynamics (Donthu et al., 2021; Zupic & Cater, 2015).

At this point, bibliometric analyses stand out as systematic approaches that aim to reveal the intellectual structure, development dynamics, and research orientations of a field by examining scientific production through quantitative indicators (Donthu et al., 2021). Publication trends over the years, citation structures, prominent journals and authors, and keyword-based conceptual networks can be made visible in a holistic manner through

bibliometric analyses. While there are bibliometric studies (Kurtuluş & Yılmaz, 2022; Orhan, 2024; Tang, 2024) focusing on various themes in science education, studies that specifically focus on argumentation and examine the field's temporal development, intellectual foundations, and thematic orientations using up-to-date data remain limited. The volume of publications and thematic diversity achieved in the field necessitate a structural mapping beyond traditional review studies. Indeed, the number of existing bibliometric studies on science education argumentation (Mulyani et al., 2024; Noris et al., 2024; Tosun, 2024; Wang et al., 2023) is relatively small, and there is a clear need for a comprehensive perspective that reveals the current development dynamics and intellectual structure of the literature, particularly including publications after 2020.

The aim of this study is to examine research focusing on argumentation in science education using bibliometric methods in response to the aforementioned methodological gap, and to provide a comprehensive overview of the field's temporal development, intellectual structure, and thematic orientations. Specifically, scientific production trends by year, distributions by country and journal, citation structures, and keyword-based thematic orientations were analyzed using publications indexed in the Web of Science database. One of the primary original contributions of this study is that it covers current data up to the end of 2025 and empirically demonstrates that the science education argumentation literature is approaching a mature stage. At this critical juncture, as the field transitions from an exploratory phase to a more established structure, identifying emerging thematic transformations and potential paradigm shifts is of strategic importance for shaping future research agendas. In this respect, the study aims not only to provide an overview of past developments but also to guide researchers by offering an analytical projection of the sub-fields in which the literature is deepening.

Preliminary analyses conducted to contextualize this expansion process in the science education argumentation literature and to support the methodological rationale of the study reveal that the field exhibits a characteristic growth pattern. The logistic growth model (Figure 1) and the cumulative growth curve (Figure 2) are mathematical approaches used to predict the growth rate, life cycle, and potential saturation level of a research field's publication output based on empirical data. These analyses indicate that research on argumentation in science education emerged between 2001 and 2010, experienced a period of rapid growth between 2011 and 2020, and has begun to approach a phase of maturity as of 2025.

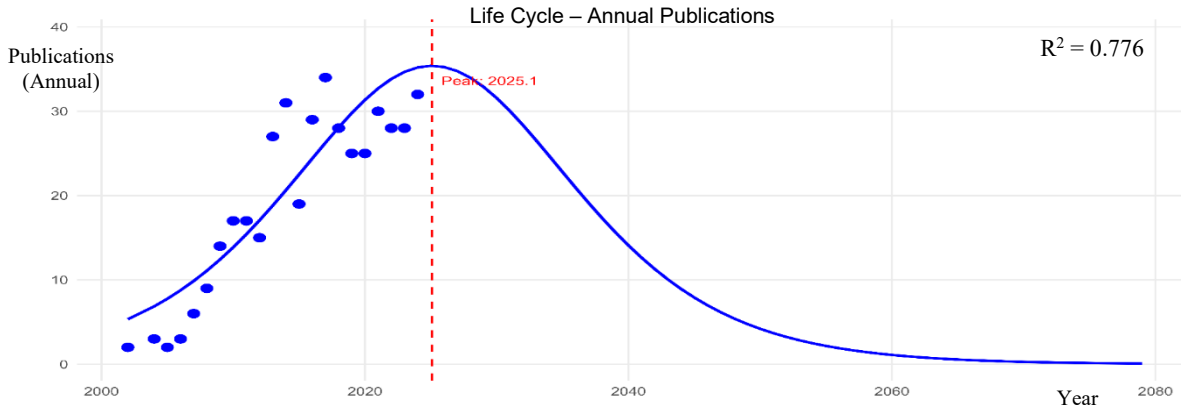


Figure 1. The life cycle of argumentation studies (logistic growth model)

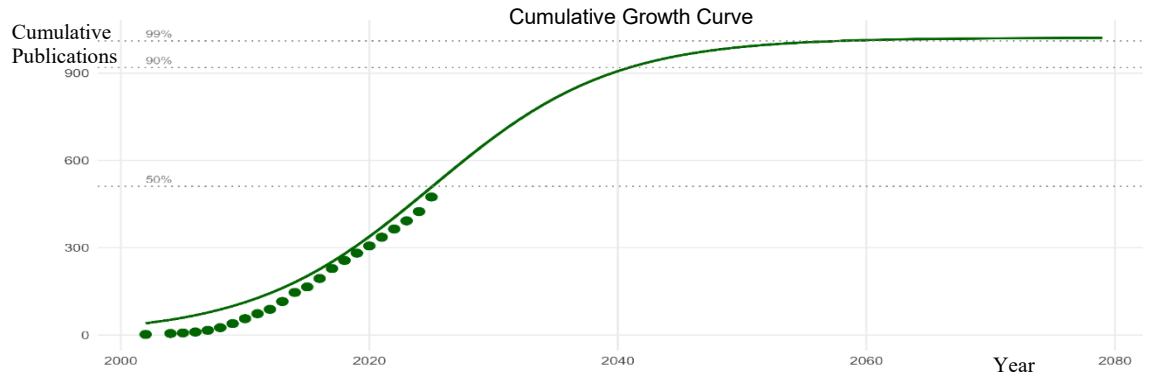


Figure 2. Cumulative growth curve of argumentation studies in science education

The model presented in Figure 2 suggests that the potential saturation level of the field may be reached around 2050, with an estimated total of approximately 1021 publications. This finding indicates that the field has largely moved beyond its exploratory phase and entered a more established stage that is open to thematic deepening and specialization. The visuals and accompanying explanations presented in this section are used descriptively to contextualize the developmental trajectory of the field; detailed quantitative analyses are presented in the Results and Discussion sections.

In this context, the present study, which aims to reveal the intellectual structure, thematic orientations, and global publication trends of argumentation research in science education from a holistic perspective, constitutes a timely and necessary contribution to the current stage of development in the field. Within this scope, the following research questions (RQ) are addressed to systematically achieve the study's objectives:

- RQ1. What are the temporal development and scientific production trends of argumentation studies in science education?
- RQ2. Which journals, authors, countries, and institutions stand out in science education argumentation studies?
- RQ3. How are the citation structures and intellectual foundations of argumentation studies in science education shaped?
- RQ4. What are the dominant thematic orientations, thematic transformations, and research trends in science education argumentation studies?

Method

Data Source and Data Set Formation

The Web of Science Core Collection (WoS) was selected as the data source for this study. WoS is widely used in literature reviews because it covers peer-reviewed journals with high impact values, provides standardized and reliable citation data, and offers structured bibliographic data that enables bibliometric analysis. The research dataset was created from publications selected from the WoS database in accordance with the specified inclusion and exclusion criteria, and the data collection process was completed and updated on January 5, 2026.

Screening Strategy and Inclusion-Exclusion Criteria

The screening process used the WoS Topic field to retrieve studies on argumentation in science education. This field was preferred for its comprehensive dataset, including titles, abstracts, author keywords, and Keywords Plus terms. The search string was "science education" AND "argumentation", aligning with common literature concepts. Using the AND operator ensured both terms co-occurred, increasing result relevance. The decision to include only the term *argumentation* in the search string was adopted to maintain the conceptual focus of the study and to capture core research addressing epistemic and pedagogical argumentation processes in the science education literature.

The retrieved publications were filtered in successive stages based on predefined inclusion and exclusion criteria to construct a dataset that was aligned with the study's purpose, methodologically comparable, and of high academic quality. Accordingly, only research articles published in peer-reviewed journals between 2001 and 2025, indexed in the SSCI, classified under the Education–Educational Research category, and written in English were included in the analysis. Book reviews, conference proceedings, book chapters, and studies outside the SSCI scope or not directly related to the context of science education were excluded from the analysis. The filtering stages applied and the number of publications obtained at each stage are summarized in Table 1.

Table 1. Data set creation process: Filtering stages and publication counts

Stage	Applied Filter / Criterion	Remaining Publication Count
1	Keywords: "Science Education" (topic) AND Argumentation (topic)	835
2	Document type: Article	729
3	Publication year: 2001–2025	724
4	WoS category: Education–Educational Research	618
5	Index: SSCI	477
6	Language: English	474

Data Cleaning Process

Prior to the analyses, the dataset underwent a cleaning process to enhance the accuracy, conceptual consistency, and reproducibility of bibliometric results. In this process, variations in author names (e.g., first name–last name order and abbreviations), institutional name formats, singular–plural usage in keywords, and spelling inconsistencies were identified using tools provided by the Bibliometrix package. Where automated procedures were insufficient, manual corrections were performed under researcher supervision to ensure consistency.

Keyword Standardization and Conceptual Integration

To increase the conceptual validity of the thematic mapping and network analyses, a comprehensive keyword standardization and conceptual integration procedure was applied following data cleaning. This stage aimed not only to correct formal inconsistencies but also to analytically harmonize concept clusters representing shared theoretical and pedagogical frameworks in the science education argumentation literature. Accordingly, keywords with high conceptual overlap were grouped under superordinate concepts by considering spelling variants, singular–plural forms, and contextual similarity across all author keywords. This consolidation was guided primarily by lexical proximity and recurrent contextual usage patterns rather than by constructing a formal ontological taxonomy. When lexical variants referred to the same pedagogical construct, practice, or theoretical framework, as indicated by their recurrent usage within similar research contexts, they were grouped under a common superordinate term. This approach prioritizes the preservation of emergent thematic structures over rigid a priori classifications. While certain terms could be conceptually distinguished at a finer theoretical level, the objective here was to reduce terminological fragmentation while preserving thematic coherence within the analytical framework.

In this process, 396 variants identified among 1,074 author keywords were consolidated under 51 superordinate conceptual terms, each comprising between 2 and 25 sub-terms. The complete list of these keyword consolidations and their corresponding variants is provided in Appendix 1 to ensure methodological transparency and reproducibility. Following consolidation, each superordinate term reached a minimum frequency threshold of five or more ($f \geq 5$), suggesting a relatively stable presence within the field. Of the remaining 678 keywords that could not be meaningfully grouped, 592 had a frequency of one ($f = 1$), while 86 appeared with frequencies ranging from two to four. To reduce conceptual noise and emphasize more robust thematic structures, keywords with a frequency of one were excluded from the thematic mapping and network analyses. Consequently, the final analytical framework consisted of 137 keywords, formed by the combination of 51 superordinate terms and 86 individual keywords with frequencies of two or higher. These units are hereafter referred to as "refined terms".

During the conceptual consolidation process, merging variant forms also increased the representational strength of the related superordinate concepts. For instance, the keyword *collaboration*, which initially appeared with a frequency of one, was merged with 12 related variants, resulting in a cumulative frequency of 19. Similarly, although the term *argument* exhibited a moderate standalone frequency ($f = 8$), it was grouped under the superordinate concept of *argumentation* together with related variants such as *arguments*, *scientific argumentation*, *socio-scientific argumentation*, and *science argumentation*, reflecting its role within a broader theoretical framework. In contrast, 11 technical variants representing structural components specific to the Toulmin Model (e.g., rebuttal, warrant, and claim) were separated from the general argument category and grouped under the umbrella term *Toulmin*. These consolidations, along with the establishment of the overarching category of *socioscientific issues* through the integration of 13 variants (e.g., socio-scientific issues, socioscientific, socio-scientific issue, local socioscientific issues), constitute the core outcomes of the standardization process. Overall, this procedure eliminated redundancy while preserving contextual meaning, yielding a thematically coherent structure that supports analytical reliability.

Data Analysis

The R-based Bibliometrix package, specifically developed for bibliometric studies, was used to analyze the dataset (Aria & Cuccurullo, 2017). Bibliometrix was selected because it enables the integration of performance analysis and scientific mapping approaches within a single analytical framework, allowing the simultaneous examination of both the conceptual and intellectual structures of the field. Two main analytical approaches were adopted: performance analysis and scientific mapping. Within the scope of performance analysis, scientific production trends by year, the most productive countries, authors, institutions, and journals, as well as citation-based indicators, were examined. Logistic and cumulative growth curve analyses were applied to reveal the temporal

development and growth dynamics of argumentation research in science education; these analyses enabled the quantitative modeling of the field's emergence, acceleration, and maturation phases.

In the scientific mapping phase, analyses were conducted on keyword co-occurrence networks, co-citation structures, thematic maps (based on the dimensions of centrality and density, including basic, motor, niche, and emerging/declining themes), trend topics (key concepts that rise or decline over time), and three-field plots (Figure 7). These analyses revealed the conceptual structure of the field and its transformation over time. Keywords Plus data were used in the three-field plot, which illustrates the intellectual structure of argumentation studies in science education, in order to reflect broad conceptual networks and terminological links with referenced sources in a comprehensive manner. In this context, while author keywords were preferred to capture micro-level research foci in thematic mapping and word cloud analyses, Keywords Plus was employed to represent the intellectual flow and macro-level conceptual bridges of the discipline. Accordingly, the three-field plot made large-scale relationships among authors, sources, and concepts visible in a holistic way.

Within the scope of the fourth research question, the analysis aimed to identify thematic shifts and research trends in the literature by reflecting researchers' theoretical and pedagogical positions. For this purpose, author keywords were used in the thematic map analysis, as they constitute the primary micro-level data source by directly reflecting how researchers position their studies within the field, demonstrating sensitivity to science education terminology, and revealing conceptual traces of thematic change. In the thematic map analysis, the minimum cluster frequency threshold was set to 20 to optimize the readability of the conceptual network derived from 474 articles and to represent the intellectual structure of the literature at a meaningful level of resolution. Preliminary tests showed that a threshold value of 10 increased bibliometric noise, dispersed thematic focus, and amplified immature or transient trends as if they were dominant. In contrast, threshold values of 30 and above excluded critical niche themes such as *dialogue*, *pedagogy*, and *decision-making*, which represent methodological and pedagogical depth in the science education argumentation literature. Therefore, a threshold value of 20 was adopted to achieve an analytical balance between basic and motor themes (e.g., argumentation and science education), which constitute the conceptual backbone of the field, and emerging themes (e.g., socioscientific issues and reasoning) that signal developmental directions. Through this threshold selection, terms with limited representation or dispersed contextual relevance were eliminated, and the dominant thematic orientations, transformation tendencies, and research strategies in science education argumentation were presented in a coherent, comparable framework based on "refined terms".

In parallel, during the visualization of conceptual networks, a minimum edge threshold of 4 was applied to balance relational strength among keywords with the overall intellectual density of the network. Experimental testing indicated that threshold values of 6 and above excessively simplified conceptual transitions, substantially reducing the visibility of secondary concepts connected to basic themes such as *socioscientific issues* and *argumentation*. By contrast, the selected threshold preserved strong connections among central concepts while enabling secondary conceptual relationships to be mapped with a higher level of analytical precision.

Within this framework, to examine the intellectual structure of the field and patterns of knowledge production in a multidimensional manner, analyses of keyword co-occurrence networks (Figure 8), thematic maps (Figure 9), and prominent research topics over time (Figure 10) were conducted. Consequently, a holistic view of argumentation research in science education was obtained at both the micro level, through author-focused conceptual representations based on author keywords, and the macro level, through intellectual flows and conceptual linkages across the literature captured by Keywords Plus. This dual perspective provided a systematic and comparable analytical foundation for addressing all research questions.

Results

RQ1. What are the Temporal Development and Scientific Production Trends of Argumentation Studies in Science Education?

When examining Figure 3, which shows the distribution of argumentation research in science education over time, it is observed that scientific production remained below 10 publications per year between 2001 and 2005. A steady increase began in 2006 and continued through 2010. Although fluctuations were observed in publication output between 2011 and 2022, the overall trend remained upward. After 2022, this upward trend strengthened again, reaching the highest publication volume during the 2024–2025 period. Overall, the results indicate that scientific production in the field has followed a consistently increasing trajectory from its early stages to the present.

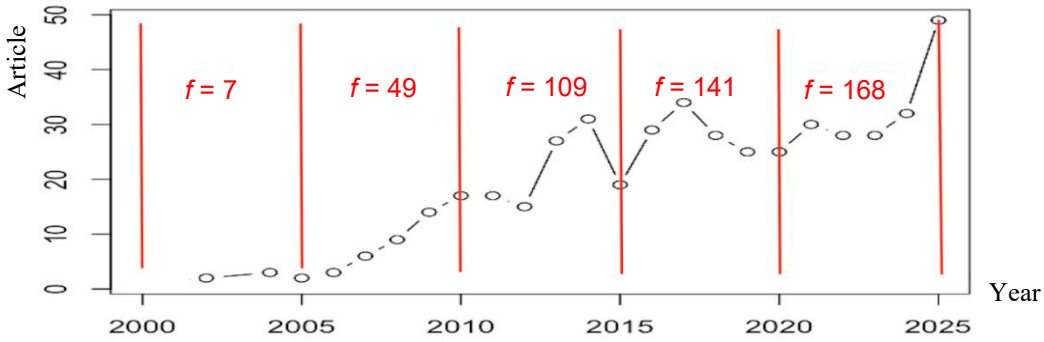


Figure 3. Annual scientific production of argumentation studies

RQ2. Which Journals, Authors, Countries, and Institutions Stand Out in Science Education Argumentation Studies?

The journals that stand out in terms of publication output in science education argumentation research are listed in Table 2 in descending order of article count. A substantial proportion of the analyzed articles (353 articles, 74.4%) were published in these 10 journals, with the *International Journal of Science Education* leading the field with 91 publications. This journal is followed by *Science & Education* (54), *Science Education* (52), and the *Journal of Research in Science Teaching* (45). Journals such as *Research in Science Education* and *Research in Science & Technological Education* also contribute a notable number of publications to the field.

Table 2. Journals with the highest number of publications in argumentation studies

Rank	Journal	Number of Articles
1	International Journal of Science Education	91
2	Science & Education	54
3	Science Education	52
4	Journal of Research in Science Teaching	45
5	Research in Science Education	38
6	Research in Science & Technological Education	19
7	International Journal of Science and Mathematics Education	18
8	Journal of Science Education and Technology	16
9	Chemistry Education Research and Practice	10
10	Instructional Science	10

Table 3 presents the authors with the highest number of publications in science education argumentation research. The analysis revealed that some authors appear under different name formats in WoS records. For example, articles by Archila were recorded as "Archila, P. A." and "Antonio Archila, P.". After consolidating these variants, Archila was identified as having contributed to 15 of the 474 articles published between 2001 and 2025, making this author the most prolific in the field. Archila is followed by Erduran (14 articles) and McNeill (14 articles). Other prominent authors include Sadler (10), Gonzalez-Howard (7), Hand (7), Molina (7), Restrepo (6), and Zeidler (6). The distribution in Table 3 represents the quantitative contribution of the most prolific authors, accounting for 86 articles in total.

Table 3. Most prolific authors in argumentation research

Rank	Author	Number of articles
1	Archila, P. A.	15
2	Erduran, S.	14
3	McNeill, K. L.	14
4	Sadler, T. D.	10
5	Gonzalez-Howard, M.	7
6	Hand, B.	7
7	Molina, J.	7
8	Restrepo, S.	6
9	Zeidler, D. L.	6
10	Kuhn, D.	5

Table 4 presents the country-level distribution of publications in science education argumentation research between 2001 and 2025. The United States of America (USA) leads the field with 159 articles (33.5%), followed by China (40 articles, 8.4%) and Turkey (33 articles, 7.0%). These are followed by Germany and Sweden (25 articles each, 5.3%) and the United Kingdom (23 articles, 4.9%). The top ten countries listed in Table 4 account for a large proportion (363 articles, 76.6%) of the articles examined, indicating a high quantitative contribution to the argumentation literature in science education.

Table 4. Countries producing the most articles in argumentation research

Rank	Country	Number of Articles	Percentage (%)
1	USA	159	33,5
2	China	40	8,4
3	Turkey	33	7,0
4	Germany	25	5,3
5	Sweden	25	5,3
6	United Kingdom	23	4,9
7	Spain	19	4,0
8	Colombia	15	3,2
9	Australia	13	2,7
10	Israel	11	2,3

Figure 4 illustrates the cumulative publication trends from 2001 to 2025 for the three most productive countries (USA, China, and Turkey). The publication curve for the USA shows a steady and uninterrupted increase throughout the period, reaching 159 articles by the end of 2025. In contrast, the publication trajectories of China and Turkey exhibit a later onset, with growth becoming particularly evident from the mid-2010s onward. By 2025, China reached 40 publications, while Turkey reached 33. Compared to the USA, both countries display lower growth slopes, with noticeable variations in growth rates across different periods.

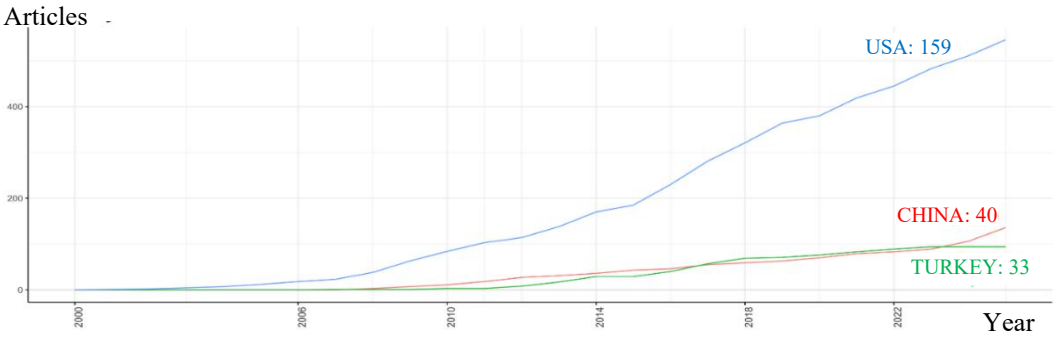


Figure 4. Countries with the highest publication output in argumentation research

Table 5 presents the institutions with the highest publication output in science education argumentation research between 2001 and 2025. According to the results, the University of Los Andes ranks first with 47 articles, followed by Boston College (26 articles). Among institutions based in Taiwan, National Taiwan Normal University and National Taiwan University of Science and Technology appear prominently. From Europe, the University of Oxford ranks among the leading institutions with 17 articles, while Recep Tayyip Erdoğan University from Turkey is also included in the top ten with 16 articles. The institutions listed in Table 5 collectively account for 201 articles (42.3%), highlighting their substantial contribution to the field.

Table 5. Most productive institutions in argumentation research

Rank	Institution	Number of Articles
1	Universidad de los Andes / University of Los Andes	47
2	Boston College	26
3	National Taiwan Normal University	19
4	National Taiwan University of Science and Technology	17
5	University of Oxford	17
6	Recep Tayyip Erdoğan University	16
7	Seoul National University	15
8	University of Iowa	15
9	University of Maryland	15
10	Florida State University	14

RQ3. How are the Citation Structures and Intellectual Foundations of Argumentation Studies in Science Education Shaped?

Figure 5 presents the most frequently cited studies in science education argumentation research. The results indicate that Zohar and Nemet (2002) is the most cited study, with over 700 citations, followed by Erduran et al. (2004) and Duschl (2008). Other highly cited works include those by Sandoval (2005), Sandoval and Reiser (2004), Sadler et al. (2007), and Sadler (2009). Most of the studies shown in Figure 5 were published between 2002 and 2009 and represent foundational contributions that shaped the early development of the field.

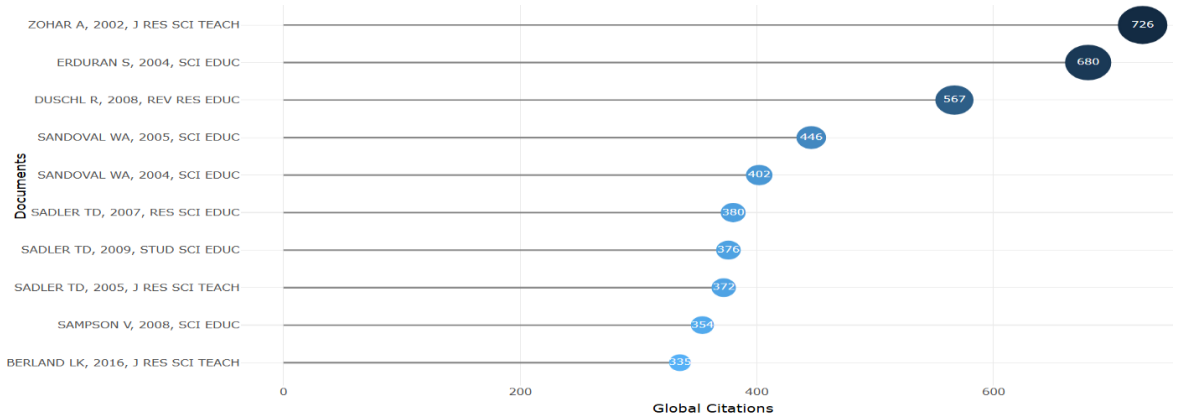


Figure 5. Most cited studies in argumentation research

The reference co-citation network for argumentation research in science education is shown in Figure 6.

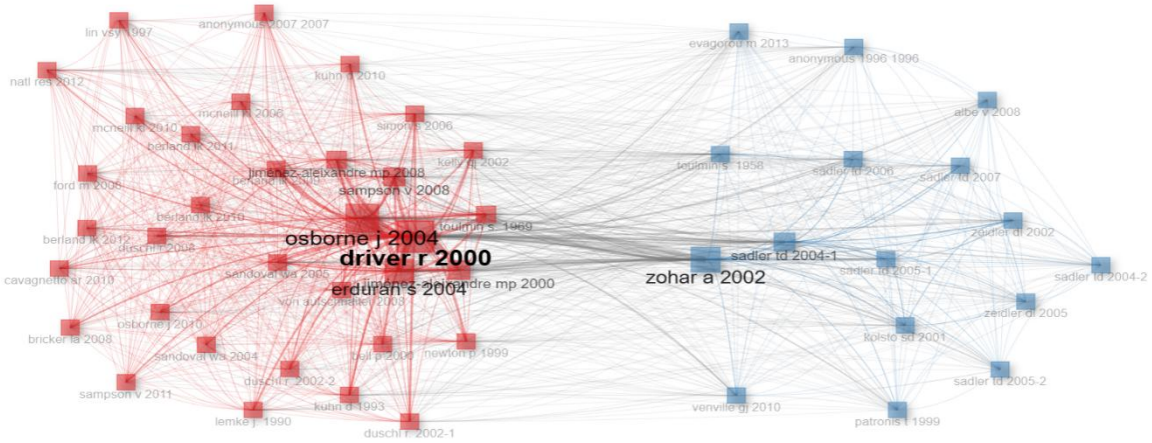


Figure 6. Co-citation network for argumentation studies in science education

The structural analysis of the co-citation network in Figure 6 reveals clear patterns in node size, link thickness, and inter-node distance. The network structure is organized around two distinct clusters. In Cluster 1, studies by Driver et al. (2000), Erduran et al. (2004), Osborne et al. (2004), and Sampson and Clark (2008) occupy central positions, characterized by large node sizes and dense interconnections. The gradual reduction in link thickness from central to peripheral nodes indicates a core–periphery structure within the network.

Cluster 2, positioned on the right side of the figure, includes studies such as Zohar and Nemet (2002), Sadler (2004), Sadler and Zeidler (2005), and Toulmin (1958). Within this cluster, Zohar and Nemet (2002) emerges as a focal node, as reflected by its node size and multidirectional connections. The distinctiveness of the distance between the clusters reveals that the two clusters in the network are spatially separated. The fact that Driver et al. (2000) establishes connections with both clusters indicates that this study is one of the bridge nodes connecting the clusters. Toulmin (1958), positioned in Cluster 2, serves as a fundamental theoretical reference, related to studies in different areas of the network. Within this distribution, the figure shows that argumentation studies are grouped under two main clusters: studies focusing on cognitive-epistemic processes and studies addressing socioscientific issues and discourse-based approaches. However, connections between these clusters are observed through some fundamental sources.

The three-field plot shown in Figure 7 illustrates the intellectual structure of science education argumentation research by depicting relationships among cited sources, authors, and keywords. Generated using Keywords Plus data, the visualization reflects conceptual consistency and clarity, with box sizes representing frequency and flow thickness indicating relational strength. In the cited sources column, Jiménez-Aleixandre and Erduran (2008) and Osborne et al. (2004) appear as dominant references. Classical works such as Toulmin (1958) and Kuhn (1993) are represented with lower frequencies but maintain multiple connections across authors and conceptual domains.

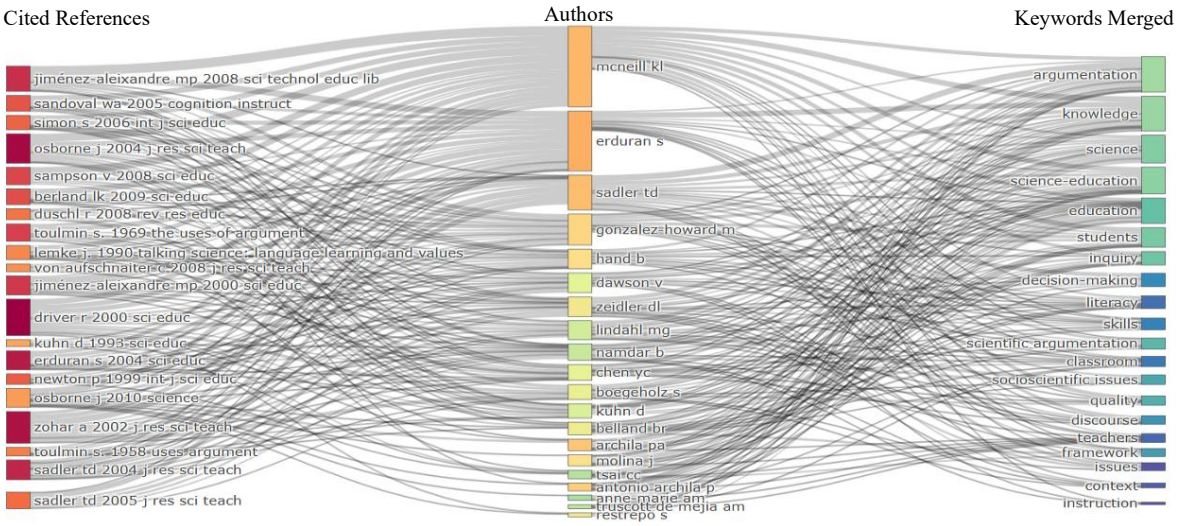


Figure 7. Reference–author–keyword relationships in argumentation studies

In the authors' field in Figure 7, McNeill occupies a central position with the highest frequency and largest box size; Erduran, Sadler, and Zeidler stand out with their multifaceted connections. In the keywords field, *argumentation*, *knowledge*, *science*, and *science education* emerge as the most prominent conceptual focal points. Examination of author–concept flows shows that McNeill and Erduran are strongly associated with *argumentation* and *knowledge*, whereas Sadler and Zeidler show stronger connections with *decision-making* and *socioscientific issues*. Authors such as Jiménez-Aleixandre, Osborne, and Sampson appear in both the cited sources and authors' fields, reflecting their role in connecting classical foundations with contemporary conceptual developments.

RQ4. What are the Dominant Thematic Orientations, Thematic Transformations, and Research Trends in Science Education Argumentation Studies?

Within the scope of this research question, the conceptual structure of science education argumentation studies (Figure 8), their thematic positioning within the field (Figure 9), and the transformation patterns related to the temporal visibility of themes (Figure 10) are examined through bibliometric indicators that complement one another while representing distinct analytical dimensions.

The co-occurrence network of keywords presented in Figure 8 visualizes the conceptual structure and thematic relationships of argumentation studies in science education. An examination of the network reveals that the keywords cluster around four main clusters with different densities and connection patterns. At the center of the network, the *argumentation* and *science education* nodes occupy a prominent position in terms of node size and multiple connection lines. The distinct thickness of the connection between these two nodes indicates that these concepts are used together with a high frequency of co-occurrence in the dataset. In the central cluster positioned around the *argumentation* node, concepts such as *discourse*, *dialogue*, *evidence*, *epistemic practice*, *methods*, and *modeling* are represented by their proximity to the center and dense interconnections. In contrast, concepts such as *translanguaging*, *elementary science education*, *earth science education*, and *Toulmin* are located within the same cluster but in more peripheral areas of the network and are associated with weaker connections.

In other sections of the network shown in Figure 8, thematic clusters diverge around distinct focal points. In the cluster centered around the *science education* node, concepts related to teaching and learning – such as *collaboration*, *inquiry*, *learning*, *teachers*, *professional development*, *critical thinking*, *argumentative practices*, and *middle school* – are positioned together. In another cluster focused on *socioscientific issues*, concepts such as *decision-making*, *reasoning*, *scientific literacy*, *epistemology*, *nature of science*, and *curriculum* converge with relatively strong internal connections, while elements such as *climate change* and *preservice teachers* are

incorporated into this structure with comparatively weaker links. Concepts such as *pedagogy* and *teaching*, located in more peripheral and sparsely connected areas of the network, are represented by lower-intensity connections in a separate cluster. These concepts are linked to the dominant clusters through thinner relational ties.

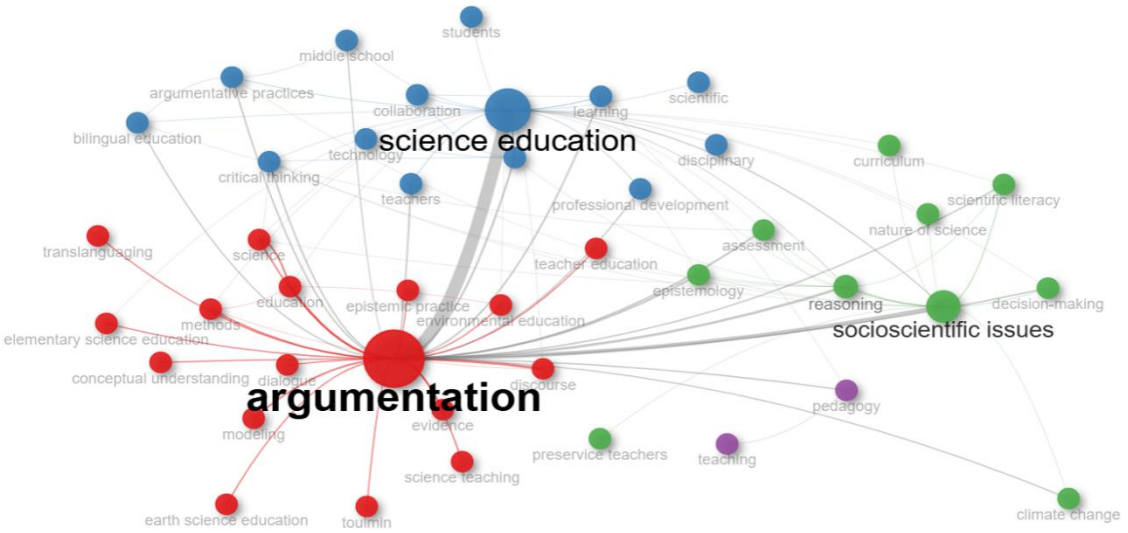


Figure 8. Keyword co-occurrence network for argumentation studies

The thematic map classifying the conceptual structure of argumentation studies in science education into four quadrants based on centrality and density criteria is presented in Figure 9. The results are presented according to their positions based on centrality and density (basic, motor, niche, and emerging/declining), rather than according to the numerical ordering of clusters. The analysis reveals that the field is structured around five principal thematic clusters. These themes are positioned as motor and basic themes representing the epistemic and theoretical backbone of the field, emerging or developing themes indicating the direction of thematic evolution, and niche themes reflecting specialized and application-oriented foci.

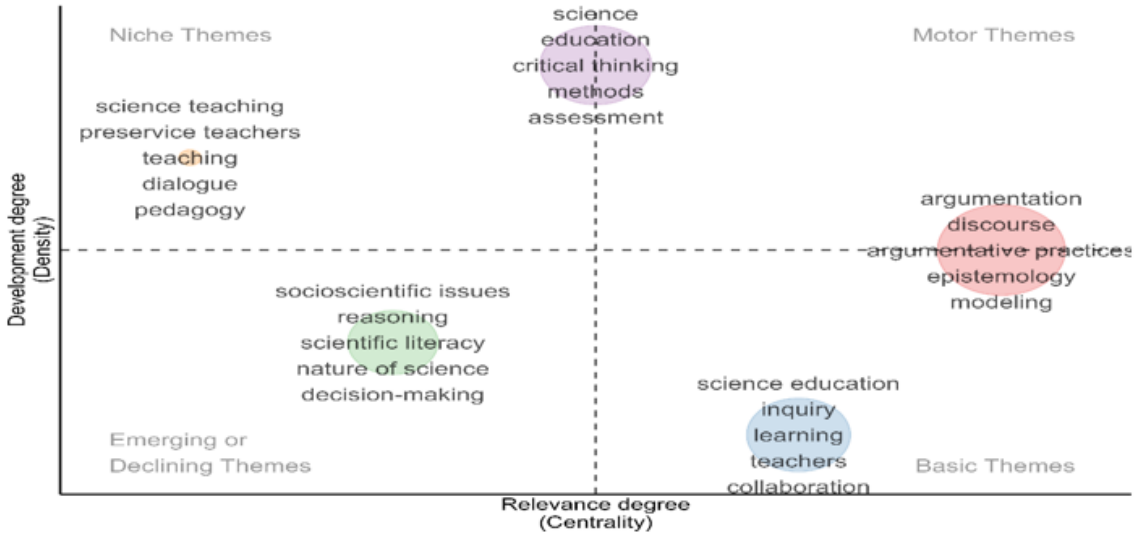


Figure 9. Thematic map of argumentation studies

Two dominant clusters emerge in the basic and motor themes region (lower right quadrant and upper right–middle transition). Cluster 1 consists of the concepts of *argumentation*, *discourse*, *argumentative practices*, *epistemology*, and *modeling* and occupies a position corresponding to the theoretical-epistemic core of the field. Cluster 2 encompasses the concepts of *science education*, *inquiry*, *learning*, *teachers*, and *collaboration* and exhibits a structure highly integrated with the broader literature. Cluster 4 (science, education, critical thinking, methods, and assessment), positioned in a transition zone between motor and niche themes, represents a methodological axis within the thematic map. Focusing on specialized applications, Cluster 5 (science teaching, preservice teachers, teaching, dialogue, and pedagogy) is located in the niche themes region and reflects a sub-area where classroom-oriented practices and pedagogical concerns are concentrated. Cluster 3, which includes *socioscientific*

issues, *reasoning*, *scientific literacy*, *nature of science*, and *decision-making*, is positioned in the emerging themes region and indicates growing thematic orientations in terms of both centrality and density.

Figure 10 reflects the temporal distribution of themes in the argumentation literature in science education, their frequency of use, and changes in research trends. In the visualization, the size of the circles represents the total frequency of each theme across the entire study period, while their position on the horizontal axis represents the median year in which this usage was most concentrated. The horizontal lines indicate the temporal span between the year (Q_1) when cumulative usage first reached 25% and the year (Q_3) when it reached 75%. The circular markers denote the median year (Q_2), corresponding to the point at which 50% of cumulative usage was achieved.

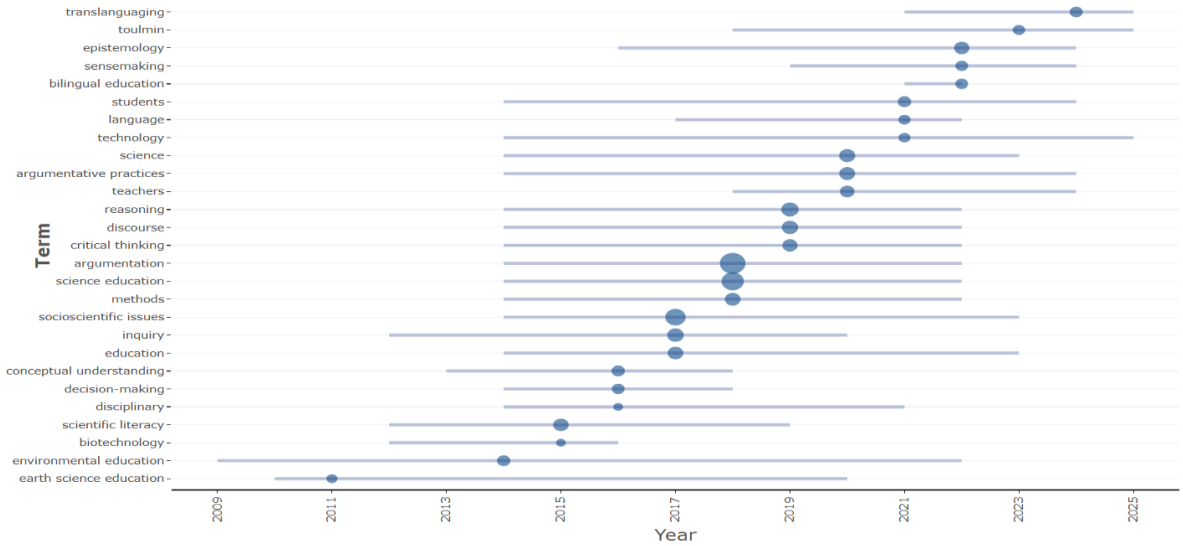


Figure 10. Research topics that have emerged over time in argumentation studies

Within the scope of the analysis, conducted using a minimum word frequency threshold of 5 and a word count of 3 per year, a total of 27 topics are visualized in Figure 10. Examination of the overall distribution shows that the topics with the highest total usage frequencies—*argumentation*, *science education*, *socioscientific issues*, *reasoning*, and *questioning*—are represented by larger circles and relatively balanced temporal distributions. In addition, the topics of *environmental education*, *technology*, *argumentative practices*, *students*, and *earth science education* exhibit the widest interquartile ranges (Q_3 - Q_1), indicating extended visibility across the literature.

With respect to the early period (2015 and earlier), the starting points (Q_1) of topics such as *environmental education*, *earth science education*, *scientific literacy*, *biotechnology*, and *inquiry* appear earlier than those of other topics. Among these, *inquiry* stands out as the topic with the highest usage frequency. Moreover, although *earth science education* and *environmental education* show early onsets, their relatively late third-quartile values (Q_3) indicate sustained prominence over an extended period. Notably, *earth science education* displays the most pronounced temporal asymmetry (Q_3 - Q_2 > Q_2 - Q_1), indicating a right-skewed distribution.

During the medium-term period (2015–2020), thematic diversity is represented by a relatively dense clustering of topics. The topics of *argumentation*, *science education*, *methods*, and *decision-making* exhibit symmetrical temporal distributions (Q_3 - Q_2 = Q_2 - Q_1). Although *decision-making* appears with a lower overall frequency, the remaining topics demonstrate both high frequency and temporal balance. During the same period, *socioscientific issues*, *reasoning*, *discourse*, *argumentative practices*, and *teachers* also show broad temporal spans and notable visibility.

In the most recent period (2020 and beyond), the median years (Q_2) of *translanguaging*, *Toulmin*, *epistemology*, *sensemaking*, and *bilingual science education* shift toward the later years of the timeline. Among these, the third-quartile values (Q_3) of *translanguaging*, *Toulmin*, and *technology* extend to the final year of analysis. The temporal distribution of *epistemology* indicates a relatively late onset and a left-skewed (Q_3 - Q_2 < Q_2 - Q_1) pattern. In particular, while the temporal visibility of *bilingual science education* concludes in 2022, *translanguaging* emerges in 2021 and continues through the end of the study period, highlighting its increasing relevance in recent research.

Discussion

This bibliometric study presents a comprehensive perspective on the development of argumentation research in science education over the last quarter century, highlighting its global trends and the thematic transformations it has undergone.

Temporal Development and Scientific Production Trends of Argumentation Studies in Science Education

The temporal distribution of argumentation studies in science education indicates that the field exhibits a dynamic structure that develops in distinct, though non-linear, phases. As shown in Figure 3, the number of publications gained significant momentum, particularly after 2010, reaching its highest value within the examined period in 2025. This quantitative increase is consistent with the logistic growth model estimates and the cumulative growth curve (Figure 2) and can be interpreted as an indication that the field is approaching maturity. Indeed, the high level of fit demonstrated by the model with empirical data ($R^2 = 0.776$) confirms that the identified growth trend reflects a statistically significant life-cycle pattern. This temporal pattern suggests that the field has followed a trajectory consistent with bibliometric life-cycle models. In this respect, the limited publication output between 2001 and 2010 corresponds to a theoretical formation phase, whereas the sharp rise observed during the 2011–2020 period reflects a phase of rapid expansion. The high and relatively stable trajectory after 2020 indicates that the field may be entering its maturation stage. Overall, the acceleration observed after 2010 suggests that argumentation has become one of the central research agendas within science education.

This marked increase appears to be driven by multiple interrelated factors. International large-scale student assessment programs (particularly PISA; OECD, 2006) and science curriculum reforms – such as the Next Generation Science Standards (NGSS; NGSS Lead States, 2013) – have explicitly emphasized scientific reasoning and argumentation, thereby increasing the institutional demand for related research. In addition, the adaptation of Toulmin's (1958) foundational theoretical framework to science education contexts (Driver et al., 2000; Osborne et al., 2004), together with the expansion of open-access publishing, has contributed to the acceleration of scientific output. The high publication volume observed in recent years suggests that scholarly activity in the field remains sustained and that research has begun to demonstrate qualitative diversification beyond mere quantitative growth. This pattern is consistent with contemporary perspectives in the literature, which increasingly conceptualize argumentation not only as a skill to be examined, but also as a core pedagogical approach shaping science teaching and learning processes (Osborne et al., 2019).

Leading Journals, Authors, Countries, and Institutions in Science Education Argumentation Research

The publication and impact patterns of argumentation studies in science education reveal that the field is concentrated around specific journals, leading authors, influential countries, and prominent institutions. This structure suggests that the literature has evolved through sustained research traditions and institutional collaborations rather than isolated individual efforts. At the same time, this structure remains dynamic, as the network continues to expand with the participation of new research communities from diverse geographical contexts.

An examination of journal distribution indicates that outlets such as *International Journal of Science Education, Science & Education,* and *Science Education* function as core publication platforms within the field (Table 2). With their long-standing traditions in science education research, these journals continue to provide central venues for scholarly discussions on foundational topics such as argumentation. At the same time, the notable presence of publications in journals such as *Research in Science & Technological Education*, *International Journal of Science and Mathematics Education*, and *Journal of Science Education and Technology* suggests that argumentation research has increasingly extended toward STEM integration, technology-enhanced learning, and interdisciplinary pedagogical contexts. This diversification indicates that argumentation has evolved into a broader research domain encompassing not only the core dimensions of science education but also contexts related to linguistic diversity and inclusive education (Lee, 2005).

In terms of author productivity, the sustained contributions of researchers such as Archila, Erduran, and McNeill have played a key role in shaping the development of the field (Table 3). In particular, the studies of Erduran and Simon (2004) and McNeill and Krajcik (2008) have become foundational references for understanding classroom-based argumentation and its assessment. Archila's publication profile, in turn, reflects the growing emphasis on linking argumentation with sociological dimensions and citizenship education. Moreover, the work of scholars

such as Sadler, Gonzalez-Howard, Hand, Molina, and Zeidler indicates that the field is not confined to a single core group, but rather continues to expand through diverse theoretical perspectives and research questions. This pattern suggests that argumentation research in science education is characterized by a vibrant and evolving academic structure shaped by center–periphery interactions.

The country-based distribution of publications highlights the dominant role of the USA in terms of research output (Table 4; Figure 4). This prominence can be attributed to the country's well-established research infrastructure in science education and its early engagement with systematic argumentation research. However, the strong representation of China and Turkey indicates that argumentation has emerged as a globally relevant research focus. The fact that Turkey ranks third worldwide with 33 publications (7.0%) is particularly noteworthy. This trend may be associated with revisions to the Turkish science curriculum in 2013 and 2018, which explicitly emphasized competencies related to argumentation—such as evidence-based explanation, discussion, and reasoning—alongside scientific process skills (MEB, 2013, 2018). The relatively balanced contribution of European countries further suggests that argumentation constitutes a shared research agenda across different educational contexts.

At the institutional level, the prominence of the University of Los Andes, Boston College, and several Taiwanese universities (Table 5) indicates that argumentation research tends to be concentrated in institutions with strong graduate programs and established research groups. The inclusion of institutions such as Recep Tayyip Erdoğan University alongside the University of Oxford underscores the presence of influential academic hubs at both global and local levels. Overall, institutional visibility suggests that research productivity in this field is shaped not only by individual scholarly efforts but also by institutional support structures and sustained research networks.

Citation Structures and Intellectual Foundations of Argumentation Studies in Science Education

The citation analyses conducted in this study indicate that argumentation research in science education is characterized by a strong degree of theoretical continuity and is structured around a relatively limited set of highly influential references. The development of the field appears to have been guided by a foundational body of early, high-impact studies that provided a shared intellectual framework for subsequent research. This pattern suggests that the literature has evolved through cumulative articulation around common references rather than through fragmented or random knowledge accumulation.

As shown in Figure 5, studies such as Zohar and Nemet (2002), Erduran et al. (2004), and Duschl (2008) have received particularly high citation counts. These works have played a central role in establishing the theoretical, pedagogical, and discursive dimensions of argumentation in science education. Their sustained citation impact indicates that they continue to function as key reference points in contemporary research rather than serving solely as historical foundations. Similarly, the prominence of studies by Sandoval and Reiser (2004) and Sadler (2004) reflects the early emergence of research linking argumentation to epistemological inquiry and socioscientific contexts.

The reference co-citation network analysis (Figure 6) makes the relational structure between studies forming the theoretical core of argumentation research in science education more visible. Studies such as Driver et al. (2000), Erduran et al. (2004), and Osborne et al. (2004), which occupy a central position in Cluster 1 of the network, are concentrated around core references representing the fundamental orientations of argumentation research in the field based on cognitive and epistemic processes. This core reflects a distinct line of research focusing particularly on the nature of arguments, forms of justification, and students' scientific reasoning processes. In contrast, sources such as Zohar and Nemet (2002), Sadler (2004), Sadler and Zeidler (2005), and Toulmin (1958), which are prominent in Cluster 2 of the network, point to a different research orientation organized around sociological issues, decision-making processes, and discourse-based approaches. The spatial separation between the clusters suggests that these two main trends form relatively independent sub-lines in the literature. However, the connections established by Driver et al. (2000) bridging both clusters and Toulmin's (1958) fundamental position within Cluster 2 through multiple relationships reflect the existence of common theoretical ground and conceptual continuity underlying this apparent divergence. This situation demonstrates that, despite thematic diversification, the field exhibits a tendency toward relational coherence through shared references.

The three-field plot presented in Figure 7 offers a holistic perspective on the intellectual structure of argumentation research in science education, illustrating the relational flows between theoretical roots, the researchers who engage with these roots, and the conceptual foci that emerge from them. The high frequency with which studies

such as Jiménez-Aleixandre and Erduran (2008) and Osborne et al. (2004) appear in the cited sources area of the diagram suggests that the current intellectual dynamism of the field is concentrated in research focused on the restructuring of the theoretical heritage in pedagogical and methodological contexts. In contrast, classic references such as Toulmin (1958) and Kuhn (1993), although with more limited absolute frequencies, continue to nourish the field's enduring epistemic ground through their multiple connections extending to different authors and concept clusters.

The relational patterns observed in the fields of authors and keywords reveal how this theoretical groundwork has been addressed in the literature through a specific focus. The strong connections established by McNeill and Erduran with the concepts of *argumentation* and *knowledge* represent the field's line of research focused on cognitive and epistemic processes, while the relationships established by Sadler and Zeidler with *socioscientific issues* and the concept of *decision-making* point to a different orientation that centers on the contextual and social dimensions of argumentation. The fact that authors such as Erduran, Sadler, and Kuhn are represented both as referenced sources and as authors, and that they occupy an important mediating position in the flow from classical references to contemporary concepts, demonstrates the establishment of a relational continuity between theoretical heritage and application-oriented research. This situation suggests that these authors, occupying both reference and producer positions in the three-field plot, serve as a bridge in transferring theoretical frameworks to current pedagogical and methodological developments. When considered alongside the citation patterns in Figures 5 and 6, this structure indicates that argumentation studies in science education have developed within an intellectual order that varies across thematic emphases yet progresses through shared theoretical foundations. Taken together, these findings address the third research question by revealing a stable yet evolving intellectual structure shaped by enduring theoretical traditions and expanding methodological orientations.

Dominant Thematic Orientations, Thematic Transformations, and Research Trends in Science Education Argumentation Studies

The results of the thematic analysis reveal that argumentation in science education is not a one-dimensional field of research, but rather a multi-layered and dynamic domain in which pedagogical, epistemic, and contextual dimensions are intertwined. This structure has expanded and deepened cumulatively over time, gaining complexity through successive layers. Indeed, argumentation research has shifted from early approaches focused primarily on cognitive processes to a more holistic orientation in which knowledge is constructed through social interaction and contextualized practices. This transformation reflects the increasing visibility of argumentation in science education as a practice that encompasses not only individual reasoning skills but also the social and dialogical construction of scientific knowledge.

The co-occurrence network of keywords presented in Figure 8 illustrates the conceptual structure of the field and the strength of relationships among key terms. The high co-occurrence density between *argumentation* and *science education* at the center of the network suggests that these concepts are not treated independently in the literature; rather, *argumentation* is positioned as a foundational component of *science education*. Concepts such as *discourse*, *dialogue*, *evidence*, and *epistemic practice*, which are concentrated in the central cluster, indicate that the field's focus has shifted from merely producing correct answers to the epistemic processes of how scientific knowledge is produced and justified. This conceptual orientation aligns with the perspective of Driver et al. (2000), who frame science learning within the social construction of scientific knowledge and its epistemic underpinnings.

The relationships established between the central core and concepts such as *socioscientific issues*, *reasoning*, *decision-making*, and *scientific literacy*, which emerge prominently in other thematic wings of the network, reflect how *argumentation* has expanded beyond its cognitive boundaries to encompass ethical, social, and citizenship-focused dimensions. This expansion is consistent with contemporary approaches that emphasize the consideration of *argumentation* in the context of socioscientific debates (e.g., Sadler, 2004; Sadler & Zeidler, 2005). In particular, the connections established between the set of socioscientific topics and epistemic concepts suggest that *argumentation* functions as a fundamental reasoning tool in making sense of current social issues. Similarly, the co-location of teaching-focused concepts such as *teachers*, *professional development*, *inquiry*, and *collaboration* indicates that the theoretical discussions in the literature are progressing towards integrating classroom practices and pedagogical transformation. This situation can be interpreted as argumentation research moving beyond being a niche pedagogical field to becoming one of the mainstream components of science education.

In contrast, the relatively low connectivity of concepts such as *translanguaging*, *climate change*, and *earth science education*, which are located in more peripheral areas of the network, indicates that these themes have not yet taken on a framing role in the argumentation literature. Similarly, the fact that the concepts of *pedagogy* and *teaching* establish more limited relationships with central clusters provides data suggesting that there is still room for development between theoretical argumentation studies and the general pedagogical literature. This view reflects that the intellectual depth of the field tends to concentrate around a specific conceptual core; however, its spread in interdisciplinary contexts and specific content areas shows signs of relative maturation. This situation can be interpreted as meaning that, while the field has a strong theoretical foundation, it has potential for development in terms of application and contextual expansion, and that the central role of critical discourse in this process remains important (Osborne, 2010).

The thematic map analysis presented in Figure 9 illustrates the strategic positions and maturity levels of the themes in the field. The positioning of the *argumentation* theme (Cluster 1) in a transitional zone between fundamental and motor themes indicates that this line of research has matured and is guiding the field. This supports the view that *argumentation* has become an established epistemic practice in science education research that guides the production, justification, and verification of knowledge, rather than merely serving as a teaching method. This positioning indicates that the view that argumentation plays a central role in the social construction of scientific knowledge (Driver et al., 2000) has become a prominent framework in the field. The inclusion of *science education*, *inquiry*, teacher-focused studies, and collaborative learning approaches within Cluster 2 reflects the field's strong pedagogical grounding. Its location among the basic themes highlights the potential for deeper integration between inquiry-based, teacher-centered approaches and argumentation research.

The positioning of Cluster 4 in a transitional space between motor and niche themes suggests that methodological and assessment-oriented studies have developed with consistent but relatively limited centrality in argumentation research. This finding aligns with prior observations that the development of robust tools for classroom-based assessment of argumentation remains an ongoing research need (Osborne et al., 2004). Cluster 5, situated in the niche themes quadrant, focuses on pre-service teacher education, pedagogical dialogue, and classroom practices, indicating that teacher education constitutes a specialized and context-sensitive area of expertise. This interpretation is consistent with studies emphasizing the importance of instructional support and professional preparation for teaching argumentation effectively (McNeill & Krajcik, 2008). Finally, the structure of Cluster 3, located in the emerging or declining themes area and centered on *socioscientific issues*, *reasoning*, *nature of science*, and *decision-making*, suggests that this thematic line is still in a developmental phase. This finding is consistent with approaches highlighted by researchers such as Sadler (2004) and Zeidler (2014) in the context of socioscientific issues, which position *argumentation* as a central component of social and ethical reasoning. Similarly, recent bibliometric results also confirm that socioscientific issues are emerging as a developing and strategically important theme in argumentation research (Noris et al., 2024).

The trend topics analysis shown in Figure 10 reveals the temporal development of argumentation research in science education and its patterns of thematic transformation. This analysis enables a combined interpretation of the conceptual continuity illustrated in Figure 8 and the historical maturation of thematic positions identified in Figure 9. In doing so, it not only identifies the themes structuring the field but also highlights the periods and intensities during which these themes gained prominence, thereby complementing the preceding analyses.

In particular, the symmetrical temporal distribution exhibited by themes with high frequency of use shows that topics such as *argumentation*, *science education*, and *methods*, which occupy a central position in Figure 8, have consistently maintained their foundational roles in the field over time. This view indicates that these concepts function as fundamental research axes that are continuous in the literature rather than being temporary areas of interest specific to certain periods (Osborne et al., 2004, Tosun, 2024).

The low first-quartile values observed for themes such as *environmental education*, *earth science education*, and *biotechnology*, more visible in the early period, suggest that these areas function as "heritage themes" that emerged during the formative phase of the field and laid the groundwork for later research. In this study, the term "heritage themes" refers to topics that were relatively prominent during the early developmental phase of the field and helped establish foundational research directions, even if their later visibility declined or diversified. In particular, the right-skewed temporal distribution of the earth science education theme indicates that this discipline-specific focus played a significant role in the early acceleration of the field but gradually yielded to more interdisciplinary and pedagogically oriented approaches. In contrast, the early emergence and sustained high frequency of use of the inquiry theme indicates that a framework focusing on the epistemic dimension of learning processes and scientific discourse practices remained dominant in the early stages of argumentation research (Duschl & Osborne, 2002).

The thematic diversification observed between 2015 and 2020 reveals that while the theoretical core of the field has been preserved, there has been a significant expansion in pedagogical and social dimensions. During this period, the high frequency and broad temporal distribution of *socioscientific issues*, *reasoning*, and *discourse* topics indicate that argumentation began to be approached as a social, ethical, and citizenship-based reasoning tool (Sadler, 2004). This trend is also mirrored in the placement of these themes within emerging and transitional clusters in Figure 9.

A particularly notable thematic transformation concerns the temporal restructuring of language-focused research. Although the *bilingual science education* theme exhibits a relatively limited temporal span up to 2022, the continued visibility of the *translanguaging* theme through the end of the analysis period suggests more than a terminological shift. Rather, it points to the growing influence of theoretical orientations that conceptualize language as a dynamic, contextual, and meaning-making resource in learning processes (Garcia & Wei, 2014). This transformation parallels the maturation of language-related themes located in emerging clusters in Figure 9 and aligns with recent synthesis studies that frame *argumentation* as an epistemic practice grounded in discourse, interaction, and multimodal representation (Tang, 2024). Similarly, the left-skewed temporal distribution of the *epistemology* topic indicates that, despite its relatively late emergence, it rapidly became a focal point of scholarly attention, underscoring the increasing centrality of epistemic concerns in argumentation research.

It is noteworthy, however, that several fundamental pedagogical concepts with high overall usage frequencies – such as *learning*, *assessment*, *modeling*, *collaboration*, and the *nature of science* – do not appear as distinct trend topics in Figure 10. This outcome is attributable to the methodological logic of trend topics analysis, which emphasizes temporal concentration rather than cumulative frequency. The widespread but temporally dispersed use of these concepts suggests that they function as pedagogical and methodological "fundamental constants" embedded throughout the field, rather than as time-bound focal topics. In this study, the term "fundamental constants" denotes core concepts that maintain a stable presence across all phases of the field's development instead of peaking within a specific period. Their stable cross-temporal presence indicates that they operate as structural anchors within the evolving thematic landscape.

Recent temporal patterns indicate that, with the resurgence of topics such as the Toulmin model, interpretation, and technology, a more holistic relationship is beginning to emerge between theoretical frameworks and application-oriented approaches. The prominence of the Toulmin model in this context indicates that the model is being revisited in line with current research questions and inquiry-based needs, rather than a return to a classical framework (Kelly, 2014). Overall, when the results obtained from Figure 10 are evaluated together with the conceptual and strategic data in Figures 8 and 9, the findings suggest that the argumentation literature in science education is organized around a solid theoretical core and is moving towards a structure that integrates contextual, epistemic, and linguistic dimensions in a more visible way over time. This view presents a framework suggesting that argumentation in science education is beginning to be addressed not only as a pedagogical tool but also within a broader context that integrates the justification of scientific thinking, epistemic processes supported by the history and philosophy of science, and a participatory understanding of science (Archila, 2015; Jiménez-Aleixandre & Erduran, 2008).

Conclusion

The bibliometric analyses conducted in this study indicate that argumentation research in science education has experienced substantial development and has moved toward a relative level of maturation over the last quarter century. The field has not only expanded in terms of quantitative output but has also acquired a multi-layered structure that strengthens its theoretical foundations, broadens its research networks, and diversifies its thematic orientations. By addressing temporal, social, intellectual, and thematic dimensions through an integrated bibliometric approach, this study has revealed the structural and conceptual development of the field from a holistic perspective. Temporal patterns demonstrate that argumentation has assumed a central position in science education research, particularly since the 2010s. This trend suggests that the field has moved beyond being a temporary research focus and is increasingly consolidating into a more stable and enduring research domain, closely linked to curriculum reforms and international assessment frameworks.

Analyses of the social and intellectual structures reveal that research in this field is shaped by strong academic networks and shared theoretical references. Common citation patterns reflect that the field is built on deep theoretical foundations; cognitive, epistemic, and discourse-based approaches have developed within a complementary relational unity. However, the intellectual geography of the field is also seen to be expanding

steadily. The prominence of countries such as Turkey and China alongside established research centers such as the United States reflects the global diffusion of interest in argumentation studies. This expansion appears to be associated with epistemic orientations embedded in local curriculum reforms and teacher education programs.

The thematic results provide a more detailed picture of the transformation experienced by the field. Rather than exhibiting a homogeneous pattern of development, argumentation research in science education demonstrates a dynamic structure in which multiple thematic strands evolve simultaneously but at different levels of maturity. Early studies focusing primarily on argument structures and individual cognitive processes have gradually become more integrated with socioscientific contexts, social decision-making processes, and classroom practices. The increasing prominence of application-oriented themes – particularly teacher education and formative assessment – has reinforced the tendency to translate the field's accumulated theoretical knowledge into pedagogical practice. This broader perspective suggests that the literature has moved beyond the question of how arguments are structured, progressing instead toward more dynamic theoretical frameworks. In particular, language-focused studies reveal a terminological and conceptual shift from a *bilingual education* orientation toward a *translanguaging* perspective. At the same time, classical argumentation models, such as Toulmin's framework, are not being displaced but rather re-contextualized and re-functionalized within digital and contemporary pedagogical environments.

Fundamental concepts such as *learning* and *assessment*, although not temporally prominent in thematic maps or represented within niche clusters, continue to function as infrastructural constants across all developmental phases of the field. This persistence can be interpreted as an important indicator of the methodological maturity and theoretical depth achieved in science education argumentation research. In this context, it can be said that the comprehensive keyword standardization and conceptual consolidation approach applied in the study strengthens the analytical consistency and validity of thematic mapping; in this respect, it provides a methodological reference framework for similar bibliometric studies. In conclusion, *argumentation* is positioned in the literature not only as a pedagogical tool in science education but also as a multi-layered and expanding epistemic ground that relates scientific reasoning to language, culture, technology, and democratic citizenship practices.

Recommendations

The bibliometric maps and thematic structure analyses presented in this study highlight current trends, structural gaps, and developmental opportunities in argumentation research in science education, suggesting several directions for future research:

- *Deepening niche areas and theoretical integration should be prioritized*. Thematic maps indicate that areas such as formative assessment, cross-linguistic transfer, teacher professional development, and curriculum integration remain niche or emerging. Future research should place greater emphasis on experimental and design-based studies that explicitly connect these themes to the core theoretical foundations of the field and to classroom practice. Systematically linking these areas to the mainstream research agenda may enhance not only thematic diversity but also theoretical coherence and pedagogical impact. Strengthening their connection with cross-temporal constants such as learning and assessment would further anchor these developments within the field's core structure.
- *Conceptual and methodological expansion should be encouraged*. Argumentation studies remain concentrated around specific theoretical frameworks. Future research may benefit from refunctionalizing classical models such as Toulmin's within contemporary digital learning environments and AI-supported meaning-making processes. Methodologically, integrating bibliometric mapping with longitudinal case studies and discourse analysis aimed at explaining how and why these trends emerge could provide a more explanatory account of the field's developmental dynamics.
- *Global and comparative perspectives should be made more visible in the research agenda*. Bibliometric production networks reveal a marked rise, particularly in countries such as China and Turkey, indicating that argumentation research is increasingly multi-centered. Comparative studies across educational systems and cultural contexts therefore represent a promising direction for future investigation.
- *Greater emphasis should be placed on impact-oriented research*. While the literature demonstrates strong connections between argumentation-based pedagogies and classroom practices, their long-term effects on student outcomes remain underexplored. Future studies should examine sustained impacts on domains such as scientific literacy, critical thinking, and citizenship education, positioning argumentation not only as a pedagogical strategy but also as an epistemic foundation.
- *Methodological transparency and standardization should be strengthened*. Future bibliometric research would benefit from more explicit reporting of data-cleaning procedures, keyword standardization

processes, and threshold selection criteria. Such transparency would contribute to greater comparability across studies and support the cumulative advancement of the field.

Limitations

Several limitations should be considered when interpreting the results of this study. First, the analyses are restricted to publications indexed in the WoS database. Although this choice ensures methodological rigor in terms of citation standards, data consistency, and bibliometric comparability, it necessarily excludes relevant studies indexed in other databases such as Scopus or Google Scholar. Consequently, the results primarily reflect high-impact publications with strong international visibility.

Second, the inclusion of only English-language publications may have limited the representation of studies produced in local languages (e.g., Turkish or Chinese) within the thematic and co-occurrence networks. Nevertheless, this decision facilitated clearer tracing of conceptual interactions and shared epistemic frameworks within the international literature. In addition, the inclusion of publications up to the end of 2025 means that some recent studies may not yet have reached full citation maturity, potentially limiting the interpretive depth of citation-based network analyses.

Finally, by their nature, bibliometric analyses do not aim to capture the qualitative depth of individual studies. The threshold values and algorithmic parameters used to enhance visual clarity and analytical focus may have led to the underrepresentation of concepts that are conceptually important but dispersed over time. Accordingly, the thematic structures identified in this study are not intended to replace in-depth pedagogical or theoretical analyses; rather, they provide a systematic and comparable analytical foundation that can inform future qualitative and mixed-method investigations.

These limitations define the scope of the macro-level perspective offered by bibliometric mapping rather than undermining the validity of the results. Future research that integrates multiple databases, includes multilingual publications, and complements bibliometric analyses with qualitative content analysis would contribute to a more comprehensive and nuanced understanding of argumentation research in science education.

Scientific Ethics Declaration

* The data used in this study were obtained from secondary bibliographic sources provided by WoS and did not involve any human participants or personal data. Therefore, this study does not require ethical committee approval. The principles of scientific research and publication ethics were strictly adhered to at every stage of the analysis process.

Conflict of Interest

* There is no conflict of interest among the authors regarding this study.

Funding

* The authors received no financial support for the research, authorship, and/or publication of this article.

Acknowledgements or Notes

* This article is derived from the first author's doctoral dissertation, supervised by the second author.

References

Archila, P. A. (2015). Using history and philosophy of science to promote students' argumentation. *Science & Education*, *24*(9), 1201–1226. https://doi.org/10.1007/s11191-015-9786-2

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, *84*(3), 287–312.

Duschl, R. (2008). Science education in three-part harmony: balancing conceptual, epistemic, and social learning goals. *Review of Research in Education*, *32*(1), 268–291. https://doi.org/10.3102/0091732X07309371

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education*, *38*(1), 39–72. https://doi.org/10.1080/03057260208560187

Erduran, S., & Simon, S. (2004). The role of argumentation in developing scientific literacy. In M. C. Linn, E. A. Davis, & P. Bell (Eds.), *Internet environments for science education* (pp. 115–140). Lawrence Erlbaum Associates.

Erduran, S., Simon, S., & Osborne, J. (2004). Tapping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education*, *88*(6), 915–933. https://doi.org/10.1002/sce.20012

Garcia, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Macmillan.

Jiménez-Aleixandre, M. P., & Erduran, S. (2008). Argumentation in science education: An overview. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from network learning, schools and computers* (pp. 3–27). Springer.

Kelly, G. J. (2014). Inquiry, signaling, and argument in science education. In J. Loughran, A. Berry, & P. Mulhall (Eds.), *Teaching science in the secondary school* (pp. 113–125). Routledge.

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education*, *77*(3), 313–337. https://doi.org/10.1002/sce.3730770306

Kurtuluş, M. A., & Yılmaz, S. (2022). STEM eğitim çalışmalarına farklı bir bakış: Bibliyometrik haritalama. *Fen Bilimleri Öğretimi Dergisi*, *10*(2), 386–405. https://doi.org/10.56423/fbod.1172514

Lee, O. (2005). Science education with English language learners: Synthesis and research agenda. *Review of Educational Research*, *75*(4), 491–530. https://doi.org/10.3102/00346543075004491

McNeill, K. L., & Krajcik, J. (2008). Scientific explanations: Characterizing and evaluating the effects of teachers' instructional practices on student learning. *Journal of Research in Science Teaching*, *45*(1), 53–78. https://doi.org/10.1002/tea.20201

MEB. (2013). *İlköğretim kurumları (ilkokullar ve ortaokullar) fen bilimleri dersi (3, 4, 5, 6, 7 ve 8. sınıflar) öğretim programı*. T.C. Milli Eğitim Bakanlığı.

MEB. (2018). *Fen bilimleri dersi öğretim programı (İlkokul ve ortaokul 3, 4, 5, 6, 7 ve 8. sınıflar)*. T.C. Milli Eğitim Bakanlığı.

Mulyani, A., Hartono, H., & Subali, B. (2024). Literature review: A snapshot of research on the argumentation of bibliometric analysis in the period 2015–2023. *International Journal of Cognitive Research in Science, Engineering and Education*, *12*(2), 451–465. https://doi.org/10.23947/2334-8496-2024-12-2-451-465

NGSS Lead States. (2013). *Next generation science standards: For states, by states.* National Academies Press.

Noris, M., Sajidan, S., Saputro, S., & Yamtinah, S. (2024). Trends and issues of inquiry and socio-scientific issue (SSI) research in the last 20 years: A bibliometric analysis. *International Journal of Education in Mathematics, Science, and Technology*, *12*(3), 773–792. https://doi.org/10.46328/ijemst.3767

OECD. (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006.* OECD.

Orhan, A. T. (2024). Eğitim bilimleri alanında STEM araştırmalarının bibliyometrik analizi. *Gazi Eğitim Bilimleri Dergisi*, *10*(3), 375–396. https://doi.org/10.30855/gjes.2024.10.03.004

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, *328*(5977), 463–466. https://doi.org/10.1126/science.1183944

Osborne, J. F., Borko, H., Fishman, E., Gomez Zaccarelli, F., & Berson, E. (2019). Impacts of a practice-based professional development program on elementary teachers' facilitation of and student engagement with scientific argumentation. *American Educational Research Journal*, *56*(4), 1067–1112. https://doi.org/10.3102/0002831218812059

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching*, *41*(10), 994–1020. https://doi.org/10.1002/tea.20035

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching*, *41*(5), 513–536. https://doi.org/10.1002/tea.20009

Sadler, T. D. (2009). Situated learning in science education: Socio-scientific issues as contexts for practice. *Studies in Science Education*, *45*(1), 1–42. https://doi.org/10.1080/03057260802681839

Sadler, T. D., Barab, S. A., & Scott, B. (2007). What do students gain by engaging in socioscientific inquiry? *Research in Science Education*, *37*(4), 371–391. https://doi.org/10.1007/s11165-006-9030-9

Sadler, T. D., & Zeidler, D. L. (2005). Patterns of informal reasoning in the context of socioscientific decision making. *Journal of Research in Science Teaching*, *42*(1), 112–138. https://doi.org/10.1002/tea.20042

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education*, *92*(3), 447–472. https://doi.org/10.1002/sce.20276

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education*, *89*(4), 634–656. https://doi.org/10.1002/sce.20065

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, *88*(3), 345–372. https://doi.org/10.1002/sce.10130

Tang, K. S. (2024). Informing research on generative artificial intelligence from a language and literacy perspective: A meta-synthesis of studies in science education. *Science Education*, *108*(5), 1329–1355. https://doi.org/10.1002/sce.21875

Tosun, C. (2024). Analysis of the last 40 years of science education research via bibliometric methods. *Science & Education*, *33*(2), 451–480. https://doi.org/10.1007/s11191-022-00400-9

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Wang, S., Chen, Y., Lv, X., & Xu, J. (2023). Hot topics and frontier evolution of science education research: A bibliometric mapping from 2001 to 2020. *Science & Education*, *32*(3), 845–869. https://doi.org/10.1007/s11191-022-00337-z

Zeidler, D. L. (2014). Socioscientific issues as a curriculum emphasis: Theory, research, and practice. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (Vol. II, pp. 697–726). Routledge.

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, *39*(1), 35–62. https://doi.org/10.1002/tea.10008

Zupic, I., & Cater, T. (2015). Bibliometric methods in management and organization. *Organizational Research Methods*, *18*(3), 429–472. https://doi.org/10.1177/1094428114562629

Author(s) Information

Esra Ergunt
Turkish Ministry of National Education,
Ankara, Türkiye
ORCID iD: https://orcid.org/0009-0007-4265-2882

Serkan Yilmaz
Hacettepe University,
Faculty of Education, Beytepe Campus, Ankara, Türkiye.
Contact e-mail: sekoyil@gmail.com
ORCID iD: https://orcid.org/0000-0003-1800-0765

Appendix 1. Consolidated Keyword List and Variant Mapping: 51 Superordinate Concepts; n = 396 Variants

The table below presents the complete mapping of 396 keyword variants consolidated under 51 superordinate conceptual terms. Consolidations were based primarily on lexical proximity and recurrent contextual usage within the dataset. The objective was to reduce terminological fragmentation while preserving thematic coherence in the analytical framework. Superordinate concepts are listed in alphabetical order.

Superordinate Concept	Included Variants
Argumentation	argumentation, scientific argumentation, socioscientific argumentation, socio-scientific argumentation, science argumentation, argument, arguments, argument structure
Argumentative Practices	argumentative practices, argumentation skills, argument skills, evidence-based arguments, argumentative writing, argumentation analysis, argumentation learning, dialectic argumentation, bilingual written scientific argumentation, individual argumentation, group argumentation, classroom argumentation, online argumentation, e-argumentation, engagement in argumentation, difficulties in writing arguments, argument evaluation, argument podcasting, argument generation skills, argument map, argument mapping, argument quality, argumentation activities, argumentation norms, argument-based instruction
Assessment	assessment, formative assessment, educational assessment, classroom assessment, self-assessment
Biotechnology	biotechnology, biotechnology education, biotechnology attitudes
Bilingual Education	bilingual education, bilingual science education, emergent bilinguals, emergent bilingual students, bilingual/bicultural, university bilingual science courses, university bilingual science education
Climate Change	climate change, climate change education
Collaboration	collaboration, collaborative argumentation, collaboration argumentation, collaboration scripts, collaborative, collaborative contexts, collaborative contributor, collaborative learning, collaborative talk, computer-supported collaborative learning, student collaboration, cognitive collaboration
Conceptual Understanding	conceptual understanding, conceptual change, conceptual development, conceptual learning, conceptual analysis
Critical Thinking	critical thinking, critical and analytical thinking, critical questioning, critical evaluation, critical evaluation skills, critique, scientific critical thinking, intellectual humility, perspective taking, higher order thinking, higher-order thinking, high-order learning skills, comparing and contrasting
Curriculum	curriculum, curriculum development, curricular tasks, curriculum enactment, educational design, educative curriculum
Decision-making	decision-making, decision making, decision-making competence, decision
Dialogue	dialogue, dialogism, dialogical argumentation, dialogic teaching, dialogic pedagogy, dialogic instruction, dialogic feeback, dialogic discourse, dialogic, dialogue theory
Disciplinary	disciplinary, disciplinary practices, disciplinary enculturation, discipline background
Discourse	discourse, discourse analysis, classroom discourse, argumentative discourse, argumentative discourse practices, classroom talk, classroom dialog, classroom discussion, classroom interaction, productive talk, interaction analysis, interactional patterns, temporal properties of teacher talk
Earth Science Education	earth science education, geology
Education	education, education for, educational policy, educational practice, educational innovation, educational questions
Elementary Science Education	elementary science education, elementary science, elementary school science, early education, elementary, elementary education
Environmental Education	environmental education, environmental, environments, environmental health, education for sustainable development, ecology, sustainability, sustainable development

Epistemic Practice	epistemic practice, epistemic practices, epistemologies in practice, epistemic discourse, epistemic tools, epistemic agency
Epistemology	epistemology, epistemological beliefs, epistemic beliefs, epistemological development, epistemological stance, epistemic cognition, epistemic criteria, epistemic goals, epistemic goal, epistemic insight, epistemic uncertainty, epistemic vigilance, scientific epistemological, personal epistemology, practical epistemology
Evidence	evidence, evidence use, evidence based, evidence evaluation, scientific evidence
Inquiry	inquiry, scientific inquiry, inquiry-based learning, inquiry learning, inquiry-based education, inquiry-based science, inquiry-based teaching, inquiry-type experiment, argument-based inquiry approach
Language	language, language use, language and literacy of science, language of classroom science, language of science, language of science and classrooms
Learning	learning, learning environment, learning preferences, learning progression, learning study, learning with social media, authentic learning, context-based learning, situated learning, cooperative learning, self-regulated learning, game-based learning, digital learning, mobile learning, flipped learning approach, intentional learning, student-centred learning
Methods	methods, methodology, mixed methods, methodologies, methodology of analysis, (dbr), design-based research, design study, case study, case studies, qualitative research, quantitative research, ethnography, content analysis, lag sequential analysis (lsa), multilevel modeling, linear mixed-effects, exploratory factor analysis (efa)
Middle School	middle school, middle school students, middle
Modeling	modeling, modelling, model, model-based learning, models, scientific modeling, modeling-based learning, modeling-based teaching, modelling-based teaching, scientific models, construct modeling
Nature of Science	nature of science, nature of science (nos), nature of scientific practice, (nos)
NGSS	NGSS, next generation science standards, (ngss)
Pedagogy	pedagogy, pedagogical issues, pedagogical content knowledge, pedagogical content knowledge (pck), technological pedagogical content knowledge, pedagogic reform, pedagogical approaches, pedagogical principles, pedagogical content
Preservice Teachers	preservice teachers, pre-service teachers, preservice science teachers, preservice elementary teachers, initial teacher education, in-service teachers, student teachers
Professional Development	professional development, teacher professional development, professional
Rasch Model	rasch model, rasch measurement, rasch analysis, rasch partial credit model
Reasoning	reasoning, socioscientific reasoning, informal reasoning, scientific reasoning, reasoning skills, evidence-based reasoning, reasoning abilities, reasoning levels, reasoning map, social scientific reasoning, flaws in reasoning, logical connectives, logical connectors, hypotheses, causal inference, induction, heuristics, bias, open-mindedness
Scaffolding	scaffolding, teacher scaffolding, argumentation scaffolding, computer-based scaffolding, hard scaffolding
Science	science, school science, sciences, science and culture, science capital, science identity, science professionals, science studies, science major
Science Education	science education, science learning, science interest, science investigation, science and technology literacy, university science education, reform-based science, community science
Science Practices	science practices, science-as-practice, scientific practices
Science Teaching	science teaching, science instruction, science classrooms, science classroom discourse, science teaching contexts, science teachers, science teachers' beliefs, science teacher education
Scientific	scientific, scientific knowledge, scientific uncertainty, scientific theory
Scientific Literacy	scientific literacy, literacy, science literacy, scientific and technological literacy (stl) teaching

Sensemaking	sensemaking, science sensemaking, meaning making, meaning-making, explanation construction, meaningful learning
Socioscientific Issues	socioscientific issues, socio-scientific issues, socioscientific, socio-scientific issue, socio-scientific, socioscientific issues-based instruction, (ssi), ssi, socioscientific issue (ssi), socioscientific issues (ssi), socioscientific issues (ssis), socio-scientific issues (ssis), local socioscientific issues
Students	students, student, student achievement, student attitudes, student beliefs, student culture, student diversity, student knowledge, student learning, student perceptions, student performance, student standpoint
Teachers	teachers, teacher, teacher beliefs, teacher perceptions, teachers' perceptions, teacher views, teachers' practices, teacher mastery, teacher's role, teachers in Saudi Arabia
Teacher Education	teacher education, teacher training, teacher development, teacher learning
Teaching	teaching, teaching practice, teaching practices, teaching strategies, teaching/learning strategies, teaching-learning sequence, teaching context, teaching materials in science education, teaching/distance education (cc)
Teacher Moves	teacher moves, teacher questions, teacher planning, teacher-student interactions, teacher collaboration
Technology	technology, technology-enhanced learning, technology-enhanced classroom, interactive technology
Toulmin	toulmin, rebuttal argument, toulmin argument pattern, toulmin argumentation pattern (tap), warrants, claims, abductive argument, sound argument, faulty argument, hypothetico-predictive argumentation, argument-critique-argument
Translanguaging	translanguaging, multilingualism, multilingual learners, multilingual students, English as a second language, English-learning, linguistically responsive teaching, content and language integration, academic language

JESEH
Journal of Education in Science, Environment and Health
Volume 12, Issue 2, 2026
https://doi.org/10.55549/jeseh.896

Conditional Effects of AI Homework Tools on Students' Academic Performance: A Systematic Synthesis of Empirical Evidence

Seyma Irmak, Kaan Bati

Article Info

Article History

Published:
01 April 2026

Received:
27 January 2026

Accepted:
05 March 2026

Keywords
Artificial intelligence in education,
Generative AI,
Systematic narrative synthesis design

Abstract

The rapid diffusion of generative artificial intelligence (AI) tools into educational contexts has fundamentally transformed how students approach homework, academic writing, and independent learning tasks. Whilst AI-assisted homework tools promise efficiency, personalization, and immediate feedback, there remains some debate over their implications for academic performance and learning quality. The present study proffers a thorough synthesis of empirical evidence, examining how students' academic performance differs when using AI homework tools compared to traditional homework methods. The review draws on experimental, quasi-experimental, and observational research conducted across secondary and higher education contexts. The findings of the study indicate that AI homework tools are associated with significantly higher grades and writing scores in most controlled comparisons, particularly in language learning contexts, with effect sizes ranging from medium to large. However, the evidence also reveals important trade-offs, including reduced knowledge retention, lower originality, and diminished critical thinking in some settings. The synthesis demonstrates that AI tools primarily optimize output quality rather than learning processes, and that their effectiveness is highly conditional on task characteristics, assessment timing, implementation fidelity, and learner characteristics.

Introduction

Homework has long been considered a core component of formal education, functioning as a mechanism through which students rehearse skills, consolidate conceptual understanding, and develop independent learning habits (Cooper et al., 2006; Dettmers et al., 2009). Classical and contemporary research on homework effectiveness has consistently emphasized that its educational value depends not merely on the quantity of tasks assigned but, on their quality, alignment with instructional goals, and the nature of feedback provided (Hattie, 2009; Trautwein & Köller, 2003). As educational technologies have evolved, homework practices have been repeatedly reshaped, from paper-based assignments to online learning platforms and adaptive systems that offer automated feedback and progress monitoring (Dede, 2014).

The emergence of generative artificial intelligence (GenAI) represents a qualitative shift in this trajectory. Unlike earlier educational technologies that delivered pre-scripted content or rule-based feedback, contemporary AI systems—such as large language models (LLMs), automated writing evaluation tools, and intelligent tutoring systems—can generate explanations, examples, and complete textual outputs in response to students' prompts (Zhou et al., 2024). These systems increasingly accompany students during homework completion, offering real-time assistance that resembles aspects of human tutoring and raising fundamental questions about authorship, cognition, and learning responsibility (Smerdon, 2024).

The swift integration of AI-driven homework assistants into educational settings has ignited a rigorous scholarly discourse. Proponents highlight the potential of these tools to democratize high-quality feedback and provide vital support for learners facing language barriers or gaps in prior knowledge, primarily through personalized and immediate interventions (Song & Song, 2023; Tamimi et al., 2024). Within this framework, AI functions as a scalable mechanism for scaffolding (Wood et al., 1976), extending instructional guidance beyond the physical classroom and operationalizing Vygotsky's (1978) socio-constructivist principles in the digital age (Luckin et al., 2016). Critics, however, caution that reliance on AI may encourage surface-level engagement, undermine critical thinking, and weaken knowledge retention by offloading essential cognitive processes to automated systems (Yavich, 2025; Yang, 2025). Concerns regarding academic integrity, authorship, and educational equity further complicate the discourse (Smerdon, 2024).

Despite the prominence of these debates, empirical evidence remains fragmented. Individual studies report divergent findings, with some documenting substantial improvements in assignment quality and grades (Chen & Gong, 2025; Song & Song, 2023) and others finding negligible or even negative effects on learning-related outcomes such as retention and independent reasoning (Yang, 2025; Yavich, 2025). Moreover, academic performance is operationalized inconsistently across studies, ranging from immediate assignment scores to delayed retention measures and qualitative indicators of engagement (Kwak, 2025). Consequently, educators and policymakers face challenges in interpreting whether observed performance gains reflect genuine learning or merely improved outputs.

To resolve this critical ambiguity and distinguish cognitive growth from superficial task completion, the present study addresses this gap by synthesizing empirical studies that directly compare AI-assisted homework with traditional homework methods. Rather than treating AI effectiveness as a binary outcome, the analysis adopts a conditional perspective, examining how task complexity, assessment timing, implementation fidelity, and learner characteristics shape observed effects (AlShibli et al., 2025; Ward et al., 2025). The research question is: How do students' academic performance outcomes differ when using AI homework tools compared to traditional homework methods?

Literature Review

AI Homework Tools and Theoretical Perspectives

AI homework tools encompass a diverse spectrum of technologies designed to facilitate out-of-class learning, ranging from long-standing intelligent tutoring systems (ITS) and automated writing evaluation (AWE) tools to contemporary generative AI (GenAI) models (Kelly et al., 2013; Zhou et al., 2024). While traditional AI tools focused on adaptive feedback within structured environments, the latest generation is distinguished by its generative capacity, enabling the dynamic creation of explanations, summaries, and complex academic drafts that go beyond static content delivery.

From a theoretical standpoint, AI homework tools intersect with multiple learning theories. Behaviorist perspectives emphasize the role of immediate feedback and reinforcement, which AI systems can deliver consistently at scale (Skinner, 1961; Kulik & Fletcher, 2016). By providing instantaneous corrections and rewards, these tools mirror the programmed instruction model, ensuring that learners consolidate correct responses before advancing to more complex tasks. Cognitive load theory suggests that AI-generated explanations and worked examples may reduce extraneous cognitive load, allowing learners to allocate more cognitive resources to germane processing (Sweller et al., 2011, pp. 15-16, 102). In contrast, constructivist and socio-cognitive frameworks highlight the importance of active knowledge construction, metacognitive monitoring, and productive struggle (Kapur, 2008; Zimmerman, 2002). Within these frameworks, AI tools may function either as scaffolds that support learning or as substitutes that bypass essential cognitive engagement, depending on how they are integrated into instructional design (Zhou et al., 2024).

Homework, Performance, and Learning Depth

Academic performance is frequently assessed using proximal indicators such as assignment grades, rubric-based scores, or standardized test results (Hattie, 2009). While such measures provide tangible evidence of achievement, they do not necessarily capture deeper learning outcomes, including conceptual understanding, transfer, and long-term retention (Bjork & Bjork, 2011; Soderstrom & Bjork, 2015). Research on surface versus deep learning demonstrates that learners can achieve high immediate performance while developing fragile knowledge structures that do not support future application (Marton & Säljö, 1976).

In AI-assisted homework contexts, this distinction becomes particularly salient. Generative AI systems are highly effective at improving surface features of academic work, such as grammatical accuracy, coherence, and organization (Chen & Gong, 2025; Song & Song, 2023). However, their influence on higher-order outcomes—critical thinking, originality, and epistemic judgment—remains uncertain and empirically contested (Yavich, 2025; Smerdon, 2024). Evaluating AI homework tools, therefore, requires attention not only to performance gains but also to the nature and durability of learning they promote, distinguishing between cognitive offloading and genuine skill acquisition.

Empirical Evidence on AI, Engagement, and Cognition

Empirical evidence regarding AI's role in education underscores a complex relationship between engagement, performance, and cognitive processing. Recent studies suggest that AI tools can bolster student motivation by offering personalized, interactive learning experiences that cater to individual needs (Bognár & Khine, 2025; Tamimi et al., 2024). In terms of measurable outcomes, performance gains of approximately 15% to 35% have been reported, particularly when adaptive learning platforms and intelligent tutoring systems provide targeted scaffolding (Kwak, 2025; Ward et al., 2025). However, these gains are not universally distributed; evidence indicates that the impact on cognitive depth varies significantly based on demographic factors, prior achievement, and distinct patterns of use—ranging from active scaffolding to passive reliance (AlShibli et al., 2025).

Cognitive impacts of AI use are similarly mixed. Some studies report positive associations between AI-supported self-regulation and problem-solving (Zhou et al., 2024), while others caution that excessive reliance on AI may reduce students' willingness to engage in independent analysis and epistemic monitoring (Yavich, 2025). Engagement effects may also change over time, with initial enthusiasm diminishing as novelty effects fade (Bognár & Khine, 2025). Collectively, these findings underscore the importance of examining AI homework tools within broader theoretical frameworks of self-regulated learning and cognitive engagement (Zimmerman, 2002).

Methodology

Research Design

This study employed a systematic narrative synthesis design (Popay et al., 2006) to integrate empirical evidence comparing AI-assisted homework tools with traditional homework methods. A narrative approach was deemed most appropriate due to the extensive methodological and conceptual heterogeneity across primary studies, including variations in educational levels, disciplinary domains, and AI tool functionalities. Unlike meta-analytic techniques that necessitate strictly commensurable effect size reporting, narrative synthesis facilitates a theory-informed integration of diverse findings while preserving the contextual nuances essential for interpreting educational interventions (Popay et al., 2006). The review process followed the PRISMA 2020 guidelines to ensure transparency and replicability (Page et al., 2021), while the overall systematic framework was informed by principles of evidence-based education research (Gough et al., 2017).

Search Strategy

A systematic search was conducted using the Elicit research engine (https://elicit.com), selected for its advanced natural language processing (NLP) capabilities, which facilitate semantic searching. Unlike traditional keyword-matching databases, this approach identifies relevant literature based on conceptual meaning, ensuring comprehensive coverage of the rapidly evolving AI nomenclature (e.g., distinguishing between 'Generative AI', 'Large Language Models', and 'Intelligent Tutoring Systems'). The search leveraged the Semantic Scholar and OpenAlex databases, which collectively index over 200 million scholarly publications. Combinations of the following keywords were used: *artificial intelligence, generative AI, homework, academic performance, student achievement, writing, problem solving,* and *learning outcomes*. Filters were applied to peer-reviewed journal articles and refereed conference proceedings published between January 2013 and January 2025.

Study Selection Process

The study selection followed a rigorous two-stage screening process (See Figure 1 for the PRISMA flow diagram). In Stage 1 (Preliminary Screening), titles and abstracts were independently screened against the research objective. Studies were excluded if they were purely theoretical, focused exclusively on teacher-facing AI, or did not involve formal educational settings. In Stage 2 (Full-Text Eligibility Assessment), the remaining publications underwent a comprehensive full-text review. The primary reason for the subsequent exclusion of 40 studies was the absence of a rigorous comparative condition. This systematic refinement resulted in a final corpus of 10 empirical studies that met all criteria. To enhance reliability, the selection was finalized through an iterative review process to ensure perfect alignment with the research question.

Inclusion and Exclusion Criteria

Studies were included if they satisfied the following conditions: (a) AI-powered tools were used directly by students during homework completion; (b) participants were enrolled in formal secondary or tertiary education; (c) academic performance outcomes were reported using quantitative or mixed-methods measures; (d) the study employed an empirical research design (experimental, quasi-experimental, or observational); and (e) a clearly defined comparison condition using traditional, non-AI homework methods was present. Studies were excluded if they were limited to conceptual discussions, opinion pieces, or examined AI use solely during supervised in-class activities.

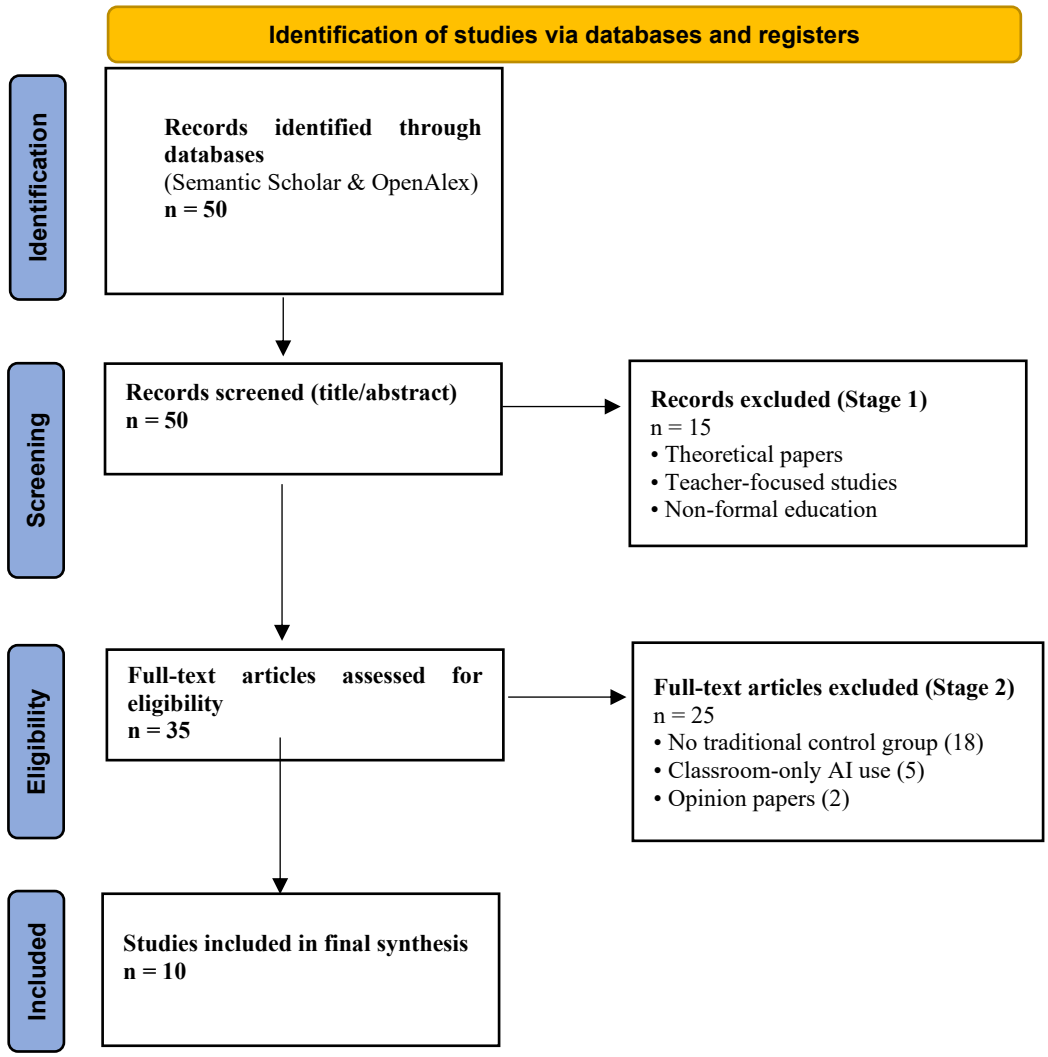


Figure 1. Centre the caption below the figure

Data Extraction and Quality Appraisal

A structured data extraction protocol was applied to record educational contexts, sample sizes, AI tool types, and performance metrics. Where effect sizes (Cohen's d) were not explicitly provided, the direction and magnitude of effects were inferred from reported statistical results. To ensure methodological rigor, a narrative critical appraisal framework was utilized (Petticrew & Roberts, 2006). Studies were assessed for internal validity, transparency of outcome measures, and the adequacy of their comparison conditions. Studies with robust experimental controls were given greater interpretive weight during the synthesis.

Data Synthesis

Findings were integrated using a thematic narrative synthesis approach (Popay et al., 2006). Results were organized according to: (1) disciplinary domain, (2) assessment timing (immediate performance vs. delayed retention), and (3) learner characteristics. This analytical strategy enabled the identification of conditional effects—circumstances under which AI-assisted homework proves beneficial, neutral, or detrimental—thereby avoiding overgeneralized conclusions regarding AI effectiveness.

Results

Overview of the Evidence Base

The synthesis drew on 10 empirical studies that explicitly compared AI-assisted homework practices with traditional approaches across secondary and higher education settings (See Table 1). Collectively, these studies provide a heterogeneous but informative evidence base regarding how AI tools influence academic performance under varying instructional conditions. However, a notable disciplinary divide was observed: while AI's impact is extensively documented in language-related tasks (e.g., EFL writing, translation), empirical evidence in STEM and pure science domains remains emerging and shows more nuanced results.

Table 1. Characteristics of included studies

Study	Study Design	Sample Size (AI / Traditional)	Duration	Educational Level	Subject Area	AI Tools Used	Traditional Method
Wale, 2024	Quasi-experimental (pretest–posttest)	92 total (groups not specified)	8 weeks	Undergraduate (third year)	EFL academic writing	Writerly, Google Docs	Paper–pencil feedback
Khaustov et al., 2024	Quasi-experimental	25/25	Not specified	Undergraduate (second year)	Translation studies	ChatGPT-4.0, Criterion, YandexGPT	Traditional instruction
Kelly et al., 2013	Quasi-experimental	63 total (8 classes /9 classes)	60-minute session	K–12 (grades 7–8)	Mathematics	ASSISTments (web-based)	Practice without feedback
Chen & Gong, 2025	Mixed methods, quasi-experimental	25/25	16 weeks	Undergraduate (third year)	Chinese as a second language	ChatGPT	Traditional teacher instruction
Smerdon, 2024	Observational	46/23 (survey respondents)	One semester	Undergraduate	Research proposal	Not specified	Not specified
Song & Song, 2023	Mixed methods, RCT	25/25	12 weeks	Undergraduate	EFL writing	ChatGPT	Traditional teacher instruction
Faridoon et al., 2025	Observational	Not specified	One semester	Higher education	Not specified	AI tutoring systems	Not specified
AlShibli et al., 2025	Quasi-experimental	32/32	4 weeks	Undergraduate (first year)	Computer science	Not specified	Textbooks, notes
Tamimi et al., 2024	Observational, mixed methods	115 total	One semester	High school	General homework	ChatGPT, StudyPool, TutorEva, OddityAI	Not specified
Boumediene & Bouakkaz, 2024	Observational survey	~1,200 students	Fall 2023 – Spring 2024	High school (grades 9–12)	Multiple subjects	ChatGPT	Not specified

Seven studies employed experimental or quasi-experimental designs with clearly defined control groups using traditional homework methods, while three relied on observational or correlational designs. Sample sizes ranged from small classroom-based interventions ($n \approx 50–80$) to large-scale institutional datasets exceeding 1,000 participants (Boumediene & Bouakkaz, 2024). The diversity of contexts allowed for cross-study comparisons of task types, disciplinary domains, and assessment practices.

Taxonomy of AI Tools Across Reviewed Studies

A critical prerequisite for interpreting the findings across the reviewed studies is recognizing that "AI homework tools" is not a monolithic category. The tools identified in the synthesis span two functionally distinct paradigms, each with fundamentally different mechanisms of action and pedagogical implications. The first category comprises *Intelligent Tutoring Systems (ITS)*, such as ASSISTments (Kelly et al., 2013), which operate within structured, domain-specific frameworks. These systems are designed primarily to *provide formative feedback*: they evaluate student responses against predefined correct answers, diagnose misconceptions, and adaptively guide learners through problem-solving sequences. Their primary function is to scaffold existing knowledge, not to generate new content on the student's behalf. The second category encompasses *Generative AI (GenAI) tools*, including large language models such as ChatGPT (Chen & Gong, 2025; Khaustov et al., 2024; Song & Song, 2023; Tamimi et al., 2024; Boumediene & Bouakkaz, 2024) and YandexGPT (Khaustov et al., 2024), as well as automated writing evaluation (AWE) platforms such as Writerly and Criterion (Wale, 2024; Khaustov et al., 2024).

It should be noted that Wale (2024) also references Google Docs as the writing environment in which the AWE tool was deployed; Google Docs itself is not an AI tool and is therefore excluded from this taxonomy. Unlike ITS, GenAI and AWE tools are designed for *direct content generation*: they can produce complete textual outputs, draft academic essays, translate texts, and provide holistic stylistic revisions in response to open-ended student prompts. This generative capacity fundamentally distinguishes them from feedback-oriented ITS. A third, less clearly defined category in the evidence base consists of *AI tutoring and homework assistance platforms* (e.g., TutorEva, OddityAI, StudyPool, and the unspecified AI tutoring systems referenced in Faridoon et al., 2025 and AlShibli et al., 2025), which blend elements of both paradigms by combining answer support with explanatory scaffolding.

Table 2. Quantitative performance results from studies with controlled comparisons

Study	Outcome Measure	AI Group Mean (*SD*)	Traditional Group Mean (*SD*)	Statistical Significance	Effect Size	Direction
Wale, 2024	IELTS writing post-test	54.68	45.85	$p < .05$	$d = 0.924$	AI better
Khaustov et al., 2024	Business letter writing	Not reported	Not reported	$p = .001$	Not reported	AI better
Khaustov et al., 2024	Contrastive/comparative essay	Not reported	Not reported	$p = .002$	Not reported	AI better
Khaustov et al., 2024	Argumentative essay	Not reported	Not reported	$p = .001$	Not reported	AI better
Khaustov et al., 2024	Film/book review	Not reported	Not reported	$p = .001$	Not reported	AI better
Kelly et al., 2013	Learning gains (pre–post)	Not reported	Not reported	$p = .1065$	$d = 0.56$ (WBH generally)	No significant difference
Chen & Gong, 2025	Academic writing post-test	89.74 (9.08)	82.15 (10.23)	$p < .05$	Not reported	AI better
Smerdon, 2024	Task performance	Not reported	Not reported	Not significant	< 1 mark / 100	No difference
Song & Song, 2023	Overall writing proficiency	59.12 (14.23)	45.18 (15.62)	$p < .001$	$d = 0.76$	AI better
Song & Song, 2023	Writing content	15.96 (3.71)	13.71 (3.12)	$p = .003$	$d = 0.65$	AI better
Song & Song, 2023	Writing organization	16.56 (3.54)	13.63 (3.63)	$p < .001$	$d = 0.84$	AI better
Song & Song, 2023	Language use	19.89 (4.82)	15.89 (4.12)	$p < .001$	$d = 0.88$	AI better
AlShibli et al., 2025	Overall average score	85.3%	79%	Not reported	Not reported	AI slightly better
AlShibli et al., 2025	Knowledge retention	20% demonstrated retention	Not reported	Not reported	Not reported	AI worse

This functional distinction is not merely taxonomic; it carries direct interpretive consequences for the reported outcomes. Studies employing ITS in mathematics contexts (Kelly et al., 2013) produced modest or non-significant effects on learning gains (d = 0.56, p = .1065), consistent with the ITS design goal of deepening procedural knowledge through targeted feedback. In contrast, studies utilizing GenAI tools in writing and language tasks reported substantially larger effects on output quality (e.g., d = 0.924 in Wale, 2024; d = 0.76–0.88 in Song & Song, 2023), which reflects the capacity of these tools to directly enhance the surface features of student-produced text. However, this performance advantage comes with a critical caveat: because GenAI tools can generate entire drafts, the observed gains may reflect the tool's output quality rather than the student's cognitive engagement. By contrast, ITS-driven gains, while more modest, are more directly attributable to student learning activity. Consequently, observed performance gains and their relationship to genuine cognitive engagement must be interpreted in light of the tool's fundamental purpose—whether it is designed to scaffold student thinking or to substitute for it.

Domain-Specific Outcomes: Language Proficiency vs. STEM Reasoning

A consistent pattern of high efficacy emerged in language-related domains (EFL, writing, translation). Studies reported statistically significant advantages for AI-assisted groups (Chen & Gong, 2025; Song & Song, 2023). As shown in Table 2, Wale (2024) reported a post-test IELTS writing score of 54.68 for the AI group compared to 45.85 for the traditional group (d = 0.924). Similarly, Song & Song (2023) documented large effect sizes across multiple dimensions, including writing organization (d = 0.84) and language use (d = 0.88). In contrast, findings from STEM and science-related domains were notably more mixed. Kwak (2025) and Kelly et al. (2013) found modest or non-significant differences in mathematics and problem-solving (p = .1065 in Kelly et al.). A striking disciplinary contrast is evident here: while AI excels in improving the "surface features" and organization of language tasks, its impact on the multi-step conceptual reasoning required in science and mathematics appears limited when the tool provides only answer verification rather than conceptual scaffolding.

The Performance-Learning Paradox: Immediate Gains vs. Retention

The temporal aspect of assessment was identified as a pivotal moderator of performance outcomes. Research focusing on immediate post-task results has consistently demonstrated a preference for AI-assisted homework (AlShibli et al., 2025; Song & Song, 2023). However, a "Performance-Learning Paradox" was identified: Proximal Success: AI users frequently attain higher grades in assignments (e.g., 85.3% vs. 79% in AlShibli et al., 2025).

Distal Failure: However, when delayed retention measures were employed, these advantages diminished. In the study conducted by Yang (2025), it was observed that students who placed significant reliance on artificial intelligence exhibited a substantially diminished capacity to retain conceptual knowledge. A seminal study by AlShibli et al. (2025) revealed a striking finding: while AI users exhibited superior report quality, only 20% demonstrated evidence of knowledge retention during independent assessments. This finding suggests that AI may encourage the development of "fragile knowledge" (Marton & Säljö, 1976), which is not conducive to long-term mastery.

Secondary Outcomes: The Trade-off Between Efficiency and Engagement

Beyond grades, AI-assisted students demonstrated higher time efficiency and motivation (Bognár & Khine, 2025; Ward et al., 2025). As detailed in Table 3, motivation gains were especially pronounced among students with lower prior achievement, who described AI tools as reducing frustration and cognitive overload (Tamimi et al., 2024).

However, the qualitative data revealed critical trade-offs. Smerdon (2024) and Yavich (2025) reported a decline in originality and independent reasoning. Instructors noted an increase in the uniformity of student work, with Boumediene & Bouakkaz (2024) observing that enhanced grammar and coherence were frequently accompanied by diminished critical thinking scores. A salient finding from Yavich's (2025) study was that less than 40% of students exhibited mastery of the content without AI assistance, thereby indicating a transition from cognitive engagement to "cognitive offloading."

Table 3. Secondary outcomes and quality indicators

Study	Completion Rates	Engagement / Motivation	Time Efficiency	Work Quality Dimensions
Wale, 2024	Not reported	Positive perceptions toward AI tools; increased motivation	Not reported	Improved task achievement, coherence, lexical resource, and grammar
Chen & Gong, 2025	Not reported	Writing motivation: AI M = 20.06 (SD = 3.33) vs. Traditional M = 18.21 (SD = 3.58), p = .001, d = 0.52	AI group completed reports faster	Enhanced ideas, coherence, lexicon, and grammatical range
Song & Song, 2023	Not reported	Higher participation, greater assignment consistency, stronger motivation	Not reported	Improved across all writing dimensions
Faridoon et al., 2025	Not reported	40% of teachers observed reduced effort	Not reported	Improved subject comprehension and problem-solving skills
AlShibli et al., 2025	Timely submissions increased from 75% to 85%	Not reported	Not reported	Better report quality but worse quiz performance
Boumediene & Bouakkaz, 2024	Not reported	Not reported	Not reported	Better grammar and flow, but lower originality and critical thinking

Note. Outcomes are reported as described in the original studies. Blank cells indicate that the indicator was not assessed or not reported.

Differential Effects by Learner Characteristics

The evidence presented indicates a correlation between the impact of AI and prior student achievement.

Strategic Use: High-achieving students employed AI tools to refine their ideas and check their understanding, achieving moderate gains without any loss of comprehension (Chen & Gong, 2025).

Dependency risk: Conversely, students with lower academic achievements exhibited a propensity to depend on direct AI-generated outputs, thereby augmenting their short-term scores while concomitantly escalating the likelihood of long-term cognitive dependency (AlShibli et al., 2025; Ward et al., 2025).

The findings, when considered collectively, reveal that AI-based homework tools do not consistently exert uniform effects on academic performance. Instead, the impact of AI in education is contingent on disciplinary context, assessment design, learner characteristics, and the degree to which AI use is pedagogically structured. These conditional patterns provide a necessary foundation for interpreting performance gains and inform the discussion of instructional implications.

Discussion

Reinterpreting Academic Performance: Productivity vs. Cognitive Change

The findings of this synthesis indicate that gains in academic performance associated with AI-assisted homework are fundamentally dependent on how performance is operationalized. Our results align with Gowtham et al. (2026) and Kwak (2025), suggesting that AI tools tend to optimize measurable outputs—such as assignment grades and completion rates—rather than the underlying cognitive processes. From a Behaviorist perspective, this optimization reflects Skinner's (1961) vision of immediate reinforcement, where students are rewarded for correct outputs. Quantitative indicators such as GPA improvements, increased completion rates, and higher standardized test scores have frequently been reported following the integration of AI tools, with gains ranging from approximately 15% to 35% in recent studies (Gowtham et al., 2026; Kwak, 2025). These results align with the present synthesis, which found significant performance advantages in language-focused assignments. However, as fundamentally argued by Soderstrom and Bjork (2015), such proximal indicators capture only a narrow slice of educational effectiveness. High performance on immediate tasks can often be a misleading proxy for actual learning, potentially obscuring significant declines in higher-order cognitive engagement and long-term retention. In the context of AI-assisted homework, this suggests that the 'polished output' may serve as a facade for what Lodge et al. (2023) describe as the erosion of assessment validity in the age of AI.

Surface Performance Gains Versus Deep Learning Outcomes

A central contribution of this study lies in clarifying the distinction between surface-level performance and durable learning outcomes. While AI tools demonstrably enhance linguistic accuracy and task efficiency, their impact on science process skills remains mixed. Our findings resonate with Zhai (2023), who argued that while generative AI can perform complex scientific tasks, it risks undermining the inquiry process if students use it as a surrogate for conceptual thinking rather than a supportive tool. This pattern is further supported by Kasneci et al. (2023), who emphasized in their comprehensive review that while large language models can enhance productivity, they pose significant risks to students' critical thinking and independent problem-solving abilities. In the context of STEM education, where multi-step reasoning and 'productive struggle' (Bjork & Bjork, 2011) are essential, this reliance creates a 'fragile knowledge' structure. Just as the Montessori method posits that intellectual independence is built through the 'mental hands-on' manipulation of concepts, AI-assisted homework risks bypassing this necessary cognitive labor, effectively replacing a scaffold (Wood et al., 1976) with a permanent cognitive prosthetic.

Engagement, Motivation, and the Illusion of Effectiveness

The reported increases in student engagement and the perceived usefulness of AI-assisted homework (Tamimi et al., 2024; Ward et al., 2025) present a psychological paradox. While AI reduces frustration by lowering the entry barrier to complex tasks, this heightened satisfaction often stems from the reduction of cognitive effort rather than a sense of mastery. As fundamentally argued by Soderstrom and Bjork (2015), high performance and positive affect during a task can lead to an "illusion of competence," where students mistake the ease of AI-facilitated task completion for genuine understanding. This "illusion of effectiveness" masks shallow processing with polished outputs, creating a facade of academic achievement that lacks conceptual depth. Furthermore, it is critical to recognize that motivation is not a constant variable. While students initially show high participation rates, this is often driven by the "novelty effect" of generative AI, which tends to diminish over time as the tool becomes a mundane utility (Bond et al., 2024). In contrast, creativity, critical thinking, and science process skills are the durable cornerstones for future problem-solving. If AI tools are used to automate the "messy" and non-linear inquiry process, the student's epistemic agency—their authority and responsibility in knowledge construction—is severely compromised.

For STEM education, which emerged as a response to the need for interdisciplinary synthesis, the risk of "cognitive offloading" is particularly high. To prevent this, AI use must be reframed through the lens of Distributed Cognition (Salomon, 1993). In this model, the AI is not a surrogate that provides answers, but a partner in a "thinking system" where the human remains the primary agent. Just as the Montessori approach advocates for the development of foundational skills through active manipulation, modern homework must require students to critique, justify, and extend AI outputs. This ensures that the technology supports the expansion of human intelligence rather than its contraction, preserving the creative struggle necessary for genuine scientific discovery.

Conditional Effects on Critical Thinking and Problem Solving

The impact of AI-assisted homework on higher-order skills, such as critical thinking and scientific reasoning, emerged as deeply conditional. Our findings suggest that AI tools do not inherently enhance or undermine cognitive development; rather, their effects are pedagogically mediated by instructional design and assessment alignment. Studies explicitly targeting problem-solving reported positive outcomes only when AI use was constrained by reflective activities and metacognitive prompts (Molenaar, 2022; Zhou et al., 2024). In contrast, overreliance on generative AI has been linked to reduced independent problem-solving and less diverse student responses, reflecting concerns that AI may supplant deeper cognitive engagement rather than support it (Qian, 2025). This underscores a critical "mediation effect": AI supports learning when it functions as a Socratic interlocutor that prompts the student to explain their reasoning, but it hinders learning when it serves as a mere content generator that provides final answers. This distinction is particularly vital for STEM education and the development of SPS. STEM inquiry is fundamentally interdisciplinary and non-linear, requiring students to engage in "messy" problem-solving that cannot be reduced to an algorithmic output. When assessments privilege surface features or final numerical answers, AI tools inadvertently discourage deeper engagement, leading to what Biggs (2003) termed "surface approaches to learning." However, when homework is designed to reward process explanation, justification, and conceptual transfer, AI can become a powerful partner in distributed cognition (Salomon, 1993). In this context, the student retains their epistemic agency by using AI to test hypotheses or

critique scientific models, rather than allowing the AI to bypass the "mental hands-on" struggle deemed essential by the Montessori philosophy.

Ultimately, the results challenge the current state of homework design. The perceived academic performance gains identified in this synthesis often reflect the AI's efficiency rather than the student's growth. To move beyond this "illusion of effectiveness," educators must align AI integration with assessment practices that value the learning process over the product. As argued by Lodge et al. (2023), the era of AI necessitates an assessment reform where students are evaluated on their ability to monitor, critique, and authorize knowledge. By forcing students to engage in higher-order epistemic monitoring, we ensure that creativity and critical thinking remain the cornerstones of scientific inquiry, preparing learners for the complex, interdisciplinary challenges of the future.

Implications for Interpreting Academic Performance Gains

Taken together, the results challenge simplistic interpretations of improved academic performance in AI-enhanced homework contexts. Performance gains reflected in grades, GPA, and completion metrics should be interpreted as indicators of enhanced productivity and output quality rather than unequivocal evidence of learning. This distinction is critical for educators and policymakers who may be tempted to equate measurable gains with educational success. The conditional patterns identified in this synthesis underscore the need to align AI integration with assessment practices that value learning processes, metacognitive regulation, and higher-order thinking. Without such alignment, AI homework tools risk reinforcing surface learning while inflating conventional performance metrics.

Implications

For educators, the findings underscore the importance of integrating AI homework tools within pedagogical frameworks that emphasize explanation, reflection, and revision. AI should be positioned as a support for learning rather than a substitute for effort. For researchers, future studies should prioritize delayed learning measures, examine interactions between AI use and metacognitive skills, and report effect sizes transparently. Policymakers should support human-centered AI integration, educator training, and long-term evaluation rather than relying on short-term performance metrics.

Limitations

This synthesis is limited by heterogeneity across included studies, incomplete reporting of effect sizes, and potential publication bias. Rapid technological change also limits the generalizability of findings to future AI systems with different capabilities.

A particularly significant limitation of this review is the pronounced imbalance in the disciplinary distribution of the evidence base. The synthesis provides robust empirical grounding for language education contexts—specifically EFL writing, academic writing in a second language, and translation tasks—where seven of the ten included studies were concentrated. In contrast, empirical research examining AI-assisted homework in STEM and pure science domains remains critically scarce. Only two studies (Kelly et al., 2013; AlShibli et al., 2025) explicitly addressed mathematics or computer science contexts, and neither reported sufficient effect size data to support strong domain-specific conclusions. No included studies examined AI homework tools in physics, chemistry, biology, or integrated STEM curricula. This disciplinary scarcity constitutes a formal limitation of the present study and must be clearly acknowledged when interpreting the generalizability of the findings.

The consequences of this "disciplinary divide" for the generalizability of the findings are substantial. The performance advantages associated with AI-assisted homework—particularly the large effect sizes documented in writing proficiency and language use—are closely tied to the surface-level features that generative AI tools are most adept at improving (i.e., grammatical accuracy, coherence, and organizational structure). These features are central to assessment in language education but are considerably less central to evaluation in STEM disciplines, where higher-order outcomes such as multi-step reasoning, mathematical proof, experimental design, and conceptual transfer are prioritized. Consequently, the predominantly positive findings of this synthesis should not be extrapolated to STEM educational contexts without direct empirical support. Future systematic reviews and primary studies should specifically target STEM domains to establish whether AI homework tools offer comparable, diminished, or qualitatively different benefits in contexts that privilege deep conceptual engagement

over linguistic output quality. Until such evidence is available, claims regarding the effectiveness of AI homework tools must be understood as domain-conditional, grounded predominantly in language and writing education rather than representing a universal finding across disciplines.

Conclusion

This study set out to examine the effects of AI-assisted homework on students' academic performance by synthesizing empirical evidence comparing AI-supported and traditional homework practices. Taken together, the findings indicate that AI tools are consistently associated with improvements in measurable academic outputs—such as assignment grades, task completion rates, and short-term assessment scores—but that these gains should be interpreted with caution when used as proxies for learning. A central conclusion emerging from this synthesis is that improvements attributed to AI-assisted homework are highly contingent on how academic performance is operationalized. Metrics commonly used in the reviewed studies tend to privilege efficiency, surface-level correctness, and linguistic or structural quality, areas in which generative AI systems are particularly effective. As a result, observed performance advantages often reflect enhanced productivity and output optimization rather than unequivocal gains in conceptual understanding, transfer, or long-term retention.

The analysis further demonstrates that the educational value of AI-assisted homework is not inherent to the technology itself but is mediated by pedagogical design, assessment alignment, and students' self-regulatory capacities. When AI use is embedded within instructional frameworks that emphasize reflection, justification, and process-oriented assessment, AI tools can function as cognitive scaffolds that support higher-order thinking. In contrast, unstructured or unrestricted AI use risks fostering cognitive offloading, homogenization of student work, and inflated performance indicators disconnected from deep learning. These findings carry important implications for both research and practice. For researchers, the results underscore the need to move beyond binary comparisons of AI versus non-AI conditions and toward more nuanced investigations that examine conditional effects, mediating variables, and the alignment between learning objectives and assessment practices. Future studies would benefit from incorporating longitudinal designs, process-based measures of learning, and explicit operationalizations of metacognitive engagement.

For educators and policymakers, the results caution against equating improvements in grades or completion rates with educational success. While AI-assisted homework can enhance accessibility, efficiency, and student engagement, its integration should be guided by clear pedagogical intentions and supported by assessment practices that value reasoning, originality, and epistemic agency. Without such safeguards, AI risks reinforcing surface learning while obscuring meaningful differences in students' understanding. In conclusion, AI-assisted homework represents neither a panacea nor a threat to learning outcomes. Its impact depends fundamentally on how it is designed, regulated, and evaluated within educational systems. By reframing academic performance as a multidimensional construct that extends beyond easily measurable outputs, this study contributes to a more balanced and theoretically grounded understanding of AI's role in contemporary education.

Scientific Ethics Declaration

* The authors declare that the scientific, ethical, and legal responsibility of this article published in the JESEH journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest

Funding

* There is no funding

References

AlShibli, A. S., Al Shibli, M., & Al Harthi, A. (2025). Exploring the impact of artificial intelligence on student academic performance. *Artificial Intelligence & Robotics Development Journal, 7*(1), 45–62. https://doi.org/10.52098/airdj.20233343

Biggs, J. (2003). *Teaching for quality learning at university* (2nd ed.). Society for Research into Higher Education & Open University Press.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *FABBS Foundation, Psychology and the real world: essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bognár, L., & Khine, M. S. (2025). The shifting landscape of student engagement: A pre–post semester analysis in AI-enhanced classrooms. *Computers and Education: Artificial Intelligence, 8*, 100395. https://doi.org/10.1016/j.caeai.2025.100395

Bond, M., et al. (2024). A systematic review of generative AI in higher education: The gap between opportunities and practice. *Applied Learning Review, 1*(1), 1-25. https://doi.org/10.1186/s41239-023-00436-z

Boumediene, H., & Bouakkaz, M. (2024). Changes in homework submission patterns with the advent of AI tools: A high school perspective. *Studies in Education Sciences, 5*(4), 112–129. https://doi.org/10.54019/sesv5n4-001

Chen, C., & Gong, Y. (2025). The role of AI-assisted learning in academic writing: A mixed-methods study on Chinese as a second language students. *Education Sciences, 15*(2), 141. https://doi.org/10.3390/educsci15020141

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of Educational Research*, *76*(1), 1-62.

Dede, C. (2014). The role of digital technologies in deeper learning. Students at the Center: Deeper Learning Research Series. *Jobs for the Future*. 1-36.

Dettmers, S., Trautwein, U., & Lüdtke, O. (2009). The relationship between homework time and achievement is not universal: Evidence from multilevel analyses in 40 countries. *School Effectiveness and School Improvement*, *20*(4), 375-405. https://doi.org/10.1080/09243450902904601

Faridoon, N., Talpur, Q., Latif, F., Naz, G., & Shahzad, T. (2025). The role of AI tutors in improving academic performance and student engagement. *Academia International Journal for Social Sciences, 4*(3), 5897–5910. https://doi.org/10.63056/ACAD.004.03.0837

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews* (2nd ed.). SAGE.

Gowtham, R., Iyer, S., & Krishnan, P. (2026). Measuring what matters: Academic performance, productivity, and learning in AI-enhanced education. *Computers & Education, 201*, 104789.

Hattie, J. (2009). The Black Box of Tertiary Assessment: An Impending Revolution. In L. H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P.M. Johnston, & M. Rees (Eds.), *Tertiary Assessment & Higher Education Student Outcomes: Policy, Practice & Research* (pp.259-275). Wellington, New Zealand: Ako Aotearoa

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379–424. https://doi.org/10.1080/07370000802212669

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kelly, K., Heffernan, N., Heffernan, C., Goldman, S., Pellegrino, J., & Soffer Goldstein, D. (2013, July). Estimating the effect of web-based homework. *In International Conference on Artificial Intelligence in Education* (pp. 824-827). Berlin, Heidelberg: Springer Berlin Heidelberg. Aied2013ws_volume8.pdf

Khaustov, O. N., Tormyshova, T. Y., & Sukhanova, N. I. (2024). Teaching students to write academic papers through the use of generative artificial intelligence tools. Tambov University Review. *Series: Humanities, 29*(5), 1194-1207. https://doi.org/10.20310/1810-0201-2024-29-5-1194-1207

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A Meta-analytic review. *Review of Educational Research, 86*(1), 42-78. https://doi.org/10.3102/0034654315581420

Kwak, M. (2025). The effectiveness of AI-driven tools in improving student learning outcomes compared to traditional methods. *Issues in Information Systems, 26*(1), 1–12. https://doi.org/10.48009/4_iis_2025_120

Lodge, J., Howard, S., Bearman, M., & Dawson, P. (2023). *Assessment reform for the age of Artificial Intelligence.* Tertiary Education Quality and Standards Agency.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed. An argument for AI in education.* London: Pearson.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. *British Journal of Educational Psychology, 46*(1), 4–11. https://doi.org/10.1111/j.2044-8279.1976.tb02980.x

Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *European Journal of Education*, *57*(4), 632-645. https://doi.org/10.1111/ejed.12527

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*(n71). https://doi.org/10.1136/bmj.n71

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell Publishing. https://doi.org/10.1002/9780470754887

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). *Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC methods programme*. University of Lancaster. https://doi.org/10.13140/2.1.1018.4643

Qian, Y. (2025). *Pedagogical applications of generative AI in higher education: A systematic review of the field*. TechTrends, 69, 1105–1120. https://doi.org/10.1007/s11528-025-01100-1

Salomon, G. (1993). *Distributed cognitions: Psychological and educational considerations*. Cambridge University Press.

Skinner, B. F. (1961). The Science of Learning and the Art of Teaching. In B. F. Skinner, *Cumulative record* (Enlarged ed., pp. 145–157). Appleton-Century-Crofts. https://doi.org/10.1037/11324-010

Smerdon, D. (2024). AI in essay-based assessment: Student adoption, usage, and performance. *Computers and Education: Artificial Intelligence, 5*, 100288. https://doi.org/10.1016/j.caeai.2024.100288

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer. https://doi.org/10.1007/978-1-4419-8126-4

Tamimi, J., Addichane, E., & Alaoui, S. M. (2024). Evaluating the effects of artificial intelligence homework assistance tools on high school students' academic performance and personal development. *Arab World English Journal, 15*(3), 45–67.

Trautwein, U., & Köller, O. (2003). The relationship between homework and achievement—still much of a mystery. *Educational Psychology Review*, *15*(2), 115-145.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press.

Wale, B. D. (2024). Artificial intelligence in education: Effects of using integrative automated writing evaluation programs on honing academic writing instruction. *Cakrawala Pendidikan, 43*(1), 273-287.

Ward, B., Bhati, D., Neha, F., & Guercio, A. (2025). Analyzing the impact of AI tools on student study habits and academic performance. In *Proceedings of the IEEE 15th Annual Computing and Communication Workshop and Conference* (pp. 1–8).

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Yang, H. (2025). Harnessing generative AI: Exploring its impact on cognitive engagement, emotional engagement, learning retention, reward sensitivity, and motivation through reinforcement theory. *Learning and Motivation*, *90*, 102136. https://doi.org/10.1016/j.lmot.2025.102136

Yavich, R. (2025). Will the Use of AI undermine students independent thinking? *Education Sciences*, *15*(6), 669. https://doi.org/10.3390/educsci15060669

Zhou, X., Teng, D., & Al-Samarraie, H. (2024). The mediating role of generative AI self-regulation on students' critical thinking and problem-solving. *Education Sciences, 14*(9), 987.

Zhai, X. (2023). ChatGPT for next generation science learning. *XRDS: Crossroads, The ACM Magazine for Students*, *29*(3), 42-46. https://doi.org/10.1145/3589649

Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Author(s) Information

Seyma Irmak
Amasya University,
Faculty of Education, Department of Mathematics and Science Education, Amasya/Türkiye
Contact e-mail: *seyma.bardak@gmail.com*
ORCID iD: https://orcid.org/0000-0003-3831-8244

Kaan Bati
Hacettepe University,
Faculty of Education, Department of Mathematics and Science Education, Ankara/Türkiye
ORCID iD: https://orcid.org/0000-0002-6169-7871

NOTE: Appendix added.

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	5-6-7
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	7-8
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	9-10
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	11-12-13
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	9-10-11
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	11-12-13
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	10-11-12-13
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	10-11-12
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	10-11-12-13
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	10-11-12
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	13
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	10-11-12-13
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	12-13
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	10-11-12
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to	14-15-16
		identify the presence and extent of statistical heterogeneity, and software package(s) used.	
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	16-24
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	16, 17, 18, 19, 20, 21, 22, 23 and 24
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	10-13
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	10-13
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	13-14
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	14
Study characteristics	17	Cite each included study and present its characteristics.	15
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	14-15
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	15-26
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	15-26
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	15-26
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	15-26
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	15-26
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	15-26
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	15-26
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	26-27
	23b	Discuss any limitations of the evidence included in the review.	32
	23c	Discuss any limitations of the review processes used.	32
	23d	Discuss implications of the results for practice, policy, and future research.	27-31
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	32
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	32
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	32
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	32
Competing interests	26	Declare any competing interests of review authors.	32
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	32

JESEH
Journal of Education in Science, Environment and Health
Volume 12, Issue 2, 2026
https://doi.org/10.55549/jeseh.905

Artificial Intelligence in Physics Education (2015–2025): Systematic Review of Trends, Applications, and Challenges

Mohammad Naser Azizi, Arafuddin Faizi, Ugur Sari

Article Info

Article History

Published:
01 April 2026

Received:
04 February 2026

Accepted:
05 March 2026

Keywords

Artificial intelligence
Physics education
Systematic literature review

Abstract

The aim of this study is to examine the role of artificial intelligence (AI) in physics teaching and learning between 2015 and 2025 through a systematic literature review. The research process was conducted in accordance with PRISMA 2020 guidelines; a total of 11,208 records were identified through searches in the Web of Science, Scopus, and ERIC databases. After removing duplicate records and applying the inclusion criteria, 40 studies were included in the final analysis. The findings indicate a marked increase in AI-supported physics education research, particularly after 2023. Most of the studies employed a mixed-methods design, with undergraduate students predominantly selected as the sample group. The results of the content analysis reveal that AI applications are most frequently concentrated in mechanics topics, followed by electromagnetism and thermodynamics. A significant proportion of the research focuses on examining the performance of generative AI systems in problem-solving, automated assessment, and personalized feedback processes. However, teacher-focused studies and long-term analyses of pedagogical impact appear to be limited. In conclusion, the field of AI-supported physics education is undergoing rapid development; however, more comprehensive research is needed in terms of methodological diversity, sample balance, and pedagogical depth.

Introduction

Over the past decade, rapid advancements in AI technologies have led to structural transformations in educational systems (Samala et al., 2025). The ability of computer systems to emulate human intelligence by demonstrating learning, reasoning, and decision-making capabilities offers significant opportunities for the personalization of instructional processes, the enhancement of learning efficiency, and the expansion of educational access (Al-Kamzari & Alias, 2025; Fayzullina et al., 2025). This transformation has become particularly evident in science education, where instructional processes have begun to be restructured through data-driven decision-making, adaptive learning environments, and automated assessment systems (He & Krajcik, 2026; Lee et al., 2025).

Due to its inherently abstract concepts, frequent reliance on mathematical modeling, and susceptibility to conceptual misconceptions, physics education has emerged as a distinctive field of inquiry for AI applications (Mahligawati et al., 2023). AI-supported adaptive learning systems, virtual laboratories, and intelligent tutoring tools make physics concepts more accessible and meaningful by delivering content tailored to students' individual learning pace and prior knowledge (Chen et al., 2020; Al-Kamzari & Alias, 2025). While natural language processing–based systems support the analysis of open-ended responses, learning analytics applications contribute to monitoring student performance and developing early intervention strategies. In this respect, AI functions not only as a pedagogical tool in physics instruction but also as an analytical framework that transforms research and assessment processes (Bralin et al., 2024; Heeg & Avraamidou, 2023; Kotsis, 2025).

Since 2015, AI-based approaches have shown a marked increase in physics education research. The systematic review conducted by Bralin et al. (2024) reveals a substantial quantitative growth in AI-focused studies in recent years, particularly in data-driven domains such as the development of assessment instruments, prediction of student achievement, and analysis of learner engagement. This trend indicates that analytical and computational methods are increasingly occupying a central position in physics education research. Similarly, Mahligawati et al. (2023) demonstrated that AI is utilized in physics instruction for concept introduction, personalized learning, social interaction, and assessment processes. Studies focusing on the secondary education level suggest that AI-supported systems have the potential to enhance conceptual understanding and motivation, particularly in abstract topics such as Newtonian mechanics, kinematics, and electromagnetism (Al-Kamzari & Alias, 2025). Broader reviews conducted within the context of science and STEM education further highlight the diversity of AI applications. Kotsis (2025) emphasizes that AI supports personalization and inquiry-based learning in STEM

education through intelligent tutoring systems, adaptive platforms, and automated assessment tools, while also raising significant ethical and political concerns such as data privacy, algorithmic bias, and equity. Almasri (2024) demonstrates that between 2014 and 2023, AI assumed functions in science education such as improving learning environments, generating examinations, evaluating student work, and predicting performance. Heeg and Avraamidou (2023) categorized AI applications in science education into nine groups and reported positive effects on learning achievement and argumentation skills. Despite this extensive body of literature, systematic and holistic syntheses specific to physics education remain limited. Most existing reviews provide a general framework within science education or STEM contexts and do not comprehensively map the pedagogical, methodological, and technological trends unique to physics education. Moreover, a substantial portion of existing studies focus on specific contexts, short time frames, or single application types, while large-scale systematic reviews conducted in accordance with standardized guidelines such as PRISMA remain scarce. This situation highlights the need for a comprehensive evaluation of AI applications in physics education in terms of their pedagogical purposes, technological approaches, methodological trends, and impacts on learning outcomes.

Accordingly, systematically examining the pedagogical contributions, methodological orientations, and limitations of AI applications in physics education is important from both theoretical and practical perspectives. In this context, the aim of the present study is to investigate how artificial intelligence has been positioned in physics teaching and learning between 2015 and 2025 through a systematic literature review, and to identify current research trends in the field, the types of AI employed, their strengths and limitations, and future directions. Additionally, the temporal and geographical distribution of the studies, their ethical and pedagogical constraints, and their recommendations for future research are analyzed. By outlining the development of AI in physics education, this review seeks to provide a conceptual and practice-oriented framework for educators, researchers, and policymakers. In line with the purpose of the study, the following research questions are addressed:

RQ1. What are the bibliometric trends in AI-supported physics instruction research?
RQ2. Which research approaches are preferred in AI-supported physics instruction studies?
RQ3. Which participant groups are predominantly examined in the reviewed studies?
RQ4. What types of AI are employed in these studies?
RQ5. For what purposes is AI used in physics instruction?
RQ6. What contributions do AI-supported applications provide to physics instruction?
RQ7. What are the main challenges encountered in AI applications?
RQ8. Based on the reviewed studies, what research gaps and future directions stand out in the field of AI-supported physics instruction?

Method

Research Design

This study was designed as a systematic literature review aimed at comprehensively and systematically examining published scientific studies on AI applications in physics education. The research process was conducted in accordance with the PRISMA 2020 guidelines (Page et al., 2021). A descriptive analysis and thematic synthesis approach was adopted, and both qualitative and quantitative studies were evaluated together. The study began with a systematic search conducted in the Web of Science, Scopus, and ERIC databases using specific keywords related to artificial intelligence and physics education. Inclusion criteria such as publication date range, language, and accessibility were applied. The studies were initially screened based on titles, abstracts, and full texts for eligibility. Subsequently, a detailed data analysis was carried out, incorporating dimensions such as the type of AI application, research design, sample characteristics, main findings, and recommendations. In structuring the review process in line with PRISMA, methodological rigor and transparency were ensured as recommended by Page et al. (2021). In addition, a model including a structured search strategy, clearly defined inclusion and exclusion criteria, and systematic data analysis was employed. This protocol provides clear, step-by-step guidance for searching, screening, and analyzing relevant literature, thereby enabling the replicability of the review process in future studies.

Data Sources and Search Strategy

The literature search was conducted in the Web of Science, Scopus, and ERIC databases. The search was limited to studies published between 2015 and 2025. Boolean operators (AND, OR) were used during the search process. The search strings used for each database are presented in Table 1. In Web of Science and Scopus, searches were

performed across all available fields (including title, abstract, and keywords), whereas in ERIC, the search was limited to the title and abstract fields.

Table 1. Search string

Databases	Keywords
Web of science	TS=("physics education" OR "physics teaching" OR "physics learning") AND TS=("artificial intelligence" OR "AI in education" OR "intelligent tutoring system" OR "adaptive learning system" OR "machine learning")
Scopus	("physics education" OR "physics teaching" OR "physics learning") AND ("artificial intelligence" OR "AI in education" OR "intelligent tutoring system" OR "adaptive learning system" OR "machine learning)
ERIC	("physics education" OR "physics teaching" OR "physics learning") AND ("artificial intelligence" OR "AI in education" OR "intelligent tutoring system" OR "adaptive learning system" OR "machine learning)

Study Selection Process (PRISMA)

In accordance with the PRISMA flow diagram, the process was conducted in four stages: identification, screening, eligibility, and inclusion. The total number of records retrieved from the databases, the removal of duplicate studies, title–abstract screening, full-text assessment, and the final number of included studies are presented in detail in the PRISMA diagram (Fig. 1). The reasons for excluding studies at the full-text stage (e.g., irrelevance to the topic, methodological inadequacy, access issues) were documented.

Identification

The literature search was conducted in the Web of Science, Scopus, and ERIC databases due to their broad scope, academic rigor, and reliable indexing of peer-reviewed publications. These databases were preferred because they include high-impact, internationally indexed, peer-reviewed journals in the fields of education and educational technology. In particular, Web of Science and Scopus provide access to current and methodologically robust research through their multidisciplinary structures and extensive citation networks. ERIC, with its specialized focus on education, supported the contextual integrity of the study. The selection of these databases aimed to systematically and comprehensively identify studies on AI applications within the context of physics education. The primary rationale for this choice was to ensure access to the international literature as comprehensively as possible, rather than limiting the review to a specific group of publications. Thus, the scope validity of the study was strengthened, and potential publication bias was minimized.

The search strategy was structured in line with the research problem and conceptual framework. The key concepts were determined along two main axes: (1) the context of physics education and (2) AI applications. These two axes were combined using Boolean operators (AND, OR). Synonyms and alternative expressions commonly used in the literature were considered to broaden the search scope. Trends in the literature, variations in conceptual usage, and current terminology in educational technology were taken into account when determining the search terms. Consequently, a comprehensive search strategy encompassing both classical AI applications and next-generation systems was developed. The applied search strings and keywords are presented in detail in Table 1 to ensure transparency and replicability. A total of 11,208 records were retrieved from the databases. After removing 158 duplicate records, 11,050 records were transferred to the screening stage.

Screening

Following the removal of duplicate records, the remaining studies were systematically evaluated in line with the purpose of the research. At this stage, 11,050 records were examined at the title and abstract level. The studies were assessed according to predetermined inclusion and exclusion criteria. As a result of the title and abstract screening, 10,117 records were excluded. Studies that were not directly related to the context of physics education, did not involve AI applications, did not meet the language criterion, or were not research articles (e.g., reviews, editorials) were excluded at this stage. After screening, 933 studies were deemed eligible for full-text evaluation and moved to the eligibility stage.

Eligibility

At this stage, the studies were evaluated in detail at the full-text level according to the inclusion and exclusion criteria presented in Table 2. The 933 studies that passed the title and abstract screening were analyzed in terms of their direct relevance to the research questions, methodological adequacy, and the extent to which they addressed AI applications within the context of physics education. Following the full-text assessment, 892 studies were excluded for various reasons. A significant portion of these studies fell outside the context of physics education or did not directly address AI applications. Additionally, studies were excluded if they were published outside the 2015–2025 time range, not published in peer-reviewed international journals, not indexed in Web of Science, Scopus, or ERIC, published in languages other than English, or did not present clearly defined methodologies (e.g., reviews, editorials, or studies lacking methodological clarity). Studies lacking accessible full texts were also excluded. As a result of this process, 40 studies were deemed suitable for final analysis and included in the systematic review (see Figure 1).

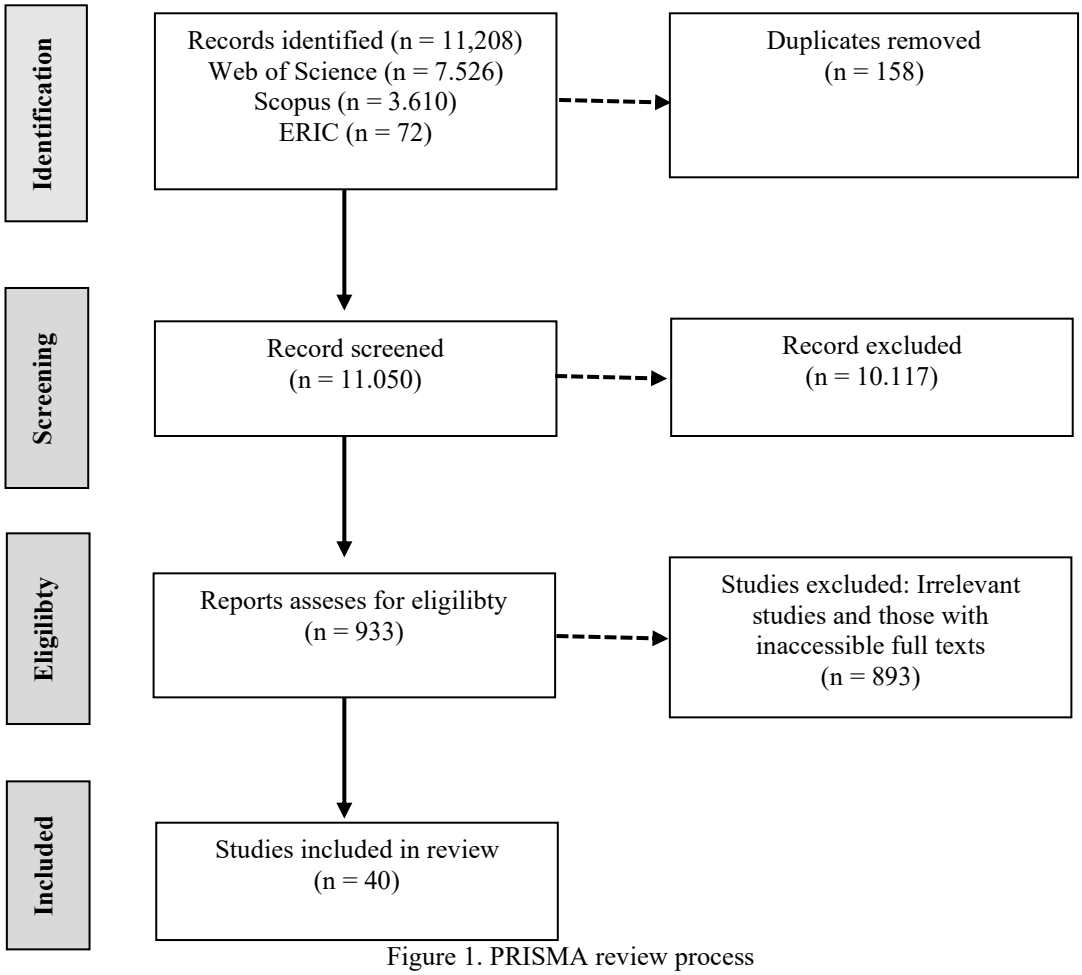


Figure 1. PRISMA review process

Table 2. Inclusion and exclusion criteria

Inclusion criteria	**Exclusion criteria**
Studies addressing AI applications in physics education	Studies outside the context of physics education
Studies published between 2015–2025	Studies published outside 2015–2025
Articles published in peer-reviewed international journals	Non-scientific publications
Studies indexed in Web of Science, Scopus, or ERIC	Studies outside the specified databases
Studies with accessible full texts	Studies without accessible full texts
Studies published in English	Publications in languages other than English
Research articles with clearly defined methodologies (qualitative, quantitative, mixed, theoretical)	Studies lacking methodological clarity; reviews; editorials

Table 3. Publication details, authors and citations

No	Year of publication	authors names	The APA citation	Country
1	2019	Santos, O. C., & Corbí, A.	(Santos & Corbí, 2019)	Spain
2	2023	Ding, L., Li, T., Jiang, S., & Gapud, A.	(Ding et al., 2023)	USA
3	2023	Bessas, N., Tzanaki, E., Vavougios, D., & Plagianakos, V. P.	(Bessas et al., 2023)	Greece
4	2023	Liang, Y., Zou, D., Xie, H., & Wang, F. L.	(Liang et al., 2023)	China
5	2023	Bitzenbauer, P.	(Bitzenbauer, 2023)	Germany
6	2023	Tong, D., Tao, Y., Zhang, K., Dong, X., Hu, Y., Pan, S., & Liu, Q.	(Tong et al., 2023)	USA & China
7	2023	Kieser, F., Wulff, P., Kuhn, J., & Küchemann, S.	(Kieser et al., 2023)	Germany
8	2023	Tschisgale, P., Wulff, P., & Kubsch, M.	(Tschisgale et al., 2023)	Germany
9	2024	Al-Kamzari, F., & Alias, N.	(Al-Kamzari & Alias, 2024)	Oman
10	2024	Jang, H., & Choi, H.	(Jang & Choi, 2024)	Korea
11	2024	Aldazharova, S., Issayeva, G., Maxutov, S., & Balta, N.	(Aldazharova et al., 2024)	Kazakhstan
12	2024	Singh, C.	(Singh, 2024)	USA
13	2024	Holme, T. A.	(Holme, 2024)	USA
14	2024	Domenichini, D., Bucchiarone, A., Chiarello, F., Schiavo, G., & Fantoni, G.	(Domenichini et al., 2024)	Italy
15	2024	López-Simó, V., & Rezende, M. F., Jr.	(López-Simó & Rezende, Jr., 2024)	Brazil
16	2024	Beltozar-Clemente, S., & Díaz-Vega, E.	(Beltozar-Clemente & Díaz-Vega, 2024)	Peru
17	2024	Monteiro, F. F., Souza, P. V. S., da Silva, M. C., Maia, J. R., da Silva, W. F., & Girard, D.	(Ferreira Monteiro et al., 2024)	Brazil
18	2024	Sirnoorkar, A., Zollman, D., Laverty, J. T., Magana, A. J., Rebello, N. S., & Bryan, L. A.	(Sirnoorkar et al., 2024)	USA
19	2024	Cho, N.	(Cho, 2024)	China
20	2024	Uğraş, H., Uğraş, M., Papadakis, S., & Kalogiannakis, M.	(Uğraş et al., 2024)	Greece
21	2025	Abdulayeva, A., Zhanatbekova, N., Andasbayev, Y., & Boribekova, F.	(Abdulayeva et al., 2025)	Kazakhstan
22	2025	Revalde, G., Zholdakhmet, M., Abola, A., & Murzagaliyeva, A.	(Revalde et al., 2025)	Latvia
23	2025	Kemouss, H., & Khaldi, M.	(Kemouss & Khaldi, 2025)	Morocco
24	2025	Guerrero-Zambrano, M., Sanchez-Alvarado, L., Valarezo-Chamba, B., & Lamilla-Rubio, E.	(Guerrero-Zambrano et al., 2025)	Ecuador
25	2025	Villegas Ch., W., Buenano Fernandez, D., Maldonado Navarro, A., & Mera Navarrete, A.	(Villegas Ch. et al., 2025)	Ecuador
26	2025	Coban, A., Dzsotjan, D., Küchemann, S., Durst, J., Kuhn, J., & Hoye, C.	(Coban et al., 2025)	Germany
27	2025	Wei, Y., Zhang, R., Zhang, J., Qi, D., & Cui, W.	(Wei et al., 2025)	China
28	2025	Bessas, N., Tzanaki, E., Vavougios, D., & Plagianakos, V. P.	(Bessas et al., 2025)	Greece
29	2025	Abdulayeva, A., Zhanatbekova, N., Andasbayev, Y., Khaimuldanov, Y., & Zhiyembayev, Z.	(Abdulayeva et al., 2025)	Kazakhstan
30	2025	Ben-Zion, Y., Einhorn Zarzecki, R., Glazer, J., & Finkelstein, N. D.	(Ben-Zion et al., 2025)	Israel & USA
31	2025	Dhitareka, A. U. P. H., Husna, H. N., & Prima, E. C.	(Dhitareka et al., 2025)	Indonesia
32	2025	Agyare, B., Asare, J., Kraishan, A., Nkrumah, I., & Adjekum, D. K.	(Agyare et al., 2025)	Ghana
33	2025	Avci, H., Lunn, S. J., & Hazari, Z.	(Avci et al., 2025)	USA
34	2025	Jufrida, K., Furqon, A., Falah, R., & Riantoni, R.	(Jufrida et al., 2025)	Indonesia
35	2025	Wattanakasiwich, P., Kaewkhong, K., & Katwibun, D.	(Wattanakasiwich et al., 2025)	Thailand
36	2025	Bravo, B., Inorreta, Y., Jara, Y., & Perez, G	(Bravo et al., 2025)	Argentina

37	2025	Meyer, A., Bleckmann, T., & Friege, G.	(Meyer et al., 2025)	Germany
38	2025	Daoudi, M.	(Daoudi ,2025)	Morocco
39	2025	Fekets, G.	(Fekets,2025)	Taiwan
40	2025	Xu, Y., Liu, L., Xiong, J., & Zhu, G.	(Xu et al., 2025)	China

Data Extraction Process

A structured data extraction form was developed for the 40 included studies. Each study was systematically analyzed in terms of publication year, country, research design (qualitative, quantitative, mixed, etc.), sample level, type of AI used, research purpose, main findings, and recommendations. The data extraction process was conducted through a database created in spreadsheet format, and all records were processed according to a standardized coding scheme. The coding scheme was developed through a combination of deductive and inductive approaches. Initially, a preliminary set of codes was derived from the research questions and relevant literature and subsequently refined based on patterns emerging from the data during the analysis process. This approach aimed to ensure data integrity and enhance the transparency of the analysis process.

Validity and Reliability

To ensure reporting validity, the review process was conducted in accordance with the PRISMA 2020 guidelines. The PRISMA checklist contributed to the transparent, traceable, and replicable reporting of the systematic review. To evaluate the methodological quality of the included studies, critical appraisal criteria appropriate to each research design were applied. Each study was examined in terms of methodological clarity, sample adequacy, consistency of data collection and analysis processes, and the grounding of findings. Studies with serious methodological deficiencies were excluded from the evaluation. To ensure the reliability of the selection and coding process, the studies were independently evaluated by two researchers. Cohen's Kappa coefficient was calculated to determine inter-rater agreement. Based on observed and chance agreement rates, the Kappa value was calculated as $\kappa = 0.85$. According to the classification of Landis and Koch (1977), this value indicates "almost perfect agreement," thereby supporting the methodological reliability of the study.

Data Analysis

A meta-analysis was not conducted in this study. The included studies were first analyzed using descriptive statistics (frequency and percentage distributions). Subsequently, a thematic content analysis was performed. The studies were classified within thematic categories such as types of AI (e.g., intelligent tutoring systems, machine learning), pedagogical purposes of AI use, benefits and challenges of AI applications, and research methods. The findings were synthesized and interpreted through a systematic and holistic approach.

Results and Discussion

In this section, the bibliometric characteristics and content-related findings of the 40 studies examined within the scope of the research are presented together. The findings were interpreted using a descriptive analysis approach based on the data provided in Table 3, including publication year, author(s), APA citation information, and country distribution. In addition, thematic evaluations were conducted in line with the content-related findings of the articles.

Figure 2 presents the distribution of the articles included in the study by publication year. An examination of the findings indicates that the number of studies has increased markedly in recent years. Considering the overall distribution, 2025 has the highest number of publications (n = 20). This is followed by 2024 (n = 12) and 2023 (n = 7), whereas only one article (n = 1) was identified in 2019. These data suggest that AI applications in physics instruction have gained significant momentum, particularly after 2023. Although one study was identified in 2019, a noticeable increase has been observed beginning in 2023. The most striking growth occurred in 2024 and 2025. When the publication trend is examined, there is an approximate 71% increase from 2023 to 2024 and an approximate 67% increase from 2024 to 2025. This pattern indicates that the field is still in a developmental phase and that its research potential is progressively expanding.

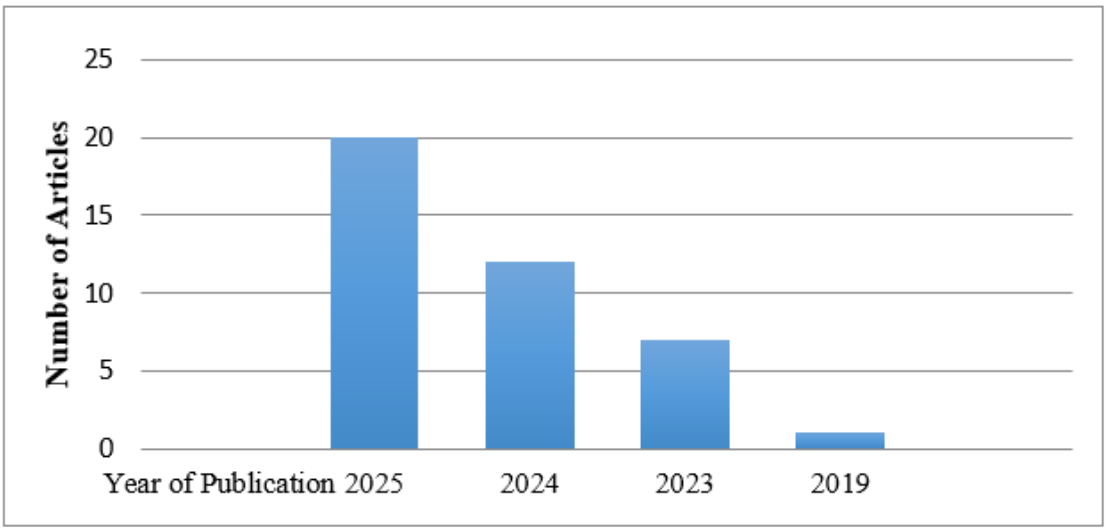


Figure 2. Distribution of articles by year

Based on the data presented in Table 3, the distribution of the 40 articles on AI-supported physics instruction was analyzed by country. The findings indicate that publications are concentrated in specific countries (Table 4). The United States (15%), China (12.5%), and Germany (12.5%) stand out as the countries with the highest publication rates. Together, these three countries account for approximately 40% of the total studies. The notable shares of Kazakhstan and Greece (7.5%) demonstrate that contributions to the literature are not limited to technologically leading countries but also include developing nations. Morocco, Brazil, Ecuador, and Indonesia (5% each) represent emerging contributions from different regions of the world, indicating a geographically diversified but still limited participation. However, a substantial proportion of the studies (25%) consists of single contributions from various other countries. This distribution suggests that while the field is beginning to expand globally, research output remains concentrated in specific centers. In other words, publication production does not exhibit a homogeneous distribution but tends to be more intensive in countries with stronger academic and technological infrastructures. Overall, the findings indicate that AI applications in physics instruction have become a global research topic; however, publication output is still concentrated in certain countries. This pattern suggests that the field is in a developmental phase and that a more balanced geographical distribution may emerge in the coming years.

Table 4. Distribution of publications by country

Country	Number of articles (n)	Percentage (%)
United States	6	15%
China	5	12.5%
Germany	5	12.5%
Kazakhstan	3	7.5%
Greece	3	7.5%
Morocco	2	5%
Brazil	2	5%
Ecuador	2	5%
Indonesia	2	5%
Other countries (each with 1 article)	10	25%

The 40 studies examined in this research were analyzed in terms of research design and participant level. The findings indicate that although methodological diversity exists, mixed methods research designs are clearly dominant (see Table 5). When the distribution presented in Figure 3 is examined, it is observed that 21 studies employed mixed methods, 10 were quantitative, 6 were qualitative, and 3 were theoretical in nature. These results suggest that the field is not limited to measurement- and comparison-based quantitative approaches; rather, qualitative and multidimensional designs aimed at understanding participant experiences, perceptions, and implementation processes have also become widespread. The fact that a significant proportion of the studies published in 2024 and 2025 adopted a mixed methods approach (e.g., Abdulayeva et al., 2025; Ben-Zion et al., 2025; Domenichini et al., 2024) indicates a shift toward a more integrative methodological orientation in the field. Quantitative studies were found to rely primarily on experimental or quasi-experimental designs and focused on measuring learning outcomes, academic achievement, or performance variables (e.g., Ding et al., 2023; Guerrero-

Zambrano et al., 2025; Xu et al., 2025). In contrast, qualitative studies tended to generate in-depth data regarding teacher perspectives, student experiences, and implementation processes (Bessas et al., 2023; Jang & Choi, 2024; Kemouss & Khaldi, 2025).

Overall, the findings demonstrate that mixed methods designs have become predominant in research on AI-supported physics education. This trend suggests that researchers aim not only to measure learning outcomes but also to evaluate implementation processes and participant experiences simultaneously. While early studies in the field primarily focused on quantitative achievement measurements, over time greater attention has been given to pedagogical impact, user experience, and classroom interaction variables.

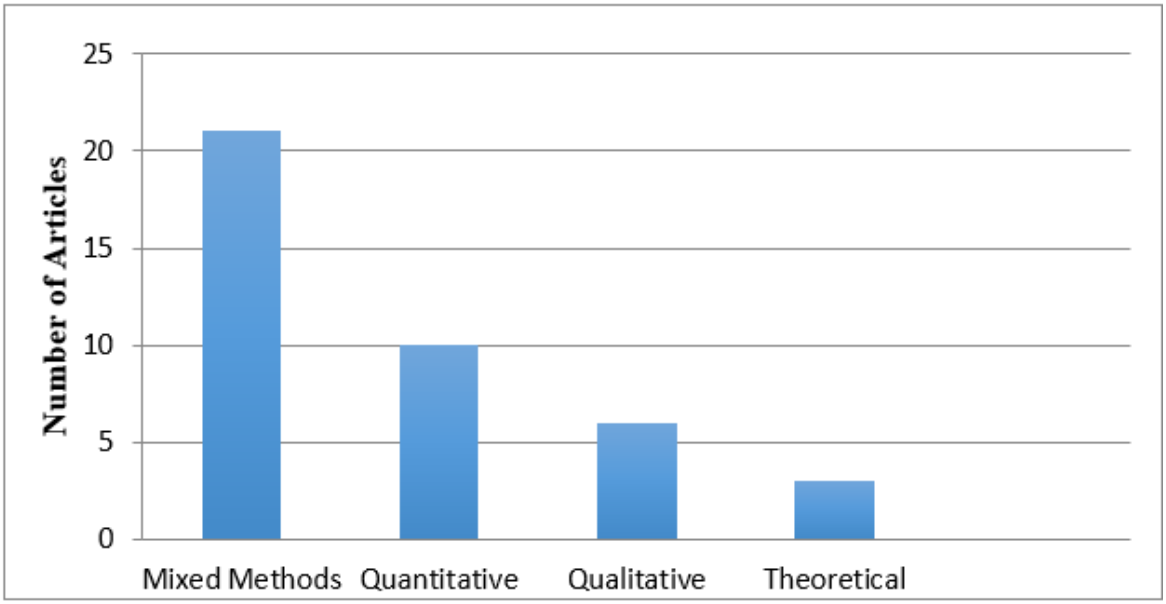


Figure 3. Distribution of publications by research type

The analyses conducted in terms of participant type indicate that research in this field is largely concentrated at the higher education level. Of the 40 studies examined, 13 were conducted exclusively with undergraduate students (e.g., Abdulayeva et al., 2025; Domenichini et al., 2024; Sirnoorkar et al., 2024). This finding suggests that the field has been predominantly shaped within the university context and that research designs have been developed primarily in higher education settings. At the high school level, 6 studies were found to involve only high school students (e.g., Abdulayeva et al., 2025; Kemouss & Khaldi, 2025; López-Simó & Rezende, 2024). In addition, 5 studies included both high school and undergraduate students (e.g., Revalde et al., 2025; Xu et al., 2025). When these two categories are considered together, a total of 11 studies involve the high school level. This indicates that secondary education also holds a significant representation within the field. Therefore, while research on AI-supported physics education is not limited to the university level, higher education remains the dominant context.

Studies involving the middle school level are more limited. Only 2 studies include middle school students, and in these studies middle school participants were examined either together with teachers (Wattanakasiwich et al., 2025) or with high school students (Bessas et al., 2025). This limited representation suggests that AI-supported physics applications for younger age groups are still in the early stages of dissemination. The relatively limited scope of physics content at the middle school level, along with the need for more pedagogically sensitive planning of implementation processes, may be among the possible reasons for this situation. The total number of studies involving teachers is 5 (e.g., Avci et al., 2025; Bravo et al., 2025; Uğraş et al., 2024). Some of these studies focus directly on teachers' perspectives, while others examine teachers alongside student groups. Representing approximately 12.5% of the total sample, this group indicates a limited distribution in terms of teacher representation. However, the sustainable and pedagogically meaningful implementation of AI systems in classroom settings is directly related to teachers' integration skills, pedagogical alignment processes, and levels of professional development. In this respect, the existing body of research appears to be structured primarily around student outcomes (e.g., achievement, performance, conceptual learning), whereas the teacher dimension has not yet been examined in sufficient depth.

Table 5. Characteristics of the studies: research design and participant type

No	Study of type	Participant type
1	Quantitative	High school students
2	Quantitative	Undergraduate students
3	Theoretical	-
4	Theoretical	-
5	Quantitative	High school students
6	Quantitative	Mixed students
7	Quantitative	Undergraduate students
8	Mixed Methods	High school and undergraduate students
9	Mixed Methods	High school students
10	Qualitative	Teachers and academics
11	Mixed Methods	High school and undergraduate students
12	Qualitative	Graduate students
13	Theoretical	-
14	Mixed Methods	Undergraduate students
15	Mixed Methods	High school students
16	Mixed Methods	Undergraduate students
17	Quantitative	K–12 Teachers
18	Mixed Methods	Undergraduate students
19	Mixed Methods	High school and undergraduate students
20	Qualitative	High School Teachers
21	Mixed Methods	Undergraduate students
22	Mixed Methods	High school and undergraduate students
23	Mixed Methods	High school students
24	Quantitative	Undergraduate students
25	Mixed Methods	Undergraduate students
26	Mixed Methods	Undergraduate students
27	Mixed Methods	Undergraduate students
28	Quantitative	Middle school students and teachers
29	Mixed Methods	High school students
30	Mixed Methods	Undergraduate students
31	Qualitative	-
32	Mixed Methods	Undergraduate students
33	Qualitative	Secondary school STEM teachers
34	Mixed Methods	Undergraduate students
35	Mixed Methods	Middle and high school students
36	Quantitative	Teachers
37	Mixed Methods	High School students
38	Qualitative	High School teachers
39	Mixed Methods	High School students
40	Quantitative	High school and undergraduate students

When the overall distribution is considered, it becomes evident that the field has not yet achieved a fully balanced structure in terms of sample diversity. The stronger technological infrastructure in higher education institutions, easier access to AI applications, and the conceptual intensity of university-level physics courses may help explain the concentration at the undergraduate level. In contrast, the limited number of studies conducted with younger age groups may be associated with the need for more careful pedagogical and ethical planning of AI-based implementations at these levels. In conclusion, the findings related to participant level indicate that research on AI-supported physics education has begun to diversify methodologically; however, there remains a need for more inclusive and balanced sample distributions. Increasing practice-based studies at the secondary education level and research focusing on teacher education may strengthen the pedagogical depth of the field and contribute to a more effective integration of technological innovations into classroom transformation.

When the 40 studies examined in this research are analyzed in terms of physics content areas, it becomes evident that the literature is concentrated around specific thematic domains (see Table 6). The findings indicate that AI-supported applications have been most frequently tested in mechanics-related topics. Studies focusing on mechanics (e.g., force, motion, energy, acceleration, centripetal force, statics, kinematics) constitute a substantial portion of the literature (e.g., Aldazharova et al., 2024; Kieser et al., 2023; López-Simó & Rezende, 2024; Xu et al., 2025). This concentration may be associated with the fact that mechanics forms the foundation of physics

instruction at both secondary and tertiary levels and, due to its problem-solving-oriented structure, provides an appropriate domain for evaluating the performance of AI applications. As one of the core pillars of physics education, mechanics encompasses concepts such as force, motion, and energy, which are frequently associated with conceptual misconceptions while also allowing for the assessment of problem-solving skills. Therefore, it may offer a suitable testing ground for evaluating the accuracy, reasoning processes, and solution steps of AI systems. Electromagnetism and electricity also exhibit a notable concentration (e.g., Cho, 2024; Revalde et al., 2025). In particular, AI performance has been tested in these domains within the context of conceptual inventories and problem-solving tasks. Studies focusing on thermodynamics and heat are also present in the literature (e.g., Jufrida et al., 2025; Putra Habib Dhitareka et al., 2025), although this domain does not appear to be represented as extensively as mechanics. Quantum physics and advanced-level topics, by contrast, are represented to a relatively more limited extent compared to mechanics and foundational areas (e.g., Bitzenbauer et al., 2023; Singh, 2024). Nevertheless, the emergence of AI- and augmented reality (AR)-supported personalized feedback applications within quantum contexts suggests that the field is gradually expanding toward more abstract and mathematically intensive content areas. This development indicates that AI systems may hold potential not only for foundational problem-solving tasks but also for the instruction of conceptually complex and abstract topics.

Studies addressing general physics or physics within a broader STEM context also occupy a significant place in the literature (e.g., Avci et al., 2025; Bessas et al., 2025; Uğraş et al., 2024). Another notable finding is that a considerable proportion of studies focus directly on testing the problem-solving, assessment, or feedback capabilities of ChatGPT or other generative AI systems (e.g., Bitzenbauer et al., 2023; Wei et al., 2025; Xu et al., 2025). This trend suggests that the field has largely concentrated on analyzing the performance of large language models in solving physics problems. While many of these studies primarily address the question, "Can AI solve physics problems?" (e.g., Aldazharova et al., 2024; Tong et al., 2023; Xu et al., 2025), issues related to pedagogical design, long-term learning effects, and conceptual change processes have been addressed to a comparatively limited extent. In conclusion, the distribution of topics indicates that the literature on AI-supported physics education is largely concentrated around fundamental concepts and problem-solving contexts, whereas advanced, interdisciplinary, and conceptually deep domains remain comparatively less explored.

Table 6. Physics topic and study title

No	Physics topics	Article title
1	Mechanics: force, torque, angular momentum, linear & circular motion	Can Aikido Help With the Comprehension of Physics? A First Step Towards the Design of Intelligent Psychomotor Systems for STEAM Kinesthetic Learning Scenarios
2	Optics	Students' perceptions of using ChatGPT in a physics class as a virtual tutor
3	Hydrostatics, fractals	Implementing AI in Physics lessons in the High School
4	Problem-solving, vectors, quantitative analysis	Exploring the potential of using ChatGPT in physics education.
5	Quantum physics: wave–particle duality, photon	ChatGPT in physics education: A pilot study on easy-toimplement activitie
6	Physics problems	Investigating ChatGPT-4's performance in solving physics problems and its potential implications for education
7	Newtonian mechanics: force, motion, acceleration	Educational data augmentation in physics education research using ChatGPT
8	Mechanics: loop-the-loop, energy, centripetal force	Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory
9	High school physics: mechanics, thermodynamics	The Essential Technology Implementations for Developing a Hybrid Module for High School Physics in the Sultanate of Oman
10	General physics	A Double-Edged Sword: Physics Educators' Perspectives on Utilizing ChatGPT and Its Future in Classrooms
11	Newtonian mechanics: force, motion, friction, inertia	Assessing AI's problem solving in physics: Analyzing reasoning, false positives and negatives through the force concept inventory
12	Graduate physics: quantum, electromagnetism, thermal, computational	2024 Jackson Award for Excellence in Graduate Physics Education lecture: Physics graduate education for the 21st century
13	Physical foundations, scientific discovery	Education Implications of Artificial Intelligence-Based Chemistry and Physics Nobel Prizes
14	Classical mechanics	An AI-Driven Approach for Enhancing Engagement and Conceptual Understanding in Physics Education

15	Mechanics: Newton's laws, acceleration, energy	Challenging ChatGPT with Different Types of Physics Education Questions
16	Statics, kinematics	Physics XP: Integration of ChatGPT and Gamification to Improve Academic Performance and Motivation in Physics 1 Course.
17	General physics	ChatGPT in Brazilian K-12 science education.
18	Mechanics: centripetal force, motion	Student and AI responses to physics problems examined through the lenses of sensemaking and mechanistic reasoning.
19	Force, electricity & magnetism, resistive circuits	An investigation of using Spark generative AI in solving physics concept inventories in English and Chinese: performance and issues.
20	General physics (STEM context)	Innovative Early Childhood STEM Education with ChatGPT: Teacher Perspectives
21	Thermodynamics, electromagnetism	Fostering AI literacy in pre-service physics teachers: inputs from training and co-variables
22	Mechanics, waves, electromagnetism	Can ChatGPT Pass a Physics Test?
23	Waves, electricity, mechanics, nuclear	Physics Teaching with Artificial Intelligence (AI): A Personalized Approach for Accommodator-Style Learners According to Kolb.
24	Mechanics, thermodynamics, waves, EM, quantum, energy, relativity	Transforming Physics Teacher Training Through ChatGPT: A Study on Usability and Impact.
25	Mechanics & electromagnetism: conceptual understanding, problem-solving	Adaptive intelligent tutoring systems for STEM education: analysis of the learning impact and efectiveness of personalized feedback.
26	Quantum cryptography	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
27	Computational physics, quantitative problems	Research on Intelligent Grading of Physics Problems Based on Large Language Models
28	General physics	The role of ChatGPT in junior high school physics education: Insights from teachers and students and guidelines for optimal use
29	Advanced physics	The Role Of Artificial Intelligence (Ai) In Personalised Physics Education
30	Ballistics, electric fields, quantum wells	Leveraging AI for Rapid Generation of Physics Simulations in Education: Building Your Own Virtual Lab
31	Heat and temperature	An exploratory study of ChatGPT in STEM teaching on heat and temperature topic
32	University-level physics education	A cross-national assessment of artificial intelligence (AI) Chatbot user perceptions in collegiate physics education.
33	Physics & STEM: general, not topic-specific	Exploring STEM educators' perspectives on the integration of AI-enabled technologies in teaching and learning.
34	Thermodynamics, heat transfer, structural equilibrium, rotational dynamics	Ai-Driven Ethnoscience Learning: Enhancing Physics Education Through Malay Cultural Insights.
35	Nuclear, fluids, EM, force & motion	Physics instructors' acceptance and implementation of generative AI.
36	Electromagnetic induction	Use of Generative Artificial Intelligence to Solve Physics Problems in Engineering.
37	Mechanics and thermodynamics	Automatic feedback on physics tasks using open-source generative artificial intelligence
38	Classical to quantum mechanics	Moroccan Teachers' Perceptions About Integrating Historical Contexts and AI in the Seamless Transition from Classical to Quantum Mechanics.
39	Physics & STEM, general, not topic-specific	Development Of An Artificial Intelligence supported Chatbot As An Interactive Learning Platform In Stem Education: Exploring Usability And Student Experience
40	Mechanics: energy, friction, elastic–plastic deformation	Graders Of The Future: Comparing The Consistency And Accuracy Of Gpt4 And Pre-Service Teachers In Physics Essay Question Assessments

Table 7 presents the publication details of the analyzed studies, categorizing each article according to journal title, volume, issue, page range, and DOI information. The reviewed studies were published in a range of internationally recognized journals in the fields of educational technology and physics education. Notable outlets include *Computers and Education* (Avci et al., 2025), *Physical Review Physics Education Research* (Kieser et al., 2023; Tschisgale et al., 2023; Wattanakasiwich et al., 2025), *Computers and Education: Artificial Intelligence* (Agyare et al., 2025; Sirnoorkar et al., 2024), and *American Journal of Physics* (Singh, 2024). This distribution indicates that research on AI-supported physics education is positioned at the intersection of educational technology and discipline-based physics education research, gaining visibility in both domains.

Table 7. Publication details of articles by journal and DOI.

No	Name of Journal	Volume	Issues	Pages	DOI
1	IEEE Access	7	-----	176458	https://doi.org/10.1109/ACCESS.2019.2957947
2	International Journal of Educational Technology in Higher Education	20	63	5-18	https://doi.org/10.1186/s41239-023-00434-1
3	International Conference on Computational Science and Computational Intelligence	....	---	-1775 1779	https:// 10.1109/CSCI62032.2023.00293
4	Smart Learning Environments	10		2-19	https://doi.org/10.1186/s40561-023-00273-7
5	Contemporary Educational Technology	15	3	2-10	https://doi.org/10.30935/cedtech/13176
6	Asia Pacific Education Review	25	-----	1379–1389	https://doi.org/10.1007/s12564-023-09913-6
7	Physical Review Physics Education Research	19	-----	020150 -13	https://doi.org/10.1103/PhysRevPhysEducRes.19.020150
8	Physical Review Physics Education Research	19	2	020123 -24	https://doi.org/10.1103/PhysRevPhysEducRes.19.020123
9	International Journal of Instruction	17	3	617-634	https://doi.org/10.29333/iji.2024.17334a
10	Journal of Science Education and Technology	34	----	267–283	https://doi.org/10.1007/s10956-024-10173-1
11	Contemporary Educational Technology	16	4	1-16	(http://creativecommons.org/licenses/by/4.0/)
12	American Journal of Physics	92	12	918–923	https://doi.org/10.1119/5.0242316
13	Journal of Chemical Education	101	---	4533−4534	https://doi.org/10.1021/acs.jchemed.4c01275
14	IEEE Global Engineering Education Conference	....	---	1-3	https://doi.org/10.1109/EDUCON60312.2024.10578670
15	The Physics Teacher	62	----	290–294	https://doi.org/10.1119/5.0160160
16	International Journal of Engineering Pedagogy	14	6	82-92	https://doi.org/10.3991/ijep.v14i6.47127
17	Frontiers in Education	9	1321547	1-8	https://doi.org/10.3389/feduc.2024.1321547
18	Computers and Education: Artificial Intelligence	7	100318	1-15	https://doi.org/10.1016/j.caeai.2024.100318
19	Discover Artificial Intelligence	4	108	1-16	https://doi.org/10.1007/s44163-024-00215-3
20	Technology, Knowledge and Learning	30	----	809–831	https://doi.org/10.1007/s10758-024-09804-8
21	Technology, Knowledge and Learning	10	1505420	1-13	https://doi.org/10.3389/feduc.2025.1505420
22	Technology, Knowledge and Learning	30	---	1-20	https://doi.org/10.1007/s10758-025-09814-0
23	International Journal of Instruction	18	3	39-58	https://doi.org/10.29333/iji.2025.1833a

24	Education Sciences	15	887	1-26	https://doi.org/10.3390/educsci15070887
25	Smart Learning Environments	12	41	1-31	https://doi.org/10.1186/s40561-025-00389-y
26	Smart Learning Environments	12	15	1-28	https://doi.org/10.1140/epjqt/s40507-025-00310-z
27	Education Sciences	15	116	1-15	https://doi.org/10.3390/educsci15020116
28	Social Sciences & Humanities	11	101610	1-13	https://doi.org/10.1016/j.ssaho.2025.101610
29	Jurnal Pendidikan IPA Indonesia	14	3	599-615	https://journal.unnes.ac.id/journals/jpii
30	Jurnal Pendidikan IPA Indonesia	63	----	424-427	https://doi.org/10.1119/5.0252343
31	Discover Education	4	309	1-25	https://doi.org/10.1007/s44217-025-00751-9
32	Computers and Education: Artificial Intelligence	8	100365	1-15	https://doi.org/10.1016/j.caeai.2025.100365
33	Computers and Education	9	100304	1-12	https://doi.org/10.1016/j.caeo.2025.100304
34	Computers and Education Open	24	----	1-27	https://doi.org/10.28945/5520
35	Physıcal Revıew Physıcs Educatıon Research	21	010155	010155-24	https://doi.org/10.1103/r2fn-kdy4
36	Physical Review Physics Education Research	43	3	73–91	https://orcid.org/0000-0002-3941-0547
37	International Journal of Science Education	….	0693	1464-5289	https://doi.org/10.1080/09500693.2025.2499220
38	Mechanics, African Journal of Research in Mathematics, Science and Technology Education	29	2	171-190	https://doi.org/10.1080/18117295.2025.2460911
39	Journal of Baltic Science Education	24	5	878–893	https://doi.org/10.33225/jbse/25.24.878
40	Journal of Baltic Science Education	24	1	187–207	https://doi.org/10.33225/jbse/25.24.187

The types of artificial intelligence employed in the 40 reviewed studies were analyzed, and the findings are presented in Table 8. Overall, the results indicate a clear dominance of generative AI–based chatbots and advanced large language models (LLMs), particularly GPT-4–based systems. ChatGPT and GPT-4 were the most frequently utilized AI tools across the studies. Research directly employing ChatGPT or GPT-4 (n = 24) accounts for more than half of the total sample, highlighting the central role of generative AI in physics education research, particularly in the post-2023 period. In contrast, classical AI-based Intelligent Tutoring Systems (ITS) were represented in only a limited number of studies. Similarly, AI-supported sensor-based systems appeared in only one study. This pattern suggests a notable shift from earlier adaptive instructional systems toward generative language models in recent research trends. In several studies, the type of AI was described using general terms without specifying the exact model or technological framework. This lack of technical detail indicates that systematic model-level comparisons remain limited in the literature. Additionally, some studies adopted a multi-tool approach. For example, one study combined ChatGPT, Google AI, and PhET simulations, demonstrating that AI tools are increasingly being integrated with existing digital learning environments rather than being used in isolation.

A broader trend analysis reveals that earlier studies primarily emphasized ITS and adaptive learning systems, whereas recent research has been largely dominated by generative AI applications. The strong presence of ChatGPT and GPT-4 suggests that the field is increasingly leveraging LLMs for problem solving, feedback generation, and conceptual explanation tasks. This trend may be associated with accessibility factors, as systems such as ChatGPT are widely available and can be readily implemented in classroom settings. In contrast, the development of ITS platforms requires extensive software design and technical infrastructure, which may explain their reduced presence in recent studies. However, the concentration of research around a single platform (ChatGPT/GPT-4) may introduce a technology-dependent orientation within the field. Moreover, given the known limitations of generative AI systems—including accuracy concerns, hallucination risks, and issues related to

pedagogical appropriateness—future studies should adopt more rigorous and comprehensive evaluation frameworks. Although data-driven modeling, simulation-based learning, and experimental practices are central to physics education, the limited representation of ITS and sensor-based AI systems points to a potential research gap. Studies employing multiple AI tools suggest that integrating generative chatbots with simulations, augmented reality, and virtual laboratory environments may offer more robust learning experiences, particularly for teaching abstract physics concepts.

Table 8. AI type used in the studies.

No	AI type used
1	AI-supported sensor-based feedback systems
2	ChatGPT
3	GPT-3.5 / GPT-4
4	ChatGPT
5	ChatGPT
6	ChatGPT-4
7	ChatGPT / GPT-4
8	ChatGPT
9	Artificial Intelligence (AI)
10	ChatGPT / GPT-4
11	GPT-4
12	Not specified
13	Artificial Intelligence (AI)
14	Artificial Intelligence (AI)
15	ChatGPT-3.5
16	ChatGPT
17	ChatGPT
18	ChatGPT-3.5 / GPT-4
19	ChatGPT-3.5 / GPT-4
20	Artificial Intelligence (AI)
21	Generative AI (GenAI) – ChatGPT
22	ChatGPT-3.5
23	Artificial Intelligence (AI)
24	Generative AI (AI-based)
25	Intelligent Tutoring System (ITS)
26	ChatGPT / GPT-4
27	Large Language Model (LLM) – GPT-4
28	ChatGPT
29	Artificial Intelligence (AI)
30	GPT-4
31	ChatGPT / GPT-4
32	ChatGPT (Generative AI Chatbot)
33	AI-supported educational technologies
34	Artificial Intelligence (AI)
35	Generative AI (GenAI)
36	ChatGPT
37	GPT-4
38	ChatGPT, Google AI, PhET Interactive Simulations
39	GPT-4
40	ChatGPT-4

Table 9 presents the applications of artificial intelligence in physics education research. The analysis of 40 studies indicates that AI has been utilized in a multidimensional manner across physics education and physics education research, including instructional support, virtual tutoring, personalized and adaptive learning systems, assessment and automated feedback, conceptual analysis, research-oriented applications, and professional development. Studies focusing on instructional use (e.g., Bessas et al., 2025; Bitzenbauer et al., 2023; Bravo & Pérez, 2025; Zou et al., 2023) demonstrate that AI systems can provide immediate feedback, guide problem-solving processes, explain concepts, and address misconceptions.

Table 9. Application of AI in physics research

No	Application of AI in studies
1	Design of psychomotor systems training through AI and sensory perception modeling and acrobatic movements (STEAM) to support the learning of physical concepts.
2	AI (ChatGPT) was used as a virtual tutor in university-level physics courses to answer physics questions on light and radioactivity and to examine students' incorrect responses.
3	AI-based ChatGPT was applied in high school physics courses to create lesson plans, answer student questions, correct misconceptions, and relate concepts to real life, especially for hydrostatic pressure and fractals topics.
4	AI-based ChatGPT was used to explain concepts, guide problem-solving, and provide immediate feedback to students in physics classes.
5	ChatGPT was used in secondary-level physics (quantum physics) courses to develop critical thinking skills.
6	ChatGPT-4 was used at middle and high school levels to solve physics problems, assess conceptual understanding, and evaluate physics reasoning performance in comparison with students.
7	ChatGPT (GPT-4 based large language model) was employed in physics education research to generate synthetic data for Force Concept Inventory (FCI) and analyze conceptual understanding and physics reasoning performance.
8	AI (NLP and machine learning) was used in physics education research to analyze large-scale qualitative text data and identify patterns.
9	AI was considered as one of the technological applications evaluated for hybrid physics modules, but ranked last in priority according to Fuzzy Delphi Method (FDM) results.
10	ChatGPT can support problem-solving and personalized learning; however, risks include dependency, educational inequality, and teachers' digital competency needs.
11	GPT-4 was tested on Force Concept Inventory (FCI) in physics education; it showed high accuracy in topics like Newton's third law but struggled with diagram interpretation and spatial reasoning, exhibiting conceptual errors.
12	Highlights the pedagogical necessity of using digital and online technologies, including large language models (LLMs), in physics education.
13	For the Physics prize, the role of physics in developing the foundations of artificial neural networks has been acknowledged.
14	Automatically generate physics learning activities in a gamified environment using Generative AI, design adaptive learning paths, and personalize the teaching of classical mechanics concepts.
15	AI is successful in definitions and simple calculations in physics; it is limited in complex and interpretive problems, thus serving as a supportive tool.
16	The combined use of ChatGPT and gamification significantly increased academic achievement and motivation in Physics 1 students (188 students, experimental–control group).
17	Most K-12 science teachers in Brazil are cautious and uncertain about whether ChatGPT can enhance educational quality or whether its use counts as plagiarism.
18	Examined "meaning-making" and "mechanical reasoning" in solving physics problems by comparing student responses with ChatGPT (versions 3.5 and 4.0); results indicate AI responses reflect a "physics definition" approach, while student responses reflect a "physics application" approach.
19	Examines the use of generative AI in physics education research and analyzes its performance in conceptual physics tests in mechanics and electromagnetism; no application for pure physics or physical simulations was considered.
20	Investigates ChatGPT's integration into early STEM education based on teacher perspectives; highlights advantages such as instant feedback, personalized content, motivation, and faster instruction while addressing technical issues and security concerns.
21	Examines the effect of a pre-service physics teacher training program to improve AI literacy and investigates the role of perceived usefulness of AI on future teaching intentions.
22	Evaluates ChatGPT's performance on multiple-choice questions and problem-solving tasks in physics education across four languages; relatively successful in theoretical questions, limited in problem-solving, with language significantly affecting performance.
23	Examines AI-supported personalized physics instruction for high school students with Kolb's accommodator learning style; AI improves learning experiences and performance via interactive simulations, immediate feedback, and adaptive activities.
24	Investigates ChatGPT's usability and impact in pre-service physics teacher training; shows AI improves educational activity design, teacher satisfaction, and effectiveness in developing adaptive, game-based physics activities.

25	AI-based intelligent tutoring systems provide real-time, adaptive, and personalized learning in mathematics, physics, and programming; significantly improved academic achievement and student satisfaction with adaptive feedback in the experimental group.
26	ChatGPT-based feedback integrated into an AR-supported quantum cryptography lab improved university students' learning outcomes and cognitive processes; AI feedback directed visual attention to task-relevant elements, enhancing learning performance.
27	Examines automated physics problem assessment based on large language models; the tree-of-thought prompt approach can score complex computational problems with high accuracy and offers strong potential for intelligent assessment in physics education.
28	AI is used in physics research, especially via language models like ChatGPT, to support conceptual explanations and investigate problem-solving processes; considered a complementary tool requiring teacher guidance for accuracy and conceptual consistency.
29	AI-based personalized learning systems adapt content based on student performance in physics education, showing significant improvements in academic achievement, problem-solving skills, and critical thinking.
30	Generative AI models (e.g., ChatGPT, Claude) enable rapid creation of interactive simulations in physics education without programming knowledge, enhancing conceptual understanding, exploratory learning, and student engagement.
31	ChatGPT's responses to STEM pedagogy questions align conceptually with existing academic literature but have limited reliability for academic and instructional purposes due to lack of sources and multiple perspectives.
32	Examines physics students' perceptions of ChatGPT using the Technology Acceptance Model (TAM); perceived ease of use and subjective norms influenced usage intention and actual use, while ethical concerns negatively affected ChatGPT use.
33	STEM teachers adopt AI as a supportive instructional tool but face institutional support and professional development challenges in classroom implementation.
34	Aims to deliver culturally sensitive and personalized STEM education in project-based learning in Indonesia's Jambi region using machine learning and educational data mining, based on Malay ethno science.
35	Investigates physics teachers' adoption and use of generative AI, highlighting knowledge gaps, language limitations, and pedagogical concerns as major barriers in the adoption process.
36	Use of ChatGPT in solving electromagnetic induction problems.
37	Automatic assessment of student responses and feedback generation in physics problem-solving tasks.
38	Supporting historical narratives, interactive learning, and facilitating conceptual transition in physics teaching.
39	AI-generated storytelling (Newton's Laws)
40	Evaluation of student responses according to cognitive levels.

However, several studies report that AI remains limited in handling complex, interpretive, and higher-order reasoning problems (López-Simó & Rezende, Jr., 2024; Sirnoorkar et al., 2024). Within the domain of personalized and adaptive learning, AI-supported systems have been shown to adapt instructional content based on student performance and contribute to improvements in academic achievement (Abdulayeva et al., 2025; Kemouss & Khaldi, 2025; Villegas Ch. et al., 2025). Adaptive feedback systems (Villegas Ch. et al., 2025), AR + ChatGPT integration (Coban et al., 2025), and gamified learning environments (Beltozar-Clemente & Díaz-Vega, 2024; Domenichini et al., 2024) have demonstrated positive effects on learning performance and student motivation.

Significant findings also emerge regarding the use of AI in assessment processes. AI systems have been employed for automated problem grading (Meyer et al., 2025; Wei et al., 2025), conceptual test performance analysis (Aldazharova et al., 2024; Cho, 2024; Revalde et al., 2025), and evaluation of student responses according to cognitive levels (Xu et al., 2025). These applications indicate the growing potential of AI-driven intelligent assessment frameworks in physics education. In addition to instructional applications, some studies have employed AI as a methodological tool in physics education research. Applications such as synthetic data generation (Kieser et al., 2023), qualitative data analysis (Tschisgale et al., 2023), and comparative analyses of student and AI-generated responses (Sirnoorkar et al., 2024) suggest that AI can contribute to research design and analytical processes. Studies focusing on teacher perspectives (Ferreira Monteiro et al., 2024; Kemouss & Khaldi, 2025; Uğraş et al., 2024) reveal both opportunities and concerns regarding AI integration. Teachers acknowledge advantages such as instant feedback and personalization (Uğraş et al., 2024). However, ethical concerns, limited digital competencies, and insufficient institutional support are identified as significant barriers (Avci et al., 2025;

Ferreira Monteiro et al., 2024; Wattanakasiwich et al., 2025). Overall, the literature positions AI as a tool with transformative pedagogical potential in physics education. Nevertheless, it also emphasizes the necessity of strengthening theoretical and pedagogical frameworks to ensure sustainable, ethical, and educationally grounded integration.

Table 10. AI benefits in studies

No	AI benefits in studies
1	Aikido movements enhanced understanding of the "moment of inertia" concept.
2	AI literacy supported students in using AI effectively in education.
3	ChatGPT sped up teachers' lesson planning and boosted students' motivation.
4	ChatGPT solved physics problems, explained solutions, and generated new exercises.
5	Intervention improved students' perceptions of ChatGPT and daily life integration.
6	Students strengthened reasoning and scientific method skills.
7	ChatGPT produced accurate conceptual answers and modeled students' prior knowledge.
8	CGT method enhanced problem-solving and explanation quality through human–AI collaboration.
9	Mobile and digital learning technologies increased student usage.
10	ChatGPT supported personalized, inquiry-based learning; curriculum and infrastructure issues remain.
11	GPT-4 provided high accuracy on Newtonian mechanics questions; some conceptual errors observed.
12	Growth mindset instructors created inclusive and motivating learning environments.
13	AI has potential to enhance instruction through data-driven strategies.
14	Gamified learning and generative AI contributed positively to physics learning.
15	ChatGPT excelled at simple calculation questions; insufficient for complex problems.
16	AI and gamification increased students' interest, self-efficacy, and engagement.
17	Teachers observed learning benefits but highlighted accuracy and ethical considerations.
18	AI responses were structured with clear assumptions; students showed richer epistemic practices.
19	ChatGPT's language performance varied; some issues in understanding physics concepts detected.
20	Teachers agreed ChatGPT is beneficial in early childhood STEM education.
21	Intervention increased students' AI literacy and intent to integrate AI.
22	ChatGPT's problem-solving success varied across languages.
23	AI-supported learning tools improved student achievement and conceptual understanding.
24	Participants found ChatGPT useful for adapting activities and reducing preparation time.
25	Intervention group showed improvements in feedback accuracy, progress, and student satisfaction.
26	AI improves learning of abstract quantum physics concepts through personalized feedback.
27	Tree-of-Thought method solved complex problems with highest accuracy.
28	Teachers used ChatGPT for lesson planning; students for rapid answers.
29	Intervention group improved in advanced problem-solving, research skills, and motivation.
30	Students experienced active learning with simulations and enjoyed the activity.
31	ChatGPT responses conceptually aligned with literature; reliability limited by missing academic references.
32	Ethical use positively guided students' ChatGPT usage.
33	Teachers used AI as cognitive and socio-emotional support; reduced routine tasks.
34	AI increased student engagement and conceptual understanding; prediction accuracy reached 85%.
35	Teachers adopted ChatGPT at different levels; used most for content creation.
36	Improved understanding of electromagnetic induction, self-regulated learning, and metacognitive awareness.
37	Student responses classified with high accuracy; feedback deemed appropriate.
38	Producing historical context enhanced critical thinking and student engagement.
39	Students generally satisfied; multilingual interaction received limited support.
40	AI (LLMs) can enhance grading consistency, support personalized learning, and assist teachers in creating and evaluating educational content.

The findings regarding the benefits of artificial intelligence (AI) in physics education are presented in Table 10. Overall, the results indicate that AI provides significant contributions across cognitive, pedagogical, motivational, and assessment dimensions. Strong evidence is particularly observed in areas such as personalization, immediate feedback, and adaptive learning (Coban et al., 2025; Kemouss & Khaldi, 2025; Villegas Ch. et al., 2025). A substantial portion of the studies report that AI directly supports the understanding of physics concepts (Abdulayeva et al., 2025; Kemouss & Khaldi, 2025; Santos & Corbí, 2019). For instance, Bravo and Pérez (2025) reported significant improvements in the understanding of electromagnetic induction topics. Some studies further

indicate that AI-supported tools enhance student performance and conceptual comprehension (Abdulayeva et al., 2025; Jufrida et al., 2025; Kemouss & Khaldi, 2025).

However, limitations have been noted in solving complex problems and tasks that require deep conceptual reasoning (Cho, 2024; López-Simó & Rezende, Jr., 2024; Revalde et al., 2025). Some studies emphasize AI's positive impact on problem-solving and scientific reasoning skills (Tschisgale et al., 2023; Tong et al., 2023; Zou et al., 2023). In Zou et al. (2023), ChatGPT was found effective in problem-solving and generating new questions. Similarly, Tschisgale et al. (2023) highlighted that human–AI collaboration improved the quality of problem-solving.

Several studies also highlight the motivational effects of AI-supported learning environments (Ben-Zion et al., 2025; Domenichini et al., 2024; Fekets, 2025). In particular, gamification and generative AI have been reported to increase students' engagement and self-efficacy (Beltozar-Clemente & Díaz-Vega, 2024; Domenichini et al., 2024). Overall student satisfaction with AI use was generally high, and teachers reported perceiving AI as beneficial in STEM education (Fekets, 2025; Uğraş et al., 2024). One of AI's strongest contributions is its support for personalized learning and feedback (Meyer et al., 2025; Villegas Ch. et al., 2025; Xu et al., 2025). According to Coban et al. (2025), personalized feedback facilitated the learning of abstract concepts. Xu et al. (2025) reported that large language models (LLMs) improved grading consistency and content creation. Villegas Ch. et al. (2025) found that adaptive feedback enhanced student satisfaction and learning progress. These findings suggest that AI can function as a supportive tool in assessment and evaluation processes.

AI's benefits for teachers are also noteworthy. Reports indicate that AI reduces lesson planning time (Bessas et al., 2023; Bessas et al., 2025; Guerrero-Zambrano et al., 2025), supports content creation (Wattanakasiwich et al., 2025), and provides cognitive and socio-emotional support by alleviating some routine tasks (Avci et al., 2025). Nonetheless, teachers highlighted limitations related to accuracy, ethical use, and technical infrastructure (Ferreira Monteiro et al., 2024; Jang & Choi, 2024; Revalde et al., 2025). Overall, literature positions AI as a transformative tool in physics education while emphasizing the need to strengthen pedagogical frameworks to ensure sustainable and ethical integration.

The challenges encountered in physics education research that incorporates AI are summarized in Table 11. These findings indicate that while AI offers significant opportunities in physics education, it also presents various limitations at pedagogical, technical, cognitive, and ethical levels. A substantial portion of the studies report that AI can make errors in understanding physical concepts and performing mathematical operations. Reported difficulties include weak performance in numerical calculations and errors in vector directions (Zou et al., 2023), high error rates in arithmetic and trigonometric calculations (López-Simó & Rezende, Jr., 2024), increased errors in multi-step problems (Revalde et al., 2025), mistakes in physics terminology and low-quality feedback in open-ended questions (Meyer et al., 2025), and lower scoring accuracy of ChatGPT-4 compared to human evaluators (Xu et al., 2025). Additionally, AI responses were found to be persuasive but not always correct (Bessas et al., 2025; Sirnoorkar et al., 2024). Some studies also highlight limitations in AI's pedagogical content knowledge (Ben-Zion et al., 2025; Putra Habib Dhitareka et al., 2025; Fekets, 2025).

Contextual limitations of LLM-based systems are frequently reported, including difficulties in converting visual questions to text (Kieser et al., 2023), hallucinations and out-of-context outputs (Ben-Zion et al., 2025; Bravo & Pérez, 2025; Kieser et al., 2023), overly simplified or biased outputs (Avci et al., 2025; Bravo & Pérez, 2025), and multilingual performance issues (Cho, 2024; Fekets, 2025; Revalde et al., 2025). Ethical concerns regarding AI use are strongly emphasized in the literature. These include the production of inaccurate or misleading information (Bessas et al., 2023; Ferreira Monteiro et al., 2024; Wattanakasiwich et al., 2025), risks of over-reliance and excessive trust (Ferreira Monteiro et al., 2024), academic integrity and privacy issues (Holme, 2024; Wattanakasiwich et al., 2025), and challenges in ethical and pedagogical integration (Abdulayeva et al., 2025).

The uncertainty surrounding AI's accuracy and reliability requires careful use in educational settings (Holme, 2024; Jang & Choi, 2024). Examined studies also highlight structural and technological barriers, such as limited digital infrastructure in public schools (Jufrida et al., 2025), lack of institutional support and cost issues (Avci et al., 2025), and participants' unfamiliarity with AI technologies (Guerrero-Zambrano et al., 2025). These findings suggest that AI integration requires not only pedagogical but also structural transformation. Some studies indicate that short-term interventions do not produce lasting effects (Bitzenbauer et al., 2023) and sample biases exist (Santos & Corbí, 2019). In conclusion, literature identifies three essential requirements for AI use in physics education: 1) Development of pedagogical frameworks (domain-specific adaptation), 2) Strengthening ethical and critical AI literacy, and 3) Improvement of infrastructure and institutional support mechanisms.

Table 11. Challenges faced by AI in studies

No	Challenges faced by AI in research
1	Differences in participants' prior knowledge and sampling bias
2	Varied perceptions of ChatGPT usage across different groups
3	Incorrect or misleading answers may cause conceptual misunderstandings
4	Weak performance in numerical calculations; errors in vector directions
5	Limitations in accuracy of AI outputs; short-term interventions reduce lasting impact
6	Errors in methodological and mathematical representations
7	Challenges converting visual questions to text; risk of bias and hallucinations
8	Issues with verifiability, reproducibility, and scalability in traditional qualitative analyses
9	Effective integration of technology in hybrid learning limited by infrastructure
10	Reliability issues, algorithmic and language constraints
11	Low success in spatial reasoning tasks; conceptual errors in physics concepts
12	Historical inequalities; negative impact of fixed-mindset instructors
13	Accuracy and reliability concerns; ethical issues in AI use
14	Lack of conceptual understanding, low motivation, insufficient personalization
15	Arithmetic and trigonometric errors; high error rates in multi-step problems
16	Gaps in mathematical knowledge and conceptual analysis at university level
17	Misinformation, ethical concerns, over-reliance on AI, low digital literacy
18	AI solutions may be convincing but not always correct; student solutions may be incomplete
19	Difficulty understanding physics concepts; language inequities
20	Technical challenges in ChatGPT integration; student-related issues
21	Difficulties in ethical, pedagogical, and behavioral AI integration
22	Low problem-solving success; language errors; inconsistent answers
23	Students struggle with abstract concepts; teacher-centered instruction; content overload
24	Limited familiarity of participants with AI technologies; technical and access limitations
25	Challenges in Intelligent Tutoring Systems (ITS)
26	the abstract and complex nature of quantum physics makes the topic difficult to understand.
27	Automatic grading of complex physics computation problems is difficult
28	Confusing AI explanations; concerns about scientific and contextual accuracy
29	AI in physics education faces limited frameworks and poor subject-specific adaptation.
30	AI hallucinations; missing pedagogical content knowledge; inconsistent outputs
31	Lack of proper academic citations; single-perspective responses; misalignment with pedagogy
32	AI research challenges: biased or oversimplified outputs, limited context, ethical concerns, and reliance on users' critical skills
33	Lack of institutional support; cost and accessibility issues; excessive AI tools
34	Limited digital infrastructure in public schools; varied teacher competency
35	Insufficient technical knowledge; AI may generate incorrect physics information; ethical and privacy concerns
36	AI challenges: biased/oversimplified answers, need for critical thinking, and context limits
37	Errors in physics terminology; low feedback quality in open-ended tasks
38	Limited analytical depth; AI cannot fully replace teacher expertise
39	Language barriers; AI cannot replace teacher; technical and system limitations
40	ChatGPT-4 scoring accuracy lower than human evaluators

The recommendations presented in Table 12 from the reviewed studies were coded using a qualitative content analysis approach, and the resulting codes were thematically classified in Table 13. The recommendations are grouped into ten main themes. One of the most prominent areas is *Pedagogical Integration and Instructional Design* (Bessas et al., 2023; Kieser et al., 2023; Singh, 2024; Sirnoorkar et al., 2024). The studies include design-oriented recommendations such as personalized learning, adaptive systems, hybrid modules, and AR–LLM integration. These studies suggest that AI should be considered not only as a content generator but also as a tool that transforms instructional design. The second dominant theme is *Teacher Education and Professional Development* (Avci et al., 2025; Bessas et al., 2025; Sirnoorkar et al., 2024; Uğraş et al., 2024). The studies emphasize that generative AI tools should be used under teacher guidance within a pedagogical framework rather than being directly provided to students. Additionally, the development of in-service training programs and discipline-specific AI professional development models is highlighted as essential. The third dominant theme is *AI Literacy and Ethical Use* (Abdulayeva et al., 2025; Ding et al., 2023; Cho, 2024; Revalde et al., 2025). Specifically, misconceptions about GenAI among students should be addressed, usage guidelines should be

clearly defined, and ethical boundaries must be established. This finding indicates that technological integration requires not only pedagogical but also ethical transformation. *Technical Development and Model Improvement* recommendations (Cho, 2024; Coban et al., 2025; Meyer et al., 2025) point to the need for enhancing LLMs' visual processing, technical language accuracy, and multilingual capabilities. This suggests that current models still lack full proficiency in physics and related disciplines. The *Critical Thinking and Scientific Reasoning* theme (Bitzenbauer et al., 2023; Sirnoorkar et al., 2024) emphasizes that students should not be passive consumers of AI outputs; instead, they should be positioned as active learners who question and evaluate AI-generated responses. The *Assessment and Evaluation* theme (Wei et al., 2025; Xu et al., 2025) focuses on the accuracy of LLM grading and human-AI collaborative assessment models. This finding highlights that hybrid models take precedence over fully automated evaluation processes. *Long-Term and Longitudinal Research* (Abdulayeva et al., 2025; Bravo et al., 2025; Guerrero-Zambrano et al., 2025) indicates a significant gap in the literature.

Table 12. Recommendation in studies

No	Research recommendation
1	The use of AI- and sensor-supported intelligent psychomotor systems in STEAM education
2	Making AI literacy a mandatory component in STEM courses (especially in physics)
3	Using ChatGPT by students with critical thinking and by teachers for planning and personalized instruction
4	Developing effective prompting strategies for LLM use in physics education and integrating them with cognitive load–reducing instructional designs
5	Using ChatGPT under teacher guidance to support critical thinking
6	Supporting scientific thinking and collaborative learning through the ethical use of AI
7	Using ChatGPT as a data augmentation tool with human expert supervision
8	Using AI as a supportive tool working alongside human analysts in qualitative research
9	Implementing hybrid physics modules across different educational levels
10	Evaluating the impact of ChatGPT in education to improve instructional practices
11	Enhancing AI models' conceptual reasoning abilities and utilizing them to improve assessment tools
12	Creating inclusive learning environments and integrating digital tools into physics education
13	Using AI tools in innovative, creative, and data-driven ways in education
14	Integrating generative AI with gamification and personalized learning approaches
15	Expanding research to other physics topics and comparing advanced models
16	Using ChatGPT and gamification to deepen learning in university-level physics education
17	Developing AI literacy and integration strategies in teacher education
18	Designing hybrid learning activities that compare AI outputs with student responses and promote critical thinking
19	Ensuring high-quality prompts, appropriate parameters, and response verification in GenAI use
20	Providing teacher support, infrastructure development, and stakeholder awareness for ChatGPT integration
21	Including AI literacy in teacher education curricula
22	Establishing AI usage policies, promoting critical evaluation, and improving multilingual performance
23	Supporting teachers in developing AI-based instructional activities
24	Investigating the long-term effects of ChatGPT on teacher education
25	Testing systems with diverse student groups and developing more adaptive designs
26	Examining the effects of AR and LLM integration with larger samples
27	Using advanced prompting strategies to improve grading accuracy in LLMs
28	Using ChatGPT as a supportive tool in teaching within an ethical and critical framework
29	Conducting longitudinal studies on the long-term effects of AI-supported learning
30	Applying AI simulations in different course contexts and examining their effects
31	Using ChatGPT as a supportive tool rather than a direct instructional material
32	Conducting more research on the role of ChatGPT in higher education
33	Developing structured AI-based professional development programs for teachers
34	Designing culturally responsive and personalized project-based learning using data mining tools
35	Developing GenAI training, infrastructure, and prompt guidelines for physics education
36	Investigating the long-term academic effects of AI tools as cognitive support systems
37	Improving LLMs through domain-specific fine-tuning in physics
38	Examining the sustainability and long-term impact of AI integration
39	Replicating studies with larger and more diverse samples
40	Developing hybrid assessment models combining AI and human evaluation

Table 13. Thematic classification of recommendations in studies

Theme	Related recommendation article numbers
Pedagogical Integration and Instructional Design	3, 7, 12, 14, 18, 28
Teacher Education and Professional Development	17, 18, 20, 28, 33
AI Literacy and Ethical Use	2, 19, 21, 22
Technical Development and Model Improvement	19, 26, 37
Critical Thinking and Scientific Reasoning	5, 18, 28
Long-Term and Longitudinal Research	24, 29, 36
Assessment and Evaluation	27, 40
K–12 and Early STEM Integration	14, 17
Cognitive and Affective Dimensions	16, 25
Infrastructure and Support Mechanisms	20, 35

Most studies are based on short-term interventions, and evidence of sustainable effects is lacking. The absence of long-term and longitudinal studies suggests that the field is still in an early developmental stage. Future research should be supported by larger samples, diverse educational levels, and sustainable impact analyses. The *K–12 and Early STEM Integration* theme (Domenichini et al., 2024; Ferreira Monteiro et al., 2024) underscores the importance of careful and guided AI use at early ages. *Cognitive and Affective Dimensions* (Beltozar-Clemente & Díaz-Vega, 2024; Villegas Ch. et al., 2025) highlight the role of AI tools as "cognitive prostheses" and the need to investigate their impact on students' motivation, engagement, and emotional experiences. Finally, the *Infrastructure and Support Mechanisms* theme (Uğraş et al., 2024; Wattanakasiwich et al., 2025) emphasizes the necessity of developing prompt-writing guidelines, technical infrastructure, and user support systems.

Conclusion

This study analyzed research on AI-supported physics instruction, revealing a notable shift in the field's trajectory. Early studies predominantly focused on Intelligent Tutoring Systems (ITS) and adaptive learning platforms, whereas recent years have seen the dominance of generative AI systems, particularly large language models based on ChatGPT and GPT-4. This shift is directly related to the accessibility of these tools and their ease of implementation in classroom settings.

The findings indicate that generative AI systems serve as significant supportive tools in problem solving, feedback generation, and conceptual explanation processes. Increased student motivation, expanded opportunities for personalized learning, and accelerated teacher preparation were among the key contributions highlighted in the literature. However, issues such as accuracy concerns, hallucination generation, limitations in spatial reasoning, pedagogical misalignment, and ethical considerations remain notable constraints. These findings suggest that generative AI should be positioned not as a replacement for teachers, but as a complementary tool supporting the instructional process. Furthermore, the relatively limited representation of simulation-based, data-driven, and experimental AI applications indicates that the field has recently gravitated toward language-based systems. Given the nature of physics education, there is significant potential in developing hybrid AI ecosystems that support modeling, experimentation, and conceptual structuring processes.

In conclusion, although generative AI systems have become a dominant component in physics education research, ensuring sustainable and pedagogically balanced development of the field requires increasing comparative studies, implementing long-term experimental designs, and developing theoretical frameworks sensitive to the epistemological structure of the physics discipline.

Recommendations

Based on the findings of this review, several recommendations can be proposed for the integration of generative AI in physics education. First, AI tools should be implemented under teacher guidance and within a clear pedagogical framework, rather than being provided directly to students. This approach ensures that AI serves as a supportive and complementary tool rather than a replacement for instructional guidance. Second, efforts should be made to enhance the reliability of AI in conceptual explanations, problem-solving, and feedback generation through model development and accuracy verification. Third, personalized and adaptive learning systems should be integrated into physics instruction to promote student motivation and improve learning outcomes. Fourth, both

teachers and students should receive education on ethical AI use, critical literacy, and responsible data practices to address potential risks related to misinformation and over-reliance. Fifth, future research should adopt long-term, discipline-specific experimental designs to evaluate the sustained impact of AI on learning and teaching in physics. Finally, hybrid AI systems that combine simulations, data-driven approaches, and language-based tools should be developed to support modeling, experimentation, and conceptual structuring, reflecting the core nature of physics education.

Scientific Ethics Declaration

* The authors declare that the scientific ethical and legal responsibility of this article published in JESEH journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest

Funding

* There is no funding

Acknowledgements or Notes

* This study was produced from the master's thesis data of the first author.

References

Abdulayeva, A., Zhanatbekova, N., Andasbayev, Y., & Boribekova, F. (2025, February). Fostering AI literacy in pre-service physics teachers: inputs from training and co-variables. In *Frontiers in Education* (Vol. 10, p. 1505420). Frontiers Media SA.

Abdulayeva, A., Zhanatbekova, N., Andasbayev, Y., Khaimuldanov, Y., & Zhiyembayev, Z. (2025). The role of artificial intelligence (AI) in personalised physics education. *Jurnal Pendidikan IPA Indonesia*, *14*(3).

Agyare, B., Asare, J., Kraishan, A., Nkrumah, I., & Adjekum, D. K. (2025). A cross-national assessment of artificial intelligence (AI) Chatbot user perceptions in collegiate physics education. *Computers and Education: Artificial Intelligence*, 8, 100365.

Al-Kamzari, F., & Alias, N. (2024). Exploring the readiness of high school physics students for project-based hybrid learning in the Sultanate of Oman. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(2), em2404.

Al-Kamzari, F., & Alias, N. (2025). A systematic literature review of artificial intelligence (AI) in secondary school physics: applications, benefits, and challenges. *Interactive Learning Environments*, 1–18.

Aldazharova, S., Issayeva, G., Maxutov, S., & Balta, N. (2024). Assessing AI's problem solving in physics: Analyzing reasoning, false positives and negatives through the force concept inventory. *Contemporary Educational Technology*, *16*(4), ep538.

Almasri, F. (2024). Exploring the impact of artificial intelligence in teaching and learning of science: A systematic review of empirical research. *Research in Science Education*, 54(5), 977–997.

Beltozar-Clemente, S., & Díaz-Vega, E. (2024). Physics XP: Integration of ChatGPT and Gamification to Improve Academic Performance and Motivation in Physics 1 Course. *International Journal of Engineering Pedagogy*, *14*(6).

Ben-Zion, Y., Zarzecki, R. E., Glazer, J., & Finkelstein, N. D. (2025). Leveraging AI for rapid generation of physics simulations in education: Building your own virtual lab. *The Physics Teacher*, *63*(6), 424–427.

Bessas, N., Tzanaki, E., Vavougios, D., & Plagianakos, V. P. (2023, December). Implementing AI in physics lessons in the high school. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 1775–1779). IEEE.

Bessas, N., Tzanaki, E., Vavougios, D., & Plagianakos, V. P. (2025). The role of ChatGPT in junior high school physics education: Insights from teachers and students and guidelines for optimal use. *Social Sciences & Humanities Open*, 11, 101610.

Bitzenbauer, P. (2023). ChatGPT in physics education: A pilot study on easy-to-implement activities. *Contemporary Educational Technology*, *15*(3), ep430.

Bralin, A., Simroorkar, A., Zhang, Y., & Rebello, N. S. (2024). Mapping the literature landscape of artificial intelligence and machine learning in physics education research. In *2024 Physics Education Research Conference Proceedings* (pp. 52–57). American Association of Physics Teachers. https://doi.org/10.1119/perc.2024.pr.Bralin

Bravo, B., Pérez, G., Lorenzo, G., Cabellos, B., de Aldama, C., & Pozo, J. I. (2025). Creencias de docentes de Argentina sobre la Inteligencia Artificial generativa en la enseñanza y el aprendizaje. *Edutec, Revista Electrónica de Tecnología Educativa*, (94), 409–431.

Chen, Z., Zhang, J., Jiang, X., Hu, Z., Han, X., Xu, M., & Vivekananda, G. N. (2020). Education 4.0 using artificial intelligence for students performance analysis. *Inteligencia Artificial*, 23(66), 124–137.

Cho, N. (2024). An investigation of using Spark generative AI in solving physics concept inventories in English and Chinese: Performance and issues. *Discover Artificial Intelligence*, 4(1), 108.

Cheung, K. K. C., Long, Y., Liu, Q., & Chan, H. Y. (2025). Unpacking epistemic insights of artificial intelligence (AI) in science education: A systematic review. *Science & Education*, 34(2), 747–777.

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), 15.

Daoudi, M. (2025). Strategic Choices for Wind Energy in Morocco: Assessment of Onshore and Offshore Wind Energy to Achieve the Low Carbon Strategy. *Next Research*, 100808.

Ding, L., Li, T., Jiang, S., & Gapud, A. (2023). Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *International Journal of Educational Technology in Higher Education*, *20*(1), 63.

Domenichini, D., Bucchiarone, A., Chiarello, F., Schiavo, G., & Fantoni, G. (2024, May). An AI-driven approach for enhancing engagement and conceptual understanding in physics education. In *2024 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1–3). IEEE.

Fayzullina, A. R., Filippova, A. A., Garnova, N. Y., Astakhov, D. V., Kalmazova, N., & Zaripova, Z. F. (2025). Artificial intelligence in science education: A systematic review of applications, impacts, and challenges. *Contemporary Educational Technology*, *17*(4), ep613.

He, P., & Krajcik, J. (2026). Artificial intelligence in science education: global insights and future directions. *Disciplinary and Interdisciplinary Science Education Research*, *8*(1), 6.

Heeg, D. M., & Avraamidou, L. (2023). The use of artificial intelligence in school science: A systematic literature review. *Educational Media International*, *60*(2), 125–150. https://doi.org/10.1080/09523987.2023.2264990

Holme, T. (2024). Education implications of artificial intelligence-based chemistry and physics nobel prizes. *Journal of Chemical Education*, *101*(11), 4533–4534.

Hu, P., Li, Y., & Singh, C. (2024). Investigating and improving student understanding of the basics of quantum computing. *Physical Review Physics Education Research*, *20*(2), 020108.

Jang, H., & Choi, H. (2025). A double-edged sword: Physics educators' perspectives on utilizing ChatGPT and its future in classrooms. *Journal of Science Education and Technology*, *34*(2), 267–283.

Jufrida, J., Kurniawan, W., Furqon, M., Anwar, K., Falah, H. S., & Riantoni, C. (2025). AI-Driven Ethnoscience Learning: Enhancing Physics Education Through Malay Cultural Insights. *Journal of Information Technology Education: Innovations in Practice*, *24*, 013.

Kemouss, H., & Khaldi, M. (2025). Physics teaching with artificial intelligence (AI): A personalized approach for accommodator-style learners according to Kolb. *International Journal of Instruction*, *18*(3), 49–58.

Kieser, F., Wulff, P., Kuhn, J., & Küchemann, S. (2023). Educational data augmentation in physics education research using ChatGPT. *Physical Review Physics Education Research*, *19*(2), 020150.

Kotsis, K. T. (2025). Artificial intelligence for physics education in STEM classrooms: A narrative review within a pedagogy technology policy framework. *Schrödinger: Journal of Physics Education*, 6(3), 204–211. https://doi.org/10.37251/sjpe.v6i3.2148

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lee, G., Yun, M., Zhai, X., & Crippen, K. (2025). Artificial intelligence in science education research: Current states and challenges. *Journal of Science Education and Technology*, 1–18.

Liang, Y., Zou, D., Xie, H., & Wang, F. L. (2023). Exploring the potential of using ChatGPT in physics education. *Smart Learning Environments*, *10*(1), 52.

López-Simó, V., & Rezende, M. F. (2024). Challenging ChatGPT with different types of physics education questions. *The Physics Teacher*, *62*(4), 290–294.

Mahligawati, F., Allanas, E., Butarbutar, M. H., & Nordin, N. A. N. (2023, September). Artificial intelligence in physics education: A comprehensive literature review. In *Journal of Physics: Conference Series* (Vol. 2596, No. 1, p. 012080). IOP Publishing.

Monteiro, F. F., Souza, P. V. S., da Silva, M. C., Maia, J. R., da Silva, W. F., & Girardi, D. (2024, February). ChatGPT in Brazilian K-12 science education. In *Frontiers in Education* (Vol. 9, p. 1321547). Frontiers Media SA.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Revalde, G., Zholdakhmet, M., Abola, A., & Murzagaliyeva, A. (2025). Can chatgpt pass a physics test?. *Technology, Knowledge and Learning*, *30*(4), 2459–2478.

Samala, A. D., Rawas, S., Wang, T., Reed, J. M., Kim, J., Howard, N. J., & Ertz, M. (2025). Unveiling the landscape of generative artificial intelligence in education: a comprehensive taxonomy of applications, challenges, and future prospects. *Education and Information Technologies*, *30*(3), 3239–3278.

Santos, O. C., & Corbi, A. (2019). Can aikido help with the comprehension of physics? A first step towards the design of intelligent psychomotor systems for STEAM kinesthetic learning scenarios. *IEEE Access*, 7, 176458–176469.

Sirnoorkar, A., Zollman, D., Laverty, J. T., Magana, A. J., Rebello, N. S., & Bryan, L. A. (2024). Student and AI responses to physics problems examined through the lenses of sensemaking and mechanistic reasoning. *Computers and Education: Artificial Intelligence*, 7, 100318.

Tong, D., Tao, Y., Zhang, K., Dong, X., Hu, Y., Pan, S., & Liu, Q. (2024). Investigating ChatGPT-4's performance in solving physics problems and its potential implications for education. *Asia Pacific Education Review*, *25*(5), 1379–1389.

Tschisgale, P., Wulff, P., & Kubsch, M. (2023). Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research*, *19*(2), 020123.

Utami, A., Dhitareka, P. H., Husna, H. N., & Prima, E. C. (2025). An exploratory study of ChatGPT in STEM teaching on heat and temperature topic. *Discover Education*, *4*(1), 309.

Uğraş, H., Uğraş, M., Papadakis, S., & Kalogiannakis, M. (2024). ChatGPT-supported education in primary schools: The potential of ChatGPT for sustainable practices. *Sustainability*, *16*(22), 9855.

Villegas-Ch, W., Buenano-Fernandez, D., Navarro, A. M., & Mera-Navarrete, A. (2025). Adaptive intelligent tutoring systems for STEM education: analysis of the learning impact and effectiveness of personalized feedback. *Smart Learning Environments*, *12*(1), 41.

Wattanakasiwich, P., Kaewkhong, K., & Katwibun, D. (2025). Physics instructors' acceptance and implementation of generative AI. *Physical Review Physics Education Research*, *21*(1), 010155.

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), 116.

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.

Author(s) Information

Mohammad Naser Azizi
Kırıkkale University
Kırıkkale, Turkiye
ORCID iD: https://orcid.org/0009-0000-2923-2081

Arafuddin Faizi
Kunduz University
Kunduz, Afghanistan
ORCID iD: https://orcid.org/0000-0001-9948-0291

Ugur Sari
Kırıkkale University, Education Faculty
Kırıkkale/Turkiye
Contact e-mail: usari05@yahoo.com
ORCID iD: https://orcid.org/0000-0002-3469-8959

JESEH
Journal of Education in Science, Environment and Health
Volume 12, Issue 2, 2026
https://doi.org/10.55549/jeseh.898

Teaching and Learning Chemistry for the 21st Century Skills Through Artificial Intelligence - A Narrative Review

Tebogo E. Nkanyani

Article Info

Article History

Published:
01 April 2026

Received:
06 February 2026

Accepted:
03 March 2026

Keywords

Generative artificial intelligence
Critical thinking
Misconceptions

Abstract

This review provides an overview of how GAI assists in chemistry learning and teaching. Aspects such as 21st century skills and the tutoring ability of GAI were examined from the learning point of view, while also covering pedagogical aspects such as Technological Pedagogical Content Knowledge (TPACK), visualization and representation levels of Chemistry, textual aspects, and problem-based learning (PBL). The GAI seemed to elicit students' 21st century skills and, to a lesser extent, display a tutoring ability, depending on whether it was a free or paid version. Furthermore, the GAI requires objectivity and sufficient TPACK owing to its hallucinations and inconsistencies in definitions, diagram interpretation, and image generation, among others. Furthermore, the chatbots seemed to struggle with representation levels, generating responses with macro-level definitions for concepts requiring definitions at the sub-micro level. These inconsistencies and hallucinations, without meticulous verification, can lead to misconceptions. Moreover, social inequalities may negatively impact meaningful learning, as paid chatbots perform better than free versions in generating and interpreting chemistry images, among other abilities.

Introduction

Artificial Intelligence (AI) has significantly impacted the education fraternity, with tools such as Chat Generative Pre-Trained Transformer (ChatGPT), notably the frequently used chatbot. Introduced late in 2022 (Haleem et al. 2022), ChatGPT has made its mark in teaching and learning. It displays many abilities, such as assisting teachers with lesson planning (Sykes, 2023), producing classroom materials (Valeri, 2025), and designing assessments (Wandi et al., 2025). However, this ability is not limited to ChatGPT and is demonstrated by other AI software. For example, Google Bard, among others, has the ability to assist students in enhancing their writing (Ling Jen & Salam, 2024). Nonetheless, it is important that students learn and improve their writing throughout the process. The advantage of AI for the learning process is that it can produce excellently written text that looks overboard and error-free. AI also provides platforms for engagement between the user and the AI through prompting engineering (Fenta, 2025). Furthermore, AI chatbots assist in reducing teachers' workload through creation, grading, and informing students (Labadze et al., 2023). Nevertheless, ethical issues must be considered. For instance, students may use AI to write assignments without actively engaging in the learning process. A story by the Independent Online (IOL) newspaper from South Africa highlights the concern demonstrated by academics, who, apart from acknowledging the significance of AI in education, continue to note the misuse of AI chatbots by university students through the 'copy and paste' technique without engaging with the chatbot in a way that will make learning meaningful (Buthelezi, 2025). Students may be taking that approach as a shortcut or due to their inability to deal with assignment loads and deadlines, consequently resorting to that unethical approach. Some take this approach because of the challenging nature of science and the complexity of some chemistry topics. Generative AI (GAI) also has other shortcomings, such as gender bias and hallucinations (Feldman-Maggor et al., 2025), consequently providing irrelevant and incorrect responses (Sedagat, 2025; Elmas et al., 2024). Therefore, users must be well conversant with the content they are interacting with on AI chatbots to avoid collecting incorrect or incomplete information.

Current Review

Understanding the use of generative artificial intelligence (GAI) in chemistry education is important as it can inform teachers and all role players about effective innovative teaching practices. Previous literature has delved into this area, but not to the extent of this review. For example, Pabuçcu-Akış (2024) conducted a bibliometric analysis of innovative technologies in organic chemistry. The author found AI as a recent tool that was implemented in organic chemistry at a rate of 16,7 %. Nonetheless, there was no clarity on how AI was used in

the reviewed studies. Similarly, Iyamuremye et al. (2024) explored the use of AI and machine learning (ML) in chemistry education. The study found opportunities for the use of AI and ML, such as individualized instruction, support from teachers, and availability of educational resources, while it also noted challenges such as reliance on available data, bias in the available models, and privacy and security issues regarding data use. It was also highlighted that the platform was used in instances of curricular development, lesson preparation, and student assessment. However, there was no clarity on how AI was used and on which components it was helpful. This review takes things further by exploring the manner in which GAI was implemented in Chemistry Education, with a focus on the learning process and pedagogical practices. The intention was to first understand how GAI influenced the learning aspects of chemistry from the students` perspective. The focus was on students' 21st century skills, and the tutoring abilities of GAI, as described in the next section. The second aspect of the review was how teachers used GAI for teaching. Instead of looking at how GAI assisted the chemistry teacher in aspects of lesson planning and assessment, this review went deeper into other critical aspects such as teachers' technological pedagogical content knowledge (TPACK), chemistry visualization, representation levels of chemistry, textual aspects of chemistry, and problem-based learning (PBL).

Learning Chemistry Through GAI

21st Century Skills

The 21st century skills are essential for the growth of any country's economy as they are consistently used in industries and other sectors. These include critical thinking, problem solving, information analysis, reasoning, and inquiry (Binkley et al., 2012). Haryani et al. (2024) expand on this by highlighting the 4Cs: "creativity, critical thinking, problem solving, communication, and collaboration" (p. 106). These skills, which are critical for the chemical industry, are emphasized in most school curricula in different countries. For example, the South African Physical Sciences Curriculum And Policy Statement (CAPS) encourages as one of its aims, "identify and solve problems and make decisions using critical and creative thinking" (DBE, 2011, p.5). On the same page (page 5), CAPS outlines aim that align with 21st century skills. In science education, there is plenty of literature on 21st century skills. For example, studies by Haryani et al. (2024), Amandi (2023), and Kennedy (2022). Nevertheless, my interest lies in the use of AI to elicit 21st century skills in chemistry education.

Critical Thinking Skills

Authors such as Cooper (2023) and Lawasi et al. (2024) have highlighted critical thinking as an important skill that students should acquire when learning science through GAI. In this context, students would ask specific questions and consequently take a critical posture in the responses generated by the chatbot (Lawasi et al., 2024). Guo and Lee (2023) explored how the use of AI chatbots, such as ChatGPT, influences students' critical thinking skills. In essence, students were not only expected to use ChatGPT for assistance with schoolwork, but also to interact with the chatbot to consequently "critique, evaluate, analyze, and draw logical conclusions" (p.4844). This is crucial for the chemistry learning process, as students would not just take information as it comes, but make judgments as they interact with the chatbot, choosing what is important and sensible. Particularly with the inaccuracies that chatbots such as ChatGPT bring through their responses (Sedagat, 2025; Elmas et al., 2024). The study (Guo & Lee, 2023) noted positive views from students on their ability to pose challenging questions, assess information, summarize it, and solve difficult content. The demonstration of critical thinking skills by students would assist them in developing objectivity rather than subjectivity when they have information at their disposal. In another study by Tassotti (2024), students noticed the incorrectness of responses from ChatGPT at a higher chemistry level. This has positive effects on the pedagogical abilities of these future teachers, as the presence of critical thinking would enable them to make better judgements on what to use and not to use in their teaching and assessment.

Problem Solving

The other 21st century skill that was explored was problem-solving. Clark (2023) used the problem-solving model initially proposed by Taasoob and Glynn (2009) to analyze ChatGPT's closed-ended responses on chemistry topics. The model would have expected the chatbot to follow the sequence: problem conceptualization, problem strategy, and then finalize with problem solution and assessment. However, the chatbot was not effective for this task. In fact, ChatGPT was only effective in problem conceptualization of both numeric and non-numeric words. Nonetheless, the chatbot did not perform well in coming up with a problem strategy, especially for numeric

questions on titration. The chatbot erred by first calculating the number of moles using the formula moles = concentration × volume, but not first converting the volume from milliliters to liters, resulting in the calculation of millimoles instead of moles (Clark, 2023). This ties in with the critical thinking skills indicated above, as students will be required to note such inaccuracies after proper assessment of responses. Further, in the example given above (Tassotti, 2024), students were able to adjust their prompts through the 5S strategy: "set the scene, simplify your language, share feedback, structure the output, be specific" (p. 2467). Through this adjustment, they were able to see improvements in the ChatGPT response and were satisfied. This notable improvement in the prompts signals a problem-solving skill, which may be key in their chemistry teaching practice. In contrast, ChatGPT 4.0 seemed to be doing better by providing an organized dissection of the interpretation of the free energy diagrams, with positive implications for critical thinking and problem-solving skills (Alasadi & Baiz, 2024). However, ChatGPT 4.0 is not a free version and requires a subscription.

Information Analysis and Reasoning

Information analysis is a critical aspect of the learning process. With this skill, students are able to critically judge what is in front of them, be it boiling points of different substances, bond energies and enthalpy values, or any other chemistry information. But how do chemistry students interact with the information provided by GAI? Ruff et al. (2024) provided a rapport of the influence of ChatGPT in student learning. Biochemistry Lab and inorganic students were able to note how ChatGPT assisted in gramma revisions. However, they also noted its inability to generate formulaic structures, and consequently suggested how ChatGPT's generative ability could be improved. A crucial aspect that ties in with their critical thinking and reasoning skills. Students in the study by Tassoti (2024) used one of the components of the five S prompting strategy, 'share feedback, ' to interact with ChatGPT to attempt to modify a certain aspect of the answer generated. They used phrases such as "please include the same facts for sulfur" (p. 2479). This type of interaction, for example, is indicative of students' ability to analyze and make judgements on information given and, more importantly, make further refinements. Other students demonstrated this by comparing the responses of ChatGPT with other sources. For example, in the same study by Tassoti (2024), a student used the following as a follow-up prompt: "I have a source saying there are 6 oxidation levels, but you say there are 4. Which statement is legit?" (p. 2479). This is indicative of a student who is willing to interact with GAI instead of copying and pasting information as it comes.

Inquiry

As a hands-on subject, science is embedded in inquiry. If used effectively, inquiry can elicit students' critical thinking skills, creativity, and innovation, among other skills. In chemistry education, students may be required to undergo different stages of inquiry, which may result in a meaningful learning experience. Pedaste et al. (2015, p.54) identify "orientation, conceptualization, investigation, conclusions, and discussion" as the general phases of inquiry. However, the phases may differ based on students' levels. Of interest to this review is whether students can use GAI for inquiry-based learning in chemistry education. Kim (2025) undertook a study wherein college students were required to produce cyclohexene from cyclohexanol. The students employed AI chatbots, such as ChatGPT, Gemini, and Microsoft Copilot, to find greener alternatives to normal reagents. The activity yielded students' digital literacy while observing sustainability and science inquiry. Students were exposed to prompt engineering, where their interaction with the chatbot improved their engagement while shaping their critical thinking skills, as demonstrated above. Although the study primarily focused on abstract chemistry, the findings have positive implications for chemistry education.

Tutoring Ability of ChatGPT Free Version

Just like any other resource, students would engage with ChatGPT in the same way they interact with textbooks, websites, YouTube videos, and other available resources. But what is crucial for this study is how the chatbot assumes a tutoring role. Leon and Vidhani (2023) explored this arena by focusing on the patterns of prompting, response coherence, and trustworthiness of chatbots. ChatGPT provided different answers for the same multiple-choice question on the number of potassium core electrons when asked repeatedly several times. This highlights its lack of reliability in answering chemistry multiple-choice questions, which can have a detrimental impact on the learning process in cases where students are dependent on it for correct answers (Leon & Vidhani, 2023). Moreover, Tassoti (2024) highlighted how ChatGPT hallucinates and struggles to provide correct responses to high-order questions. This shortcoming may have detrimental effects on advanced students who always want to challenge themselves with more complex content.

Furthermore, in addition to being biased, hallucinations may lead students, especially those who are overly reliant on resources, to take in incorrect information (Feldman-Maggor et al., 2024), attracting misconceptions in the process. ChatGPT 4.0 was also found to demonstrate tutoring abilities by answering questions that require graphical analysis. For example, in the study by Alasadi and Baiz (2024), the chatbot was asked to analyze the free energy diagram, calculate the thermodynamic favorability of the reactant or product using the free energy change for both plots provided, and evaluate how the amount of activation energy affects the rate of reaction. Consequently, the chatbot was able to correctly distinguish between the two plots, indicating the first plot as that of that of a negative $\triangle G$ which is in contrast to the second plot with positive $\triangle G$. Moreover, the chatbot was able to indicate the limitations of the plots, for example: "these diagrams do not provide quantitative kinetic data but allow for the comparative understanding of the relative activation energies (E_a) and the potential rate of reaction," (p. 2718). This sense of transparency has positive consequences for the learning process, as students would be able to understand what each image represents and what it cannot. Nonetheless, the authors felt it could do better in highlighting possible misconceptions regarding the interpretation of plots and determination of results, which could be useful for first-time learning (Alasadi & Baiz, 2024). Despite these shortcomings, ChatGPT4.0 still demonstrated tutoring abilities similar to those of "personal tutors" through its ability to provide quick and sufficient clarity of different via a number of follow-up inquiries (Alasadi & Baiz, 2024, p. 2719). Its tutoring capabilities in analyzing challenging chemistry diagrams are indicative of knowledge advancement in chemistry concepts, which may assist in the student's conceptual understanding through engagement with the chatbot on scientific information and diagrams (Alasadi & Baiz, 2024).

The tutoring capability of ChatGPT 4.0 was again tested in organic chemistry diagrams of different qualities, where the first one was assisted through hand annotations, while the second and third diagrams were hand drawn, with the third adjusted to show double bonds. The chatbot was asked to solve the synthesis problem, provide details of how to approach similar problems, provide reagents required for each step, and make judgements on the order of reagents versus the expected outcomes (Alasadi & Baiz, 2024). ChatGPT 4.0 was able to provide clear step-by-step instructions for the synthesis of meta-bromo nitrobenzene and provided clear justification for the positioning of bromination, showing its tutoring efficiency by allowing students to mentally sequence organic reaction transformation, yielding a deeper interaction with the content (Alasadi & Baiz, 2024). Moreover, ChatGPT 4.0 was able to maintain its objective of explaining concepts despite a change in the quality or orientation of the image. However, its shortcoming was that it misinterpreted the images with "reduced resolution and inversion," incorrectly labelling them as bromo-nitrobenzene, while its correct identification is a meta-isomer (p. 2720). Furthermore, the chatbot's interpretation of the two drawn diagrams mentioned above was reported to provide different and incorrect solutions to the synthesis problem (Alasadi & Baiz, 2024). This would require a high level of objectivity from students, rather than subjectively relying on ChatGPT responses.

Teaching Chemistry Through GAI

It is also crucial to understand how GAI assists in chemistry teaching. However, teaching is complex and can be viewed from different perspectives. This review focuses on TPACK, visualization of chemistry, representation level of chemistry, textual aspects of chemistry, and problem-based learning (PBL).

TPACK

Feldman-Maggor et al. (2024) explored the type of knowledge required for teachers to efficiently use GAI. TPACK was used as a lens because of its nature of assessing teacher knowledge and pedagogies from a technological point of view. Even though authors such as Lorenzo and Romeike (2023) suggested an extension of TPACK with an AI component to yield DTPACK, disregarding the invalidity of TPACK alone to measure the use of GAI, the authors Feldman-Maggor still felt its relevance in their study. Nonetheless, they echoed the sentiments of Lorenzo and Romeike. Their intention was to understand how teachers CK and PCK influence their prompting ability and how they would ultimately assess and judge the quality of the prompts generated. This is a crucial aspect of teaching and has strong implications for science lesson planning, as teachers are expected not only to acquire different sources but also to sort what is important from what is not relevant to their lesson. For example, they used a teacher's dialogue with ChatGPT to assess whether the teacher would be able to use PCK to evaluate the response generated by the chatbot and ultimately follow up with content-specific prompts. The dialogue was about the teacher asking for the difference between ionic and molecular materials and ultimately seeking guidance on relevant teaching strategies. The teacher sought questions that would assist in eliminating students' misconceptions about NaOH and CH_3CH_2OH, regarding the role of "OH" as an ion and -OH as a

functional group. Although the chatbot did not go where the teacher wanted it to, the teacher modified and improved the prompts using their own PCK.

The chatbot provided a satisfactory answer, although it was still insufficient. The chatbot also hallucinated by providing incorrect references that could not be identified after the teacher's verification. Therefore, the teacher demonstrated sufficient PCK and CK through dialogue with ChatGPT prompts. The implication is that the 'T' part of TPACK aligns to the teachers' PCK. This underscores the importance of teachers' PCK when seeking clarity from ChatGPT in their lesson preparation, which would allow them to be objective and critical throughout the process. The absence of such a framework may lead to teaching lessons that are fully characterized by misconceptions.

Visualization of Chemistry

Visualization is important in chemistry because it has a bearing on students' understanding of chemistry through its different levels of representation: macro, sub-micro, and symbolic (Talanquer, 2011). I am interested in the image production capabilities of generative AI in chemistry education. A chatbot, such as ChatGPT 4.0, was tested by Alasadi and Baiz (2024) to assess its ability to engage with and interpret visual chemistry representations. In this context, the study explored the tutoring abilities of the chatbot through engagement with students, the chatbot's ability to provide understanding of complex figures, analysis of chemical structure, and ability to read tables and graphs, among others (Alasadi & Baiz, 2024). These aspects are key to the conceptual understanding of chemistry, which is embedded in visual representations. However, even though ChatGPT was able to interpret basic chemical structures, it struggled with complex diagram analysis when images were inverted or diagrams were of poor quality. Nonetheless, the paid version, ChatGPT 4.0, was efficient in reading hand-drawn tables and graphs, students' handwritten calculations, and spectral analysis with favorable effects on organic chemistry – reading molecular formulas and molecular structures where it showed variation in isomers and identifying relevant structures (Alasadi & Baiz, 2024). Nonetheless, the implications of these abilities for teaching chemistry still require further exploration. Nascimento Júnior et al. (2024) went a step further by exploring the text and image capabilities of seven GAI bots such as ChatGPT 3.5; ChatGPT 4.0, Google Bard, Bing Chat, Adobe Firefly, Leonardo AI, DALLE-E in chemistry education. The focus was on how these GAIs generate textual and image content of chemistry. In this subsection, I focus on the implications for the visual aspects of teaching chemistry. In this context, the GAI chatbots were exposed to text through prompting and were expected to generate images; they also used the same prompts to recognize images from organic chemistry reaction mechanisms, resonance, and energy diagrams. The image creation aspects were chemical bonds, Lewis structures, and atomic models. However, the authors noted that chatbots such as Copilot/BingChat and Google Bard/Gemini were unable to respond effectively to the prompts compared with ChatGPT 4.0, which appeared to be responding well to text despite some notable errors. ChatGPT 4.0 was able to generate correct chemical responses, accurately recognizing molecules and forms of arrows compared to Copilot/BingChat and Google Bard/Gemini. Nascimento Júnior et al. (2024) further indicated pointed that Google Bard, unlike its counterparts, failed to identify alkyl bromide as the reaction substrate. The GAI chatbots also seemed to struggle with generating covalent bond representations (Nascimento Júnior et al., 2024). The representation had a narrow nucleus closely attached to electrons that seemed to share orbits, which could create misconceptions. The GAI also uses the concept of electron transfer to represent ionic bonds rather than demonstrating ionic bonding itself. Nonetheless, the GAI performed well when the user changed from DALL-e to Python code while prompting for Lewis structures. Furthermore, Akaygun and Kilic (2025) conducted a study in which chemistry PSTs were expected to use chatbots to create visualizations that could be used in their lessons. PSTs were able to apply their prompting abilities influenced by their PCK, which they adapted to achieve desirable outcomes.

Representation Levels of Chemistry

Chemistry is represented as either or a combination of symbolic, macroscopic, and/or sub-microscopic levels. However, students still find it difficult to integrate the levels in chemistry learning (Murni et al., 2022). They confuse one level as the other level and vice versa (Nascimento Junior et al., 2024). The GAI prompts responses seem to be of less help because of their definitions of covalent, metallic, and ionic bonds. They would provide a definition for one type that focuses on one type of representation, while the definition for the other type of bond will focus on other representation types. This may create more confusion for students when one teaches them with unverified content from chatbots, where they may end up seeing the same representations in different definitions, while that is not supposed to be the case, building a myriad of misconceptions and misunderstandings in the process. Furthermore, Nascimento Júnior et al. (2024) found that GAI responses failed to show an understanding

of different representations on more than one occasion. For example, when illustrating metallic bonds, the GAI used macroscopic characteristics such as color and texture instead of representations from a sub-microscopic level.

Textual Aspects of Chemistry

The GAI seems to perform well in generating text from chemistry content. In their study, Nascimento Junior et al. (2024) applied prompting engineering to seek a definition for covalent bonding. Even though the chatbot provided a well-written definition, it contained some inaccuracies that may have led to students' misconceptions. Phrases such as "covalent bonds typically form between non-metal atoms, as these elements tend to gain, loose or share electrons…." (p. 3770) may lead to students mistaking ionic and covalent bonds (Nascimento Junior et al., 2024). Furthermore, all the chatbots under study seemed to "personify" atoms in their definition of covalent bonds by using words such as "atom seek" (Nascimento Junior et al., 2024, p. 3770). The language produced by the chatbots may, in such cases, lose its scientific nature, as chemistry involves non-living particles. Moreover, the chatbots such as Google Bard, Bing chat, ChatGPT 3.5, and ChatGPT 4.0 gave responses that were common with students' misconceptions on chemical bonding, irrespective of the level of complexity of the prompts generated – from beginner, intermediate, then advanced prompts (Nascimento Junior et al., 2024). Therefore, science teachers must be meticulous when using chatbots to prepare teaching content. Teachers would have to use different sources to cross-compare the responses from the chatbots.

Problem Based Learning (PBL)

As an active learning strategy, PBL is a highly encouraged teaching strategy in science. This is because it complements the inquiry nature of science. Furthermore, this aligns well with the aims of the science curriculum. For example, in the South African curriculum and policy statement (CAPS) for physical sciences, skills such as problem-solving and critical thinking are over-emphasized under the assessment taxonomy' and 'aims and principles' sections. Equally, it is prudent to understand how the use of GAI assists in the implementation of those skills in the science classroom. The literature shows that through the use of AI simulations, students are more interactive, and learning objectives are effectively attained (Ramos & Condotta, 2024). In this context, students were able to simulate the drying curves and particle size distributions using ChatGPT. Even though the study involved chemical engineering students, there are key highlights that the study makes which has implications for pedagogical aspects of chemistry education. For example, in one of the ChatGPT prompts, the participants indicated their intention to design a PBL task and asked the chatbot to suggest a topic (p. 3249). However, this was not the end of the story. The participants asked the chatbot to suggest extra 7 topics in order to accommodate all eight groups. In summary, not only did ChatGPT assist practitioners in the conceptualization part of the PBL task, but also shared pedagogical responsibilities, such as producing first drafts, generating rubrics, inquiry stages, preliminary simulation settings, and team aspects (Ramos & Condotta, 2024). The implication is that chemistry teachers would be able to use ChatGPT in the design and assessment of PBL tasks, whether in the acid and bases topic, organic compounds, or even separation of substances topics at the senior high school level. The ability of ChatGPT in the conceptualization point will come in handy in helping the teacher select and allocate different subtopics for each group. However, teachers would still have to apply their PCK and skills, using ChatGPT as a co-teacher rather than a complete replacement of their involvement. This step would be useful during the prompting stage.

Discussion

Learning Chemistry Through GAI

Learning is one of the most critical components of education, and scholars have always defined and redefined how different teaching strategies and platforms influence the learning process. This review highlighted how GAI influences 21st century skills in the learning of chemistry. Important skills, such as critical thinking skills, have been highlighted in this study, where students, through their expressions, are able to assess the information provided by the chatbots, challenge it with questions, summarize the information, and use it to solve complex subject matter (Guo & Lee, 2023). This skill is also demonstrated when students are able to notice the inaccuracy of the information a chatbot like ChatGPT would bring in the higher chemistry level (Tassotti, 2024). Inaccuracies, if not noted and addressed in the learning process, may lead to misconceptions, which is ineffective for the learning process. Therefore, critical thinking skills are key in the use of GAI for learning chemistry, as they would allow students to be objective in the information they receive. However, objectivity also requires students to have

information analysis and reasoning skills. Similar to critical thinking, students are expected to be critical of the information at their disposal, which in this case would be AI responses. Students noted that even though ChatGPT assisted in their revision of grammar, it could not produce formulaic structures (Ruff et al., 2024). Formulae are important in chemistry learning as they play a role in understanding reactions and mechanisms. It is also critical in the representation of chemistry, which is embedded in the particle and molecular nature. Through their information analysis and reasoning, students were able to suggest how the chatbot's generative ability could be improved, which has positive pedagogical implications for student teachers who would use that skill during their teaching practices. Chemistry students can demonstrate their information analysis and reasoning skills through the 5S prompting strategies (Tassoti, 2024). In the process, the students were able to modify their prompts through their interaction with ChatGPT while verifying the responses from ChatGPT through other sources. Interestingly, they used the information they received from other sources to create a dialogue with the chatbot, which is crucial for meaningful learning that encourages student-centeredness. Teachers should teach chemistry in a way that encourages the development of these crucial critical thinking, information analysis, and reasoning skills, which have an effect on learning on platforms other than the classroom. More importantly, environments which would require students to be self-directed. Moreover, chemistry learning should also take into cognizance that science is a hands-on subject that is embedded in inquiry and problem-solving part the key skills. Students can use GAI for the conceptualization of the inquiry, where they use AI chatbots to suggest alternative reagents (Kim. 2025). These have positive implications for primary and secondary school students who may use GAI to conceptualize their inquiry projects, where they are required to, for example, find recyclable material and do water quality projects, among others. However, they may require guidance from their teachers, depending on the classroom level. Nonetheless, students can integrate their problem-solving strategies during inquiry through GAI (Clark, 2023). However, a chatbot like ChatGPT has some limitations in that it is only able to conceptualize the problem but is unable to generate a problem strategy for topics such as titration when it involves numeric words (Clark, 2023). Therefore, students need to be sharp in their problem solving when using GAI to conceptualize chemistry inquiry. Learning through the GAI also has implications that may require chatbots to assume the tutor role, especially when students learn in their own space without the presence of their teachers. The chatbot, such as ChatGPT, was found to be inconsistent in its responses when asked the same question a couple of times (Leon & Vidhani, 2023) and gave incorrect answers for high-order questions, hallucinating in the process (Tassoti, 2024). This shortcoming negatively impacts chemistry learning, especially for students who challenge themselves with higher-order questions during the learning process. Moreover, hallucinations may add to the misconceptions that students already have regarding the topics they are learning. Nonetheless, ChatGPT 4.0 seems to perform better than its predecessor. For example, it can analyze graphical information from the free energy diagram, calculate the thermodynamic favorability, and measure the effect of activation energy on the rate of reaction (Alasadi & Baiz, 2024). Nonetheless, it could not highlight possible misconceptions in the data. The chatbot demonstrated its tutoring abilities through organic chemistry diagrams of different annotation qualities and asked to provide reagents, order, and outcomes, which it did not disappoint (Alasadi & Baiz, 2024). Nonetheless, it struggled with low-quality and low-resolution diagrams. These shortcomings would require students to have a combination of the 21st century skills mentioned above, tied with objectivity. Moreover, even though ChatGPT 4.0 provides a better tutoring ability than ChatGPT 3.5, it is a paid chatbot that requires users to subscribe, limiting access. In particular, students from more disadvantaged backgrounds are found.

Teaching Chemistry Through GAI

PCK and CK remain among the most important components of teaching. It also contributed to the skeleton of the TPACK framework, which focuses on how teachers use technology, such as AI, to teach. The key aspect of this review was how the use of GAI influenced chemistry teachers PCK and CK. The teacher would be able to demonstrate this through dialogue with the AI chatbot. Through the interaction, teachers were able to ask the chatbot for assistance in addressing the misconception about the -OH group for a base and the one for the alcohol, where consequently the responses were not what the teacher was anticipating. However, the teachers were able to adapt their prompts until the chatbot provided a satisfactory answer, even though it was insufficient. The teacher could not verify the references provided by the chatbots. This highlights the significance of sufficient PCK and CK when teachers use GAI for lesson planning, as their absence is likely to bring along misconceptions. GAI chatbots seem to be of less significance in addressing students' confusion regarding representation levels when learning chemistry. For instance, it would provide a definition of a certain type of bond (metallic, ionic, covalent bonds) using a particular representation while using another representation to provide a definition of another type of bond. This may create difficulties for teachers who use GAI for lesson planning, as they will have to explain different bonds in different representations, resulting in more confusion and misconceptions during their teaching of chemistry.

Another key aspect of teaching chemistry is the visualization of chemistry content, which has massive consequences for the understanding of chemistry, a subject that is highly characterized by the molecular and particle nature of matter. Furthermore, ChatGPT 4.0 demonstrated tutoring abilities when students engaged with it. The students were able to understand complex figures, analyze chemical structures, interpret tables, and analyze graphs. Nevertheless, the chatbot could not assist when the diagrams were disoriented or of poor quality. Chatbots also have the ability to generate images after being prompted through text. However, this was applicable to ChatGPT 4.0, despite minimal errors, while its counterparts, such as Copilot/BingChat and Google Bard/Gemini, seemed to underperform on that front. Furthermore, Google Bard/Gemini was found to be more wanting, as it struggled to identify a substrate in a reaction. Moreover, GAI struggles when it is expected to represent covalent bonding, where it misplaces and misrepresents the position and orientation of atomic particles. These misrepresentations have dire implications for the conceptual understanding of chemistry during the teaching process if teachers decide to rely on them without cross-checking with other sources. It is clear that, through its hallucinations, the GAI chatbots would use the macro level to represent sub-microscopic aspects such as ionic bonds. However, the use of Python coding instead of DALL-e during the prompting strategy enhanced the responses generated by the chatbot. For example, when the prompting involved the generation of Lewis structures. Furthermore, another positive aspect of GAI is its ability to produce text. This is expected from a large language model that is trained more on producing and refining text. Nonetheless, the chatbots seem to provide incorrect definitions, such as those of covalent bonding. The chatbot seems to demonstrate poor content knowledge, which may result in misconceptions if the teacher uses them without confirming from other sources. The other challenge observed was that the chatbots personified atoms, again with implications for misconceptions. The final aspect that is important for the teaching of chemistry is the approach the teacher uses to teach. PBL is a strategy that aligns well with the nature of science, and its use cannot be easily disregarded. Teachers may use chatbots to design PBL tasks. So does students during the conceptualisation, data collection or any step. They can also use the chatbot to suggest topics and even suggest additional topics or subtopics for their group members. Students can use ChatGPT, for example, to simulate drying curves and particle distributions. Nonetheless, the level of PCK should be at the top, which may require students to continuously consult with their teachers throughout the process.

Limitations

This study focused on the use of GAI in chemistry education and did not involve other science education subjects such as physics education and life sciences education. Moreover, the review focused on the teaching and learning aspects of chemistry, but could not go deeply into other aspects such as motivation, self-directed learning, self-regulation, achievements, or technology acceptance, which also have positive implications for the learning process. Furthermore, artificial intelligence is a very broad area that may include subareas such as machine learning, deep learning, and robotics. This study focused only on GAI. Above all, the papers included in the review did not focus on all types of GAI chatbots. Only ChatGPT 3.5, ChatGPT 4.0, Google Bard, Bing Chat, Adobe Firefly, Leonardo AI, DALLE-E were used in the reviewed studies. Furthermore, improved versions of the indicated chatbots currently exist, which did not exist when the reviewed studies were undertaken.

Conclusion and Future Recommendations

This paper provides an overview of the use of GAI in teaching and learning chemistry. The review noted that, with the use of the GAI, students were able to demonstrate critical thinking by posing questions, evaluating the responses from the chatbots, and consolidating the information through summaries. Through this, students are able to notice the inaccuracy of the information through their assessment of the responses from the chatbot. Furthermore, a chatbot like ChatGPT demonstrates limitations in the problem-solving stage unless the user utilizes a more adapted prompting strategy, such as the 5Cs. Furthermore, students can use AI chatbots for conceptualization and throughout the inquiry process in chemistry education. Nonetheless, students need to be more objective and wary when using ChatGPT for learning. For example, an AI chatbot such as ChatGPT was found wanting in playing the tutoring role, where it showed inconsistencies in the answers that it generated, while on the other hand, it struggled with high order questions –hallucinating in the process. This shortcoming may lead to misconceptions for students who would rely only on the chatbot without verifying information from other sources. Nonetheless, despite struggling with interpreting diagrams that have poor resolution and orientation, ChatGPT 4.0 seems to be a better tutor than its predecessor. Nonetheless, this version is not free and requires a subscription. This may be a challenge for students from disadvantaged backgrounds. Furthermore, the use of GAI has pedagogical implications for chemistry education. The teacher showed elements of TPACK during the prompting stage, such as CK and PCK. While prompting itself relies on the teachers' CK, the teacher would use their PCK during their dialogue with ChatGPT in asking for the difference between ionic and molecular materials

while also seeking relevant teaching strategies, including those that eliminate a misconception on the 'OH' from NaOH and CH3CH2OH. After noting the inability of the chatbot to provide satisfactory responses, the teacher modified their prompts. Another aspect of chemistry teaching is the representation and visualization of chemistry using GAI. The free version of ChatGPT performed well in interpreting diagrams but struggled with poor-quality images, whereas ChatGPT 4.0 was able to read tables and graphs, student handwritten texts, and spectra analysis.

Recommendations

However, future research should explore this further. When used effectively, teachers can integrate GAI into their teaching. Another point to note is that GAI chatbots produce the desired responses when users employ Python code compared to DALL-e. Furthermore, GAI chatbots struggle with representation levels, where they provide correct-looking responses with incorrect representation. This may require a more knowledgeable user. Moreover, although GAIs does well in producing texts, it personifies particles during definitions. This may lead students to view particles as living beings, creating more misconceptions. Lastly, the chatbots demonstrated the ability to aid in the conceptualization of PBL tasks by suggesting topics that can be assigned to different groups, producing drafts and rubrics, and structuring inquiry stages and preliminary simulations, among others. Nonetheless, what is crucial is that despite all these GAI abilities, the teacher would still have to demonstrate their objectivity, PCK, and CK throughout the prompting stages. Furthermore, teachers should not only rely on the responses from chatbots without verifying them with other sources. Moreover, teachers should not just use the responses from the chatbots as final drafts, but as initial drafts that would still be modified after verification with other sources and the user's own assessment.

Scientific Ethics Declaration

* The author declares that the scientific ethical and legal responsibility of this article published in JESEH journal belongs to the author.

Conflict of Interest

* The author declares that there is no conflict of interest.

Funding

* This research did not receive any funding.

Acknowledgements or Notes

* The author acknowledges the use of software such as Paperpal for language editing.

References

Akaygun, S., & Kilic, I. (2025). Generative Artificial Intelligence (GenAI) as the artist of chemistry visuals: Chemistry preservice teachers' reflections on visuals created by GenAI. *Journal of Chemical Education, 102 (7),* 2549–2564. https://doi.org/10.1021/acs.jchemed.4c00775

Alasadi, E.A., & Baiz, C.R. (2023). Generative AI in education and research: Opportunities, concerns, and solutions. *Journal of Chemical Education, 100* (8), *2965-2971.* https://doi.rg/10.1021/acs.jchemed.3c00323

Binkley ,M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In Griffin, P., McGraw, B., & Care, E. (Eds). *Assessment and teaching of the 21st century skills*. New York: Springer

Buthelezi, S. (2025, April 14). Ethical debate surrounding AI use by students in South African universities. *Independent Online (IOL).* https://iol.co.za/mercury/2025-04-14-ethical-debate-surrounding-ai-use-by-students-in-south-african-universities/

Clark, T. M. (2023). Investigating the use of an artificial intelligence chatbot with general chemistry exam questions. *Journal of Chemical Education*, *100*(5), 1905–1916. https://doi.org/10.1021/acs.jchemed.3c00027

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*, *32*(3), 444–452. https://doi.org/10.1007/s10956-023-10039-y

Elmas, R., Adiguzel-Ulutas, M., & Yılmaz, M. (2024). Examining ChatGPT's validity as a source for scientific inquiry and its misconceptions regarding cell energy metabolism. *Education and Information Technologies*, *29*(18), 25427–25456. https://doi.org/10.1007/s10639-024-12749-1

Fenta, A. A. (2025). A review on enhancing education with AI: exploring the potential of ChatGPT, Bard, and generative AI. *Discover Education*, *4*(1). https://doi.org/10.1007/s44217-025-00426-5

Guo, Y., & Lee, D. (2023). Leveraging ChatGPT for enhancing critical thinking skills. *Journal of Chemical Education*, *100*(12), 4876–4883. https://doi.org/10.1021/acs.jchemed.3c00505

Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, *2*(4). https://doi.org/10.1016/j.tbench.2023.100089

Haryani, E., Cobern, W. W., Pleasants, B. A-S. & Fetters, M. (2024). Exploring pedagogical strategies: Integrating 21st-century skills in science classrooms. *Journal of Education in Science, Enviroment and Health, 10*(2), 106-119. https://doi.org/10.55549/jeseh.697

Iyamuremye, A., Niyonzima, F. N., Mukiza, J., Twagilimana, I., Nyirahabimana, P., Nsengimana, T., Habiyaremye, J. D., Habimana, O., & Nsabayezu, E. (2024). Utilization of artificial intelligence and machine learning in chemistry education: a critical review. In *Discover Education, 3*(1). https://doi.org/10.1007/s44217-024-00197-5

Kim, J. (2025). Integrating Artificial Intelligence (AI) chatbots and green chemistry principles in the synthesis of cyclohexene. *Journal of Chemical Education, 102*, 3058-3064. https://doi.org/10.1021/acs.jchemed.5c00212

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education, 20, (1).* https://doi.org/10.1186/s41239-023-00426-1

Lawasi, M. C., Rohman, V. A., & Shoreamanis, M. (2024). The use of AI in improving student's critical thinking skills. *Proceedings Series on Social Sciences & Humanities*, *18*, 366–370. https://doi.org/10.30595/pssh.v18i.1279

Ling Jen, S., & Rahim Hj Salam, A. (2024). Using Google bard to improve secondary school students' essay writing performance. *Journal of Creative Practices in Language Learning and Teaching, 12* (1). https://doi.org/10.24191/cplt.v12i1.24999

Leon, A.J., & Vidhani, D. (2023). ChatGPT needs a chemistry tutor too. *Journal of Chemical Education,100 (10)*, 3859–3865. https://doi.org/10.1021/acs.jchemed.3c00288

Lorenz, U., Romeike, R. (2023). What is AI-PACK? – Outline of AI competencies for teaching with DPACK. In: Pellet, JP., Parriaux, G. (eds) Informatics in schools. beyond bits and bytes: Nurturing informatics intelligence in education. ISSEP 2023. *Lecture Notes in Computer Science*, *14296.* Springer, Cham. https://doi.org/10.1007/978-3-031-44900-0_2

Murni, H.P.,Azhar, M.,Ellizar, E., Nizar, U. K. & Guspatni, G. (2022). Three levels of chemical representation-integrated and structured inquiry-based reaction rate module: its effect on students' mental models. *Journal of Turkish Science Education, 19(3),* 758-772. https://doi.org/10.36681/tused.2022.148

Nascimento Júnior, W. J. D., Morais, C., & Girotto Júnior, G. (2024). Enhancing AI responses in chemistry: Integrating text generation, image creation, and image interpretation through different levels of prompts. *Journal of Chemical Education, 101*(9), 3767–3779. https://doi.org/10.1021/acs.jchemed.4c00230

Pabuçcu-Akış, A. (2024). Using innovative technology tools in organic chemistry education: bibliometric analysis. *Chemistry Teacher International, 7*, 141 - 156. https://doi.org/10.1515/cti-2024-0055

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review, 14,* 47–61. https://doi.org/10.1016/j.edurev.2015.02.003

Ramos, B., & Condotta, R. (2024). Enhancing learning and collaboration in a unit operations course: Using AI as a catalyst to create engaging problem-based learning scenarios. *Journal of Chemical Education, 101(8),* 3246–3254. https://doi.org/10.1021/acs.jchemed.4c00244

Ruff, E. F., Engen, M. A., Franz, J. L., Mauser, J. F., West, J. K., & Zemke, J. M. (2024). ChatGPT writing assistance and evaluation assignments across the chemistry curriculum. *Journal of Chemical Education, 101(6)*, 2483-2492. https://doi.org/10.1021/acs.jchemed.4c00248

Sedagat, S. (2025). Plagiarism and wrong content as potential challenges of using chatbots like ChatGPT in medical research. *Journal of Academic Ethics, 23*, 185-188. https://doi.org/10.1007/s10805-024-09533-8

Sykes, D. (2023). Exploring the use of ChatGPT in lesson planning: Possibilities, experiences, and limitations. *Literacies and Language Education: Research and Practice*, 39-51. English Language Institute, KUIS.

Talanquer, V. (2011). Macro, Submicro, and Symbolic: The many faces of the chemistry "triplet". *International Journal of Science Eduction, 33*:2, 179-195. https://doi.org/10.1080/09500690903386435

Tassoti, S. (2024). Assessment of students use of generative artificial intelligence: Prompting strategies and prompt engineering in chemistry education. *Journal of Chemical Education*, *101*(6), 2475–2482. https://doi.org/10.1021/acs.jchemed.4c00212

Valeri, F., Nilsson, P., & Cederqvist, A. M. (2025). Exploring students' experience of ChatGPT in STEM education. *Computer and Education: Artificial Intelligence,* https://doi.org/10.1016/j.caeai.2024.100360

Wangdi, T., Rigdel, S.K., Dawa, T., & Tshering, T. (2025). Using ChatGPT as an assessment tool in education: A systematic literature review of practices and limitations. *Issues in Educational Research, 35(2),* 818-837. http://www.iier.org.au/iier35/wangdi.pdf

Author(s) Information

Tebogo E. Nkanyani
University of Pretoria,
George Storrar Dr &, Leyds St, Groenkloof, Pretoria, 0027,
South Africa
Contact e-mail: tebogo.nkanyani@up.ac.za
ORCID iD: 0000-0003-4924-3882