

CURRENT DEVELOPMENTS AND THEMATIC REVIEWS IN EDUCATION SCIENCES

Editor

Prof. Mehmet Cem Odacioğlu, PhD.

Current Developments and Thematic Reviews in Education Sciences

Editor

Prof. Mehmet Cem Odacıoğlu, PhD.

Publisher
Platanus Publishing®

Editor in Chief
Prof. Mehmet Cem Odacıoğlu, PhD.

Cover & Interior Design
Platanus Publishing®

The First Edition
October, 2026

ISBN
978-625-91314-4-3

©copyright
All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, or any information storage or retrieval system, without permission from the publisher.

Platanus Publishing®
Address: Natoyolu Cad. Fahri Korutürk Mah. 157/B, 06480, Mamak,
Ankara, Turkey.

Phone: +90 312 390 1 118
web: www.platanuspublishing.com
e-mail: platanuskita@gmail.com



Platanus Publishing®

CONTENTS

CHAPTER 1	5
The Ripple Effect of Team Teaching: Enhancing Learning and Engagement Through Collaboration	
<i>Müge Kalıpcı</i>	
<i>Aslı Nur Önem</i>	
<i>Emrah Muyan</i>	
CHAPTER 2	41
The Effect of Computer-Assisted Language Learning on English Achievement: A Literature Review Based on Experimental Evidence	
<i>Doğuş Parçal</i>	
<i>Ferhat Bahçeci</i>	
CHAPTER 3	73
Artificial Intelligence in Mathematics Education: A Systematic Synthesis of Research on Affective Outcomes (2020–2025)	
<i>Hatice Nur Erbay</i>	
<i>Dilara Çakır</i>	
CHAPTER 4	95
Research Trends and Intellectual Structure of Chain-of-Thought Studies in Education: A Bibliometric Analysis	
<i>Jale Ipek</i>	
<i>Birol Can Sürmeli</i>	
<i>Melih Askar</i>	
CHAPTER 5	119
Turkish EFL Students’ Conceptualizations of Generative AI Use in Language Learning: A Metaphor Analysis	
<i>Esin Saylan</i>	
<i>Neslihan Keleş</i>	
<i>Elif Aydın Yazıcı</i>	
CHAPTER 6	133
Science Teachers' Curriculum Literacy Self-Efficacy Regarding The 2024 Türkiye Century Maarif Model Curriculum	
<i>Büşra Özcan</i>	
<i>Hasene Esra Yıldırım</i>	

CHAPTER 7.....157

**Mathematical Problem Solving in the Age of Artificial Intelligence: A
Proposed Cognitive Responsibility-Based Model**

Sezer KÖSE Biber

Mahir Biber

CHAPTER 1

The Ripple Effect of Team Teaching: Enhancing Learning and Engagement Through Collaboration

*Müge Kalıpcı¹
Aslı Nur Önem²
Emrah Muyan³*

¹ Asst. Prof. Dr., Erciyes University, 0000-0002-9932-5408

² Res. Asst., Erciyes University, 0009-0009-9253-5400

³ Lect., Erciyes University, 0000-0001-9651-0280

INTRODUCTION

With the rise of alternative approaches in the field of education, as noted by Cook and Friend (1996), scholars have recently focused on exploring the opportunities afforded by collaborative teaching models, especially those include the cooperation of multiple instructors sharing a single classroom. Therefore, it has been increasingly vital to understand these innovative methods, and their effects on both teachers and students.

According to Haag et al. (2022), co-teaching is a collaborative effort in which two or more instructors from the same discipline participate actively in all aspects of delivery, including course design, instruction, and assessment before, during, and after the course. However, Napoles (2025) describes it as a collaborative teaching practice that can occur either within a hierarchical structure or through a partnership model. Cook and Friend (1996) define co-teaching as “two or more professionals delivering substantive instruction to a diverse, or blended group of students in a single space” (p. 156). As their definition of co-teaching outlines, four key components are identified; (a) two educators, (b) delivery of meaningful instruction, (c) diverse groups of students, and (d) common settings. Barron and Friend (2024) revealed that for an effective co-teaching, the supportive elements are logistics including matters of time, professional roles, specific co-teaching approaches, and administrative support.

There are five distinct approaches in co-teaching as identified by Cook and Friend (1996): One Teach One Assist, Station teaching, Parallel teaching, Alternative teaching, and “Team teaching. Team teaching, among these models, is particularly noteworthy as a collaborative approach that promotes a richer educational experience for learners. Although often used interchangeably with co-teaching, Sanders-Smith et al. (2021) characterize team-teaching as a method that entails teachers to collaborate in the classroom as supportive and equal partners, creating a productive relationship and learning environment in the classroom that can often benefit students beyond what a single classroom teacher can provide. This model allows educators coordinate their teaching efforts, sharing responsibilities for lesson planning, instruction, and assessment. By facilitating the job division for lesson planning, instruction, and assessment, team teaching allows teachers to organize their teaching efforts collaboratively. The integration of different teaching styles and expertise not only provides enhanced learning experiences for different groups of students but also contributes to teachers’ own professional development (Canaran & Mirici, 2020; Hoa, 2022).

Team teaching has been effectively implemented across several disciplines, including special education, primary and secondary education, higher education, and foreign language instruction (Cook & Friend, 1996; Bayram & Canaran, 2018). Team teaching promotes a sense of community within the classroom, improving student satisfaction and creating a supportive and engaging learning

environment that increases active participation as students feel more comfortable expressing their ideas in a cooperative environment (Campbell et al., 2018; Tichenor & Tichenor, 2019; Yan et al., 2024;). Bacharach and Heck (2012) found the benefits of team teaching for the students as “increased opportunities to work in small groups, more individual attention, faster answer to their questions, quickly returns of papers and grades, better behavior by fellow students, eagerness to come to school and fewer classroom disruptions” (p.57). As evidenced by López-Hernández (2023), co-teaching may also effectively support formative and learning-oriented assessment.

In addition, team teaching can provide a platform for teachers to develop and experiment with new pedagogical strategies because they can receive immediate feedback from colleagues and promote a culture of continuous improvement and professional learning that ultimately benefits students’ educational experiences (Tichenor & Tichenor, 2019). By participating in collaborative practices, educators have the opportunity to share innovative ideas and approaches, which can lead to more effective teaching methodology and curriculum designs that meet a wider range of student needs.

While team teaching features with its numerous advantages, there might be pedagogical challenges in team teaching in relation to anxiety among students with the presence of two teachers in a single course about grading or the authority designation (Plank, 2023; Ling et al., 2023). Implementing team teaching requires great effort and time for its preparation during the whole semester (Hurd & Weilbacher, 2018). Nevertheless, the researchers strongly argue that the benefits outweigh the difficulties.

While co-teaching has been gaining significance in the literature, it remains underexplored and rarely practiced in many higher education contexts in Türkiye, where traditional, single-instructor models are dominant. Given the relevance and significance of the topic, and building on the existing literature, this study aims to further investigate the impact of collaborative teaching practices within higher education, exploring students’ perceptions and instructors’ reflections on practical application.

Statement of the Problem

Information Package of the Department where the study was conducted states that the Program Qualifications aim for graduates to possess the skills to work effectively both individually and in groups, develop and carry out projects, innovations, research & development activities related to their profession, and maintain lifelong learning abilities and competencies in their professional and personal domains. Yet, it was observed that many students in the department seem to lack essential skills in research and teamwork. These deficiencies were evident in student performance across research assignments, oral presentations, and written work submitted in various courses, and has been recognized as a

common issue by the faculty members. Although two interconnected courses are offered in the second year for students to learn the basics of conducting academic research, writing, and presentation skills, it is also observed that students often struggle with the intensive content and limited class hours of these courses, which are scheduled for two hours per week. Additionally, students frequently report struggles in making a smooth transition from the basic English skills courses they completed in their first year to more advanced academic research and group work tasks. The issue is further compounded by the fact that instructors also mention challenges in delivering intensive content in a limited session to a large group of students that exhibit low engagement. It was apparent that there was a need to make these courses more engaging for students to effectively teach essential academic research, writing and presentation skills, while at the same time setting an example of effective teamwork and collaboration. This research, therefore, recognized the necessity of implementing innovative teaching strategies to meet the diverse needs of students and address these issues. This study was mainly grounded in the principles of team teaching defined by Cook and Friend (1996), and was complemented by the integration of parallel and station teaching models to enhance pedagogical diversity and collaboration since there might be different needs in the classroom environment. Building upon this theoretical foundation, the present study aims to investigate the effects of team teaching in higher education on students' learning experiences from the perspective of students, and also to explore instructors' reflections on its implementation.

Purpose of the Study

The present study aims to investigate how students in tertiary education view the contribution of team teaching to their learning experiences. It is hypothesized that team teaching, as a collaborative approach that brings together two instructors, will stimulate students' interest and positively impact their learning by enabling them to benefit from diverse instructional experiences, multiple perspectives and teaching styles. Additionally, the study also seeks to gather opinions of instructors who are involved in team teaching to see how they perceive the effectiveness of such collaborative teaching practices and how they reflect on their experiences. It is anticipated that instructors will recognize the value of team teaching for their professional development, though they may also express specific challenges inherent in the approach.

METHODOLOGY

This qualitative research study that aims to examine the efficiency of team teaching approach in a course at a tertiary level originated from a research project on collaborative teaching conducted at a state university in Türkiye. This research was found ethically appropriate with the decision of This research was found ethically appropriate with the decision of Erciyes University Social and Human Sciences Ethics Committee, as per their decision dated 28/03/2023 with reference

to the application number 83. The researchers employed qualitative approach in order to obtain more detailed information on the research topic, and to gain insights into the experiences, feelings and ideas of the participants. The research was initiated by three researchers interested in innovative teaching strategies that move beyond traditional teaching to improve students’ learning experience and also contribute to the instructors’ professional development. The research team involved three faculty members from the Department of English Language and Literature (hereafter ELL), each with extensive experience in teaching English as a foreign language. Two of the researchers took active role in team teaching throughout the research period, while the third was responsible for coordination tasks and classroom observations during the team-taught lessons.

The present study, which lasted for one academic year, employed the team teaching method to introduce a relatively underexplored teaching approach within the department. Throughout the research period, the selected student group received both theoretical and practical training in academic research, writing, and presentation skills. Since students are expected to work collaboratively as teams in these courses and the courses themselves are demanding both for students and instructors in terms of time management, content delivery, and assessment, it was believed that team teaching could be an effective tool in addressing these challenges. Table 1 provides an overview of the instructors and their co-teaching involvement throughout the research.

Table 1
Overview of Instructors’ Co-Teaching Involvement

Instructor/Researcher identifier	Teaching activities in the study
R1	Parallel teaching Team teaching Station teaching
R2	Parallel teaching Team teaching Station teaching
R3	Classroom observation Coordination

Participants

The sample consisted of 110 second-year ELL students. To streamline data collection, a convenient sampling method was used to select participants who were readily accessible and suitable for the research. The courses selected for team teaching were *C1* for the first semester, and *C2* for the second semester (See Table 2). These two interconnected courses were offered both in daytime and evening classes, with each teaching program further divided into two sections:

Section A and Section B. Instructors, namely R1 and R2, were assigned to one class section in each of the daytime and evening sessions. The instructors synchronized their content and assessment before, during, and after the team teaching process to build unity.

Table 2
Courses included in the study

Course code	Course type	Undergraduate degree	Year	Semester
C1	Compulsory	ELL	2	First
C2	Compulsory	ELL	2	Second

Implementation

In line with the purpose of the study, the researchers planned to team teach two courses: *C1* and *C2* (See Table 3). The goal was to provide role models for students, contribute to the professional and personal development of the instructors, encourage mutual sharing of knowledge and experiences among instructors, make learning more engaging for students and positively impact the overall quality of teaching. The team teaching method was employed to introduce a relatively unknown teaching approach within the department. During one academic year consisting of two semesters, each lasting 14 weeks, students were given both theoretical and practical training in academic research, writing, and presentation skills in team-taught lessons.

At the beginning of the new academic year, the instructors developed a joint lesson plan aligned with the course content and learning outcomes. This plan was shared with students ahead of time in the first week of the semester. An informational meeting was held to brief students on the process and provide details about the courses to be team-taught. All participants were informed that data collected during this study would be used for academic purposes only and their identities would be kept confidential. Accordingly, Informed Consent Forms were obtained. Students were also notified that in these courses, they were expected to form their own project teams, select a specific topic, conduct a literature review in line with academic research standards, and write a project report that summarized their research, findings, and recommendations. Literature review constituted the first part of their project in *C1* which they were to build on the following semester in *C2*. In addition to the final report, all student groups were required to create informative and engaging presentations to be delivered in class. Last but not least, they were asked to prepare an additional end-product such as a video, infographic, website, game, or app. At the end of the second semester, each student team was assigned a specific week to showcase their project work.

To help facilitate the teamwork process among students and communicate timely any team-related issue with the instructors, students were reminded to assign a group leader within each team. It was emphasized that project topics should be tailored to students' own interests and abilities to enable meaningful learning experiences, enhance critical thinking, and promote collaboration. Students were provided with a list of ideas from which they could choose a topic or draw inspiration for their project. The team leaders were also to assess the work of themselves and their teammates by grading their performance on a scale from 1 (*little or weak effort*) to 5 (*excellent work*) in areas such as involvement in developing ideas, willingness to take on responsibility, cooperation with other team members, engagement and enthusiasm, roles of leadership and facilitator, and appreciation of others' contributions. They were assured that their feedback would be kept anonymous and used only by the researchers to evaluate the group dynamics.

The class was split into two sections, A and B, with each instructor (R1 and R2) assigned to one class section. Both groups were taught the same content on the same day of the week. In other words, R1 and R2 conducted parallel teaching and delivered the same course content on the same day to separate class sections: Daytime Group A and Daytime Group B, as well as Evening Group A and Evening Group B. During the weeks specified in the syllabus, the Daytime and Evening groups were merged within their respective time slots, allowing for the implementation of the team teaching method, where both instructors co-taught the lessons. There was a total of eight team teaching sessions conducted throughout the academic year, with four sessions in the first semester and another four in the second semester. Student work was also monitored, provided feedback and assessed collaboratively by the two instructors. Table 3 presents the schedule for team-taught classes.

Table 3
Team Teaching Schedule

Course code	Semester	TT No	Corresponding week in the study	Period	Program	Instructor
C1	First	1	3	Oct 2023-Jan 2024	Daytime Evening	R1, R2
		2	6	Oct 2023-Jan 2024	Daytime Evening	R1, R2
		3	8	Oct 2023-Jan 2024	Daytime Evening	R1, R2
		4	10	Oct 2023-Jan 2024	Daytime Evening	R1, R2
C2	Second	5	1	Feb-June 2024	Daytime Evening	R1, R2
		6	2	Feb-June 2024	Daytime Evening	R1, R2
		7	3	Feb-June 2024	Daytime	R1, R2

			Evening	R1, R2
8	4	Feb-June 2024	Daytime	R1, R2

The instructors held regular meetings to review and assess each team-taught session and to make sure that necessary modifications and improvements were made for the subsequent ones. Each session was designed at least one week in advance. The instructors coordinated tasks by dividing responsibilities such as turn-taking in lecturing, defining roles for hands-on activities, determining where to stand in class, ascertaining interaction styles and methods for observing the class. For instance, while one of the instructors was responsible for preparing the documents and extra materials to be delivered to the students, the other was responsible for preparing games and engaging activities to be used in class.

Based on positive feedback from the students by the end of the first semester, C2, in which team teaching was to be implemented in the second semester, was designed to be team-taught throughout the entire term. The first four weeks of C2 were team-taught by the two instructors (R1 and R2). Collaboration between the two instructors included co-preparation of lessons, team teaching, pre and post evaluations and shared student assessments. In Week 5 of the second semester, after the completion of the eighth team teaching session of the year, station teaching was introduced. This session, during which students were given feedback and evaluation on the progress of their final group project, was conducted in separate classrooms in two groups.

First, a survey link was shared with students to indicate their preferred instructor to work with that particular week. Depending on their research interests and specializations, R1 and R2 specified topics on which they could provide further assistance. The students were instructed to consult with their group members and, in relation to their project topic, have their group leaders respond to the survey on behalf of the team to express their preference. The students were expected to attend the session led by their chosen instructor with their respective group members to ask questions and get feedback on their progress. They were reminded that they could also visit other instructor’s session during any remaining time to receive additional feedback. Table 4 displays the assessment of student performance throughout the study.

Table 4
Evaluation of students' group projects

Semester	Task	Evaluation	Setting	Shared Assessment Grid
First	Literature Review	Single instructor	Out of class	Used
Second	Final project paper Oral presentation End-product	Team of instructors (R1, R2)	In-class	Used

Data Collection Tools

Based on an extensive literature review, the researchers developed Feedback Survey, Self-reflection Journal, and Focus Group Interview Form and used to collect data (See Table 5). For all data collection tools used in the study, expert opinions were solicited. The experts in the field were requested to evaluate each question in the Feedback Survey, Self-reflection Journal, and Focus Group Interview Form regarding their suitability for the study’s purposes. Revisions and further refinements were made in accordance with experts’ suggestions for editing. The tools were validated by experts in the field of Educational Sciences, Language Teaching and Linguistics.

Table 5
Data collection methods and focus areas

Feedback Survey	Self-reflection journal	Focus group interview
Structure of team teaching	Implementation challenges	Team teaching vs individual instruction
Teacher collaboration and interaction	Impacts of teaming on student learning	Identified positive impacts on learning
Instructional techniques and activities	Personal reflections on team teaching	Identified negative effects on learning
Learning environment and student engagement	Strengths of team teaching	Strengths of team teaching
	Weaknesses of team teaching	Weaknesses of team teaching
	Key takeaways	Preferred team size
	Improvements for future implementations	Student demand
		Reasons for teacher resistance

Feedback Survey

After selected team-taught classes, the students were asked to participate in an anonymous online survey designated to assess their satisfaction with the team-taught sessions and elicit detailed responses as to most valued aspects and possible drawbacks of the method from the perspective of the learners.

Focus Group Interview

Focus groups interviews are valuable tools in collecting qualitative data since they allow for in-depth understanding of an issue by eliciting multiple opinions, more natural responses and enabling discussion among participants (Kuckartz & Rädiker, 2019). For this reason, the students who participated in team-taught classes were sent an online link to decide whether they would be interested in being interviewed. 22 participants responded positively and joined the focus group interviews held in 3 consecutive sessions to fit the availability of the student groups based on their class hours. The purpose of the focus group interviews was to understand the students' perceptions of the team-taught lessons and their contributions to the overall learning experience. The questions aimed to gather feedback on the effectiveness of team-taught classes in comparison to traditional single-instructor classes. To gain insight into the strengths and potential drawbacks of team teaching model from the students' perspective, the participants were asked to offer suggestions for improving and broadening its implementation.

Self-Reflection Journal

Following the team teaching practice, each researcher wrote a self-reflection paper that aimed to gather detailed feedback on their experiences with the approach, focusing on both the practical aspects and the impact on teaching and learning. Reflective journals are effective resources that enable educators to keep a record of their practices, review and assess their implementation and document their reflections in concrete form (Agustin, 2019). The self-reflection journals were written by the instructors for all eight team-taught lessons. These reflections were intended to evaluate the effectiveness of the team teaching model, identify areas for improvement, and provide insights for enhancing this instructional approach for future implementations. For lessons in the last 5 weeks of the term during which students presented their final projects, which were also assessed and provided feedback by both instructors together in class, no self-reflection journals were kept. In post-meetings, however, the researchers discussed team-assessed student work and collaboratively identified the benefits and challenges of this model.

Data Analysis

Data were collected from October 2023 to June 2024. Each focus group interview lasted about 60 minutes and was audio-recorded for systematic analysis. Two researchers also took notes for further discussion while one interviewer asked the questions. Audio-recorded texts were transcribed verbatim and proofread carefully by the researchers. Participants' responses to the interview questions, together with all data collected from the feedback surveys and self-reflection journals were analyzed qualitatively using MAXQDA24. Thematic coding was applied to categorize key themes and identify patterns in

the collected data. This was accomplished through a systematic review of all responses, coding segments and grouping them into coherent themes.

FINDINGS

While comprehensive data were collected to discuss the implementation of team teaching approach, this paper specifically focuses on the impacts of team teaching on students’ learning experiences in higher education context, the participants’ perceptions of the model, and the instructors’ reflection on its implementation. In line with this, the following data are presented and discussed accordingly.

Focus Group Interview

The focus group interviews aimed to gain understanding of students’ perceptions of team-taught classes and their views on the efficiency of the approach for enhancing their learning experiences. Table 6 presents a thematic analysis of the data collected from these interviews.

Table 6
Thematic Analysis of Focus Group Interviews

Theme	Category	f	%
Team Teaching vs Individual Instruction	team teaching	69	100,0
Identified Positive Impacts on Learning	comprehension and engagement	10	90,9
	learning techniques	1	9,1
Identified Negative Effects on Learning	environmental distractions	3	60,0
	adjustment challenges	2	40,0
Strengths Of Team Teaching	increased engagement and curiosity	13	32,5
	diverse teaching approaches	11	27,5
	professional development and enhanced performance	7	17,5
	better comprehension	6	15,0
Weaknesses of Team Teaching	coordination issues	5	35,7
	distractions	5	35,7
	management challenges	4	28,6
	adjustment issues	2	14,3
Preferred Team Size	two instructors	15	93,8
	two+ instructors	1	6,3
Student Demand	team-taught classes in different courses	15	100,0
Reasons for Teacher Resistance	challenges in collaboration	3	50,0
	desire for sole authority	4	37,5
	personality conflicts	1	12,5

As highlighted in Table 6, the participants’ perspectives on team teaching outlined several aspects of the approach, including its strengths, weaknesses, and

impact on learning experiences. The key themes that emerged from the analysis are discussed in detail below.

Team Teaching vs Individual Instruction

When inquired about the ways in which this experience of team-taught lessons differs from their experience in other classes with a single instructor, and their preference regarding single-instructor classes or team-taught lessons, 100% of respondents expressed a preference for team-taught instruction.

“I prefer team-taught classes with two instructors. The reason is clear. Firstly, it captures my attention more in terms of focus. Secondly, I don't think there's any situation that would provide better education than having two teachers who are harmonious and integrated. Thirdly, it's an effective opportunity that saves time when teachers complement each other's shortcomings.” (P 12)

“I believe having two teachers in class is beneficial because it divides the energy into two parts. One teacher begins teaching and providing detailed explanations, and if the energy starts to decrease, the other teacher steps in. They also sometimes complement each other with examples. So, for me, having two teachers is better.” (P 20)

Such a unanimous preference for team-taught highlights the appreciation of the model among students for the advantages it brings into classroom, such as diversified explanations, increased attention and lively atmosphere.

Identified Positive Impacts on Learning

Upon being asked to reflect on their learning experiences in team-taught classes to discuss specific positive outcomes, the participants highlighted significant advantages of team teaching, particularly in enhancing comprehension and engagement. 90,9% noted that the combination of varied teaching methods employed by two instructors allows students to gain a better understanding of the topics. Additionally, it was mentioned that maintaining attention and interest for a longer time was facilitated in such settings since the collaborative nature of team teaching encouraged students to participate more actively. The following quotes selected from participant responses underscore these findings:

“Having two teachers working as a team in the classroom has enhanced my learning experience, indeed. Two different teaching styles and various examples from each teacher contribute significantly to seeing things from diverse

perspectives. This makes understanding the lesson more effective.” (P 10)

“For instance, each teacher employs a distinct teaching style. One teacher tends to provide more examples to ensure comprehension, whereas the other simplifies explanations for us to understand. I believe this greatly enhances our learning process. I've observed that comprehending the same lesson becomes more straightforward when both teach together.” (P 19)

“Because your teaching methods vary, we can learn something valuable from both of you simultaneously. Given that you have different ways of maintaining our attention, having two teachers is more advantageous.” (P 17)

“Explaining the topic feels like putting together a puzzle. There are no deficiencies in your explanations, but when there are aspects that we don't grasp, it feels like completing the whole picture. You complement each other like pieces of a puzzle, allowing us to understand the topic more comprehensively and securely. Overall, I find this approach very positive.” (P 12)

These student reflections underscore the value of team teaching in fostering a meaningful learning environment in which increased comprehension and active participation is encouraged by the collaborative and complementary instruction.

Identified Negative Effects on Learning

Although the participants acknowledged positive impacts of team teaching, they also discussed some drawbacks. Regarding any adverse effects on students' learning experience, environmental factors such as noise, classroom setting, and dominance of certain students seem to feature as the most commonly mentioned issues. Participants, however, also reported that these factors were not directly related to the team teaching method itself but were rather attributed to common problems observable in any class, regardless of the teaching approach. Another point that emerged was adjustment challenges experienced on the first days that stemmed from meeting a new teacher and adapting to a new teaching style.

“With the increase in student numbers when classes merged, I did find my focus occasionally disrupted by the noise. However, I want to emphasize that this wasn't due to any

actions on your part. Initially, I was somewhat hesitant because I was accustomed to Instructor A, which made it a bit challenging for me to adjust to Instructor B's teaching style at first." (P 13)

"Initially, for instance, because I didn't know you, I approached the first class with caution. I wondered how you would react if I made mistakes or if you would respond with something like, 'How is it possible that you do not know this?' There was some initial hesitation, but as I learned more and got to know the instructors better, I became more comfortable, and my participation increased. It turned out to be a positive experience for me." (P 3)

Based on these reflections, it can be asserted that students recognize the challenges experienced in team-taught classes as typical issues present in any lesson or classroom setting, rather than inherent obstacles that might undermine the effectiveness of the approach in practice.

Strengths Of Team Teaching

While the findings identified some negative effects, such as environmental distractions and adjustment challenges, the study highlighted significant benefits of team teaching. The participants shared their views on the strengths of employing such a collaborative approach in teaching. The responses, ranked by frequency, were grouped as follows: increased engagement and curiosity (32,5%), diverse teaching approaches (27,5%), professional development and enhanced performance (17,5%), and better comprehension (15,0%).

There appears to be strong recognition among participants that team-taught lessons could be beneficial in breaking the monotony of the lessons and increasing student engagement. The opportunity to learn from two different instructors increased curiosity among students as it fostered anticipation of what each educator would bring to the discussion.

"It contributes significantly to focusing because having only one instructor can become somewhat monotonous. Constantly hearing a single voice can cause the brain to tune it out over time. Our ability to maintain focus has decreased over time as we've grown up, influenced by technology. Therefore, this method greatly contributes to our ability to focus. Having a different voice is stimulating. Moreover, from a behavioral perspective, instructors serve as role models to us during group work. While educators often discuss how group work should be conducted, your demonstration of teamwork behavior allows us to observe firsthand how you contribute to each

other and communicate effectively. As a result, we learn how to collaborate as a team more effectively.” (P 13)

“I feel compelled to listen more attentively. I sense a greater responsibility to participate actively and prove myself. I believe this aspect strengthens the learning experience, at least for me. It encourages student engagement, in my opinion.” (P 13)

“Sometimes, as educators, you may overlook or forget certain aspects while teaching us. Another instructor stepping in to add supplementary sentences or points to your explanations is actually advantageous for you as professionals. And for us students, it's also a positive aspect.” (P 10)

Responses indicate that students also perceived the potential benefits of team teaching for instructors' professional development. 17,5 % of the respondents noted improved performance and increased comfort and energy among instructors in the presence of a colleague. Sharing the workload in class enables instructors to use their energy more effectively. Moreover, working alongside a colleague may enable instructors to learn from each other's practices and develop their teaching skills.

“When two educators are present in the same classroom, they tend to demonstrate a higher level of performance. Generally, people tend to perform better when they are being observed by someone in a similar role. For instance, a guitarist often plays exceptionally well when watching or being watched by another guitarist. It's somewhat analogous. This might keep you more alert. I believe it's something we commonly observe in classrooms with two educators.” (P 5)

“At the end of the lesson, instructors can discuss "What could I have changed?" and develop themselves. This can bring educators to different perspectives. So, it's a method that not only impacts students but also motivates educators towards improvement and encourages self-critique.” (P 12)

Diversity of perspectives provided by multiple instructors is recognized by participants as a strong means of deepening comprehension and critical thinking. Exposure to varied teaching styles enables students to engage with course content in multiple ways. A chance to have alternative explanations from another instructor appears to be valued by students since it is essential in aiding learners who struggle with understanding content from the perspective of a single instructor.

Weaknesses of Team Teaching

While team teaching received wide appreciation, the participants addressed certain weaknesses the approach may present. Upon being inquired about the potential drawbacks this collaborative teaching approach may bring, 35,7% of the respondents identified coordination issues as a notable challenge. Students acknowledged that the effectiveness of the lesson would be hindered if the instructors failed in collaboration while forming a team. Lack of communication and great differences in terms of personality may cause challenges in collaboration. In a similar vein, 35,7% of the respondents perceived noise in crowded classrooms, information overload caused by the presence of two instructors, and physical positioning of each instructor in the classroom to be possible issued in negatively impacting students' focus.

"I think that if instructors who choose to team-teach are incompatible and cannot find common ground, it could reduce the effectiveness of the lesson. This isn't the case for you because you two work well together, making our lessons with two instructors go smoothly. However, if there were two individuals who didn't understand each other, constantly interrupted, or tried to dominate each other, it wouldn't be as successful. It could create problems."
(P 1)

"As long as there is integration and cohesive collaboration between educators, there are no inherent weaknesses in a method involving two different instructors. If there are two individuals who are incompatible, and one interrupts frequently while the other speaks, it can confuse students about whom to focus on. Moreover, even physical positioning within the classroom holds significance. When both instructors are centrally located simultaneously, students might question whom to listen to or observe. However, when one instructor speaks, the other can step aside or move to the back, minimizing these distractions. Other aspects that I consider are purely personal and student-related factors, such as assertiveness, seeking opportunities to speak, discomfort in crowded environments, or complaints about noise, which I don't attribute to the educators themselves." (P 4)

Additionally, 28,6% of the respondents underscored management challenges such as difficulties in effectively allocating time during classes in which two instructors would have a say, confusion that could be caused by two diverse teaching philosophies unless instructors manage to create a cohesive approach, and frequency of the implementation. Fewer participants (14,3%) mentioned possible difficulties some students may initially experience in adjusting to a new teaching approach.

Preferred Team Size

Students were also asked to share their views on the ideal team size for team teaching implementations in class. 93,8% of the respondents agreed that two instructors were sufficient for an effective lesson. All participants expressed interest in seeing the implementation of team teaching in many other courses and with different instructors. They unanimously demanded an increase in the number of team-taught classes within the department.

Reasons For Teacher Resistance

Even though students articulated a strong demand for the expansion of collaborative teaching approaches, they also acknowledged that teacher reluctance could be a significant barrier. The respondents primarily attributed teacher resistance to team teaching to several factors such as challenges in collaboration, a desire for sole authority, and personality conflicts. While discussing possible reasons behind instructors' hesitance to participate in team teaching implementations, the respondents discussed the challenges of finding a suitable teaching partner and establishing an effective collaboration. They recognized that personality conflicts may challenge instructors to cooperate with other colleagues, which in turn can hinder effective teamwork. Not only is the practice demanding in itself, but instructors may also wish not to share the authority with another instructor, feeling more comfortable in taking control of their lessons independently.

Feedback Survey

Many research studies worldwide use feedback from students to reflect on teaching practices and improve the quality of education (Mandouit, 2018), and it has become increasingly common in higher education settings (Floden, 2016; Harvey, 2003). In this study, student feedback surveys aimed to explore and analyze students' perceptions of team-taught classes. Table 7 presents the thematic analysis of student surveys, highlighting key areas such as the structure of team teaching, teacher collaboration and interaction, instructional techniques and activities, and learning environment and student engagement. Selected quotes from participants are presented without identifiers (e.g., P1, P2) as the survey was conducted anonymously.

Table 7

Thematic Analysis of Student Surveys

Theme	Category	f	%
Structure of Team Teaching	having two teachers in class	71	28,0
	team teaching	10	3,9
	merging two classes	7	2,8
Teacher Collaboration and Interaction	collaboration of teachers	13	5,1
	teaching styles	58	22,8
	perspectives	37	14,6
	attitudes	8	3,1
Instructional Techniques and Activities	presentations	11	4,3
	in-class activities	11	4,3
	feedback	7	2,8
Learning Environment and Student Engagement	effective learning	11	4,3
	lesson flow	3	1,2
	low affective filter	2	0,8
	peer support	3	1,2

When asked about what they enjoyed most about team teaching, the respondents mainly highlighted the structure of team teaching (34,6%) and teacher collaboration and interaction (45,7%). Students appreciated having two teachers present in the classroom simultaneously, which provided a better and more effective learning experience.

Structure of Team Teaching

The present study offers insights into why students perceive team teaching favourably. In a traditional classroom, a single teacher typically assumes the role of the sole authority in their subject area. For many students, having two teachers in the same classroom stands out as a new and engaging experience. The following participant statements that were collected anonymously through surveys after team-taught sessions prove students' appreciation for this approach:

"I enjoyed having two instructors teach at the same time and two classes combined."

"I liked the idea of having two instructors teaching together."

"I think I enjoyed most two instructors teaching as a team."

"It is good to listen to the topics from two instructors together."

"Having two instructors explaining the subject is what I enjoyed the most."

Teacher Collaboration and Interaction

In this study, 45,7% of participants indicated that the team teaching experience enriched their learning through exposure to diverse perspectives and collaborative instruction. When asked what they enjoyed most about team teaching, students frequently mentioned ‘different perspectives’, highlighting the value they placed on this aspect. By offering multiple viewpoints, team teaching fosters a more meaningful and effective learning experience. Participants also appreciated the collaborative efforts of the instructors, as reflected in their comments:

“Taking classes with two instructors and listening to the topic from different perspectives was beneficial.”

“I was interested in the presentation of the lectures together; they offered different perspectives.”

“Two experienced teachers presented the main points from two different perspectives.”

“Two teachers working together and giving various examples was highly rewarding.”

“Teachers collaborate very well with each other. Different perspectives were good.”

“It was excellent to see different perspectives on different subjects.”

These quotes reveal that students not only appreciated the variety of perspectives but also acknowledged the positive impact of instructor collaboration, which was crucial for their learning experience.

Instructional Techniques and Activities

The study found out that students highly enjoyed the techniques and activities used by their instructors, accounting for 11,4% of the responses, as well as the classroom environment cultivated through team teaching, which comprised 7,5% of the feedback. The implementation of a team teaching approach enables two instructors to share their unique individual strengths, so increasing the overall effectiveness of instruction and facilitating student learning. Participants emphasized the benefits of different teaching styles employed during team teaching. Different teaching styles were helpful in terms of offering preferences to students. For instance, one respondent stated,

“It was more beneficial for me to understand lessons through various examples because the questions in my mind were resolved through the variety and abundance of these examples.”

Similarly, another participant noted,

“Being exposed to different teaching methods is an asset.”

It is clear that diversity in instructional strategies boosted student engagement. Moreover, collaboration of the teachers was enjoyed as it provided diversity in terms of presentation techniques and made the lesson more engaging. Statements such as the ones below prove how different instructional materials and presentation styles can boost student interest and make learning easier.

“I enjoyed the slide show the teachers prepared; it was visually good.”

“Different presentations delivered by the instructors catch my attention.”

The presence of two instructors also provided multiple perspectives on the subject matter, which students found beneficial. For example, the following quote from a student underlines how various perspectives can ease understanding and foster critical thinking:

“I was interested in the presentation of the lectures together; they offered different perspectives. This may not always be the case in conventional classes taught by a single instructor.”

Additionally, the synergy between the instructors was evident in the flow of lessons and the supportive classroom environment. Comments below selected from participant responses reflect the positive impact of team teaching on creating an inviting and effective learning atmosphere.

“Both instructors’ teaching techniques were very beneficial and good. I was satisfied with the lesson.”

“I enjoyed seeing two instructors’ harmonious and energetic teaching styles.”

In summary, the findings indicate that team teaching is crucial as it brings different teaching styles together and makes the classroom environment more vibrant. By integrating diverse teaching styles and fostering a collaborative learning space, team teaching improves instruction effectiveness and increases

student engagement and satisfaction. These results show the value of team teaching as a strategy for creating a dynamic and supportive educational experience.

Self-Reflection Journals

Self-reflection journals kept by the instructors after each team-taught lesson aimed to gain in-depth insight into their experiences with collaborative teaching strategies, and to capture their perceptions of teamwork, effectiveness of the approach, specific challenges faced, and the impacts of team teaching on students’ learning experiences. Table 8 presents the thematic analysis of the self-reflection journals, highlighting the categories that emerged.

Table 8
Thematic Analysis of Self-Reflection Journals

Theme	Category	f	%
Implementation Challenges	Lesson planning and time management	12	48
		5	20
	Collaboration and teacher interaction	5	20
	Teaching environment and student-related issues	3	12
	Adapting a new approach		
Impacts of Team teaching on Students’ Learning	Enhanced understanding and learning	25	58,15
	Positive classroom dynamics and engagement	12	27,90
		6	13,95
	Improved interaction and collaboration		
Personal Reflections on Team Teaching	Professional development	17	53,13
	Reflective teaching practices	13	40,63
	Exploring teacher dynamics	2	6,25
Strengths of Team Teaching	Effective learning and classroom dynamics	33	55,0
	Teacher collaboration and development	19	31,7
	Reduced workload and role modelling	8	13,3
Weaknesses of Team Teaching	Challenges in time and workload	13	43,3
	Needs for institutional and observational support	9	30,0
	Pedagogical and implementation challenges	8	26,7
Key Takeaways	Effective team collaboration	15	51,72
	Professional development	9	31,03
	Planning	3	10,34
	Fostering student engagement	2	6,90
Improvements for Future Implementations	Efficient lesson planning	6	66,67
	Observation and feedback	2	22,22
	Diversification	1	11,11

Implementation Challenges

Although team teaching is widely regarded as beneficial, it is not easily implemented in many educational settings. Firstly, as an innovative approach, it is uncommon in traditional teaching environments for both educators and students. Successful team teaching requires close collaboration between instructors at every stage of the process. In reflective journals following each practice, instructors reported challenges, particularly during the planning phase and in-class implementation. Key concerns are discussed below.

To ensure a successful team-taught class, lesson planning is essential. Effective lesson planning should extend beyond merely dividing tasks and encompass various aspects, such as selecting teaching materials, aligning teaching styles, and organizing content. After engaging in team teaching, instructors emphasized these points in their reflections. Here are some quotes from their journals:

“Both instructors need to dedicate ample time to planning, ensuring a smooth lesson flow and considering their respective interests and expertise.” (R2)

“The preparation stage required detailed and careful planning. Classes typically proceed in parallel, but differences in teaching styles necessitate thorough coordination and readiness for unexpected situations.” (R2)

When more than one instructor is present in the same classroom, time management issues may arise due to the increased explanations and examples provided by each. With a single instructor, decisions about the lesson’s pace or skipping certain parts can be made more freely. However, in a team-taught setting, one may hesitate to interrupt or override the other. Instructors highlighted these challenges in their reflections. Below are some quotes from their journals:

“I can say that there is a need for planning for effective use of class time. We ran out of time for one great group activity we had planned for the end of the lesson.” (R1)

“Despite detailed planning, we slightly exceeded the lesson duration. Having two instructors enriched discussions but extended the allocated time.” (R2)

Classroom size and students’ misbehaviour may not be directly related to team teaching; however, they emerged as notable concerns in self-reflection journals.

“The most fundamental issue is the overcrowded class size (for daytime classes, I mean). If we fail to maintain effective classroom management, noise within the classroom could make it difficult for the instructor's voice to reach students sitting at the back.” (R1)

These insights underline the need for meticulous planning, adaptable time management, and strategies to address classroom dynamics. The success of the approach would necessarily depend on carefully attended preparations to effectively stimulate collaboration between instructors, organize lesson planning and regulate teacher autonomy in class.

Impacts of Team teaching on Students' Learning

Instructors highlighted several positive effects of team teaching on students' learning. It is not surprising that in a team-taught class, the presence of two instructors allows for the inclusion of more examples and detailed explanations, which significantly facilitate student learning. Here are some quotes from the journals:

“I felt that my partner and I interacted well and delivered the lesson more effectively. We supported each other by providing alternative examples to aid students' understanding of the topic. I believe it was quite valuable for students to be able to get the opinions of two instructors and to consult on points that were unclear to them. They had the chance to address potential issues in their research process and make revisions where necessary.” (R1)

The instructors' remarks revealed that team teaching has a notable impact on classroom dynamics, fostering student engagement and capturing students' attention.

“It has been observed that students approach the lesson with curiosity due to not being accustomed to seeing two instructors in the classroom working together. It can be said that the applied method generally increased students' interest in the lesson. I observe their engagement remains active for a longer period when there is intermittent interaction between instructors on stage, highlighting interactive discussions and different perspectives, which facilitates students' understanding of the topic” (R1)

One of the strengths of team teaching, as noted in reflections, is fostering interaction and collaboration among teachers and students which creates a more dynamic and engaging learning environment.

“Peer-learning opportunities are higher compared to traditional classroom settings.” (R1)

“I guess teaming positively affected student learning in two ways. Firstly, it provided two sources of information for students. There were two professionals who were knowledgeable about the topic and available to assist them when they needed help.” (R2)

These reflections emphasize the value of team teaching in fostering a dynamic, interactive, and effective learning experience for students.

Personal Reflections on Team Teaching

When asked how team teaching influenced their teaching, instructors highlighted its profound impact on reflective teaching practices, professional development, and teacher dynamics. They emphasized that team teaching is one of the most effective professional development methods, as it enables teachers to observe each other without judgment, learn from diverse teaching styles, and apply new strategies to their own instruction. Team teaching encourages instructors to compare their teaching approaches, identify strengths and areas for improvement, and incorporate effective strategies observed in their partners.

“When I compared my teaching with my partner's, I saw both our common points and aspects where we differ, which I liked. I observed that my partner elaborated on the lesson more extensively, with more repetitions and examples. I can incorporate more question-answer and review exercises in my own lessons to gauge whether the subject is understood.” (R1)

“It was a great opportunity to observe another colleague while teaching the same subject. I focused on our differences and similarities. When I noticed the similarities, it boosted my confidence as I felt I was doing the right thing in the classroom. On the other hand, when I found out some differences, I was happy to make my teaching richer.” (R2)

By observing their colleagues, instructors gain fresh perspectives and insights that enrich their instructional techniques. Sharing responsibilities also eases the workload and creates opportunities for professional growth.

“Particularly in the crowded daytime classroom, we were able to share the load in terms of observing students' work and providing immediate answers to their questions. In this regard, we had a very successful class in terms of classroom management. I felt more comfortable and less fatigued in the presence of my partners.” (R1)

“Seeing the same lesson taught by a different colleague constantly brings new insights to my mind. Each instructor brings their own experiences and beliefs to the teaching-learning process. Having a second teacher in the classroom allows me to observe and learn from their methods, which I can then apply to other classes.” (R2)

Team teaching fosters collaboration and mutual support, which enhances not only the learning environment but also instructors’ job satisfaction and enthusiasm for teaching.

“Working as a team not only enhanced my enjoyment of the task and the class but also allowed me to derive great pleasure, especially when teaching with my partner in the evening class. Student performance and participation exceeded our expectations.” (R1)

These reflections highlight that team teaching is not only a means of improving classroom instruction but also a powerful tool for teacher development and collaboration.

Strengths of Team Teaching

The strengths of team teaching align closely with its key impacts, including effective learning, positive classroom dynamics, teacher collaboration and development, and reduced workload. While the workload in the classroom is minimized due to shared responsibilities, the planning phase poses challenges, requiring detailed coordination before each lesson.

Team teaching enhances the classroom environment by introducing diverse teaching styles and fostering a dynamic atmosphere.

“I believe team teaching adds a distinct dynamic to the class. I am of the opinion that the energy of the lesson increases and students are interested in learning from both educators. It allows students to experience different teaching styles and perspectives.” (R1)

“Observing two different teaching styles helps keep students alert and facilitates their engagement in following the lesson. Our lesson is inherently challenging in terms of content, and students encounter a variety of new terms in each class. Both my partner and I use numerous examples to ensure students understand the topic and remain active during the lesson. While one of us explains the topic, the other provides additional examples to enhance clarity. This approach inevitably benefits students.” (R2)

Team teaching provides opportunities for professional growth through peer observation, feedback, and collaborative teaching strategies.

“One of its greatest benefits, in my opinion, is the opportunity for educators to observe a colleague's teaching and gain insights into alternative approaches. Additionally, receiving feedback on one's own teaching from another educator who observes the lesson is invaluable for professional development. I also believe that the coordinated efforts of two educators, demonstrating supportive, respectful, and confident attitudes towards each other, set a positive example for students during group work.” (R1)

Sharing responsibilities helps manage large classrooms effectively while modelling teamwork for students. However, this benefit contrasts with the increased workload during the planning phase.

“Sharing the workload collaboratively in the presence of large groups helps maintain high energy levels and ensures the efficient use of class time.” (R1)

“Teaching is demanding, but sharing the responsibility thanks to team teaching is amazing. The instructor feels that when they fall short in one part of teaching, the other colleague will be there to help.” (R2)

These reflections highlight the balance between the benefits of shared responsibilities and the challenges of meticulous planning. Team teaching not only enhances the learning environment but also serves as a powerful tool for teacher growth and collaboration.

Weaknesses of Team Teaching

Although team teaching offers numerous advantages, it also presents certain challenges. These can be categorized into time and workload management, needs for institutional and observational support, and pedagogical and implementation challenges. Effective time management during lessons and the increased workload during the planning phase are commonly cited issues.

“There is a need for planning for effective use of class time. I believe one of the biggest weaknesses is the timing issue we experienced.” (R1)

“One of the fundamental challenges of team teaching is the additional workload it brings. Allocating time for the planning stage could sometimes be demanding. The additional workload brought by team teaching planning can be exhausting within our current workload and sometimes creates stress.” (R1)

“One should never assume that two teachers divide the lesson materials, and while one of the teachers covers some pages, the other covers the rest. It is not the case. The team should take into consideration all the factors which may affect the flow of the lesson. First and foremost, both teachers should be ready to spend a lot of time to design a very simple lesson.” (R2)

Researchers emphasized the importance of institutional backing to address the logistical demands of team teaching.

“It is crucial to also receive institutional support to ensure that the necessary conditions are met.” (R1)

Implementing team teaching effectively requires careful planning to address in-class dynamics, such as the positioning and movement of educators to avoid distractions.

“I have seen the importance of where and how a second educator positions themselves during the lesson. One educator moving around the classroom during the lecture can be distracting. Positioning at a point where both students and educators can be seen but without distracting either's attention could resolve this issue.” (R1)

These challenges highlight the need for strategic planning, institutional support, and mindful classroom practices to ensure the success of team teaching while minimizing its weaknesses.

Key Takeaways

When reflecting on their team teaching experiences, teachers emphasized the salient role of collaboration. By its nature, team teaching requires at least two educators to work closely together, making effective collaboration essential for success. Researchers highlighted planning as the most critical stage of team teaching. This step involves coordinating lesson flow, scheduling meetings, and preparing materials; tasks that demand a shared commitment and teamwork. Beyond collaboration, team teaching was also noted as a powerful tool for professional development. Researchers found that working together not only enhanced their teaching skills but also positively impacted student engagement. By leveraging their combined expertise, educators discovered new ways to captivate students, even with challenging or less appealing content.

“Help and support from another teacher is invaluable as we cannot save the world alone. I have learned how valuable teacher collaboration is. Despite teaching for many years, I have seen that learning never stops”. (R2)

Instructors have acknowledged the importance of careful planning, collaboration and shared responsibilities in successfully implementing team teaching and maximizing its benefits for students' learning.

“Someone who dares to implement team teaching should know that they need to work hand in hand. I have observed how sharing responsibilities can greatly relieve a teacher.” (R2)

They also highlighted the positive effects of team teaching on both students' learning and their professional development.

“What I learned from this experience is that students are not necessarily resistant to innovation and can enjoy different techniques, even if the lesson content is challenging or doesn't directly appeal to their interests. Leading such a practice, despite all its challenges, is valuable not only for the advantages it brings to students but also for providing reflective feedback for us educators about our teaching.” (R1)

Improvements for Future Implementations

Team teaching is an innovative approach, and its implementation at a state university in Türkiye where the research took place marked a first-time experience for both teachers and students. While this practice is not widely used in higher education, its introduction aimed to make teaching more effective and learning more accessible. The initiative demonstrated promising results, but there are areas for further refinement to enhance its impact on both student learning and teacher collaboration.

“If I were to utilize this practice again, I would consider shortening some parts of the lesson content. I would discuss with my partner in a pre-implementation meeting about how long the lecture would last and to what extent we would interact with each other at various points.” (R1)

“We should decide in advance how the lesson will be conducted and where educators will be positioned within the classroom.” (R1)

It is obvious that clear and structured planning, including time allocation and teacher roles, is crucial for the smooth delivery of team-taught lessons.

“The course content is very intense, and the topics taught each week are ones that students see for the first time. I would like to see this practice implemented in a course with less and simpler content.” (R2)

Adapting teaching methods to match course requirements and student needs is essential for keeping lessons engaging and relevant. This feedback suggests exploring the feasibility and effectiveness of team teaching in courses with varying content complexity to identify its broader applicability.

“We try to diversify our planning and activities based on the course content and priorities.” (R1)

The reflections underscore the importance of careful planning, adaptability, and ongoing collaboration to maximize the benefits of team teaching. As this innovative approach gains traction, its refinement can pave the way for more dynamic and impactful teaching practices in higher education. Instructors who actively participated in this practice expressed great satisfaction with team teaching and its positive impact on both student learning and their own professional development. They unanimously agreed that they would recommend this approach to other educators who are interested in fostering close collaboration with colleagues and exploring innovative methods to broaden their perspectives.

DISCUSSION & CONCLUSION

Based on the findings of the study, the following observations were made: The students displayed enthusiasm toward co-teaching activities and requested more team-taught classes to be organized and delivered in the department. They acknowledged that learning in lessons taught collaboratively by two instructors had a considerable impact on increasing their motivation and interest in participation, which in turn improved their comprehension. They reported being satisfied with the team-taught classes since they could receive more attention in the presence of more than one instructor. This aligns with the literature highlighting more teacher attention as one of the principal benefits of the approach (Villa et al., 2024).

Another positive result obtained from the findings is that team teaching sparked a great interest among students, who were all used to traditional single-instructor education. The students enjoyed getting feedback from two instructors at a time, which enabled them to receive additional support and provided them

with diverse perspectives and teaching styles. This finding aligns with the work of Laughlin et al. (2011), who emphasize the educational benefits of varied perspectives in team teaching. The participants were of the idea that having two instructors teach as a team in the class was beneficial and adequate. Concerning whether more co-instructors would be advantageous, while few participants were curious to know how it would work with three instructors, they mostly believed that anything more than two might be challenging and confusing. This suggests that while two instructors are generally favourably received, for a more cohesive learning environment, the number of instructors to team teach a class should be taken into account depending on the context and needs of students.

The study revealed promising results for students in increased support and learning gains. The findings suggest that team teaching is an engaging and effective strategy that offers students diverse perspectives, increases their engagement, and promotes comprehensive learning, thereby positively influencing their overall learning experiences. This outcome corroborates the conclusions of Zach and Avugos (2024), who observed high levels of engagement and satisfaction among students as well as a boost in their confidence and interactions in class. The participants recognized the value of co-teaching approaches in bringing dynamism and variety to the learning environment, thereby enhancing the overall quality of the education. Similar to the findings of Crow and Smith (2006), this study further confirms that students perceive team-taught classes to be beneficial and effective. As an innovative teaching approach that enables practitioners to pool their knowledge and expertise, team teaching provides learners with broad opportunities for their various learning styles to be catered to and needs to be met.

Furthermore, self-reflection journals underscore the benefits of team teaching for professional development of instructors. Collaborative-teaching practices equip instructors with opportunities to leverage their strengths by bringing together their knowledge and expertise. This aligns with the findings of Hoa (2022) that acknowledges the positive outcomes of team teaching in assisting educators to improve their professional skills by fostering professional collaboration, sharing of knowledge and communication. Consistent with the findings of Rytivaara et al. (2023), the journals reveal that collaborative teaching provides opportunities for teachers to observe one another, share knowledge and ideas about classroom practices, and later apply these new insights in their individual teaching.

It is no surprise that each teacher brings their own experiences, which make teaching and learning more effective and enable teachers to learn from one another's experiences. Parallel to Gast et al. (2017), whose study underlined that focusing on team based professional development activities fosters teacher learning, teacher collaboration, and reflective practices in higher education, the

findings of the present study highlighted that this collaborative approach to teaching also cultivated instructors' professional development through promoting learning from different perspectives, observing colleagues, and reflecting on their own teaching. As Lock et al. (2016) point out, collaborative teaching creates a supportive environment in which multiple ways of thinking, learning and teaching are encouraged. Insights from self-reflection journals indicate that instructors appreciated the opportunity to observe and learn from each other's teaching, assess their own strengths and areas for further development, and improved dynamism brought to the lesson.

Having said that, the reflections documented in the journals demonstrate that extra time needed for co-planning and effective coordination brings with it additional workload and pressure in terms of time management. This time investment needed to coordinate the process in sharing responsibilities and forming collegiality may render the practice as time consuming (Lock et al., 2016). Researchers suggested that receiving institutional support is a must to manage such challenges and to ensure that essential conditions for an effective teaming among instructors and logistic requirements for practice are in place. This underlines that the success of team teaching not only relies on strong partnership and successful coordination between instructors but also on solid administrative support to arrange instructors' schedules in a way to allow them find additional time for collaboration and effective planning.

Most students also seemed to appreciate the challenges of designing an effective team teaching plan, which was likely to bring extra workload for instructors. The complaints or negative aspects mentioned by students were found not to be associated with team teaching itself but rather with the facilities and physical aspects of the classroom settings. One major challenge the approach posed for some participants was adjusting to shifts between single-instructor and team-taught lessons. Several respondents suggested that, even though they got used to it over time, it might be more comfortable to design and deliver the entire course as team-taught from the outset.

Through findings of the present study, it is evident that the benefits of team teaching outweigh the possible challenges that could be faced by students and educators. The collaboration and communication between instructors also set an example for students, and assist their social development and guide them in their group projects. Increased participation and enthusiasm stand out as one of the most salient features of team teaching practices. Whilst acknowledging the difficulty of team teaching, The presence of more than one instructor not only enables educators to shoulder the workload together but also helps students sustain their interest and focus on the lesson for longer. Students are exposed to multiple perspectives, instruction strategies, various examples, and explanations presented by more than one instructor at a time.

To conclude, this research study investigated the multifaceted role of team teaching as a collaborative teaching approach in the higher education context, and its impact on students' learning experience and their perceptions of the approach. Although team teaching is not yet widely practiced in higher education, it has been gaining popularity across various educational settings (De Weerd et al., 2024). This study involved a sample size from one department in a single university, which could possibly limit the generalizability of the findings. Further studies in various higher education contexts and comparative studies are suggested to strengthen and generalize the findings of this study. Future research should include a broader range of courses from multiple universities and several degree programs to capture a more diverse set of participants and consider including more team-taught courses and exploring different co-teaching models to observe their effectiveness.

ACKNOWLEDGMENTS

This study is based upon a project funded by Erciyes University Scientific Research Projects Coordination Unit within the framework of Research University Support Program (ID: SBAU-2023-12519).

REFERENCES

- Agustin, A. (2019). Reflective journal as a self-directed and sustainable professional development tool for pre-service teachers: A case study in English language education study program. *Scholaria: Jurnal Pendidikan dan Kebudayaan*, 9(2), 103-110.
- Bacharach, N., & Heck, T. W. (2012). Voices from the field: Multiple perspectives on a co-teaching in student teaching model. *Educational Renaissance*, 1(1), 49-62.
- Barron, T., & Friend, M. (2024). Co-Teaching: Are We There Yet? *Journal of Educational and Psychological Consultation*, 35(2), 193–218. <https://doi.org/10.1080/10474412.2024.2422895>.
- Bayram, İ., & Canaran, Ö. (2018). Reflections on a teacher-led CPD model for EFL teachers, *International Online Journal of Education and Teaching (IOJET)*, 5(2), 419-431. <http://iojet.org/index.php/IOJET/article/view/406/244>
- Campbell, E. L., Reedy, A. R., Baird, M. J., & Baird, D. M. (2018). Better Together: Co-Teaching in Undergraduate Applied Psychology Courses. *Psychology Teaching Review*, 24(2), 3-11.
- Canaran, Ö. & Mirici, İ. (2020). A New Model of Team Teaching for Teacher Professional Development: A Case Study of In-Service English Teachers. *TED EĞİTİM VE BİLİM*. 10.15390/EB.2020.8430.
- Cook, L., & Friend, M. (1996). Co-teaching: Guidelines for creating effective practices. In E.
- L. Meyen, G.A. Vergason, & R. J. Whelan (Eds.), *Strategies for teaching exceptional children in inclusive settings*, 155-182. Love.
- Crow, J., & Smith, L. (2005). Co-teaching in higher education: Reflective conversation on shared experience as continued professional development for lecturers and health and social care students. *Reflective Practice*, 6(4), 491-506.
- De Weerd, D., Simons, M., Struyf, E., & Tack, H. (2024). Studying the Effectiveness of Team Teaching: A Systematic Review on the Conceptual and Methodological
- Credibility of Experimental Studies. *Review of Educational Research*, 00346543241262807.
- Flodén, J. (2016). The impact of student feedback on teaching in higher education. *Assessment & Evaluation in Higher Education*, 42(7), 1054–1068. <https://doi.org/10.1080/02602938.2016.1224997>

- Gast, I., Schildkamp, K., & van der Veen, J. T. (2017). Team-based professional development interventions in higher education: A systematic review. *Review of Educational Research*, 87(4), 736–767.
- Haag, K., Pickett, S. B., Trujillo, G., & Andrews, T. C. (2023). Co-Teaching in Undergraduate
- STEM Education: a Lever For Pedagogical Change Toward Evidence-Based Teaching?. *CBE—Life Sciences Education*, 22(1), es1.
- Harvey, L. (2003). Student Feedback. *Quality in Higher Education*, 9(1), 3–20. <https://doi.org/10.1080/13538320308164>
- Hurd, E., & Weilbacher, G. (2018). Developing and Using a Co-Teaching Model within a Middle Level Education Program. *Current Issues in Middle Level Education*, 23(1), n1.
- Hoa, L. H. (2022). Team -Teaching as a Tool for Professional Development. *International Journal of Language, Literature and Culture*, 2(3), 24–31. <https://doi.org/10.22161/ijllc.2.3.3>
- Kuckartz, U., Rädiker, S. (2019). Analyzing Focus Group Data. In: Analyzing Qualitative Data with MAXQDA. Springer, Cham. https://doi.org/10.1007/978-3-030-15671-8_15
- Laughlin, K., Nelson, P., & Donaldson, S. (2011). Successfully applying team teaching with adult learners. *Journal of Adult Education*, 40(1), 11–18.
- Ling, L. Y., Naidu, M., & Bai, K. (2023). Online Co-Teaching in Higher Education: Prospects, Drawbacks, and the Way Forward. *Journal of Institutional Research South East Asia*, 21(1).
- Lock, J., Clancy, T., Lisella, R., Rosenau, P., Ferreira, C., and Rainsbury, J. (2016). The Lived Experiences of Instructors Co-teaching in Higher Education. *Brock Education Journal*, 26(1), 22–35. doi:10.26522/brocked.v26i1.482
- López-Hernández, A., Buckingham, L. R., & Strotmann, B. (2023). Enhancing Learning-Oriented Assessment Through Co-Teaching in Higher Education. *Studies in Educational Evaluation*, 79, 101307.
- Mandel, K., & Eiserman, T. (2016). Team teaching in high school. *Educational Leadership*, 73(4), 74-77.
- Mandouit, L. (2018). Using student feedback to improve teaching. *Educational Action Research*, 26(5), 755–769. <https://doi.org/10.1080/09650792.2018.1426470>
- Nápoles, J. (2025). Co-teaching: A review of the literature. *Update: Applications of Research in Music Education*, 43(2), 34-40.
- Plank, K. M. (Ed.). (2023). *Team teaching: Across the disciplines, across the academy*. Taylor & Francis.

- Rytivaara, A., Pulkkinen, J., & Palmu, I. (2023). Learning about students in co-teaching teams. *International Journal of Inclusive Education*, 27(7), 803-818.
- Sanders-Smith, S. C., Lyons, M. E., Yang, S. Y. H., & McCarthy, S. J. (2021). Valuing relationships, valuing differences: Co-teaching practices in a Hong Kong early childhood center. *Teaching and Teacher Education*, 97, 103230.
- Tichenor, M., & Tichenor, J. (2019). Collaboration in the Elementary School: What Do Teachers Think?. *Journal of Curriculum and Teaching*, 8(2), 54-61.
- Villa, R. A., Thousand, J. S., & Nevin, A. I. (2004). *A guide to co-teaching: Practical tips for facilitating student learning*. Corwin Press.
- Yan, Y., Zuo, M., & Luo, H. (2024). Investigating co-teaching presence and its impact on student engagement: A mixed-method study on the blended synchronous classroom. *Computers & Education*, 222, 105153.
- Zach, S., & Avugos, S. (2024). Co-teaching in higher education: implications for teaching, learning, engagement, and satisfaction. *Frontiers in Sports and Active Living*, 6, 1424101.

““CHAPTER 2

The Effect of Computer-Assisted Language Learning on English Achievement: A Literature Review Based on Experimental Evidence

Doğuş Parçal¹
Ferhat Bahçeci²

¹ Fırat University, Graduate School of Educational Sciences, Department of Educational Sciences, Division of Curriculum and Instruction, Elazığ, Türkiye
ORCID: 0009-0001-0287-1665

² Assoc. Prof. Dr., Fırat University, Graduate School of Educational Sciences, Department of Educational Sciences, Division of Curriculum and Instruction, Elazığ, Türkiye, ORCID: 0000-0001-6363-4121

1. INTRODUCTION

Like almost every field of education, foreign language teaching has been reshaped in fundamental ways by the swift progress of information and communication technologies. As computers have become cheaper, smaller, and easier to use, their role in education has grown, and the fear of falling behind has accelerated the spread of technological tools in language learning. Against this backdrop, computer-assisted language learning (CALL) has emerged as one of the most well-grounded and wide-ranging ways of bringing technology into language instruction, supported by a body of research that has expanded and matured steadily across the past four decades (Mohsen, 2025).

Now that English serves as the shared language of global communication, teaching it to foreign- and second-language learners has taken on unprecedented importance. In many countries higher education—especially in engineering, informatics, and the natural sciences—is conducted in English, making a strong foundation essential. Where traditional instruction falls short, computers, mobile devices, internet resources, and multimedia software have become natural complements to language teaching, and the literature generally agrees that, through their appeal, effectiveness, and usefulness, digital tools positively influence student motivation and achievement (e.g., Almekhlafi, 2006; Mahmoudi et al., 2012; Vadivel et al., 2022).

A key advantage of computer-assisted instruction over traditional methods is individualized learning: CALL programs let students progress at their own pace, navigate the content freely, and review material as needed. Immediate feedback fosters self-confidence and self-regulation, while multiple on-screen stimuli—animation, visuals, color, sound, and music—capture attention and enhance retention. CALL thus creates an interactive environment that puts ‘learning by doing’ into practice.

Nevertheless, the extent and quality of technology’s contribution have long been debated. Specialists note that directly comparing technology-supported and traditional instruction is inherently limited, because technology forms an integrated whole with content, method, and context. Even so, policymakers and teachers keep asking whether technology-using classrooms are more successful than those without, and this paper aims to give a coherent, evidence-based answer within the sources reviewed.

Drawing on the body of sources examined here, this study primarily sets out to survey and bring together the evidence on how computer- and technology-supported language instruction affects English achievement. It seeks answers to five questions: (1) Do CALL applications significantly improve students’ overall English achievement? (2) How does this effect vary across skill areas such as vocabulary, grammar, listening, speaking, pronunciation, writing, and reading? (3) Which moderator variables (proficiency level, duration, device, software type,

gender, etc.) determine CALL's effectiveness? (4) What are students' attitudes and motivation toward CALL, and how do these relate to achievement? (5) What implementation challenges arise, and which technologies stand out for the future?

The article then presents its theoretical framework and method, organizes the findings by skill area and key variables, discusses the factors and moderators behind CALL's effectiveness, evaluates implementation barriers and emerging technologies, and closes with conclusions and recommendations.

2. THEORETICAL FRAMEWORK

2.1. Definition and Scope of Computer-Assisted Language Learning

At its most general, computer-assisted language learning can be understood as a technology-mediated process, spanning both cognitive and sociocultural interaction, through which learners organize their acquisition of language by means of a computer or a comparable digital device. The device is treated not as a determinant of achievement in itself but as a cultural tool that mediates learning in the Vygotskian sense (Vygotsky, 1978). This flexible definition accommodates the constantly changing nature of the field, since CALL is continually reshaped by evolving hardware and software. CALL is therefore best understood not as a single tool or method but as an umbrella concept covering a wide array of technological tools used for language teaching.

The programs and materials within CALL are generally grouped into four categories. The first consists of software built specifically for language learning, including interactive drills, quizzes, and multimedia CD-ROM programs. The second is general-purpose software—word processors, presentation tools, spreadsheets—used to support learning. The third gathers web-based materials: wikis and blogs, online dictionaries and encyclopedias, news portals, electronic texts, and web quests. Computer-mediated communication (CMC) forms the fourth category, operating either in real time (online chat) or with a delay (e-mail, discussion forums, and message boards).

As mobile technologies proliferated, mobile-assisted language learning (MALL) took shape as a branch of CALL that relies on smartphones, tablets, and wireless links so that learners can practice regardless of time or location. More recently still, technology-enhanced language learning (TELL) has gained prominence, reaching past computers and mobile devices to take in online platforms together with virtual and augmented reality (VR and AR). This expansion shows that the terminological and theoretical boundaries of the field are continually redrawn alongside technology.

2.2. Historical Development of Computer-Assisted Language Learning

Efforts to bring technology into language teaching date back to the 1960s. The roughly forty-year development of CALL is widely divided into three phases—

behavioristic, communicative, and integrative (Mohsen, 2025). This classification, proposed by Warschauer and Healey (1998), shows that theoretical trends in CALL have progressed in parallel with broader trends in applied linguistics.

The initial phase, behavioristic (structural) CALL, was first envisioned during the 1950s and put into practice through the 1960s and 1970s. Rooted in the behaviorist theories that prevailed at the time, this approach cast the computer as a mechanical instructor, or a deliverer of content, supplying repetitive drills paced by the learner. ‘Drill and practice’ was its defining feature, technology’s main function was presenting material, and the objective was accuracy; language was treated as a structural whole, and the grammar-translation and audio-lingual methods prevailed.

The second phase, communicative CALL, grew out of the communicative language teaching movement of the 1970s and 1980s and gained ground as microcomputers proliferated in the early 1980s. It responded to the criticism that earlier drill-and-practice programs did not allow authentic communication. Grounded in cognitive-constructivist and information-processing theories, it viewed the computer as a tool for developing learners’ mental models through the target language, with activities designed for meaningful peer interaction and fluency; language was treated as a cognitive system constructed in the mind.

The third phase, integrative CALL, rests on multimedia technology and the Internet. Multimedia formats like the CD-ROM merged sound, animation, video, graphics, and text on one device, while hypermedia allowed learners to set their own routes by connecting resources. Computer-mediated communication, which spread with the World Wide Web in the early 1990s, has arguably had the greatest impact on language teaching. In this phase sociocultural theories, collaborative learning, and authentic discourse came to the fore, and the computer became a ‘toolbox’ and integrated information-management system, with the learner an active, collaborating participant with agency over their own products.

As Internet access expanded, CALL moved from stand-alone workstations toward networked machines, and communicative methods yielded to socio-collaborative ones, with discussion turning to questions of learner autonomy, the design of collaborative projects, and suitable forms of assessment. This history shows that CALL is not merely a technological innovation but a reflection of paradigm shifts in language learning theories (Mohsen, 2025).

Table 1. The Three Developmental Phases of Computer-Assisted Language Learning According to Warschauer

Feature	Structural/Behavioristic CALL (1970–1980)	Communicative CALL (1980–1990)	Integrative CALL (21st century)
Role of the computer	Carrier of information (tutor)	Workstation (student)	Toolbox / integrated system
Learning theory	Behaviorism	Cognitive-constructivist	Sociocultural theories
Dominant teaching approach	Grammar-translation and audio-lingual	Communicative language teaching	Content-based / specific purposes
Main use	Drill and practice	Communicative exercises	Authentic discourse and agency
Primary learning objective	Accuracy	Fluency	Agency and collaboration

Source: Adapted from Warschauer and Healey's (1998) phase model and organized on the basis of the current literature (Mohsen, 2025).

Table 1 summarizes this shift across the three phases: the computer's role moved from carrier of information (tutor) to integrated toolbox, and the learning objective from accuracy to fluency and, ultimately, learner agency.

2.3. Learning Theories Supporting Computer-Assisted Language Learning

The theoretical underpinnings of CALL are informed by a range of fields, including artificial intelligence, psychology, instructional technology, computational linguistics, and human-computer interaction. Programmed instruction and cognition research inform its learning design, while computational linguistics feeds work on machine translation and natural language processing. This interdisciplinary character requires that CALL's technological and pedagogical dimensions be considered together.

Constructivism is among the frameworks most often invoked in CALL. On this view, learners construct knowledge by building on prior learning, and technology-rich classrooms support active, exploratory, autonomous participation. Since CALL settings enable learners to build their own mental representations by working through the target language, they map closely onto constructivist principles.

Motivation theories are also central to explaining CALL's effectiveness. Computer environments motivate many learners, who take more risks and behave more spontaneously because they feel less threatened; access to authentic materials, e-mail, and collaborative activities further nurtures motivation. In

computer-assisted writing settings, favorable attitudes have been reported as a key force behind the motivation students bring to producing high-quality writing (Mahmoudi et al., 2012).

To account for the link between attitude and behavior, CALL research also turns to cognitive frameworks, notably the Theory of Reasoned Action put forward by Fishbein and Ajzen (1975). By linking beliefs, attitudes, subjective norms, intentions, and behavior, the theory explains how attitudes predict behavioral intention. In CALL, students' attitudes toward technology are treated as an important influence on their intention to use it and thus on their learning outcomes (Almekhlafi, 2006). This theoretical plurality indicates that CALL's effect should be understood in cognitive, affective, and social dimensions alike.

3. METHOD

This article is a descriptive-integrative literature review (narrative synthesis) drawing solely on the reviewed body of sources. It drew together experimental and quasi-experimental research on how computer and mobile technologies affect English achievement, examining each study by context (EFL/ESL), educational level (primary, secondary, and higher education), target skill area, research design, and principal findings—dimensions in line with the criteria recommended for synthesis work in second language research. The findings were then organized and synthesized around skill areas and key variables. The reviewed sources also include experimental studies and large-scale surveys from the Turkish context, addressed under a separate heading.

In the synthesis, the means, standard deviations, and significance values reported for the experimental and control groups of individual studies were taken as the basis, as these show CALL's concrete effect in specific contexts. For effect sizes, Cohen's (1988) reference points were applied, with roughly 0.20 taken as small, about 0.50 as medium, and 0.80 or greater as large. No original statistical pooling was performed; findings were conveyed and interpreted only as presented in the sources. This approach makes it possible to view both context-specific details and general trends together.

The majority of the studies reviewed relied on pretest/posttest control-group designs of an experimental or quasi-experimental type, gauging effectiveness through achievement tests, proficiency examinations, and skill-specific measures (e.g., Naba'h et al., 2009; Gambari et al., 2014). Initial group equivalence was generally tested with pretest scores and controlled through analysis of covariance or matching before posttest differences were interpreted (e.g., Naba'h et al., 2009). Affective variables such as attitudes and motivation were examined through scales, questionnaires, and interviews (e.g., Mahmoudi et al., 2012; Vadivel et al., 2022). This methodological diversity allows CALL's effect to be evaluated consistently from different perspectives.

4. FINDINGS

4.1. Overall English Achievement

The reviewed experimental studies show that computer-assisted instruction has a positive effect on overall English achievement (e.g., Wang, 2021; Yanwar & Sitthiworachart, 2022). Jahansouz Shahi (2016), measuring overall proficiency with the PET exam in Iran, found that after controlling for the pretest the experimental group's mean (16.95) significantly exceeded the control group's (13.06), a large effect with $F(1,67) = 229.03$, $p = .000$ and partial eta-squared of .774. Similarly, Kazu and Kuvvetli (2026), measuring the four core skills together, found through ANOVA that the experimental group's posttest mean (77.59) clearly exceeded the control group's (48.91), contributing substantially to the total variance. Experimental groups thus consistently outperformed technology-free control groups; the distribution of effect sizes across skill areas is examined in the following subsections.

4.2. Vocabulary

Vocabulary is one of the most intensively studied skill areas in CALL research, with consistently positive results. Experimental studies from primary to higher education show that computer environments presenting visual, auditory, and contextual cues together significantly enhance vocabulary learning and retention.

In the Iranian context, Nejati et al. (2018) reported that an experimental group working with vocabulary software clearly surpassed a control group taught in the traditional way: while the control group showed no meaningful shift ($t = 1.203$, $p > .05$), the experimental group made significant gains ($p < .001$). Male and female learners did not differ significantly ($t = .308$), which led the authors to conclude that the effect held regardless of gender (Nejati et al., 2018). Technology's multiple-representation affordances thus strongly support vocabulary acquisition.

Publishing in CALL-EJ, Ali (2018) found the CALL group's posttest mean (35.03) clearly above the traditional group's (22.83), a significant difference ($t = -6.649$, $p < .05$); the CALL group also outperformed the traditional group on the delayed posttest (34.40 vs. 22.37), indicating better long-term retention (Ali, 2018). Similarly, Namaziandost et al. (2021) found that mobile- and computer-based applications significantly enhanced vocabulary acquisition, while studies of asynchronous CALL reported that high-achieving students benefited in both recall and retrieval.

Examining attitude and achievement in vocabulary learning, Mahmoudi et al. (2012) found that Iranian graduate students developed positive attitudes toward computer-assisted vocabulary learning that were positively related to achievement, reporting that computers made learning easier and more enjoyable.

This decisive role of attitude is consistent with work identifying positive attitude as one of the strongest predictors of language learning outcomes (Getie, 2020).

Recent studies reinforce the value of digital game-based and multimedia approaches to vocabulary. Using Quizizz with 8th-grade students in Türkiye, Kazu and Kuvvetli (2023) found that digital game-based learning significantly enhanced vocabulary acquisition and fostered positive perceptions. Zhang, Zou, and Cheng (2025) showed that a self-regulated digital game-based program improved both motivation and EFL vocabulary knowledge, and Far and Taghizadeh (2024) found that digital gamification increased collocation knowledge more than non-digital, non-gamified conditions. In the multimedia dimension, Teng (2025) showed that working memory played a decisive role in incidental vocabulary learning from captioned videos, indicating that audiovisual support interacts with individual differences.

4.3. Grammar

In grammar, CALL's effect is generally reported as positive but more modest than in other skills—arguably because grammar is largely acquired through implicit learning and communicative use, so software exercises alone may be insufficient. Even so, strong experimental evidence shows that well-designed grammar software produces tangible gains.

In one of the most detailed studies, Naba'h and colleagues (2009) taught the passive voice to 212 secondary students randomly assigned to four experimental and four control groups, comparing computer software with traditional instruction via ANCOVA. Groups were equivalent at pretest (experimental 22.09; control 21.66), but at posttest the experimental mean (26.21) exceeded the control (23.95), with a significant method effect ($F = 75.47$, $p < .05$), showing that the software markedly increased grammar achievement (Naba'h et al., 2009).

The same study also found significant differences by gender and academic track, with higher posttest means for male students (25.70 vs. 24.46; $F = 15.353$) and science-track students (25.56 vs. 24.60; $F = 15.968$) (Naba'h et al., 2009). The researchers attributed the effect to self-paced work, anytime/anywhere access, easy feedback, and the attention- and retention-enhancing power of multisensory screen elements. This tendency for computer-assisted grammar instruction to be at least as effective as—and usually more effective than—traditional instruction is echoed by other studies (Pirasteh, 2014).

Grammar findings are also sensitive to cultural and contextual factors: some studies report computer-assisted instruction as effective as traditional instruction, others as superior (Pirasteh, 2014), with differences partly explained by contextual variables. Recent syntheses note that grammar has been researched less than the output-oriented skills of speaking, writing, and vocabulary, although interest in technology-supported grammar has grown.

Recent experimental work highlights technology-supported feedback and gamification in grammar. Liu and Hwang (2024) found that metalinguistic corrective feedback embedded in digital games increased novice EFL learners' grammar achievement and engagement. Within a Vygotskian sociocultural framework, Mardian and Nafissi (2022) showed that synchronous computer-mediated corrective feedback supported grammatical development, and El Morabit and Manegre (2025) found that a video-based, learner-centered application improved production and comprehension over the grammar-translation method.

4.4. Listening

Listening is relatively difficult to teach, and students often remain passive in traditional classrooms; by combining auditory and visual stimuli, CALL can turn them into active listeners. Using a one-semester pretest/posttest design with first-year university students in Algeria, Nachoua (2012) demonstrated that CALL enhanced motivation and, in turn, listening performance, with learners in the experimental group substantially surpassing their control-group peers as exposure to visual alongside auditory input fostered active engagement.

Examining multimedia software, Barjesteh and Isaee (2023) found the experimental group's listening scores rose significantly from a pretest mean of 29.00 to a posttest mean of 36.12 ($p = .000$), while the traditionally taught control group barely changed (28.20 to 28.30). The experimental group's gain of about 7.12 points against the control's 0.10 points shows the clear superiority of technology-supported instruction, suggesting that enriching listening materials with visual and contextual elements directly affects comprehension (Barjesteh & Isaee, 2023).

Recent listening research highlights flipped designs, mobile applications, and visual enrichment. Jia, Hew, and Li (2025) found that a flipped decoding model significantly improved foreign-language listening proficiency. With a large sample, Jiang, Peters, Plonsky, and Pajak (2024) showed that learners using Duolingo as their sole tool made measurable gains in reading and listening. Liu and Elola (2025) found that nonverbal sensitivity and modality preferences affected comprehension in visually enriched listening tasks, underscoring the interaction of listening technologies with individual differences.

4.5. Speaking and Pronunciation

Speaking and pronunciation often receive too little attention in traditional EFL classrooms despite being critical for communication. Pronunciation stands out as one of the skills that benefits most from technology, while gains in speaking generally remain more limited: pronunciation can be improved measurably through technologies such as automatic speech recognition (ASR) and visual

feedback, whereas speaking is a more complex, interactional skill that is harder to assess.

In pronunciation, ASR-based mobile applications have been studied frequently. One found that an experimental group receiving instant written visual feedback via an ASR-based app improved more in producing the target sound than other groups (Liakin et al., 2015). With software that visualizes the speech signal (e.g., Praat), even short, guided practice has proven effective for acquiring segmental features. Mobile-assisted pronunciation training (MAPT) has significantly improved pronunciation performance (Lan, 2022), and one app-based program improved not only pronunciation but also attitudes toward the training and the application (Bashori et al., 2024).

Demonstrating computer-assisted pronunciation teaching (CAPT), Gambari and colleagues (2014) found in Nigeria that groups equivalent at pretest (experimental 20.13; control 20.60; $p = .461$) diverged at posttest, with the experimental mean (81.60) exceeding the control (68.80), a significant difference ($t = 6.335$, $p = .000$). Differences by verbal ability were significant ($F = 44.653$), while the absence of a male-female difference indicated the effect was independent of gender (Gambari et al., 2014). This strong effect is consistent with studies reporting positive outcomes for mobile-assisted training and ASR-based feedback (Lan, 2022; Liakin et al., 2015).

MALL's contribution to speaking has also been shown in primary studies: smartphone-based vlog projects raised speaking performance from 57.62 to 67.54 ($t = 29.38$, $p < .001$) (Huang, 2021), and studies of AI chatbots found that mind-map-supported artificial intelligence outperformed the traditional approach (Lin & Mubarak, 2021).

Recent speaking and pronunciation research documents online interaction, digital storytelling, and augmented reality. Rahimi and Fathi (2024) found that e-tandem learning, pairing students online with native speakers, significantly improved EFL speaking skills and willingness to communicate (WTC). Huang (2023) showed that digital storytelling enhanced speaking proficiency, WTC, and group cohesion, and Zhu, Zhang, and Li (2024) found that augmented reality filters advanced L2 pronunciation practice. At primary level, Hsieh, Yeh, and Chen (2025) found that a robot-and-tangible-object system improved pronunciation and WTC.

4.6. Writing

Writing is one of the skill areas where CALL exerts its strongest effects. These effects are well explained by Cognitive Load Theory: the automatic feedback of word processors, the ease of revision, and collaborative online tools reduce the extraneous cognitive load of writing's mechanical aspects, freeing the learner's capacity for the germane load of planning and higher-order text structuring

(Sweller, 1988). Writing classrooms are among the environments most receptive to the influence of computers.

In a quasi-experimental study in Iran, Zaini and Mazdayasna (2014) found that an experimental group taught in a computer laboratory outperformed a control group doing traditional writing tasks in appropriate article and tense use, attributing this to the automatic feedback from Microsoft Word combined with teacher comments, which helped students correct their own errors (Zaini & Mazdayasna, 2014).

Two recent studies reinforce CALL's writing effect with concrete figures. Working with 240 students at the University of Tabuk, Alshaikhi (2025) found the mean writing score rose from 3.74 (pretest) to 4.41 (posttest), a significant relationship with a strong correlation of 0.801 ($p = .000$), with higher initial scores yielding greater gains. Xie (2022), in China, found that computer-assisted technology reduced errors in English composition and raised the total score from 58.6% to 69.6%; the experimental class's cumulative errors fell from 194 to 97, a sharper decline than the control's 195 to 116, indicating rapidly improved writing accuracy (Xie, 2022).

Recent work highlights technology-supported feedback and collaborative online tools in writing. Zou, Xie, and Wang (2023) reported that peer, teacher, and self-feedback delivered through technology strengthened collaborative writing, dispositions toward critical thinking, and learner engagement. Adhami and Taghizadeh (2024) showed that integrating inquiry-based and computer-supported collaborative learning into a flipped classroom improved academic writing, and Hoang and Hoang (2024) found that Google Docs-based collaboration in a fully online course significantly advanced EFL students' academic writing.

4.7. Reading

Although reading has been studied less than other skills in CALL research, effect sizes are generally positive. Content analyses show researchers have favored output-oriented skills such as speaking, writing, and vocabulary, leaving reading the least attended-to area and its evidence base more limited.

Still, several studies report positive reading results. Online dictionaries, e-texts, electronic glosses, and data-driven learning support reading comprehension; mobile-assisted synchronous collaborative translation-annotation systems have been reported to aid EFL reading comprehension, and podcasts and blogs hold potential for advanced EFL learners (Azizi et al., 2022). By individualizing the reading process and easing access to texts, these tools raise comprehension.

Some findings suggest reading and writing benefit less from technology than other skills. Examining the four skills together, Ikonta and Ugonna (2015) found

significant gains in listening and speaking after a technology-supported intervention ($t = 2.64$ and $t = 2.66$; combined $t = 3.74$) but no significant pretest-posttest difference for reading ($t = 0.620$) or writing ($t = 0.912$), although the combined reading-writing scores did improve significantly ($t = 3.66$) (Ikonta & Ugonna, 2015). This complex picture indicates that technology's effect varies with the skill, the tool, and the intervention design.

Recent reading research focuses on mobile designs and the reading medium. Yang and Hu (2024) showed that cognitive load mediates the difference between mobile-assisted and paper-based EFL reading, so the medium's effect is not direct. Likewise, Chen and colleagues (2022) improved reading achievement, behavior, and engagement through a mobile strategy designed to increase reading engagement—evidence that technology's effect on reading depends on design and mediating variables.

4.8. Mobile-Assisted Language Learning (MALL)

As mobile devices have multiplied, MALL has grown into one of the liveliest and most rapidly expanding subfields within CALL. The portability, anytime access, and out-of-class practice afforded by smartphones and tablets contribute powerfully to language learning; recent work such as Pyo and Lee (2025) shows that mobile environments' positive effect persists and that they also support learner autonomy.

MALL's contribution to vocabulary is especially well documented. In Iran, a comparison of mobile and computer versions of Rosetta Stone found that both platforms significantly improved vocabulary knowledge (Namaziandost et al., 2021). Smartphone studies show that apps presenting words with multiple features—spelling, pronunciation, meaning, synonyms, antonyms, example sentences—increase acquisition, and that delivering words or idioms via SMS and e-mail outperforms traditional paper-based instruction (Thornton & Houser, 2005). Because such apps can be used across a semester, they offer a longer, more cumulative period for vocabulary learning.

There is also strong evidence of long-term effects. Working with a digital flashcard app (NGSL builder), Zakian and colleagues (2022) divided 86 students between an experimental and a control condition; those in the experimental group learned core vocabulary outside class through the app, while the control group used conventional word lists. Across three measurements (pretest, posttest, delayed posttest), the app contributed significantly to vocabulary development, with the experimental group outperforming the control (Zakian et al., 2022)—evidence of effectiveness for long-term retention. Similarly, Hasan et al. (2022) reported that a WhatsApp-based program positively affected tertiary-level EFL vocabulary learning.

Within the Technology Acceptance Model, research on MALL acceptance centers on how useful and how easy to use learners judge a tool to be, together with their intention to adopt it. One found that users of a bilingualized mobile dictionary held the most positive perceptions (Zhang, Hennessy, & Pérez-Paredes, 2025) and recommended making mobile devices more personalized, engaging, and effective. Taken together, these findings suggest MALL will occupy a central place in the future of CALL.

Recent MALL research also addresses student acceptance and the teacher dimension. Zhang, Hennessy, and Pérez-Paredes (2025) documented Chinese EFL students' acceptance of mobile dictionaries as varying by dictionary type, and Hafour (2022) found that MALL training reshaped, in a positive direction, how both trainee and practicing EFL teachers viewed and made use of mobile technology.

4.9. Attitudes and Motivation

CALL's effect on achievement is not limited to cognitive gains; students' attitudes toward technology and their motivation are integral to it. The literature underscores that how students perform provides valuable evidence for judging CALL programs, and that a positive attitude is the single strongest influence on language learning outcomes (Getie, 2020); attitude is accordingly handled as both an outcome and a predictor variable.

Working with primary and middle school students in the United Arab Emirates, Almekhlafi (2006) assigned 83 students to experimental (43) and control (40) groups and found, via ANOVA, a significant difference favoring CALL users ($p < .05$); the experimental students developed positive attitudes, found CALL useful, intended to use it in future, and benefited more as their computer skills and experience grew (Almekhlafi, 2006). Examining teacher perceptions in Türkiye, Aydın (2013) likewise reported generally positive attitudes alongside infrastructure and professional-development limitations, confirming this affective dimension in another context.

The positive attitude-performance link has been confirmed repeatedly. Mahmoudi et al. (2012) found that Iranian graduate students' positive attitudes toward computer-assisted vocabulary learning were significantly related to performance. Vadivel et al. (2022), with 50 participants at Cihan University, reported that most students said the computer system supported their English learning, raised motivation, and—through online self-assessment—helped them monitor their progress; students' active participation revealed the approach's popularity and its acceleration of learning, though limitations such as system failure stood out as a principal disadvantage (Vadivel et al., 2022).

Considered through Krashen's Affective Filter Hypothesis, computer environments—especially those allowing asynchronous, individual study—can

weaken the affective filter by lowering anxiety, encouraging risk-taking and spontaneous self-expression (Krashen, 1982). Access to authentic materials, e-mail, and collaboration likewise nurtures motivation, whereas activities that isolate learners, stifle creativity, or cause frustration may lower it. CALL's contribution to motivation thus depends on careful, learner-centered design, underscoring that technology must be paired with a pedagogy attentive to students' affective experience.

Recent research deepens understanding of the affective and self-regulatory dimensions of technology-enhanced environments. Applying structural equation modeling, Shen, Wang, Yang, and Yu (2026) identified two-way links between the academic emotions of Chinese university EFL learners and the self-regulated learning strategies they drew on. Lo (2025) showed that dual-subtitled videos affected behavioral, emotional, and cognitive engagement differently across conditions, and Cui and Wang (2024) found that the social annotation tool Perusall increased pre-class engagement and learning outcomes, with even low-proficiency students achieving comparable results.

4.10. Research Conducted in the Turkish Context

Research in the Turkish context adds important evidence on CALL's effect on English achievement. In a higher-education experiment in Türkiye, Kazu and Kuvvetli (2026) compared an experimental group using a mobile app (Duolingo) with a control group receiving traditional instruction across the four core skills. The experimental group's posttest mean (77.59) was markedly higher than the control's (48.91), and repeated-measures ANOVA showed the strongest gains in speaking ($\eta^2 = 0.570$), followed by listening ($\eta^2 = 0.431$) (Kazu & Kuvvetli, 2026)—evidence that MALL significantly improves Turkish EFL students' overall achievement, especially oral skills.

Studies of teacher and student views in Türkiye complete the picture. One found that EFL teachers generally viewed technology positively, though infrastructure, time, and professional development limited effective use (Aydin, 2013). In a large survey of 1,855 middle and high school students in Van, the factors making English difficult were examined: as learners' achievement grades declined, they endorsed these obstructing factors more strongly, and such endorsement ran higher for male learners than for female ones and for high-school students than for middle-school students (Tunç & Kozikoğlu, 2022). This suggests that technology-enhanced, motivating environments may be especially valuable for low-achieving students.

Türkiye-based studies also address emerging technologies. Examining university students' views on augmented reality (AR) in foreign language teaching, one found generally positive attitudes (Akçayır & Akçayır, 2016). A correlational study of high school students found a significant positive relationship between language learning strategies and English achievement, with

strategies significantly predicting achievement (Baş, 2014)—suggesting that technology’s contribution grows when it is designed to support self-regulated, strategic learning rather than serving merely as a tool.

Table 2. Summary of Selected Experimental Studies

Study / Context	Skill / Sample	Design	Key finding
Naba’h et al. (2009) – Jordan	Grammar; 212 students	Pretest/posttest, ANCOVA	Experimental 26.21 > Control 23.95; $F=75.47$, $p<.05$
Almekhlafi (2006) – UAE	General EFL; 83 students	Experimental-control, ANOVA	Significant in favor of the experimental group ($p<.05$); positive attitudes
Ali (2018) – CALL-EJ	Vocabulary	Pretest/posttest, t-test	CALL 35.03 > Traditional 22.83; $t=-6.649$
Alshaikhi (2025) – Saudi Arabia	Writing; 240 students	Paired pretest/posttest	Mean 3.74 → 4.41; $p=.000$
Gambari et al. (2014) – Nigeria	Pronunciation/oral English	Experimental-control, t-test	Experimental 81.60 > Control 68.80; $t=6.335$
Jahansouz Shahi (2016) – Iran	General proficiency (PET)	ANCOVA	Experimental 16.95 > Control 13.06; $\eta^2=.774$ (large)
Xie (2022) – China	Writing/general	Experimental-control	Total score 58.6% → 69.6%; errors decreased
Barjesteh and Isaee (2023) – Iran	Listening	Pretest/posttest	Experimental 29.00 → 36.12; control unchanged
Kazu and Kuvvetli (2026) – Türkiye	Four skills (Duolingo)	Repeated-measures ANOVA	Experimental 77.59 > Control 48.91; speaking $\eta^2=.570$
Rahimi and Fathi (2024) – Iran	Speaking, WTC; 42 students	Experimental-control, mixed methods	Significant improvement in speaking and WTC in the e-tandem group
Liu and Hwang (2024) – Taiwan	Grammar; 76 students	Experimental-control (digital game)	Metalinguistic feedback improved grammar achievement
Jia et al. (2025)	Listening; 80 students	Flipped, experimental-control	Experimental (44) > Control (36); significant advantage in listening/decoding

Study / Context	Skill / Sample	Design	Key finding
Jianget al. (2024)	Reading, listening; 245 students	Longitudinal assessment	Reading-listening gains with Duolingo exceeding CEFR expectations
Zou et al. (2023) – China	Writing; 90 students	Experimental (3 feedback types)	Peer feedback most effective for engagement and critical thinking
Yang and Hu (2024) – China	Reading; 191 students	Repeated measures, ANOVA/SEM	Cognitive load higher in mobile reading; full mediator of the medium effect
Far and Taghizadeh (2024) – Iran	Vocabulary/collocation; 75 students	Experimental-control, ANCOVA	Digital gamification significantly improved collocation knowledge
Liao et al. (2024)	General English; 50 students	Experimental-control (rural/urban)	AR game increased performance and motivation in both groups
Nejati et al. (2018) – Iran	Vocabulary; experimental-control	Pretest/posttest	Significant in favor of the experimental group ($p < .001$); no gender difference
Zaini and Mazdayasna (2014) – Iran	Writing; quasi-experimental	Experimental-control	Experimental group superior in the use of articles and tenses
Nachoua (2012) – Algeria	Listening	Pretest/posttest	Experimental group markedly higher than the control
Mardian and Nafissi (2022) – Iran	Grammar; 40 students	Experimental-control, mixed methods	Synchronous corrective feedback improved grammar
El Morabit and Manegre (2025) – Spain	Grammar; 76 students	Experimental-control	Video-based instruction superior to grammar-translation
Zhang, Zou, and Cheng (2025) – China	Vocabulary; mixed methods	Experimental-control	Self-regulated digital game improved vocabulary and motivation
Teng (2025)	Vocabulary; 125 students	Between-groups design	Caption type and working memory influenced vocabulary learning

Study / Context	Skill / Sample	Design	Key finding
Liu and Elola (2025)	Listening; 160 students	Experimental	Visual richness and modality preference affected comprehension
Huang (2023) – Taiwan	Speaking; digital storytelling	Experimental-comparison	Speaking proficiency and group cohesion increased significantly
Zhu et al. (2024) – China	Pronunciation; AR filters	Mixed methods	AR improved pronunciation production skills
Hsieh et al. (2025) – Taiwan	Pronunciation; 29 students	Pretest/posttest	Robot system improved pronunciation and WTC
Adhami and Taghizadeh (2024) – Iran	Writing; 61 students	Three-group experiment	Integrated flipped design improved academic writing
Hoang and Hoang (2024) – Vietnam	Writing; 24 students	Pretest/posttest	Google Docs significantly improved academic writing
Chen et al. (2022) – Taiwan	Reading; quasi-experimental	Experimental-control	REPS strategy increased reading achievement and engagement
Zhao et al. (2024)	General; VR	Experimental-control	Gamified VR improved performance and attitudes
Chu et al. (2023) – Taiwan	Gamification	Quasi-experimental	Expert-system-guided self-regulation increased EFL achievement

Note: The table summarizes a portion of the experimental studies addressed in the paper and does not cover all of the sources.

Table 2 shows that experimental studies from Jordan to Türkiye, across skills from grammar to the four core skills, consistently favored the experimental groups, whose posttest means clearly exceeded those of controls—indicating that CALL’s effect is relatively context-independent. The post-2020 additions (digital games, mobile applications, and virtual/augmented reality) reinforce this pattern.

Table 3. Distribution of the Recent (Post-2020) Technologies Included in the Review by Skill Area

Technology / Approach	Primary skill	Representative studies	Key contribution
Digital game-based learning and gamification	Vocabulary, grammar, general	Zhang, Zou, and Cheng (2025); Far and Taghizadeh (2024); Liu and Hwang (2024); Chu et al. (2023); Kazu and Kuvvetli (2023)	Enhances vocabulary/grammar gains through motivation, engagement, and self-regulation
Virtual and augmented reality (VR/AR)	Pronunciation, writing, general	Zhao et al. (2024); Chen et al. (2023); Zhu et al. (2024); Liao et al. (2024)	Raises performance and motivation through contextualized, immersive experiences
Mobile-assisted language learning and applications	Reading, vocabulary, general	Yang and Hu (2024); Chen et al. (2022); Zhang, Hennessey, and Pérez-Paredes (2025); Hafour (2022); Jiang et al. (2024)	Provides access at any time; its effect varies according to the medium and cognitive load
Online collaboration and peer/automated feedback	Writing	Zou et al. (2023); Adhami and Taghizadeh (2024); Hoang and Hoang (2024)	Improves collaborative writing, feedback, and engagement
Digital storytelling and e-tandem interaction	Speaking, WTC	Huang (2023); Rahimi and Fathi (2024)	Increases speaking proficiency and willingness to communicate
Robot/ASR and multimedia-supported instruction	Pronunciation, listening	Hsieh et al. (2025); Liu and Elola (2025); Teng (2025)	Provides interactive, audiovisual support for pronunciation and listening
Flipped and blended learning	Listening, grammar	Jia et al. (2025); El Morabit and Manegre (2025); Mardian and Nafissi (2022)	Supports skill development through in-class/out-of-class design
Affective and self-regulatory technology environments	Attitudes, motivation	Shen et al. (2026); Lo (2025); Cui and Wang (2024)	Positively affects engagement, emotions, and self-regulation

Note: The table groups the post-2020 primary (experimental/empirical) studies included in the review by technology type; meta-analyses and systematic reviews are excluded.

Table 3 groups the post-2020 primary studies by technology and approach. Recent research spans digital games and gamification, virtual/augmented reality,

mobile applications, and online collaboration and automated feedback, mostly yielding positive results across skills—showing that CALL is increasingly turning toward more personalized, immersive, and self-regulation-supporting technologies.

5. DISCUSSION

5.1. Factors Explaining the Effectiveness of Computer-Assisted Language Learning

Considered together, the findings show that CALL's positive effect on English achievement rests not on a single cause but on several mutually reinforcing pedagogical features. Foremost is individualization: CALL lets each student work at their own pace, navigate content freely, and review as needed, making it possible to tailor learning to the individual rather than to the homogenizing traditional classroom.

A second factor is immediate, continuous feedback, which builds self-regulation and confidence by showing scores and correct answers instantly; in writing, automated feedback helps students notice and correct their own errors (Zaini & Mazdayasna, 2014). A third is multiple representation and multisensory appeal—text, images, sound, animation, and video together increase attention and retention, largely explaining the large effect sizes in vocabulary.

A fourth factor is access to authentic materials and collaborative learning: the Internet exposes students to authentic discourse while CMC tools support peer interaction. The fifth and most critical is motivation and positive attitudes, as learners who feel less threatened take more risks. That these factors work jointly suggests CALL's effect stems less from technology itself than from the pedagogical design it supports.

An important caveat follows: the literature stresses that technology's effect often arises not from technology alone but from its synergistic interaction with method and content, which are hard to disentangle in practice. Rather than asking whether technology or pedagogy is effective, it is more productive to ask under which pedagogical conditions technology is effective.

5.2. Moderator Variables: Proficiency Level, Duration, Device, and Software Type

CALL's effectiveness is not fixed but varies with moderator variables. One is proficiency level: experimental studies suggest advanced and intermediate learners tend to benefit more than beginners, implying that the autonomous study CALL demands presupposes a baseline of linguistic and digital competence.

Treatment duration is another variable: CALL groups tend to outperform technology-free groups regardless of duration, but gains increase as interventions

lengthen—so technology’s full pedagogical benefit emerges only through sufficient, sustained use.

Device type also matters: the skill area and context determine device effectiveness. In short, repetition-heavy tasks such as vocabulary, the portability of mobile devices is an advantage, whereas in tasks needing subtitles and visual detail, such as listening, the computer’s large screen is superior.

Software type and feedback form are likewise decisive: there is no single ‘right’ software, effectiveness depends on fit to purpose and context, and monitoring and regulatory feedback are especially useful. Together these findings show that attention to detail in CALL design directly determines its effectiveness.

5.3. The Gender Variable

Gender’s role in CALL effectiveness is inconsistent across the literature—some studies find significant differences, others none. In the Jordanian grammar study, male students scored significantly higher, which the researchers attributed to boys’ strong interest in computers and multimedia (Naba’h et al., 2009); the Cihan University study likewise noted greater male interest in learning with machines (Vadivel et al., 2022). Such explanations, tying the gap to men’s ‘natural’ interest in technology, rest on an essentialist assumption; contemporary approaches explain the variance not through biological predisposition but through gender roles in access to and use of devices and socialization grounded in gender schema (Bem, 1981).

By contrast, many studies find no significant gender difference. In the Iranian vocabulary study, none was detected in CALL-based vocabulary learning ($t = .308$) (Nejati et al., 2018); in the Nigerian pronunciation study, none was found between male and female students exposed to the CAPT package (Gambari et al., 2014); and studies of digital audio- and video-based packages likewise report none (Gambari et al., 2012). This mixed picture suggests gender’s effect may vary with cultural context, technology, and targeted skill, so generalization should be avoided; the literature acknowledges that findings on gender are not yet conclusive and need further research.

6. CHALLENGES ENCOUNTERED IN IMPLEMENTING COMPUTER-ASSISTED LANGUAGE LEARNING

Despite its opportunities, CALL faces significant implementation barriers. One of the most cited is limited access to technological infrastructure and resources: students in disadvantaged or rural areas struggle to access reliable internet, up-to-date hardware, and software. This concerns not only the first-level digital divide (lack of devices and connectivity) but also the second-level divide—differences in students’ and teachers’ ability to use digital tools effectively—which structurally prevents equitable benefit from technology (van Dijk, 2020). High costs further exacerbate the problem.

A second barrier is the digital literacy of teachers and students. Students may use smartphones and computers for entertainment yet lack the skills to use them effectively for learning, while many teachers lack the skill, knowledge, and confidence to integrate technology into practice. Without adequate professional development, teachers cannot use digital tools to full potential, harming the learning experience (Tafazoli, 2021).

Teacher workload and time constraints are among the barriers most cited in qualitative research. In one study, all teachers (100%) cited workload and heavy teaching loads, and most cited difficulty of access (86%), lack of time (86%), and inability to keep pace with technology (86%); cost (71%) and inadequate facilities (86%) were also prominent (Rintaningrum, 2023). Working with blogs, for instance, requires much teacher time to review and comment on each student's writing, and larger classes lengthen assessment—showing that technology integration requires attention to teachers' working conditions, not just tools.

Teacher attitudes and pedagogical mindsets are also decisive. Drawing on Hong's (2010) model, a qualitative study of CALL adoption in higher education sorted the barriers into three clusters: teacher preparation for CALL, factors residing in individual teachers, and contextual conditions (Tafazoli, 2021). Some teachers still hold a transmissive mindset, doubt teaching with technology, and resist change (Aydin, 2013), and incomplete technological pedagogical content knowledge (TPACK)—especially in the technological content and technological pedagogical dimensions—hampers integration (Tafazoli, 2021). Recommended remedies include restructuring teacher education, providing continuous professional development, and incorporating CALL into in-service training at the policy level.

Several strategies for overcoming these barriers stand out. Foremost is training teachers in technology integration, which builds their skills, their capacity to keep knowledge current, and their confidence. Emphasizing technology's value, setting clear goals, encouraging a mindset shift, and supporting lifelong learning are also recommended. Successful integration thus depends on addressing technical, institutional, and human dimensions together.

7. FUTURE DIRECTIONS AND EMERGING TECHNOLOGIES

As technology evolves rapidly, new possibilities in English teaching are inevitable. The innovations singled out most often are three: artificial intelligence, augmented reality, and virtual reality (AI, AR, and VR); together these are anticipated to reshape the way language is taught and learned, with VR settings, for instance, placing students inside lifelike simulations for authentic practice.

Augmented reality enriches perceived reality by integrating digital information with the real environment. One study found an AR-based ubiquitous

learning system improved EFL learners' performance (Chen et al., 2023), and another found AR-based games supported contextual vocabulary learning, fostered presence and immediacy, developed problem-solving and information-management skills, and increased motivation (Liao et al., 2024). Robots have served as learning partners, especially for young learners, supporting word-picture matching tasks (Hsieh et al., 2025). A Turkish study likewise found university students generally positive toward AR in foreign language teaching (Akçayır and Akçayır, 2016).

Artificial intelligence applications are spreading quickly. Intelligent personal assistants (e.g., Alexa) and AI chatbots are being examined for supporting L2 listening and speaking; mind-map-supported AI has outperformed the traditional approach (Lin & Mubarak, 2021), and online video pronunciation dictionaries (e.g., YouGlish) have improved word usage and pronunciation (Fu & Yang, 2019). These developments point toward a CALL future of increasingly personalized, interactive, and autonomous learning.

The literature also urges caution with new technologies, whose rapid pace renders some tools obsolete quickly. Although gamification raises mainly extrinsic motivation through badges, points, and leaderboards, it remains debatable, within Self-Determination Theory, whether such external rewards become internalized regulation and produce lasting learning, since students may focus on points rather than deep understanding (Ryan & Deci, 2000). New technologies should therefore balance motivational appeal against meaningful learning, and syntheses suggest that maturing tools such as VR and AR warrant detailed analysis in future reviews.

Recent experimental evidence for these projections is accumulating. In virtual reality, Zhao, Panjaburee, Hwang, and Wongkia (2024) found a self-regulation-based gamified VR system improved English learning performance and affective attitudes. In augmented reality, Chen and colleagues (2023) found that in-context learning with smart glasses in a science museum enhanced learning and motivation. In gamification, Chu, Hwang, Chien, and Chang (2023) found that expert-system-guided self-regulated digital games raised EFL achievement, and Liao, Wu, Gunawan, and Chang (2024) reported that an AR-based game produced significant gains in English performance and motivation for primary students in both rural and urban settings—suggesting that emerging technologies add to achievement when paired with sound pedagogical design rather than adopted for their novelty.

8. CONCLUSION AND RECOMMENDATIONS

This chapter has reviewed and synthesized CALL's effect on English achievement on the basis of the sources examined. The findings support a strong, consistent conclusion: compared with traditional instruction, CALL applications mostly yield statistically significant achievement gains favoring the experimental

groups, being at least as effective as technology-free instruction and more effective when rigorous designs are used.

By skill area, the effects form a spectrum: strongest in pronunciation, writing, and vocabulary; near-large in listening and reading; moderate in grammar; and comparatively small but generally positive in speaking. Technology thus contributes more to measurable, repetition-based, feedback-amenable skills and less to complex, interactional ones. Mobile-assisted language learning stands out as the most dynamic subfield, with strong results especially in vocabulary and long-term retention.

CALL's effectiveness is also context-dependent, varying with moderators such as proficiency level, treatment duration, device and software type, and feedback form; intermediate and advanced learners, longer interventions, purpose-appropriate software, and monitoring-type feedback produced stronger effects. Gender findings were inconsistent, depending on cultural and contextual factors. The positive link between attitudes and motivation and achievement shows CALL's effect is affective as well as cognitive. Turkish-context studies confirm this trend—mobile and computer-assisted applications improve Turkish EFL students' achievement, especially oral skills, and teachers and students generally view technology positively—though larger samples and longer durations are needed.

Overall, CALL is a powerful tool for enhancing English achievement, but technology alone is not a solution: its effect stems largely from the surrounding pedagogical design, method, and content. A successful implementation therefore combines sound pedagogy, competent teachers, and appropriate technological infrastructure.

Four practical recommendations follow. First, CALL activities should be interactive, engaging, aligned with specific language areas, and adapted to students' initial proficiency. Second, teachers should receive continuous, high-quality professional development in technology integration and technological pedagogical content knowledge, beyond basic computer skills. Third, equitable policies on infrastructure access and digital literacy should narrow the digital divide. Fourth, a motivating environment supporting self-regulated learning strategies and student interests should be created.

This study has several limitations. First, it was designed not as a systematic review or PRISMA-protocol search but as an integrative, narrative review of empirical (experimental) studies selected purposively; this prioritizes depth and conceptual coherence but limits any claim to fully represent the field's literature, so confirmation through comprehensive database searches (e.g., WoS, Scopus) is recommended. Relatedly, only high-citation-impact empirical studies were included; about 800 relevant but lower-impact studies from the database searches (WoS, Scopus, IEEE, Springer, Taylor & Francis, ScienceDirect) were excluded

to keep the scope manageable and quality-focused, which further limits representation. Second, the studies differ considerably in sample size, duration, instruments, and design, making direct comparison of effect sizes difficult. Third, most are short, small-sample, and concentrated in higher education, limiting generalizability to primary and preschool levels. Fourth, some could not fully control confounds such as publication bias and the Hawthorne effect. Finally, because the review focuses on quantitative effect sizes and achievement scores, qualitative and contextual details of learning were addressed only to a limited extent.

Several avenues emerge for future research: tracing the long-term effects of CALL and emerging technologies; undertaking systematic analysis of maturing tools, among them artificial intelligence, augmented reality, and virtual reality; and giving fuller attention to under-researched areas such as speaking, reading, and grammar. Research designs that reveal the synergy among teachers, content, and technology and that account for cultural and contextual variables are also needed. In conclusion, when carefully designed and effectively integrated, computer-assisted language learning can make English instruction more engaging, accessible, and effective.

ACKNOWLEDGMENTS

This chapter is derived from the first author's master's seminar titled "The Effect of Computer-Assisted Language Learning on English Achievement," completed under the supervision of the second author at the Graduate School of Educational Sciences, Firat University. The authors thank everyone who contributed to the study.

REFERENCES

- Adhami, N., & Taghizadeh, M. (2024). Integrating inquiry-based learning and computer supported collaborative learning into flipped classroom: Effects on academic writing performance and perceptions of students of railway engineering. *Computer Assisted Language Learning*, 37(3), 521–557. <https://doi.org/10.1080/09588221.2022.2046107>
- Akçayır, M., & Akçayır, G. (2016). University students' opinions about use of augmented reality in foreign language teaching. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi*, 18(2), 1169–1186. <https://doi.org/10.17556/jef.86406>
- Ali, Z. (2018). The effect of using computer-assisted language learning (CALL) on EFL learners. *CALL-EJ*, 19(1), 43–59.
- Almekhlafi, A. G. (2006). The effect of Computer Assisted Language Learning (CALL) on United Arab Emirates English as a Foreign Language (EFL) school students' achievement and attitude. *Journal of Interactive Learning Research*, 17(2), 121–142.
- Alshaikhi, T. (2025). The influence of Computer-Assisted Language Learning (CALL) on writing skills and attitudes of first-year EFL students at the University of Tabuk. (Conference paper). <https://doi.org/10.1109/ICCIT63348.2025.10989337>
- Aydin, S. (2013). Teachers' perceptions about the use of computers in EFL teaching and learning: The case of Turkey. *Computer Assisted Language Learning*, 26(3), 214–233. <https://doi.org/10.1080/09588221.2012.654495>
- Azizi, Z., Namaziandost, E., & Rezai, A. (2022). Potential of podcasting and blogging in cultivating Iranian advanced EFL learners' reading comprehension. *Heliyon*, 8(5), e09473. <https://doi.org/10.1016/j.heliyon.2022.e09473>
- Barjesteh, H., & Isace, H. (2023). Technology is an asset: Enhancing EFL learners' vocabulary knowledge and listening comprehension through CALL. *Research Square* (Preprint). <https://doi.org/10.21203/rs.3.rs-3136533/v1>
- Baş, G. (2014). Lise öğrencilerinin dil öğrenme stratejileri ile İngilizce dersindeki akademik başarıları arasındaki ilişki [The relationship between high school students' language learning strategies and their academic achievement in the English course]. *Buca Eğitim Fakültesi Dergisi*, 37, 166–179.
- Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2024). 'Look, I can speak correctly': Learning vocabulary and pronunciation through websites equipped with automatic speech recognition technology. *Computer Assisted Language Learning*. <https://doi.org/10.1080/09588221.2022.2080230>
- Bem, S. L. (1981). Gender schema theory: A cognitive account of sex typing. *Psychological Review*, 88(4), 354–364. <https://doi.org/10.1037/0033-295X.88.4.354>

- Chen, H.-R., Lin, W.-S., Hsu, T.-Y., Lin, T.-C., & Chen, N.-S. (2023). Applying smart glasses in situated exploration for learning English in a national science museum. *IEEE Transactions on Learning Technologies*, 16(5), 820–830. <https://doi.org/10.1109/tlt.2023.3276702>
- Chen, M.-R. A., Hwang, G.-J., Lin, Y.-H., Abou-Khalil, V., Li, H., & Ogata, H. (2022). A reading engagement-promoting strategy to facilitate EFL students' mobile learning achievement, behaviour and engagement. *International Journal of Mobile Learning and Organisation*, 16(4), 489–506. <https://doi.org/10.1504/ijmlo.2022.125968>
- Chu, S.-T., Hwang, G.-J., Chien, S.-Y., & Chang, S.-C. (2023). Incorporating teacher intelligence into digital games: An expert system-guided self-regulated learning approach to promoting EFL students' performance in digital gaming contexts. *British Journal of Educational Technology*, 54(2), 534–553. <https://doi.org/10.1111/bjet.13260>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cui, T., & Wang, J. (2024). Empowering active learning: A social annotation tool for improving student engagement. *British Journal of Educational Technology*, 55(2), 712–730. <https://doi.org/10.1111/bjet.13403>
- El Morabit, S., & Manegre, M. (2025). Comparing student-centered learning in a video-based app to the grammar-translation method in an EFL class. *Computer Assisted Language Learning*, 38(7), 1592–1610. <https://doi.org/10.1080/09588221.2023.2301447>
- Far, F. F., & Taghizadeh, M. (2024). Comparing the effects of digital and non-digital gamification on EFL learners' collocation knowledge, perceptions, and sense of flow. *Computer Assisted Language Learning*, 37(7), 2083–2115. <https://doi.org/10.1080/09588221.2022.2146724>
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Addison-Wesley.
- Fu, J. S., & Yang, S.-H. (2019). Exploring how YouGlish facilitates EFL learners' speaking competence. *Educational Technology & Society*, 22(4), 47–58. <https://www.jstor.org/stable/26910184>
- Gambari, A. I., Gbodi, B. E., & Olumba, R. N. (2012). Effects of audio and video compact disc instructional packages on students' performance in senior secondary schools phonetics in Minna, Niger State. *Journal of Science, Technology, Mathematics and Education*, 9(1), 250–261.
- Gambari, A. I., Kutigi, A. U., & Fagbemi, P. O. (2014). Effectiveness of computer-assisted pronunciation teaching and verbal ability on the achievement of senior secondary school students in oral English. *Gist Education and*

- Getie, A. S. (2020). Factors affecting the attitudes of students towards learning English as a foreign language. *Cogent Education*, 7(1), 1738184. <https://doi.org/10.1080/2331186X.2020.1738184>
- Hafour, M. F. (2022). The effects of MALL training on preservice and in-service EFL teachers' perceptions and use of mobile technology. *ReCALL*, 34(3), 274–290. <https://doi.org/10.1017/s0958344022000015>
- Hasan, M. K., Fakih, A.-H., Seraj, P. M. I., & Hasmirati. (2022). The effect of technology-assisted language programme on vocabulary learning among EFL students at the tertiary level. *Heliyon*. <https://doi.org/10.1016/j.heliyon.2022.e10313>
- Hoang, D. T. N., & Hoang, T. (2024). Enhancing EFL students' academic writing skills in online learning via Google Docs-based collaboration: A mixed-methods study. *Computer Assisted Language Learning*, 37(7), 1504–1526. <https://doi.org/10.1080/09588221.2022.2083176>
- Hong, K. S. (2010). CALL teacher education as an impetus for L2 teachers in integrating technology. *ReCALL*, 22(1), 53–69. <https://doi.org/10.1017/S095834400999019X>
- Hsieh, W.-M., Yeh, H.-C., & Chen, N.-S. (2025). Impact of a robot and tangible object (R&T) integrated learning system on elementary EFL learners' English pronunciation and willingness to communicate. *Computer Assisted Language Learning*, 38(4), 773–798. <https://doi.org/10.1080/09588221.2023.2228357>
- Huang, H.-T. D. (2023). Examining the effect of digital storytelling on English speaking proficiency, willingness to communicate, and group cohesion. *TESOL Quarterly*, 57(1), 242–269. <https://doi.org/10.1002/tesq.3147>
- Huang, H.-W. (2021). Effects of smartphone-based collaborative vlog projects on EFL learners' speaking performance and learning engagement. *Australasian Journal of Educational Technology*, 37(6), 18–40. <https://doi.org/10.14742/ajet.6623>
- Ikonta, N. R., & Ugonna, N. C. (2015). The effect of Rosetta Stone (computer-assisted language learning) software on English as second language students' proficiency in English language. *African Educational Research Journal*, 3(1), 69–79.
- Jahansouz Shahi, M. A. (2016). The impact of e-learning on improving Iranian EFL learners' language skills: Decreasing learning anxiety. *Journal of Fundamental and Applied Sciences*, 8(3S), 261–275. <https://doi.org/10.4314/jfas.v8i3s.180>

- Jia, C., Hew, K. F., & Li, M. (2025). Towards a flipped SEF-ARCS decoding model to improve foreign language listening proficiency. *Computer Assisted Language Learning*, 38(3), 369–396. <https://doi.org/10.1080/09588221.2023.2191655>
- Jiang, X., Peters, R., Plonsky, L., & Pajak, B. (2024). The effectiveness of Duolingo English courses in developing reading and listening proficiency. *CALICO Journal*, 41(3), 249–272. <https://doi.org/10.1558/cj.31612>
- Kazu, İ. Y., & Kuvvetli, M. (2023). A triangulation method on the effectiveness of digital game-based language learning for vocabulary acquisition. *Education and Information Technologies*, 28(10), 13541–13567. <https://doi.org/10.1007/s10639-023-11756-y>
- Kazu, İ. Y., & Kuvvetli, M. (2026). Digital language learning with Duolingo: Assessing its impact on listening, speaking, reading, and writing skills. *Journal of Computers in Education*, 13(1), 165–191. <https://doi.org/10.1007/s40692-025-00355-0>
- Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon Press.
- Lan, E.-M. (2022). A comparative study of computer and mobile-assisted pronunciation training: The case of university students in Taiwan. *Education and Information Technologies*, 27(2), 1559–1583. <https://doi.org/10.1007/s10639-021-10647-4>
- Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal*, 32(1), 1–25. <https://doi.org/10.1558/cj.v32i1.25962>
- Liao, C.-H. D., Wu, W.-C. V., Gunawan, V., & Chang, T.-C. (2024). Using an augmented-reality game-based application to enhance language learning and motivation of elementary school EFL students: A comparative study in rural and urban areas. *Asia-Pacific Education Researcher*, 33(2), 307–319. <https://doi.org/10.1007/s40299-023-00729-x>
- Lin, C.-J., & Mubarak, H. (2021). Learning analytics for investigating the mind map-guided AI chatbot approach in an EFL flipped speaking classroom. *Educational Technology & Society*, 24(4), 16–35. [https://doi.org/10.30191/ETS.202110_24\(4\).0002](https://doi.org/10.30191/ETS.202110_24(4).0002)
- Liu, L.-A., & Hwang, G.-J. (2024). Effects of metalinguistic corrective feedback on novice EFL students' digital game-based grammar learning performances, perceptions and behavioural patterns. *British Journal of Educational Technology*, 55(2), 687–711. <https://doi.org/10.1111/bjet.13400>
- Liu, Y.-T., & Elola, I. (2025). Seeing beyond the words: Modality preferences, nonverbal sensitivity, and the role of visual dynamic richness in L2 listening. *System*, 134, 103834. <https://doi.org/10.1016/j.system.2025.103834>

- Lo, S. (2025). Viewing dual-subtitled videos under different learning conditions: Effects on learners' behavioural, emotional, and cognitive engagement. *Computer Assisted Language Learning*, 38(4), 742–772. <https://doi.org/10.1080/09588221.2023.2219711>
- Mahmoudi, E., Abdul Samad, A., & Abdul Razak, N. Z. (2012). Attitude and students' performance in Computer Assisted English Language Learning (CAELL) for learning vocabulary. *Procedia – Social and Behavioral Sciences*, 66, 489–498. <https://doi.org/10.1016/j.sbspro.2012.11.293>
- Mardian, F., & Nafissi, Z. (2022). Synchronous computer-mediated corrective feedback and EFL learners' grammatical knowledge development: A sociocultural perspective. *Iranian Journal of Language Teaching Research*, 10(2), 115–136. <https://doi.org/10.30466/ijltr.2022.121186>
- Mohsen, M. A. (2025). Mapping the evolution of computer-assisted language learning research: A 44-year bibliometric overview. *European Journal of Education*, 60(2), e70051. <https://doi.org/10.1111/ejed.70051>
- Naba'h, A. A., Hussain, J., Al-Omari, A., & Shdeifat, S. (2009). The effect of computer-assisted language learning in teaching English grammar on the achievement of secondary students in Jordan. *The International Arab Journal of Information Technology*, 6(4), 431–439.
- Nachoua, H. (2012). Computer-assisted language learning for improving students' listening skill. *Procedia – Social and Behavioral Sciences*, 69, 1150–1159. <https://doi.org/10.1016/j.sbspro.2012.12.045>
- Namazandost, E., Alekasir, S., Saberi Dehkordi, E., & Tilwani, S. A. (2021). An account of EFL learners' vocabulary learning in a mobile-assisted language environment: The case of Rosetta Stone application. *Computer-Assisted Language Learning Electronic Journal*, 22(1), 80–110. <https://callej.org/index.php/journal/article/view/323>
- Nejati, E., Jahangiri, A., & Salehi, M. R. (2018). The effect of using computer-assisted language learning (CALL) on Iranian EFL learners' vocabulary learning: An experimental study. *Cypriot Journal of Educational Science*, 13(2), 351–362. <https://doi.org/10.18844/cjes.v13i2.752>
- Pirasteh, P. (2014). The effectiveness of Computer-Assisted Language Learning (CALL) on learning grammar by Iranian EFL learners. *Procedia – Social and Behavioral Sciences*, 98, 1422–1427. <https://doi.org/10.1016/j.sbspro.2014.03.561>
- Pyo, J., & Lee, C. H. (2025). Developing learner autonomy and EFL listening skills through mobile-assisted blended learning. *Innovation in Language Learning and Teaching*, 19(5), 465–480. <https://doi.org/10.1080/17501229.2024.2372068>

- Rahimi, M., & Fathi, J. (2024). Employing e-tandem language learning method to enhance speaking skills and willingness to communicate: The case of EFL learners. *Computer Assisted Language Learning*, 37(4), 924–960. <https://doi.org/10.1080/09588221.2022.2064512>
- Rintaningrum, R. (2023). Technology integration in English language teaching and learning: Benefits and challenges. *Cogent Education*, 10(1), 2164690. <https://doi.org/10.1080/2331186X.2022.2164690>
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78. <https://doi.org/10.1037/0003-066X.55.1.68>
- Shen, B., Wang, Y., Yang, Y., & Yu, X. (2026). Relationships between Chinese university EFL learners' academic emotions and self-regulated learning strategies: A structural equation model. *Language Teaching Research*, 30(1), 317–342. <https://doi.org/10.1177/13621688221144832>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- Tafazoli, D. (2021). Affordances of computer-assisted language learning in higher education: A qualitative inquiry. (Qualitative study).
- Teng, M. F. (2025). Effectiveness of captioned videos for incidental vocabulary learning and retention: The role of working memory. *Computer Assisted Language Learning*, 38(1–2), 206–234. <https://doi.org/10.1080/09588221.2023.2173613>
- Thornton, P., & Houser, C. (2005). Using mobile phones in English education in Japan. *Journal of Computer Assisted Learning*, 21(3), 217–228. <https://doi.org/10.1111/j.1365-2729.2005.00129.x>
- Tunç, M., & Kozikoğlu, İ. (2022). İngilizce öğrenmeyi zorlaştıran faktörlere ilişkin öğrenci görüşleri [Student views on the factors that make learning English difficult]. *e-Kafkas Eğitim Araştırmaları Dergisi*, 9, 279–298. <https://doi.org/10.30900/kafkasegt.986233>
- Vadivel, B., Khalil, N. R., Asif, M., & Ajanil, B. (2022). Computer-assisted language learning and English learning in Cihan University: A mixed-methods study. *Education Research International*, 2022, Article 8700068. <https://doi.org/10.1155/2022/8700068>
- van Dijk, J. A. G. M. (2020). *The digital divide*. Polity Press.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

- Wang, X. (2021). Evaluation of the effectiveness of computer-assisted language teaching in foreign language learning. *Mathematical Problems in Engineering*, 2021. <https://doi.org/10.1155/2021/7143815>
- Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching*, 31(2), 57–71. <https://doi.org/10.1017/S0261444800012970>
- Xie, Y. (2022). Effectiveness of computer-aided technology for teaching English courses. *Scientific Programming*, 2022. <https://doi.org/10.1155/2022/2133028>
- Yang, X., & Hu, J. (2024). Distinctions between mobile-assisted and paper-based EFL reading comprehension performance: Reading cognitive load as a mediator. *Computer Assisted Language Learning*, 37(7), 2051–2082. <https://doi.org/10.1080/09588221.2022.2143527>
- Yanwar, A., & Sitthiworachart, J. (2022). Technologies in enhancing English language skills. *Journal of Higher Education Theory and Practice*, 22. <https://doi.org/10.33423/jhetp.v22i14.5542>
- Zaini, A., & Mazdayasna, G. (2014). The effect of computer assisted language learning on the development of EFL learners' writing skills. *Procedia – Social and Behavioral Sciences*, 98, 1975–1982. <https://doi.org/10.1016/j.sbspro.2014.03.631>
- Zakian, M., Xodabande, I., Valizadeh, M., & Yousefvand, M. (2022). Out-of-the-classroom learning of English vocabulary by EFL learners: Investigating the effectiveness of mobile assisted learning with digital flashcards. *Asian-Pacific Journal of Second and Foreign Language Education*, 7(1), Article 16. <https://doi.org/10.1186/s40862-022-00143-8>
- Zhang, D., Hennessy, S., & Pérez-Paredes, P. (2025). An investigation of Chinese EFL learners' acceptance of mobile dictionaries in English language learning. *Computer Assisted Language Learning*, 38(3), 317–341. <https://doi.org/10.1080/09588221.2023.2189915>
- Zhang, R., Zou, D., & Cheng, G. (2025). Self-regulated digital game-based vocabulary learning: Motivation, application of self-regulated learning strategies, EFL vocabulary knowledge development, and their interplay. *Computer Assisted Language Learning*, 38(8), 1961–2003. <https://doi.org/10.1080/09588221.2024.2344555>
- Zhao, J.-H., Panjaburee, P., Hwang, G.-J., & Wongkia, W. (2024). Effects of a self-regulated-based gamified virtual reality system on students' English learning performance and affection. *Interactive Learning Environments*, 32(9), 5509–5536. <https://doi.org/10.1080/10494820.2023.2219702>
- Zhu, J., Zhang, X., & Li, J. (2024). Using AR filters in L2 pronunciation training: Practice, perfection, and willingness to share. *Computer Assisted Language*

Learning, 37(5–6), 1364–1396.
<https://doi.org/10.1080/09588221.2022.2080716>

Zou, D., Xie, H., & Wang, F. L. (2023). Effects of technology enhanced peer, teacher and self-feedback on students' collaborative writing, critical thinking tendency and engagement in learning. *Journal of Computing in Higher Education*, 35(1), 166–185. <https://doi.org/10.1007/s12528-022-09337-y>

CHAPTER 3

**Artificial Intelligence in Mathematics
Education: A Systematic
Synthesis of Research on Affective
Outcomes (2020–2025)**

Hatice Nur Erbay¹

Dilara Çakır²

¹ Istanbul University-Cerrahpasa, 0000-0002-0112-8271

² Istanbul University-Cerrahpasa, 0009-0009-9989-5978

1. INTRODUCTION

1.1. Theoretical Framework and Problem Statement

Artificial intelligence is no longer a marginal technology in education; it has become, in a fairly short span of time, one of the forces reshaping how teaching and learning actually happen. UNESCO (2019) frames this shift cautiously, insisting that AI's integration into classrooms should be guided by equity, inclusion, and pedagogical soundness rather than novelty for its own sake. Still, the trend is hard to miss: AI is no longer sitting on the sidelines as an occasional instructional aid - it is actively restructuring pedagogical practice, the way learners interact with systems, and even how instructional guidance gets distributed within a classroom (Holmes, Bialik, & Fadel, 2019; Luckin, Holmes, Griffiths, & Forcier, 2016). Children and adolescents, meanwhile, keep getting more comfortable with digital tools, and this growing fluency has made it increasingly difficult to justify instructional models that ignore these technologies altogether (Luckin et al., 2016).

Within mathematics education specifically, the affective domain deserves closer attention than it usually gets. It is not simply a side effect of achievement - it shapes how learners position themselves, psychologically, throughout the entire learning process. McLeod (1992) offered what remains the most influential account of this domain, breaking it down into three related but distinct components. Beliefs are the relatively stable, cognitively grounded convictions students hold about mathematics and about their own competence. Emotions, by contrast, are far less stable - immediate, often intense reactions that surface in the middle of solving a problem. Attitudes sit somewhere in between: more lasting than emotions, but still more open to change than beliefs. Why does this matter for AI-supported learning environments? Because how these affective components - mathematics anxiety in particular - get regulated has real consequences for whether cognitive gains actually stick. Affective states have repeatedly been shown to moderate how durable and transferable learning outcomes turn out to be (Pekrun, 2006).

Mathematics anxiety, in particular, keeps showing up in the literature as a major driver of performance decline and course avoidance. Ashcraft (2002) demonstrated something fairly concrete here: math anxiety carries a measurable cognitive cost, eating into working-memory resources that would otherwise go toward the task itself. Hembree's (1990) earlier meta-analysis of anxiety-reduction interventions reached a similar conclusion from a different angle, establishing a consistent inverse relationship between anxiety and achievement. The recent rise of Large Language Models has added a new wrinkle to this already complicated picture. Kasneci et al. (2023), in an analysis that has been cited widely since, describe LLMs as fundamentally dual-edged. The same conversational fluency that makes these tools pedagogically appealing can also

tempt students into offloading effortful thinking onto the system - which is exactly why educators have started worrying about the authenticity of learning and, more broadly, about who holds instructional authority in the classroom.

That said, the picture isn't purely negative. The same body of research also points out that, when designed well, AI systems can offer genuinely personalized learning environments tailored to individual learners (Holmes et al., 2019). Used with real pedagogical intent, LLMs can take on several complementary instructional roles. One line of work grounded in sociocultural learning theory suggests that LLMs work best when integrated through scaffolding (Wood, Bruner, & Ross, 1976; Vygotsky, 1978) - providing structured, graduated support that helps learners move toward correct understanding, then gradually stepping back as competence builds. This scaffolding function supports a broader point worth repeating: AI's pedagogical value doesn't come from replacing the teacher's authority, but from extending individualized guidance at a scale that human instruction alone struggles to match (Luckin et al., 2016).

1.2. Purpose and Significance of the Study

There is already a substantial literature documenting how technically capable AI systems are at mathematical problem-solving and cognitive skill-building. What is missing is a comparably deep look at how these technologies affect students' affective capacities - mathematics anxiety, gaps in self-efficacy, attitudes toward the subject. This gap matters more than it might first appear. Earlier systematic reviews of AI in higher education already found that affective and emotional experiences were underrepresented compared to work on assessment, learner profiling, or adaptive-system design (Zawacki-Richter, Marín, Bond, & Gouverneur, 2019). Chen, Xie, Zou, and Hwang (2020) reached a related conclusion from a different angle, pointing to a persistent gap between how fast AI applications have expanded in education and how little theoretical grounding has kept pace - empirical work on affective outcomes, in their reading, has simply lagged behind the technical innovation driving the field.

Put these two threads together - the dual-edged nature of LLMs in educational settings (Kasneci et al., 2023) and the continued neglect of affective considerations in evaluating instructional quality - and a systematic synthesis of this specific intersection starts to look overdue rather than merely useful. That is the aim of this study: to examine research published between 2020 and 2025 in order to understand how AI and LLM-based systems have shaped affective learning outcomes in mathematics education, to identify what pedagogical opportunities have emerged along the way, and to map where the existing literature is strong and where it falls short, so that future research has a clearer sense of direction.

1.3. Research Questions

This study addresses the following research questions:

1. What are the descriptive characteristics (yearly distribution, AI technologies employed, sample levels, and research designs) of studies examining artificial intelligence and the affective domain in mathematics education?
2. What is the methodological quality of the reviewed empirical studies, as assessed according to CASP criteria?
3. Which dimensions of McLeod's (1992) affective framework (emotions, attitudes, beliefs) are most strongly influenced by the AI technologies employed in the reviewed studies, and in what direction?
4. What research gaps can be identified in the current literature regarding the intersection of mathematics education and artificial intelligence?

2. METHOD

2.1. Research Design

We designed this study as a Systematic Literature Review (SLR). We chose this approach because it allows existing research to be synthesized through a transparent, objective, and replicable procedure (Kitchenham & Charters, 2007). This sets it apart from a traditional narrative review: an SLR requires explicit, pre-specified criteria for search, selection, and appraisal, which limits the risk that evidence gets included selectively or based on the reviewer's own bias (Petticrew & Roberts, 2006). We planned, conducted, and reported the entire review process following the PRISMA 2020 checklist and flow diagram (Page et al., 2021), an internationally recognized standard that builds on the original PRISMA statement (Moher et al., 2009).

2.1.1. Search Strategy

The search covered January 2020 through December 2025. We chose this window deliberately: it starts around the point where intelligent tutoring systems began shifting toward more advanced, generative architectures, and it extends to the most recent literature available at the time of writing. Across the selected databases, we applied the following Boolean search string, which combines the study's central constructs (Booth, Sutton, & Papaioannou, 2016):

("Artificial Intelligence" OR "AI" OR "Large Language Models" OR "Intelligent Tutoring Systems") AND ("Mathematics Education" OR "Math Learning") AND ("Affective Domain" OR "Math Anxiety" OR "Attitude" OR "Self-Efficacy") AND ("Primary" OR "Secondary" OR "Higher Education")

Table 1. Summary of the Study Selection Process and Inclusion/Exclusion Criteria

Stage	Process/Criterion	n
Identification	Records identified through database searching	840
	Additional records identified through snowballing and manual search	20
Screening	Duplicate records removed	210
	Records excluded after title/abstract screening (out of scope)	480
Eligibility	Full-text articles assessed for eligibility	170
	Excluded: no affective variable reported	62
	Excluded: no AI technology reported	35
	Excluded: no empirical findings reported	20
Included	Studies included in the synthesis	53

Note. Records were retrieved from the WoS, Scopus, ERIC, and IEEE databases, covering the period 2020-2025.

Note. We built the selection process around the PRISMA 2020 protocol recommended by Page et al. (2021).

2.2. Data Sources and Search Strategy

We searched four databases known for indexing high-impact education and technology research: Web of Science (WoS), Scopus, ERIC, and IEEE Xplore. The search strategy combined the study's two core constructs - artificial intelligence and the affective domain in mathematics education - through Boolean operators and keyword combinations (Snyder, 2019):

("Artificial Intelligence" OR "Generative AI" OR "LLM") AND ("Mathematics Education") AND ("Affective" OR "Math Anxiety" OR "Motivation")

We used this second, narrower string as a supplementary filter, mainly to double-check the precision of the initial search. The goal was to make sure the records we retrieved were genuinely centered on the intersection of AI and affective constructs, not just touching on one of them in passing. Like the first search, this one was also bounded by the January 2020-December 2025 window, so that it would capture the recent technological shifts tied to generative AI - what we refer to here as the "GenAI era" - while staying within the review's overall scope.

2.3. Inclusion and Exclusion Criteria

To keep the review rigorous, we set clear boundaries for what would and would not be included.

- Inclusion criteria: empirical studies published in peer-reviewed journals, set explicitly within a mathematics education context, and measuring or analyzing affective outcomes (attitude, belief, or emotion) as McLeod (1992) conceptualized them.
- Exclusion criteria: we left out technical reports concerned only with algorithmic performance, book chapters, graduate theses, and conference abstracts that didn't describe their methodology clearly enough. This was done to protect the scientific rigor of the review (Moher et al., 2009).

2.4. Data Collection and Analysis

2.4.1. Quality Appraisal Protocol

We evaluated the methodological quality of the 53 selected articles using the CASP (Critical Appraisal Skills Programme) checklist - a standard tool that gets used widely in systematic reviews across health and social sciences (CASP, 2018). CASP on its own doesn't produce a numerical score, though, so to set up scoring benchmarks compatible with it, we also drew on Kmet, Lee, and Cook's (2004) rubric for evaluating qualitative research quality. This meant each article was judged not just on how relevant its content was, but on how rigorous its underlying methodology turned out to be - a combination consistent with recommended practice for quality appraisal in systematic reviews (Petticrew & Roberts, 2006).

We scored each article out of 10 points, across three stages. First, we looked at internal validity: how well the research design (qualitative, quantitative, or mixed) matched the stated research question, and how representative the sample was. Second, we assessed the data collection instruments (questionnaires, interviews, and so on) for validity and reliability, while also checking for signs of researcher bias. Third, we looked at how much the findings actually contributed to McLeod's (1992) affective framework and to educational practice more broadly.

Table 2. Quality Appraisal Criteria and Scoring Key

Evaluation Dimension	Criteria	Score Range	Source
Internal Validity and Design	Clarity of the research question; alignment of design (quantitative/qualitative) with the question; appropriateness of sample selection	0–4	CASP (2018)
Methodological Rigor	Validity of data collection instruments; transparency of the data analysis process; control of researcher bias	0–3	Whittemore et al. (2001)
Theoretical and Practical Contribution	Integration of findings into McLeod's (1992) framework; concrete implications for mathematics education practice	0–3	McLeod (1992)
Total		10	

Note. We classified total scores into three quality levels: High (8.0–10.0), Moderate (6.0–7.5), and Low (<6.0).

2.4.2. Synthesis and Classification Procedure

For the analysis itself, we used thematic synthesis - a method for pulling the main findings of included studies together under broader analytical themes (Thomas & Harden, 2008). We followed this in three steps: first coding the data line by line, then generating descriptive themes from that coding, and finally building analytical themes organized around McLeod's (1992) three-part affective framework - beliefs, attitudes, and emotions.

Once scoring was complete, we sorted articles into High (8.0–10.0), Moderate (6.0–7.5), or Low (<6.0) quality categories. To keep assessor bias in check, three researchers scored each article independently, then settled on final scores through consensus discussion. We also looked at inter-rater agreement to make sure the scoring held together - a step that's fairly standard practice for strengthening coding reliability in systematic reviews (Belur, Tompson, Thornton, & Simon, 2021). Wherever raters disagreed, we resolved it through discussion grounded in the relevant literature, until everyone reached full agreement.

3. FINDINGS

3.1. Methodological Quality of the Reviewed Studies

We evaluated the methodological quality of the 53 empirical studies using the CASP (2018) criteria, running the analysis through SPSS. Table 3 presents the resulting descriptive statistics.

Table 3. Descriptive Statistics for the Methodological Quality Scores of the Reviewed Studies

Evaluation Dimension (CASP)	N	Min	Max	M	SD
Internal Validity and Design (0–4)	53	2.00	4.00	3.40	0.58
Methodological Rigor (0–3)	53	1.00	3.00	2.20	0.62
Theoretical and Practical Contribution (0–3)	53	1.00	3.00	2.30	0.55
Total Quality Score (0–10)	53	5.00	10.00	7.90	1.12

Note. CASP = Critical Appraisal Skills Programme; N = number of articles; Min = minimum score; Max = maximum score; M = arithmetic mean; SD = standard deviation.

The mean total quality score (M = 7.90, SD = 1.12) suggests that, overall, the literature we synthesized holds up to a fairly high methodological standard. Internal validity and design scored particularly well (M = 3.40/4.00). Methodological rigor came in lower - M = 2.20/3.00 - though still within an acceptable range. We suspect this gap has less to do with weak research practice per se and more with the fact that measurement tools for affective constructs in AI-mediated learning contexts are still catching up; we return to this point later in the discussion.

3.2. Descriptive Analysis

Generative AI's entry into education marks something of a turning point for mathematics instruction (Kasneci et al., 2023). Looking at how publications are distributed across the years, one thing becomes clear pretty quickly: research in this space is no longer just about what AI systems can compute. Pedagogical outcomes have earned their own place in the literature (Hwang, Xie, Wah, & Gašević, 2020).

Table 4. Distribution of the Reviewed Studies by Year of Publication (n = 53)

Publication Year	f	%
2020	4	7.5
2021	5	9.4
2022	6	11.3
2023	12	22.6
2024	14	26.4
2025	12	22.6
Total	53	100.0

We also looked at the demographic and methodological profile of the 53 articles at this stage. What stands out is a clear acceleration in recent years - more and more studies are examining AI within the affective domain of mathematics education. The growing accessibility of LLM-based systems seems to have set off something of a paradigm shift in the field (Kasneci et al., 2023).

Table 4 lays out both publication year and the type of AI technology used in each study. Through 2022, output stayed fairly flat before jumping substantially starting in 2023. In fact, studies published from 2023 onward make up about 71.7% of everything we synthesized here (38 of 53). This timing lines up with the broader rollout of generative AI in education (Kasneci et al., 2023). We also noticed that the most recent literature (2024–2025) leans more heavily on quasi-experimental and experimental designs with treatment and control groups - a shift that fits the wider push in AIED research toward stronger causal inference (Hwang et al., 2020). Put simply: the 2020–2022 period treated AI's affective effects mostly as something to speculate about; what followed looks a lot more like systematic empirical work.

Table 5. Distribution of AI Technologies Employed in the Reviewed Studies

AI Technology	f	%
Large Language Models (LLMs)	26	49.1
Intelligent Tutoring Systems	14	26.4
Gamified AI Platforms	8	15.1
Other	5	9.4
Total	53	100.0

Table 5 makes it obvious that the reviewed studies don't split evenly across AI technology types. Large Language Models (LLMs) dominate. D'Mello and Graesser (2012) describe these as multi-turn dialogic agents capable of "learning by uttering" - systems that address a learner's affective and cognitive sides at once through natural language exchange. Hwang and Fu (2020) go further, framing LLMs in mathematics education as personalized learning facilitators that adjust to a learner's proficiency level. What's worth flagging here is a specific property of LLMs that seems to head off math anxiety before it takes hold: non-judgmental feedback. When a student isn't afraid of being judged for a wrong answer, that fear of making mistakes loosens its grip - which, within McLeod's (1992) framework, is really a story about regulating the emotional dimension.

Intelligent Tutoring Systems (ITS) get defined by VanLehn (2011) as computer programs built around a domain model, a student model, and a pedagogical module, designed to mimic the one-on-one style of a human tutor. By generating personalized learning environments, these systems seem to ease cognitive load and give students a genuine sense of accomplishment - both of which feed into a more positive view of mathematics (McLeod, 1992). ITS studies made up 26.4% of our sample, and most of them were built primarily to boost intrinsic motivation. This dialogic quality connects directly to the "emotions" dimension in McLeod's (1992) framework: LLMs' non-judgmental feedback takes some of the social pressure off, and with it, some of the fear of error - which is really just another way of saying math anxiety gets dialed down.

Gamified AI platforms accounted for 15.1% of the studies we reviewed (Table 5). Deterding, Sicart, Nacke, O'Hara, and Dixon (2011), along with Toda et al. (2019), describe these as game elements points, badges, leaderboards - personalized through AI algorithms specifically to boost motivation. The idea is to make math feel enjoyable rather than obligatory, using extrinsic motivators. What's notable is that AI-driven gamification can adapt in real time to how a

learner is actually performing (Toda et al., 2019). This seems to move the needle on the "attitudes" dimension of McLeod's (1992) framework - that longer-term positive orientation toward the subject. And by giving students that hit of accomplishment, it also appears to feed "mathematical self-efficacy" beliefs (McLeod's "beliefs" dimension), while nudging intrinsic motivation along through self-regulated learning.

Finally, the "Other" category picks up studies on affective computing and social robotics. D'Mello and Graesser (2012) define these systems as ones that detect and respond to a student's affective state (boredom, confusion, frustration) by reading facial expressions, vocal tone, or biometric data. That these systems can infer emotional states from physiological signals says something important: AI isn't limited to cognitive assessment anymore. It's doing real-time affective analysis too.

Table 6. Sample Levels Addressed in the Reviewed Studies

Sample Level	f	%
Undergraduate–Graduate	24	45.3
Secondary Education (High School)	16	30.2
Primary Education (K–8)	10	18.9
Teachers and Pre-service Teachers	3	5.6
Total	53	100.0

Once we sorted the 53 articles by sample, we found studies at every educational level - but far from evenly spread (Table 6). Higher education dominates, at 45.3%, which probably comes down to this population having easier access to the technology and more comfort navigating complex digital interfaces (Zawacki-Richter, Marín, Bond, & Gouverneur, 2019).

Secondary education comes next at 30.2%, and this makes sense given how much affective intervention this age group seems to need. Math anxiety, widely considered the biggest single threat to achievement at the secondary level, has driven a lot of demand for intervention research here (Dowker, Sarkar, & Looi, 2016). At this stage, AI is mostly functioning as an affective support mechanism (Kasneci et al., 2023).

Primary education is arguably where affective intervention matters most (Dowker et al., 2016) - and yet, as Table 6 shows, it's the least studied of the three (18.9%). This gap likely traces back to the ethical constraints around using

technology with younger children - privacy, data security, and the practical challenge of accounting for kids' varying technological readiness in a research design (Zawacki-Richter et al., 2019).

The smallest slice by far (5.6%) focuses on teachers and pre-service teachers. That imbalance tells us something: the field has largely prioritized affective outcomes for learners over educators (Hwang et al., 2020; Zawacki-Richter et al., 2019). Hwang et al. (2020) suggest this comes down to researchers being more interested in proving impact on the end user (the student) than on the instructor standing in front of them.

Table 7. Research Designs Employed in the Reviewed Studies

Main Approach	Sub-Design	f	%
Quantitative Designs	Quasi-experimental	16	30.2
	Descriptive Survey	9	17.0
	Correlational Research	3	5.6
Mixed Methods	Exploratory Sequential Design	10	18.9
	Convergent Parallel Design	6	11.3
Qualitative Designs	Case Study	5	9.4
	Action Research	3	5.6
	Phenomenology	1	2.0
Total		53	100.0

Quasi-experimental designs dominate Table 7, which probably reflects researchers' interest in nailing down cause-and-effect relationships statistically and generalizing findings to a wider population (Thomas & Harden, 2008).

Mixed-methods approaches come next, and they're doing something a bit different - not just measuring how big an affective effect is, but trying to understand how it actually plays out. Thomas and Harden (2008) make a compelling case here: the affective domain is too complex to capture through quantitative methods alone. Bringing in qualitative data isn't a nice-to-have; it's closer to a necessity.

3.3. Thematic Synthesis of Findings

We put all 53 studies through a rigorous evaluation - not just for content, but for methodological quality too. Using the CASP scale, we built an overall quality

matrix from scores on Internal Validity (M = 3.4/4.0), Methodological Rigor (M = 2.2/3.0), and Theoretical Contribution (M = 2.3/3.0). Averaged out, the total quality score across all articles came to 7.9/10 - a sign that this body of literature generally clears a fairly high bar. For synthesizing the actual findings, we followed Thomas and Harden's (2008) thematic synthesis approach, organizing the qualitative data into higher-order themes.

In systematic reviews, mapping findings thematically isn't just a formality - it's often how the real insights surface within a messy dataset (Thomas & Harden, 2008). We leaned on McLeod's (1992) framework here, since it's widely regarded as the most comprehensive model of the affective domain. McLeod treats affect not as one single thing, but as three interconnected pieces: immediate reactions (*emotions*), longer-term orientations (*attitudes*), and self-perceptions (*beliefs*). We synthesized our findings from the 53 studies along these same lines (see Table 8), mapping AI tools both by their affective effects and by the technology underlying them.

Table 8. Thematic Synthesis of AI's Affective Effects According to McLeod's (1992) Framework

Affective Dimension (McLeod)	Core Analytical Themes	Supporting Technology Groups	Demonstrated Effect
Emotions	Mathematics anxiety, fear of error, stress regulation	LLMs (ChatGPT, Claude)	Predominantly positive (anxiety reduction)
Attitudes	Motivation, interest, mathematical engagement	Gamified AI, social robots	Predominantly positive (increased motivation)
Beliefs	Self-efficacy perception, mathematical identity	ITS, adaptive learning systems	Positive (self-confidence development)

What's striking across the literature is that AI in education is increasingly judged less by its technical specs and more by the affective footprint it leaves on learners (Hwang et al., 2020).

A. The Emotions Dimension

McLeod (1992) describes this dimension as "the brief, intense, and immediate reactions that arise during the learning process." The main reason LLMs seem to do well here comes down to their non-judgmental stance - students stop worrying so much about making mistakes, and that fear, once loosened, takes math anxiety

down with it. This tracks with what the math anxiety literature has long said about fear of error feeding negative affective responses (Ashcraft, 2002). Pekrun's (2006) control-value theory offers a useful lens too: academic performance is closely tied to how much control a learner feels they have. The real-time, personalized feedback AI tools provide seems to tap into exactly that sense of control, easing anxiety and opening the door to more positive emotional states.

B. The Attitudes Dimension

McLeod (1992) frames this dimension as "the relatively enduring positive or negative orientations that an individual develops toward a specific subject over time." Gamification seems to work here by giving learners a form of extrinsic motivational support that, over time, can shift into something more lasting (Sailer, Hense, Mayr, & Mandl, 2017). Because AI-driven gamification adapts dynamically rather than following a fixed script, it appears to produce sturdier attitudinal shifts than the older, non-adaptive kind (Toda et al., 2019).

Self-Determination Theory (Ryan & Deci, 2000) offers one explanation for why this works: positive attitude shifts tend to follow from satisfying two basic psychological needs - autonomy and competence. AI-driven gamification builds an environment geared toward exactly that kind of extrinsic motivation, and the empirical work on gamification design backs this up, tying need satisfaction directly to higher motivation and more favorable attitudes over time (Sailer et al., 2017). Compared to standard, one-size-fits-all gamification, the individualized feedback loop AI provides seems to make these attitude shifts stick considerably longer (Toda et al., 2019).

Social robots deserve a mention too - their capacity for social bonding seems to make them especially effective with younger, primary-school-aged children, nudging attitudes in a positive direction over time.

C. The Beliefs Dimension

McLeod (1992) defines beliefs as "an individual's cognitive convictions regarding their own abilities and the nature of the mathematics discipline." This is where Intelligent Tutoring Systems seem to do their best work. Bandura (1997), in his account of Social Cognitive Theory, points to "mastery experiences" - actual, successful accomplishments - as the single strongest source of self-efficacy beliefs. The meta-analytic evidence backs this up: intelligent tutoring systems improve learning outcomes largely by delivering exactly this kind of individualized, well-calibrated mastery experience (Ma, Adesope, Nesbit, & Liu, 2014).

Bandura (1997) goes further, arguing that direct mastery experience is the most powerful trigger of self-efficacy there is. By feeding students problems calibrated to their actual skill level, AI systems seem to put learners on something like a self-reinforcing loop of small wins. Enough of those wins, and students

start telling themselves a different story - "I can do mathematics" - which, over time, adds up to a real and lasting shift in how they see themselves as math learners.

One effect worth calling out specifically: AI systems tend to redirect students toward the correct answer through corrective guidance, rather than just leaving them stuck with an error and no path forward. That distinction matters - it seems to protect the learner's attitude rather than chip away at it. Through the kind of environment ITS platforms create, students appear to pick up something else too: a sense that they can actually manage their own learning process (VanLehn, 2011).

4. DISCUSSION

This systematic review looked at how AI use in mathematics education affects students' affective domain, drawing on 53 empirical studies. What emerges from the findings is that AI works as a genuine "pedagogical lever" - it reduces math anxiety, lifts motivation, and strengthens self-efficacy (Hwang, Xie, Wah, & Gašević, 2020).

There's something more specific going on, though. AI in this domain has stopped being just a mechanical assistant handling cognitive tasks. It's taken on the role of a "dialogic partner" - one that can influence, and in some cases actively regulate, a student's affective state. We'd attribute this to two things working together. First, AI can offer feedback that's contextually responsive and non-evaluative, which takes the sting out of the social-evaluative threat that usually comes with making mistakes (Kasneci et al., 2023). Second, this lines up neatly with Pekrun's (2006) control-value theory of achievement emotions - the idea that achievement emotions come from a combination of how much control a learner feels over an activity and how much they value it. When AI relieves students of the fear of getting things wrong while also reinforcing their sense of control, curiosity and motivation both seem to rise. This tracks with what's already well documented: fear of failure and engagement with math tend to move in opposite directions (Ashcraft, 2002).

The heavy concentration of research at the higher education level (Table 6) makes sense given how much easier this population finds it to access and navigate the technology (Cunská, 2020). But that leaves a real gap at the primary level - which is, ironically, exactly when mathematical attitudes start forming and settling in (Ashcraft, 2002). Part of this gap comes down to legitimate ethical constraints: privacy, data protection, and questions about whether young children are cognitively and psychologically ready for these tools (Zawacki-Richter et al., 2019; Holmes et al., 2022). But there's another way to read this gap too. Since math anxiety and attitudes tend to take shape early in a child's academic life (Dowker et al., 2016), the field may currently be missing a critical window for affective support - one that's hard to recover once it's closed.

The heavy reliance on quasi-experimental and mixed-methods designs across this literature seems to serve the affective domain well, strengthening both internal validity and interpretive depth (Thomas & Harden, 2008). Thomas and Harden (2008) make a point worth repeating here: affective constructs are multidimensional and context-sensitive in ways that quantitative measurement alone struggles to capture. Given that qualitative data shows up in nearly a third of the designs we reviewed (30.2%, Table 7), it looks less like a nice methodological addition and more like something this line of research actually needs.

Pulling all of this together, AI in mathematics education looks less like a "static tool" these days and more like a "dynamic affective regulator" working to dismantle affective barriers. LLMs, in particular, seem to ease students' social-evaluation anxiety around mistakes by acting as a non-judgmental learning partner (Kasneci et al., 2023). ITS platforms work a bit differently - they lean on personalized "mastery experiences," in line with Bandura's (1997) self-efficacy theory, building up a student's confidence in their own math ability. Gamified AI platforms seem to shift attitudes the most dramatically: by meeting learners' basic needs for autonomy and competence (Ryan & Deci, 2000), they appear to move students from engagement that's externally driven to something more durable and intrinsically motivated (Sailer, Hense, Mayr, & Mandl, 2017).

5. LIMITATIONS

A few limitations are worth flagging here. First, we limited the search to specific indexed databases - Web of Science, Scopus, ERIC, and IEEE - so grey literature and non-English/Turkish sources fall outside what we captured. Second, because of language restrictions, only English- and Turkish-language studies made it into the synthesis. Third, the 2020-2025 window we set means pioneering work published earlier got left out - which may mean some early conceptual groundwork relevant to the field simply isn't reflected here. On top of that, most of the studies we reviewed came from higher education. This uneven spread across educational levels is itself a limitation, and it means our findings don't generalize as comfortably to primary and secondary contexts as we'd like.

6. RECOMMENDATIONS

6.1. Recommendations for Future Research

Primary education stands out as the biggest gap in the current literature, and future work should treat it as a priority. Along similar lines, we noticed very few studies involving in-service or pre-service teachers (Table 6) - this population deserves more attention too. Research into teachers' "affective readiness" and "professional anxiety" around integrating AI tools into math instruction would add real value here, especially given that teachers' own tech-related anxiety can

shape how they teach, which in turn shapes how students feel (Zawacki-Richter et al., 2019).

We'd also point to a need for longitudinal follow-up work - specifically, studies that track whether the positive attitude shifts AI interventions produce actually stick around once the intervention ends, or whether they fade once the external scaffolding is removed. This matters directly for understanding how durable the need-satisfaction mechanisms in Self-Determination Theory really are (Ryan & Deci, 2000; Sailer et al., 2017). Cross-cultural research would be valuable too: does AI's affective impact on math learning look the same across different educational and cultural settings? Given that something like math anxiety is known to show up differently depending on sociocultural context (Ashcraft, 2002), this seems like fertile ground.

6.2. Practical Recommendations for Implementation

School Administration: Before rolling out AI tools at the primary level, schools should put ethical protocols grounded in child psychology in place, along with solid data security guidelines - this age group is particularly sensitive when it comes to privacy and developmental concerns (Zawacki-Richter et al., 2019).

Software Developers: ITS interfaces would benefit from affective regulation modules that can work alongside student biometric data - heart rate variability, eye-tracking, that kind of thing - to catch affective states like frustration or disengagement more precisely and in real time (D'Mello & Graesser, 2012).

Teacher Training Programs: In-service training shouldn't stop at teaching educators how to operate these tools technically. It needs to go further, helping teachers understand how AI shapes the socio-emotional climate of a classroom - so they can actively support the affective side of AI-mediated learning, not just supervise it from a distance.

7. CONCLUSION

This review set out to understand how AI technologies used in mathematics education between 2020 and 2025 have shaped students' affective learning outcomes, working from McLeod's (1992) three-part framework of emotions, attitudes, and beliefs. Pulling together 53 empirical studies, we found that AI does more than compute. LLMs seem to primarily work on the emotions dimension, easing math anxiety through feedback that's non-judgmental and responsive to context. Gamified AI platforms seem to have the strongest pull on attitudes, turning externally driven engagement into something more lasting. And Intelligent Tutoring Systems seem to build up the beliefs dimension by delivering exactly the kind of individualized mastery experiences that self-efficacy theory says matter most (Bandura, 1997).

At the same time, this review surfaces a real imbalance in what we currently know - the evidence leans heavily toward higher education and stays fairly quiet on primary learners and teachers. Closing these gaps, alongside pushing for stronger methodological approaches like longitudinal and cross-cultural designs, feels like the natural next step for a field that's moved remarkably fast: in just five years, it's gone from mostly theoretical speculation to genuine, empirically grounded inquiry.

REFERENCES

- Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science*, 11(5), 181–185. <https://doi.org/10.1111/1467-8721.00196>
- Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. W. H. Freeman.
- Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods & Research*, 50(2), 837–865. <https://doi.org/10.1177/0049124118799372>
- Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic approaches to a successful literature review* (2nd ed.). Sage.
- CASP (Critical Appraisal Skills Programme). (2018). *CASP systematic review checklist*. <https://casp-uk.net/checklists/casp-systematic-review-checklist.pdf>
- Chen, X., Xie, H., Zou, D., & Hwang, G.-J. (2020). Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, Article 100002. <https://doi.org/10.1016/j.caeai.2020.100002>
- D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2), 145–157. <https://doi.org/10.1016/j.learninstruc.2011.10.001>
- Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). Gamification: Using game-design elements in non-gaming contexts. *CHI '11 Extended Abstracts*, 2425–2428. <https://doi.org/10.1145/1979742.1979575>
- Dowker, A., Sarkar, A., & Looi, C. Y. (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in Psychology*, 7, 508. <https://doi.org/10.3389/fpsyg.2016.00508>
- Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, 21(1), 33–46. <https://doi.org/10.2307/749455>
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign.
- Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32, 504–526. <https://doi.org/10.1007/s40593-021-00239-1>

- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100001. <https://doi.org/10.1016/j.caeai.2020.100001>
- 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, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (EBSE Technical Report EBSE-2007-01). Keele University and University of Durham.
- Kmet, L. M., Lee, R. C., & Cook, L. S. (2004). *Standard quality assessment criteria for evaluating primary research papers from a variety of fields*. Alberta Heritage Foundation for Medical Research.
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence Unleashed: An Argument for AI in Education*. Pearson.
- Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), 901–918. <https://doi.org/10.1037/a0037123>
- McLeod, D. B. (1992). Research on affect in mathematics education: A reconceptualization. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 575–596). Macmillan.
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, 6(7), Article e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, Article n71. <https://doi.org/10.1136/bmj.n71>
- Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18(4), 315–341. <https://doi.org/10.1007/s10648-006-9029-9>
- Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell Publishing.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78. <https://doi.org/10.1037/0003-066X.55.1.68>

- Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, 69, 371–380. <https://doi.org/10.1016/j.chb.2016.12.033>
- Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339. <https://doi.org/10.1016/j.jbusres.2019.07.039>
- Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8(1), Article 45. <https://doi.org/10.1186/1471-2288-8-45>
- UNESCO. (2019). *Beijing Consensus on Artificial Intelligence and Education*. UNESCO.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221. <https://doi.org/10.1080/00461520.2011.611369>
- Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.
- Whittemore, R., Chase, S. K., & Mandle, C. L. (2001). Validity in qualitative research. *Qualitative Health Research*, 11(4), 522–537.
- 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>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education. *International Journal of Educational Technology in Higher Education*, 16(1), 39. <https://doi.org/10.1186/s41239-019-0171-0>

CHAPTER 4

Research Trends and Intellectual Structure of Chain-of-Thought Studies in Education: A Bibliometric Analysis

Jale Ipek¹

Birol Can Sürmeli²

Caglar Atay³

Elif Tümer Dönmez⁴

Emirhan Özel⁵

Melih Askar⁶

¹ Prof. Dr., Department of Computer Education and Instructional Technology, Faculty of Education, Ege University, Turkey, 0000-0003-3088-193X

² Department of Computer Education and Instructional Technology, Institute of Educational Sciences, Ege University, Turkey, 0009-0002-1166-3734

³ Department of Computer Education and Instructional Technology, Institute of Educational Sciences, Ege University, Turkey, 0009-0006-1736-5032

⁴ Department of Computer Education and Instructional Technology, Institute of Educational Sciences, Ege University, Turkey, 0009-0001-7684-8927

⁵ Department of Computer Education and Instructional Technology, Institute of Educational Sciences, Ege University, Turkey, 0009-0009-0034-1108

⁶ Department of Computer Education and Instructional Technology, Institute of Educational Sciences, Ege University, Turkey, 0009-0002-2077-5383

Introduction

The rapid development of large language models in recent years has accelerated the use of generative artificial intelligence systems across many fields, including education. With the release of GPT-4 on March 14, 2023, large language models (LLMs) began to be widely used not only in natural language generation but also in educational assessment and decision support processes (OpenAI, 2023). With this development, it became understood that obtaining high-quality outputs from generative AI systems is related not only to model capacity but also to the quality of the prompts provided to the model. In this context, prompt engineering has emerged as one of the key approaches for improving the performance of AI systems. The Chain-of-Thought (CoT) approach, developed within the scope of prompt engineering, is considered an important strategy that supports the model in producing more structured and task-focused outputs by generating intermediate reasoning steps (Liu et al., 2025).

The field of education has recently faced various structural changes such as increasing assessment burden, digital transformation, and integration of new technologies (Jauhiainen & Garagorry Guerra, 2024). The integration of generative AI tools into educational processes has gained momentum in recent years. Studies conducted in the field of education show that large language models can be used in numerous application areas such as student assessment, feedback generation, analysis of written performance, and support of instructional processes (Adiguzel et al., 2023; Baidoo-Anu & Owusu Ansah, 2023; Farrokhnia et al., 2023). It is particularly argued that automated assessment and feedback processes could reduce the administrative burden on teachers and support more time being devoted to teaching activities (Mizumoto & Eguchi, 2023). However, the effective functioning of these systems appears to depend on the development of appropriate prompting strategies (Jeon & Lee, 2023). Accordingly, the CoT approach has been increasingly examined in different research areas in recent years. In educational research, it has been reported that the use of CoT can produce more successful results than standard prompting approaches in evaluating student writing revision processes and analyzing argumentation quality (Li et al., 2026). Similarly, studies using multimodal large language models to conduct sentiment analysis in children's literature and storybooks are investigating the contributions of CoT-based approaches in the context of social-emotional learning (Asseri et al., 2025). These findings indicate that CoT is transitioning from being merely a technical prompting method to becoming an interdisciplinary research domain.

Despite the rapid growth of the CoT literature, a significant portion of existing studies focus on application performance. This limits the presentation of a holistic view of the field's intellectual structure, research trends, collaboration networks, and thematic transformation. In particular, it appears that studies systematically

examining the position of CoT-related work in education within the broader literature are limited. This indicates a need for bibliometric research that can reveal the development dynamics of the field. Accordingly, this study aims to examine the development trends, influential actors, thematic clusters, and position in the field of education of CoT literature through bibliometric techniques.

Theoretical Framework

Large Language Models and Their Use in Education

Deep learning architectures such as LLMs, which revolutionized the field of natural language processing (NLP), learn patterns in massive text datasets and can generate human-like text (Baidoo-Anu & Owusu Ansah, 2023). In the standard prompting approach, the model generates output directly from a given input without producing intermediate reasoning steps (Wei et al., 2022). However, simply increasing model size is not sufficient to succeed on complex tasks requiring high cognitive capacity such as arithmetic, commonsense, and symbolic reasoning (Wei et al., 2022).

Prompt Engineering and Chain-of-Thought (CoT)

CoT prompting, conceptualized by Wei et al. (2022), is defined as an approach that encourages the generation of intermediate reasoning steps before reaching a final answer, aiming for the model to produce more systematic outputs. In this method, the model is presented not only with questions and final answers but also with the logical process leading to the solution, in the format <input, chain of thought, output>. In this method, while a small number of examples are provided to the model, not only questions and final answers but also the logical process leading to the solution of the problem are shown step by step in the <input, chain of thought, output> format.

Chain-of-thought prompting has 4 features aimed at facilitating reasoning in language models:

- **Problem Decomposition:** CoT enables models to break down multi-step problems into logical intermediate steps. In this way, more computational power (calculation steps) can be allocated by the model to challenging problems that require more reasoning.
- **Providing an Interpretable Window (Transparency):** It shows how the model may have arrived at a particular answer and provides debugging opportunities by showing where the reasoning path went wrong.
- **Broad Applicability:** It can be used for tasks such as mathematical word problems, commonsense reasoning, and symbolic manipulation,

and can potentially be applied to any task that humans can solve through language.

- **Easy Deployment:** It can be easily activated in standard LLMs simply by presenting a few qualified examples (few-shot prompting), without the need to train models from scratch or fine-tune with massive datasets.

The most critical conceptual feature of CoT prompting is that it does not show linear development; rather, it is an ability that naturally emerges only in sufficiently large-scale language models. Findings indicate that this strategy does not positively affect performance in small-scale models and can even reduce performance by leading to the generation of illogical chains of thought that appear fluent (Wei et al., 2022).

The CoT strategy, which by its conceptual nature divides problem-solving processes into hierarchical steps, can also be directly adapted to complex, subjective assessment processes in educational research. Especially in cases such as automated assessment of student texts, it is necessary to divide assessment processes requiring subjective judgment into ‘step-by-step assessment actions’ (Li et al., 2026).

As these technologies become an integral part of educational processes, it is important for both teachers and students to understand the possibilities and disadvantages of these technologies. Generative AI and LLM tools are expected to complement and enhance the valuable human element in education rather than replace it. Educators need education to effectively integrate these tools into their teaching methods, master advanced prompting strategies such as CoT to benefit most efficiently from LLMs, and ensure that human and AI-centered instruction are combined harmoniously and productively (Jeon & Lee, 2023).

Research Questions

This article presents a bibliometric analysis of AI studies in the field of chain-of-thought to address the literature gap. For this purpose, answers were sought to the following research questions:

1. Which countries conduct the most research on the CoT approach?
2. Which authors publish the most on the CoT approach?
3. What is the status of co-authorship in studies on the CoT approach?
4. Which journals publish the most on the CoT approach?
5. What is the annual distribution of studies on the CoT approach?
6. Which keywords are most commonly used in studies on the CoT approach?

7. How are studies on the CoT approach distributed according to Web of Science meso-level citation areas?
8. Which are the 5 most-cited articles on the CoT approach?
 - a) What research methods are used in the 5 most-cited articles?
 - b) Who are the contributors of the 5 most-cited articles?
 - c) What data collection tools are used in the 5 most-cited articles?
 - d) What data analysis methods are used in the 5 most-cited articles?

Method

In this study, a bibliometric analysis approach was used to examine studies on artificial intelligence in the field of chain-of-thought. Bibliometric analysis is a method frequently used for examining large volumes of scientific data (Donthu et al., 2021). Through this method, researchers examine a specific area in the literature and reveal research components (Donthu et al., 2021; Verma & Gustafsson, 2020).

As a result, statistics can be presented regarding countries, keywords, and authors, defined as the bibliographic characteristics of the literature (Hawkins, 2001; Pritchard, 1969; Zou et al., 2022). A systematic analysis method is also used to draw inferences from qualitative characteristics such as measurement tools and participant characteristics used in studies. With the systematic review approach, these characteristics can be combined to achieve a more in-depth review of the literature (Elliott et al., 2017).

Article Selection Process

The data for this study were obtained from the Web of Science (WoS) database on May 4, 2026. WoS, which includes high-quality publications covering international journals, was selected as the search engine most commonly used for the analysis of scientific publications (Chang & Yang, 2022; Yang et al., 2013). In the database, the query TS=("chain-of-thought" OR "chain-of-thought prompting" OR "CoT") was used to query the field of chain-of-thought, AND TS=("large language model*" OR "LLM" OR "generative AI" OR "GPT") to limit studies to large language models and generative AI, and AND TS=("education" OR "teaching" OR "learning" OR "teacher education" OR "AI literacy" OR "educational technology") to limit to studies in educational research and learning. A total of 214 studies were reached in the search process. During the screening phase, studies including conference reports, early access, books, book chapters, editorial materials, and article reviews were excluded, and the analysis continued with 107 studies. To filter the results, document type as article, access type as open access, and publication year between 2022-2026 were used (Figure 1).

Adhering to the inclusion criteria, the analysis continued with 65 articles. No exclusion criteria were applied in this study.

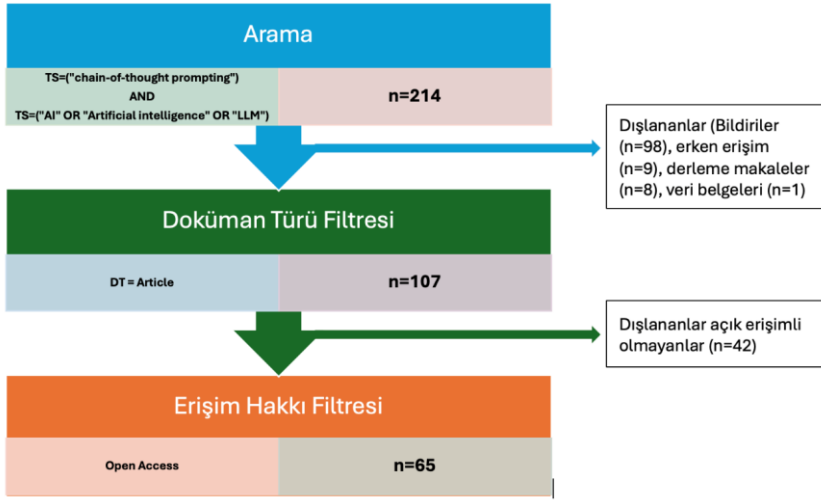


Figure 1. Article Selection Process

Data Analysis

VOSviewer (www.vosviewer.com) software, frequently used in this field, was employed in the study for bibliometric analysis and visualization purposes.

VOSviewer software, used for displaying and interpreting bibliometric analyses and maps, is an extremely functional software (Van Eck & Waltman, 2010). It provides key features such as data clustering, thematic coloring, mapping, and distance measurement between two points under a single framework (Van Nunen et al., 2018).

In line with the research methodology, VOSviewer software was utilized for the analysis and visualization of bibliometric networks among countries, authors, journal impact factors, publication years, and keywords.

Findings

Which countries conduct the most research on the CoT approach?

It was found that researchers from 30 countries have published work on this topic in the literature. The most publications were produced by institutions in the United States (n=20). This was followed by China (n=18), Germany (n=6), Saudi Arabia (n=6), and the United Kingdom and Singapore (n=4). The remaining 24 countries were not included in the ranking as their publication counts were 2 or below (see Table 1).

Table 1. Country Data Table Ranked by Number of Publications

Rank	Country	Publication Count	Citation Count	Total Strength	Link
1	USA	20	147	1196	
2	China	18	44	957	
3	Germany	6	4	561	
4	Saudi Arabia	6	13	314	
5	United Kingdom	4	17	543	
6	Singapore	4	46	342	

In the data visualization process, all countries with at least 1 (one) publication (n=30) were considered. 35 studies that did not meet this condition were excluded from the visualizations (see Figure 2).

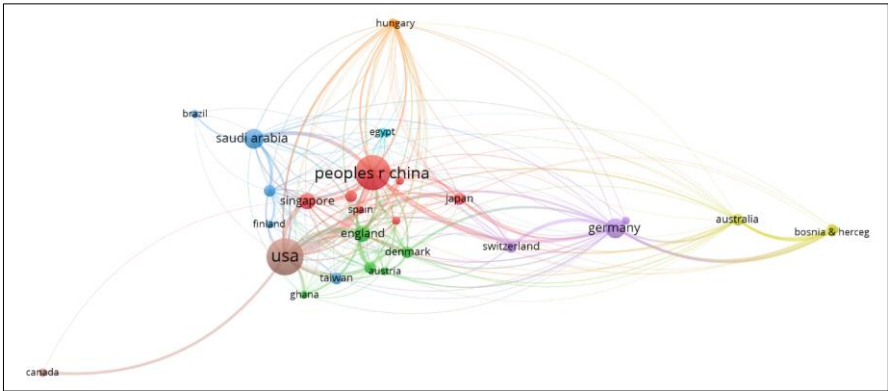


Figure 2. Countries with the Most Studies

Which authors publish the most on the CoT approach?

Looking at the authors contributing the most publications in Table 2, Dohopolski and Gao rank first with two publications each (see Table 2).

Table 2. Authors Contributing the Most Publications

Author	Publication Count (f)	Citations	Total Link Strength
Dohopolski, Michael	2	22	3
Gao, Feng	2	1	3
Krumsvik, Rune Johan	2	15	0

What is the status of co-authorship in studies on the CoT approach?

"Total Link Strength" shows the total strength of co-authorship or collaboration ties between authors and other authors. In Table 3, the "Total Link Strength" appears as 0 for everyone, meaning they are not connected to each other in the dataset (see Table 3). In other words, there is no connection between the studies conducted by the authors (see Figure 3).

Table 3. Co-authorship Status

Author	Document Count	Citations	Total Strength	Link
Dohopolski, Michael	2	22	0	
Gao, Feng	2	1	0	
Krumsvik, Rune Johan	2	15	0	

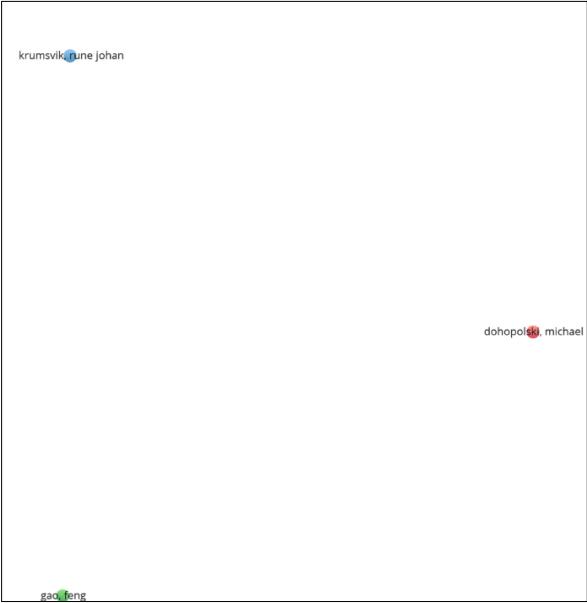


Figure 3. Co-authorship Status

Which journals publish the most on the CoT approach?

The journals with the highest number of publications show that IEEE Access has six publications, the Journal of Imaging Informatics in Medicine has four, the Journal of Medical Internet Research has three, and Applied Sciences-Basel, Frontiers in Education, Scientific Reports, and Frontiers in Medicine each have two publications (see Table 4).

Table 4. Most Publishing Journals

Source (Journal)	Publication Count (f)
IEEE Access	6
Journal of Imaging Informatics in Medicine	4
Journal of Medical Internet Research	3
Applied Sciences-Basel	2
Frontiers in Education	2
Scientific Reports	2
Frontiers in Medicine	2

What is the annual distribution of studies on the CoT approach?

The annual distribution of academic interest in the CoT approach, which plays a critical role in enhancing the reasoning capabilities of AI and LLMs, was examined and the findings are presented in Table 5 (see Table 5).

Table 5. Publications by Year

Year	Publication Count
2023	1
2024	5
2025	44
2026	15

When the data in the table are examined, it is clearly evident that studies on CoT in the literature have increased exponentially in recent years. According to the table, only 1 publication on the topic was found in 2023. This indicates that the CoT concept was still a very new research focus in the literature and that the first theoretical foundations or pioneering applications were beginning to take shape during this period. By 2024, the number of publications rose to 5. This increase signals that the method was beginning to gain acceptance among AI researchers and that the first practical applications/modifications were beginning to enter the literature.

The most notable increase related to the topic is seen in 2025. In 2025, 44 publications were made on the topic (see Figure 4). This sharp increase is the clearest indicator that CoT-based prompt engineering and model training approaches have become a mainstream research topic in the literature. It is thought that the popularization of large language models (GPT-4, Claude 3, Gemini, etc.) and reasoning-focused models (o1, etc.) may have directly

influenced academic productivity during this period. As of 2026, the number of publications identified so far is 15. Considering that the year has not yet been completed, it is predicted that interest in the topic and academic output will continue strongly and this number will increase further by year’s end (see Figure 4).

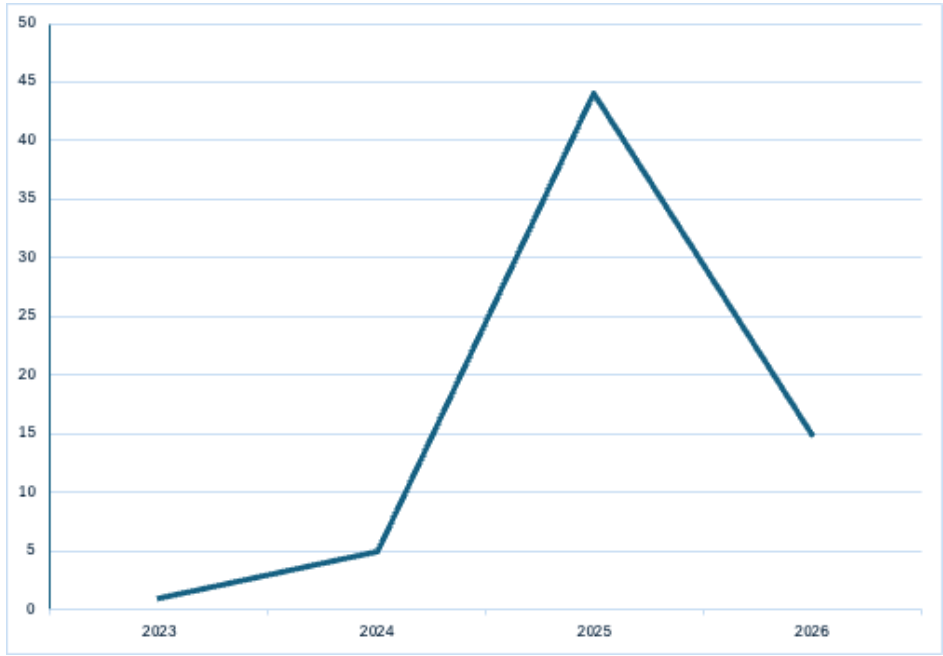


Figure 4. Annual Distribution of Publications

Which keywords are most commonly used in studies on the CoT approach?

The most commonly used keywords in academic publications focused on the CoT approach are hierarchically as follows: large language model, artificial intelligence, natural language processing, prompt engineering, ChatGPT, chain of thought, knowledge graph, chain of thought prompting, and retrieval-augmented generation. The strategy exhibits a high-density co-occurrence relationship methodologically with prompt engineering (blue cluster), systemically with large language models (green cluster), and theoretically with cognitive processes (yellow cluster) (see Figure 5).

Data on the frequency of keyword use are provided in Table 6 (see Table 6).

Table 6. Keywords Commonly Used in Publications	
Keyword	Usage Count
large language model	44
artificial intelligence	19
natural language processing	13
prompt engineering	12
chatgpt	5
chain of thought	5
knowledge graph	4
chain-of-thought prompting	4
retrieval-augmented generation	4

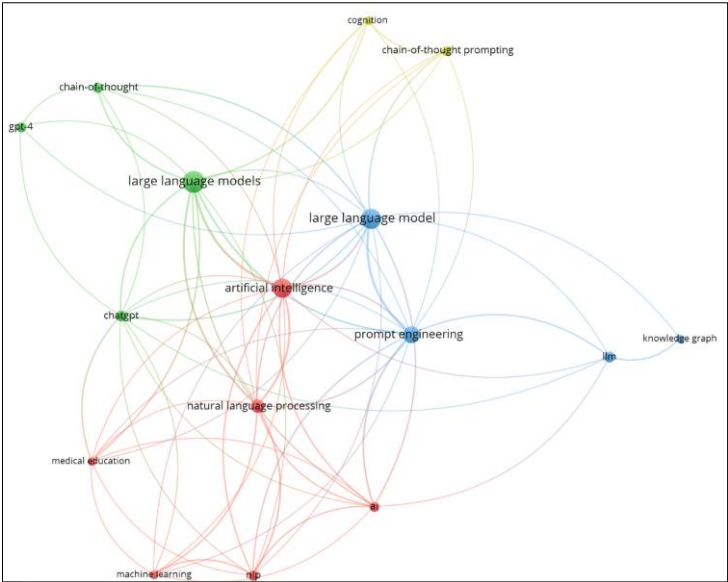


Figure 5. Most Frequently Used Keywords in Studies

How are studies on the CoT approach distributed according to Web of Science meso-level citation areas?

The Web of Science database algorithmically clusters studies according to the frequency with which authors cite each other. WoS makes macro, meso, and micro level definitions from largest to smallest. The distribution of the accessed set of studies by meso-level headings is given in the table below. Figure 6 shows fields covering at least two studies (see Figure 6).

Studies from Communication Sciences (n=19), Software Engineering (n=4), Knowledge Engineering and Representation (n=4), Nutrition and Dietetics (n=3), Nursing (n=2), Medical Ethics (n=2), Neuroimaging (n=2), and Computer Vision and Graphics (n=2) are represented.



Figure 6. Web of Science Meso Citation Areas

Which are the 5 most-cited articles on the CoT approach?

When examining the most highly cited studies, the first place belongs to Martinez’s 2025 study titled Re-evaluating GPT-4’s bar exam performance, attracting attention with a total of 51 citations. This is followed by Englhardt et al.’s 2023 study titled From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models, focusing on the analysis of mobile and behavioral health data using large language models, with 31 citations. In third place is Jauhiainen and Garagorry Guerra’s 2024 study examining the use of generative AI in education and investigating ChatGPT-4’s role in evaluating students’ written responses, with 29 citations. In fourth place is a study by Ott et al., published in 2023, titled ThoughtSource: A central hub for large language model reasoning data, aiming to centralize reasoning data for large language models, with 21 citations. Finally, Beattie et al.’s 2024 study addressing the use of large language models in clinical trial matching and automation in patient screening processes ranks fifth with 18 citations (see Table 7).

Table 7. Five Most-Cited Articles on the CoT Approach

Rank	Title	Author(s)	Total Citations	Avg. Annual Citations
1	Re-evaluating GPT-4's bar exam performance	Martinez (2025)	51	17.00
2	From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models	Englhardt et al. (2023)	31	10.33
3	Generative AI in education: ChatGPT-4 in evaluating students' written responses	Jauhiainen & Garagorry Guerra (2024)	29	9.67
4	ThoughtSource: A central hub for large language model reasoning data	Ott et al. (2023)	21	5.25
5	Utilizing Large Language Models for Enhanced Clinical Trial Matching: A Study on Automation in Patient Screening	Beattie et al. (2024)	18	6.00

What research methods are used in the 5 most-cited articles?

Primarily empirical and experimental research methods were adopted to test the performance and reliability limits of LLMs in the fields of education, law, and medicine. Martinez (2025) conducted an empirical analysis using official exam data to re-evaluate GPT-4's bar exam performance, while Jauhiainen and Garagorry Guerra (2024) used a mixed approach combining qualitative and quantitative methods to measure AI's ability to assess student compositions. Ott et al. (2023) developed an experimental meta-dataset (ThoughtSource) infrastructure integrating seven scientific/medical, three general domain, and five mathematical QA datasets to standardize complex reasoning processes. In the context of clinical research, Beattie et al. (2024) tested the accuracy of LLMs in patient matching automation through longitudinal patient records and structured clinical criteria using a retrospective cohort design, while Englhardt et al. (2024) applied an empirical and rule-based evaluation method measuring models' reasoning capacity in analyzing longitudinal behavioral data collected through passive sensing from mobile devices and clinical insights (see Table 8).

Table 8. Research Methods Used in the 5 Most-Cited Articles

Rank	Title	Research Method
1	Re-evaluating GPT-4's bar exam performance	Methodological Re-evaluation
2	From Classification to Clinical Insights...	Experimental Study and Qualitative
3	Generative AI in education: ChatGPT-4 in evaluating students' written responses	Quantitative / Comparative Experimental Study
4	ThoughtSource: A central hub for large language model reasoning data	Dataset Development / Quantitative
5	Utilizing Large Language Models for Enhanced Clinical Trial Matching...	Experimental Study

Who are the contributors of the 5 most-cited articles?

The sample and data source structures of the examined studies contain methodologically significant differences. In a clinical context, Beattie et al. (2024) relied on 288 patient records labeled by medical experts, while Ott et al. (2023) relied on thousands of QA examples compiled from 15 existing NLP datasets; no individual human participants were used in either study. In contrast, Jauhiainen and Garagorry Guerra (2024) evaluated 54 written responses from 18 university students through both a faculty member and ChatGPT-4, while Enghardt et al. (2024) adopted a three-tiered structure, working with a sensor dataset consisting of undergraduate students, 15 lay raters, and 8 clinicians. Martinez (2025) preferred to re-analyze existing statistical distribution data rather than collecting new data, thus relying on NCBE bar exam statistics and previously published GPT-4 responses as primary data sources (see Table 9). This diversity among sample sizes and source types reflects the yet-unstandardized methodological approaches to evaluating the performance of large language models, drawing attention to the difficulty of comparative studies.

Table 9. Contributors in the 5 Most-Cited Articles

Article	Data Source / Study Group	n (Sample)	Description
Utilizing Large Language Models for Enhanced Clinical Trial Matching... Beattie et al. (2024)	n2c2 2018 Challenge dataset (Harvard University - NLP Clinical Challenges)	288 patient records; 202 publicly available for training; 20 for prompt engineering, 182 for testing	Longitudinal patient records; 13 selection criteria including diabetes, cardiovascular disease. Each record labeled by 2 medical experts.
Re-evaluating GPT-4's bar exam performance Martinez (2025)	NCBE official data; Illinois Bar Exam percentile charts (2018-2019); Katz et al. (2023) GitHub repository (GPT-4 responses)	Human: first-time takers + licensed attorneys (Number not specified; calculated from statistical distribution)	GPT-4 and GPT-3.5 UBE exam scores compared with July and February exam participants, first-time takers only, and licensed attorneys.
Generative AI in education: ChatGPT-4... Jauhainen & Garagorry Guerra (2024)	18 university students (A1-A18); 1 faculty member (human evaluator)	54 written responses (3 reading texts x 18 students); 3,240 evaluations (ChatGPT-4, 10 repetitions x 5 criteria x 54 responses)	English written responses (24-256 words). Students participated voluntarily with ethics board waiver.
ThoughtSource: A central hub... Ott et al. (2023)	15 existing NLP and QA datasets (WorldTree V2, EntailmentBank, OpenBookQA, MedQA, MedMCQA, PubmedQA, MMLU, CommonsenseQA, StrategyQA, QED, AQUA-RAT, ASDiv, GSM8K, MAWPS, SVAMP)	Thousands of QA examples (MedMCQA largest dataset; ThoughtSource-33 subset: 198 items)	No human participants; data compiled from sources. Medical and general domain questions; human expert annotations + CoT data generated with 6 different LLMs.
From Classification to Clinical Insights... Enghardt, Ma et al. (2024)	GLOBEM dataset (single university, undergraduate students, 3-year longitudinal data); 15 lay raters (university	GLOBEM: undergraduate students; Raters: 15 people (\$20 gift card); Clinicians: 8	GLOBEM: smartphone and wearable (Fitbit) sensor data; depression/anxiety label with PHQ-4. 15 raters scored LLM outputs. 8

Article	Data Source / Study Group	n (Sample)	Description
	campus, 18+); 8 mental health experts (6 PhD, 2 master's)	people (\$50 gift card)	clinicians had 1-hour Zoom sessions.

What data collection tools are used in the 5 most-cited articles?

When evaluating studies examining the performance and reasoning processes of large language models in different disciplines, it is evident that the data collection processes are largely based on structured data sources that are aligned with research objectives and mostly derived from natural settings. In a study conducted in the field of education, Jauhiainen and Garagorry Guerra (2024) used students’ written responses to exam questions and expert evaluation criteria for these responses as the primary data source, analyzing the effectiveness of generative AI systems in assessment processes. In another study conducted in the legal field, Martinez (2025) re-examined GPT-4’s bar exam performance, using official exam guides and state-level reported success percentiles as the primary dataset.

In the context of health and clinical psychology, Englhardt et al. (2024) conducted a comprehensive data collection process by integrating passive sensing data obtained from smartphones and wearable technologies with standard clinical scales to evaluate the analysis and reasoning competencies of large language models. Similarly, in another study conducted in the medical field, Beattie et al. (2024) focused on the automation of clinical trial matching processes, evaluating laboratory findings, pathology reports, and eligibility criteria for clinical trials from electronic health records as data sources. On the other hand, Ott et al. (2023), aiming to systematically examine the reasoning and inferential skills of large language models, incorporated into the analysis process an integrated meta-data structure called “ThoughtSource” that brings together multiple QA datasets covering different disciplines, along with accompanying human-generated explanatory texts. In this respect, these studies demonstrate that large language models are evaluated not only on the basis of performance outputs but also on the basis of data structures and contextual reasoning processes (see Table 10).

Table 10. Data Collection Tools Used in the 5 Most-Cited Articles

Article Name	Data Collection Tools / Data Sources
Generative AI in education: ChatGPT-4 in evaluating students' written responses	Written responses of students to open-ended exam questions in a university-level AI course, scoring keys (rubrics) based on explicit criteria, and independent evaluation notes of two instructors teaching the course.
Re-evaluating GPT-4's bar exam performance	Uniform Bar Exam (UBE) official guides published by the National Conference of Bar Examiners (NCBE), state-level bar passing score statistics, and official achievement percentile data of exam participants.
From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models	Longitudinal behavioral data collected through passive sensing from smartphones and wearable devices (sleep duration, step count, location mobility, call/messaging logs) and standard clinical survey data.
ThoughtSource: A central hub for large language model reasoning data	Standardized QA data banks consisting of seven scientific/medical, three general domain, and five mathematical word problems compiled under the "ThoughtSource" meta-dataset, along with human-authored chain-of-thought explanations.
Utilizing large language models for enhanced clinical trial matching: A study on automation in patient screening	Longitudinal patient data from electronic health records (EHR), structured clinical texts, laboratory results, pathology reports, and official eligibility/inclusion criteria lists for oncology clinical trials.

What data analysis methods are used in the 5 most-cited articles?

In the five articles examined, different quantitative, qualitative, and mixed data analysis methods were used to answer the research questions. In their educational research, Jauhiainen and Garagorry Guerra (2024) used descriptive statistics, qualitative content analysis, and inter-rater reliability analysis to compare AI and human evaluator scores. In a study conducted in a legal context, Martinez (2025) conducted empirical performance analysis, distributional analyses, and percentile comparisons with human participants to evaluate the language model's bar exam results. Englhardt et al. (2024), focusing on the analysis of health data, used qualitative coding and thematic analysis methods to examine clinical insights alongside rule-based performance metrics to test model outputs. Similarly, Beattie et al. (2024), examining medical clinical trial matching, calculated standard classification statistics such as sensitivity, specificity, accuracy, precision, and F1 score with a retrospective analysis design

to measure automation performance. Finally, Ott et al. (2023), researching AI’s complex reasoning skills, conducted zero-shot and few-shot prompt performance evaluations and statistical accuracy analyses to measure response accuracy rates of models on their created dataset (see Table 11).

Table 11. Data Analysis Methods Used in the 5 Most-Cited Articles

Article Name	Data Analysis Methods
Generative AI in education: ChatGPT-4 in evaluating students' written responses	Descriptive statistics, qualitative content analysis, inter-rater reliability analysis, quantitative comparison of human and AI scores.
Re-evaluating GPT-4's bar exam performance	Empirical performance analysis, distributional analyses, success percentile comparison with human test-takers.
From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models	Performance metrics calculation, rule-based evaluation of outputs, qualitative coding and thematic analysis of clinical insights.
ThoughtSource: A central hub for large language model reasoning data	Prompt performance analysis, zero-shot and few-shot learning performance evaluation, statistical accuracy analysis.
Utilizing large language models for enhanced clinical trial matching: A study on automation in patient screening	Retrospective performance analysis, standard classification metrics calculation (sensitivity, specificity, accuracy, precision, F1 score).

Conclusion and Discussion

The rapid development of LLMs has enabled these systems to be widely used in different fields such as education, health, law, and engineering. In this process, the CoT approach, which aims to improve the reasoning and problem-solving performance of models, has emerged as one of the most notable strategies in prompt engineering. In this research, studies on the CoT approach were examined through bibliometric analysis, and the current structure, development trends, and methodological characteristics of the field were revealed.

Research results show that the CoT literature has largely developed around the United States (f=20) and China (f=18) at the country level. This finding suggests that these countries’ investments in AI research and their leading positions in LLM development are also reflected in academic output. The US and China are followed by Germany (f=6), Saudi Arabia (f=6), the United Kingdom (f=4), and Singapore (f=4). However, the concentration of publication output in a limited number of countries indicates that CoT research has not yet achieved a

balanced global spread. Furthermore, the fact that the US has the highest number of citations with 147 reveals that it holds an important position not only in quantitative production but also in directing the field and creating impact.

When author contributions are examined, Michael Dohopolski, Feng Gao, and Rune Johan Krumsvik have been identified as the most productive researchers with two publications each. However, the finding that total link strength values of authors in the co-authorship network analysis are zero is noteworthy. This indicates that research in the field has not yet been shaped around strong collaboration networks and that researchers work more independently or in small-scale teams. Considering that AI research requires high computational power, data access, and interdisciplinary expertise, it is evaluated that national and international research networks to be established in the future could make significant contributions to the development of the field.

When the journals in which the studies were published are examined, engineering and medical informatics-focused journals such as IEEE Access ($f=6$), Journal of Imaging Informatics in Medicine ($f=4$), and Journal of Medical Internet Research ($f=3$) stand out. In the field of education, Frontiers in Education ($f=2$) is noteworthy. These results indicate that the CoT approach is not only addressed as a technical research topic but is also being intensively tested in application areas such as health, education, and decision support systems. The density of publications in the medical field in particular suggests that the CoT strategy has become an important research focus in clinical decision support processes requiring explainability and reasoning.

When the annual distribution of publications is examined, the CoT literature shows a very rapid growth trend. While there was only one publication in 2023, this number rose to five in 2024 and forty-four in 2025. The fact that fifteen publications have been reached as of 2026, despite the year not yet being completed, shows that academic interest in the topic continues strongly. This increase can be said to be related to the acceleration of research on prompt engineering and reasoning processes with the spread of advanced language models such as GPT-4, Claude, and similar. The findings obtained indicate that the CoT approach has quickly moved from being a specific research topic to becoming one of the core research areas of generative AI research.

Keyword analysis results also support this assessment. The most frequently used keywords in studies were determined as “large language model” ($f=44$), “artificial intelligence” ($f=19$), “natural language processing” ($f=13$), and “prompt engineering” ($f=12$). The dense co-occurrence relationships among keywords indicate that research is largely shaped around model performance, prompt design, and natural language processing processes. However, the more limited presence of concepts representing educational, cognitive, and pedagogical dimensions suggests that CoT research is still primarily addressed

from a technical and methodological perspective. This points to an important research gap, especially from an educational sciences perspective.

The five most-cited studies within the research scope were examined in detail. Findings show that these studies predominantly conduct empirical research. Bar exam results, university students' open-ended responses, electronic health records, smartphone sensor data, and datasets such as ThoughtSource were used in the examined studies. In data analyses, different methods such as F1 score, sensitivity, accuracy rates, zero-shot prompt performance evaluations, and content analysis were utilized. The diversity of data sources and analysis techniques used demonstrates that CoT research has broad application areas across different disciplines. However, the absence of a common criterion or standard methodological framework in performance evaluations reveals that the field is still in a developmental stage and the need for methodological standardization continues.

In conclusion, this research revealed that studies on the CoT approach have shown rapid growth in recent years, that research is concentrated in certain countries, that it extends across interdisciplinary application areas, and that it does not yet have strong collaboration networks and standard evaluation frameworks. The findings obtained indicate that CoT research plays an important role not only in technical performance enhancement but also in the development of explainable and reliable AI systems. However, there is a need for more comprehensive research examining the pedagogical effects of CoT in the context of educational sciences, its contributions to learning processes, and its potential in assessment applications.

Recommendations

Based on the research findings, the following recommendations can be developed:

- Due to the weakness of co-authorship networks, partnerships and research networks can be established to increase interdisciplinary and international collaborations among researchers.
- Considering the diversity of methods and criteria used in evaluating CoT performance, standardized measurement tools and evaluation frameworks for assessing the reasoning processes of large language models can be developed.
- In the field of educational sciences, more empirical research examining the effects of CoT on student assessment, feedback generation, learning analytics, and instructional design processes can be conducted.

- The majority of existing studies concentrate on LLMs. In the future, research examining the effectiveness of the CoT approach in open-source and lower-cost small language models can be conducted.
- Research using qualitative and mixed methods can be emphasized to evaluate the CoT approach not only in terms of performance outputs but also in terms of the quality, explainability, and educational value of the intermediate reasoning steps it produces.

Statements and Declarations

Acknowledgments/Notes: Not applicable.

Supplementary Materials: Not applicable.

Author Contributions: All authors contributed equally. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Data Availability: Not applicable.

Ethics Approval: Not applicable.

Informed Consent: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Adiguzel, T., Kaya, M., & Cansu, F. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*, 15(3), 429. <https://doi.org/10.30935/cedtech/13152>
- Asseri, B., Abaker, E., Al Mogren, M., Alhefdhi, T., & Al-Wabil, A. (2025). Deciphering Emotions in Children's Storybooks: A Comparative Analysis of Multimodal LLMs in Educational Applications. *AI*, 6(9), 211. <https://doi.org/10.3390/ai6090211>
- Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1), 52-62. <https://doi.org/10.61969/jai.1337500>
- Beattie, J., Neufeld, S., Yang, D., Chukwuma, C., Gul, A., Desai, N., Jiang, S., & Dohopolski, M. (2024). Utilizing large language models for enhanced clinical trial matching: A study on automation in patient screening. *Cureus*, 16(5), Article e60044. <https://doi.org/10.7759/cureus.60044>
- Chang, C. Y., & Yang, J. C. (2022). Concept mapping in computer-assisted learning environments: Bibliometric analysis. *Interactive Learning Environments*, 31(10), 6678-6695.
- Donthu, N., Kumar, S., Pandey, N., & Lim, W. M. (2021). Research components, intellectual structure, and collaboration patterns, *Journal of International Marketing: An analytical retrospective*. *Journal of International Marketing*, 29(2), 1-25.
- Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., ... & Pearson, L. (2017). Living systematic review: 1. Introduction-the why, what, when, and how. *Journal of Clinical Epidemiology*, 91, 23-30.
- Englhardt, Z., Ma, C., Morris, M. E., Chang, C.-C., Xu, X. O., Qin, L., McDuff, D., Liu, X., Patel, S., & Iyer, V. (2024). From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2), Article 56. <https://doi.org/10.1145/3659604>
- Farrokhnia, M., Banihashem, S., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, 61(3), 460-474. <https://doi.org/10.1080/14703297.2023.2195846>
- Hawkins, D. T. (2001). Bibliometrics of electronic journals in information science. *Information Research*, 7(1), 7-1.

- Jauhiainen, J. S., & Garagorry Guerra, A. (2024). Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innovations in Education and Teaching International*, 62(4), 1377-1394. <https://doi.org/10.1080/14703297.2024.2422337>
- Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, 28(12), 15873-15892. <https://doi.org/10.1007/s10639-023-11834-1>
- Li, T., Matsumura, L. C., Wang, E., Litman, D., Liu, Z., Correnti, R., & Hong, M. (2026). AI-assisted evaluation of revision patterns in young students' argument writing. *Frontiers in Education*, 11. <https://doi.org/10.3389/educ.2026.1736489>
- Liu, J., Wang, C., & Liu, S. (2025). Prompt Engineering in Clinical Practice: A Tutorial for Clinicians (Preprint). *Journal of Medical Internet Research*. <https://doi.org/10.2196/72644>
- Martinez, E. (2025). Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*, 33, 581-604. <https://doi.org/10.1007/s10506-024-09396-9>
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essays scoring. *Research Methods in Applied Linguistics*, 2(2), 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- OpenAI. (2023). GPT-4 technical report. arXiv. <https://doi.org/10.48550/arXiv.2303.08774>
- Ott, S., Hebenstreit, K., Lievin, V. et al. ThoughtSource: A central hub for large language model reasoning data. *Sci Data* 10, 528 (2023). <https://doi.org/10.1038/s41597-023-02433-3>
- Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation*, 25, 348.
- Van Eck, N., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538.
- Van Nunen, K., Li, J., Reniers, G., & Ponnet, K. (2018). Bibliometric analysis of safety culture research. *Safety Science*, 108, 248-258.
- Verma, S., & Gustafsson, A. (2020). Investigating the emerging COVID-19 research trends in the field of business and management: A bibliometric analysis approach. *Journal of Business Research*, 118, 253-261.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837. <https://doi.org/10.48550/arXiv.2201.11903>

- Yang, L., Chen, Z., Liu, T., Gong, Z., Yu, Y., & Wang, J. (2013). Global trends in solid waste research from 1997 to 2011 using bibliometric analysis. *Scientometrics*, 96, 133-146.
- Zou, D., Huang, X., Kohnke, L., Chen, X., Cheng, G., & Xie, H. (2022). Bibliometric analysis of trends and research topics of empirical research on TPACK. *Education and Information Technologies*, 27(8), 10585-10609.

CHAPTER 5

Turkish EFL Students' Conceptualizations of Generative AI Use in Language Learning: A Metaphor Analysis

*Esin Saylan¹
Neslihan Keleş²
Elif Aydın Yazıcı³*

¹ Dr. Lect., Trabzon University, School of Foreign Languages, Trabzon, Türkiye
ORCID: 0000-0002-7629-8283

² Lect., Trabzon University, School of Foreign Languages, Trabzon, Türkiye
ORCID: 0000-0002-7614-9317

³ Dr. Lect., Trabzon University, School of Foreign Languages, Trabzon, Türkiye
ORCID: /0000-0001-6634-1025

Introduction

Artificial intelligence has become a topic increasingly discussed from multiple perspectives within foreign language education research. Existing literature has investigated the experiences, beliefs, and perceptions of various stakeholders such as teachers, students, school administrators, and educational institutions regarding the use of GenAI across diverse contexts (Derakhshan et al., 2024; Lo et al., 2024; Teng, 2024; Yang & Li, 2024). Similarly, various studies addressing the role and prominence of AI in foreign language teaching and learning have been conducted within the Turkish context (Hınız, 2024; Karataş et al., 2024; Orhan & Çınar-Yağcı, 2026; Eşki & Eryılmaz, 2026).

The use of metaphors has long served as a research approach to reveal how a certain phenomenon is perceived by different agencies in the realm of foreign language teaching and learning. Since metaphors enable individuals to express complex and abstract concepts in more concrete, intuitive, and relatable ways, they offer a powerful analytical tool for examining students' perceptions and experiences regarding artificial intelligence (Zirenko et al., 2025). An examination of existing metaphor-based studies reveals that a significant number of recent research on artificial intelligence focuses specifically on certain generative AI chatbots, especially ChatGPT (Azap, 2025; Bayraktar-Çepni & Erbay-Çetinkaya, 2026; Selvi et al., 2025; Vo & Huynh, 2025). For example, Azap (2025) and Selvi et al. (2025) investigated EFL students' metaphorical perceptions of ChatGPT, focusing on how learners conceptualize the role and functions of ChatGPT in their language learning experiences. Similarly, Bayraktar-Çepni and Erbay-Çetinkaya (2026) examined students' metaphorical conceptualizations of ChatGPT, further highlighting learners' interpretations regarding their exam routines. In contrast, Vo and Huynh (2025) focused on EFL teachers' perspectives, examining how teachers metaphorically conceptualize ChatGPT in relation to language teaching.

Foreign language learning experiences are not limited to the use of a single AI tool. Particularly in learning processes outside the classroom, students can utilize a variety of AI-based tools for purposes such as text generation, translation, grammar and vocabulary support, speaking practice, feedback, information access, and other learning objectives. Consequently, addressing students' perceptions of AI solely through the lens of a specific chatbot may lead to overlooking the multidimensional experiences arising from their interactions with diverse AI tools. In other words, while metaphor studies focusing on a specific tool may elucidate students' perceptions of a singular technology like ChatGPT, they offer only a limited explanation of how the meaning of AI as a whole is constructed within foreign language learning processes. However, studies adopting a broader focus on AI or GenAI limited in number. One notable metaphorical study was conducted by Yan et al. (2024) with Chinese university

EFL learners which examines how they metaphorically conceptualized their use of GenAI in English language learning.

Rather than focusing on a single AI tool or chatbot, the present study examines the broader metaphorical conceptualizations held by EFL students regarding the various AI tools they utilize throughout their foreign language learning journeys. Conducted within the Turkish context, the study aims to shed light on how students position GenAI in their language learning processes, the roles they attribute to it, and the meanings through which they structure their experiences with these technologies through elicited metaphors. Accordingly, the following research question guided the study:

RQ1: How do Turkish EFL learners conceptualize the use of GenAI in language learning through metaphors?

Methodology

This study was conducted within a qualitative research framework to examine how students learning English as a foreign language in the context of Turkish higher education conceptualize the use of GenAI. Elicited Metaphor Analysis (EMA) was employed to reveal the participants' interpretations regarding the role of GenAI in the language learning process and their implicit conceptual structures. EMA is a widely used approach in educational discourse for examining students' and teachers' beliefs, experiences, and conceptualizations of teaching and learning processes (Seung et al., 2015).

Setting and Sample

The sample consisted of students enrolled in the English preparatory program at the School of Foreign Languages of a public university in Türkiye during the 2025–2026 academic year. Participants were recruited using convenience sampling based on their accessibility and willingness to take part in the study (Creswell & Creswell, 2018). A total of 51 prep students enrolled either in ELT or Translation programs participated in the questionnaire. However, the demographic characteristics of the participants who provided valid metaphors (n = 33) are presented in Table 1.

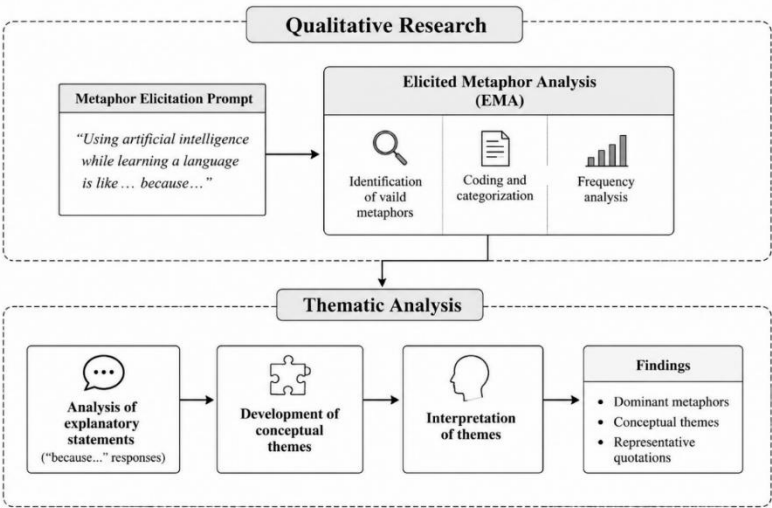
Table 1. Demographics for Participants

	Variable	Value
Gender	Female	25
	Male	8
Age Range	17-18	14
	19-20	17
	21 and over	2
Program	English Language Teaching	21
	Translation and Interpretation	12
Total Participants		33

Data Collection and Analysis

Data were collected through an online fill-in-the blanks metaphor elicitation questionnaire via Google Forms The questionnaire comprises two sections. The first section elicited participants’ demographic characteristics, including age, gender, and department. The second section was designed to elicit participants’ metaphorical representations of GenAI use in the context of language learning. The metaphor elicitation task was administered in Turkish. Before completing the questionnaire, participants were informed that their involvement was voluntary and that their responses would remain independent of their course grades and academic evaluation.. Figure 1 summarizes the whole research orientation of the present study:

Figure 1. Research Design



To elicit metaphorical conceptualizations, participants were asked to complete the following prompt: “Using artificial intelligence while learning a language is like ____ because ____”. All responses were compiled and systematically organized in an Excel spreadsheet, which was used to facilitate the manual examination, organization, and frequency calculation of the data. In line with the principles of EMA, participants’ metaphorical expressions were manually identified and examined, and metaphors conveying similar conceptual meanings were grouped into broader conceptual categories. The frequency of individual metaphors and conceptual categories was subsequently calculated. In a second stage of analysis, participants’ explanations and justifications in the “because” component were examined thematically to identify patterns and broader themes underlying their metaphorical conceptualizations. The findings were then

interpreted by considering the metaphorical expressions, conceptual categories, frequencies, and emerging themes together.

Findings and Discussion

Fifty-one participants responded to the metaphor elicitation question; however, 33 produced valid metaphors accompanied by explanatory statements and were included in this metaphor analysis. 18 responses were excluded from the analysis because participants either did not provide a metaphorical expression or provided responses that were irrelevant to the elicitation, such as “I don’t know” or blank responses. The relatively high number of excluded responses may suggest that some participants experienced difficulty in spontaneously generating an appropriate metaphor to express their conceptualizations of GenAI use in language learning. Most surprisingly, one participant stated that “When you can’t come up with a metaphor, you even find yourself wanting to ask AI for help with that.”

Conceptual Categorization of Metaphors

The elicited metaphors were first organized into conceptual categories, followed by a thematic analysis of the participants’ explanatory statements. The metaphor analysis revealed that participants conceptualized GenAI in language learning through a range of roles, objects, and experiential domains. There are eleven conceptual categories of reported metaphors as shown in Table 2.

Table 2. Conceptual Categorization

Conceptual Category	Codes (Metaphors)	f
Teacher / Tutor	teacher (n = 5), online/virtual teacher (n = 2), private tutor (n = 3), incompetent teacher (n = 1)	11
Information Source	dictionary (n = 2), library (n = 1), Internet search (n = 1)	4
Learning Companion	friend (n = 1), friendship (n = 1), assistant (n = 1), human (n = 1)	4
Education / Learning	game (n = 1), recipe tutorial (n = 1), online course (n = 1)	3
Guide	guide (n = 1), navigation system (n = 1)	2
Objects	mirror (n = 1), scale (balance) (n = 1)	2
Nature	forest (n = 1), ocean (n = 1)	2
Buildings	bridge (n = 1), factory (n = 1)	2
Food	packaged food (n = 1)	1
Vehicles	boat (n = 1)	1
Psychological States	addiction (n = 1)	1

The ‘Teacher/Tutor’ category was the most prominent encompassing 11 metaphorical expressions. The predominance of teacher-related metaphors suggests that participants primarily conceptualized GenAI in terms of an instructional or pedagogical role which positions it as a source of explanation and language-learning support. At the same time, the occurrence of the metaphor

‘incompetent teacher’ indicates that this conceptualization was not positive, and some learners perceived limitations in GenAI’s instructional capacity.

The second most frequent categories are ‘Information Source’ and ‘Learning Companion’. Each category includes four metaphors. The Information Source category comprised *dictionary* (n = 2), *library* (n = 1), and *Internet search* (n = 1). These metaphors position GenAI within a resource-oriented conceptual domain, which emphasize its connection to information access in the language learning process. On the other hand, ‘Learning Companion’ category included *friend*, *friendship*, *assistant*, and *human* (n = 1 each), which conceptualize GenAI through interpersonal and supportive roles.

The third most dominant category is ‘Education/Learning’ (*game*, *recipe tutorial*, *online course*) (n = 1 each). These are followed by four categories with two metaphorical expressions each: ‘Guide’ (*guide*, *navigation system*), ‘Objects’ (*mirror*, *scale/balance*), ‘Nature’ (*forest*, *ocean*), and ‘Buildings’ (*bridge*, *factory*) (n = 2 each). Finally, ‘Food’ (*packaged food*), ‘Vehicles’ (*boat*), and ‘Psychological States’ (*addiction*) were each represented by a single metaphor (n = 1 each).

Overall, this conceptual categorization demonstrates that participants drew on diverse domains when conceptualizing GenAI in language learning. While the Teacher/Tutor category was clearly dominant, the remaining metaphors extended beyond conventional instructional roles to encompass information resources, interpersonal relationships, guidance, educational activities, physical objects, natural environments, structures, food, transportation, and psychological states. Thus, at the conceptual level, GenAI was represented through a multidimensional set of metaphorical frames rather than a single dominant conceptual domain.

Thematic Patterns in Learners’ Metaphorical Conceptualizations of GenAI

The thematic analysis of the elicited metaphors revealed that EFL students’ conceptualizations of GenAI can be categorized into five main themes: AI as an Instructional Resource, AI as a Companion in the Learning Journey, AI as a Learning Platform for Creation, AI as a Reflective Learning Tool, and AI as a Double-Edged Technology. While the first four themes indicate that students primarily view GenAI as a functional, supportive resource that facilitates learning, the fifth theme reflects a critical perspective regarding GenAI’s limitations and potential risky consequences. The overall distribution of metaphors under each emerging theme is listed below:

Table 3. Thematic Orientation

Emerging Themes	Codes (Metaphors)	f
AI as an Instructional Resource	teacher(n=5), online/virtual teacher (n=2), private tutor (n = 3), dictionary (n = 2), library (n = 1), Internet search (n = 1), boat (n = 1), factory (n = 1)	16
AI as a Companion in the Learning Journey	friend (n = 1), friendship (n = 1), assistant (n = 1), human (n = 1), guide (n = 1), navigation system (n = 1), bridge (n = 1)	7
AI as a Creative Learning Platform	game (n = 1), recipe tutorial (n = 1), online course (n = 1), forest (n = 1), ocean (n = 1),	5
AI as a Reflective Learning Tool	mirror (n = 1), scale/balance (n = 1)	2
AI as a Double-Edged Technology	incompetent teacher (n = 1), packaged food (n = 1), addiction (n = 1)	3
		TOTAL:33

The theme with the highest frequency was ‘AI as an Instructional Resource’ (f = 16). Table 4 presents the metaphors, participants’ justifications, and illustrative quotes associated with Theme 1. Metaphors such as ‘teacher’ .‘online/virtual teacher’, and ‘private tutor’ indicate that students perceive AI as a teaching and learning resource capable of partially assuming traditional teaching roles. In particular, the students view GenAI not solely as a general technology, but as an instructor that directly provides information, offers explanations, and supports individual learning during the foreign language learning process. Meanwhile, metaphors within these theme, such as ‘dictionary’, ‘factory’, ‘library’, ‘crossing a river by boat’, and ‘Internet search’ complement GenAI tools’ instructional role by highlighting the dimension of information access. These metaphors suggest that GenAI is viewed as a reference source, enabling students to access the information they need quickly and conveniently. Consequently, this theme reveals that, in the eyes of students, GenAI is positioned as a learning resource that both ‘teaches’ and ‘provides information.’

Table 4. AI as an Instructional Resource (Theme 1)

Metaphors	Participants' Rationales	Example Participant Expressions
teacher	explaining difficult topics reinforcing grammar providing feedback error correction	P19: <i>"Using artificial intelligence while learning a language is like having a virtual teacher because it provides help and feedback, but it can never replace a teacher."</i>
online teacher		
private tutor		
		P10: <i>"Using artificial intelligence while learning a language is like looking up a word in a dictionary because it enables learners to access and learn information in a similar but faster way."</i>
dictionary		P17: <i>Using artificial intelligence while learning a language is like a factory because it can provide us with information about almost anything in a short amount of time."</i>
factory	quick access to information	
library	learning new vocabulary conducting research	
Internet search	obtaining additional knowledge	P21: <i>"Using artificial intelligence while learning a language is like crossing a river by boat rather than swimming because AI provides us with the answer we are looking for quickly."</i>
boat		P24: <i>"Using artificial intelligence while learning a language is like a library because you can access the information you need with just one click."</i>

The second theme has been identified as ‘AI as a Companion in the Learning Journey’ (f = 7) as illustrated in Table 5. Metaphors like ‘friend’, ‘friendship’, ‘assistant’, and ‘any human being’ indicate that the relationship students establish with AI is not just instrumental. Instead, some students perceive GenAI as an accessible and supportive entity that accompanies them throughout the learning process. Furthermore, metaphors like ‘guide’ and ‘navigation system’ highlight AI’s role in providing direction and guidance to students during their learning journey. The metaphor ‘bridge’ reinforce this perspective according to which AI is conceptualized as an element that facilitates the transition from the students’ current state to their intended learning goals. Consequently, within this theme, GenAI emerges not only as a technology that delivers information but also as a companion that stands by the student and assists their progress along the learning journey.

Table 5. AI as a Companion in the Learning Journey (Theme 2)

Metaphors	Participants' Rationales	Example Participant Expressions
friend	providing support giving feedback or ideas helping when needed interacting	P28: <i>"Using artificial intelligence while learning a language is like having a close friend because it can inform us about many different topics and support us while answering our questions, but it also has its own limitations and can make mistakes."</i>
friendship		
human		
assistant		
guide	providing directions helping learners make progress	P17: <i>"Using artificial intelligence while learning a language is like a bridge because it connects what I already know with what I want to learn."</i>
navigation system		P18: <i>"Using artificial intelligence while learning a language is like using a navigation system because it shows us the way, but we have to keep our hands on the wheel to reach our destination."</i>
bridge		

The third theme emerged as ‘AI as a Creative Learning Platform’ (f = 5). The metaphors within this theme such as ‘game’, ‘recipe tutorial’, ‘online course’, ‘forest’ ,and ‘ocean’ identify that students perceive GenAI as a learning space offering extensive opportunities for creation and exploration. The ‘forest’ and ‘ocean’ metaphors evoke AI’s vastness and limitless possibilities. The ‘recipe tutorial’ metaphor reveals that GenAI can be viewed as a platform that provides step-by-step guidance for tasks and transforms learning into practice. Similarly, the ‘game’ metaphor suggests that learning can become more facilitated thanks to GenAI tools. This theme suggests that GenAI is accepted as space where new ideas, creative content, and learning experiences can be generated. Table 6 summarizes participants’ rationales and entailment within Theme 3.

Table 6. AI as a Creative Learning Platform (Theme 3)

Metaphors	Participants' Rationales	Example Expressions	Participant
game recipe tutorial online course	practicing through generated examples and activities	P6: "Using artificial intelligence while learning a language is like following a video tutorial for making a favorite recipe rather than simply following the written instructions because reading the instructions alone does not show whether the process is being carried out correctly, whereas watching the video allows us to understand how the process should be performed." P20: "Using artificial intelligence while learning a language is like playing a game because, just as we learned things through games when we were children, what we learn through artificial intelligence stays in our minds."	
forest ocean	exploring new learning possibilities creating new things	P13: "Using artificial intelligence while learning a language is like an endless ocean because we can sometimes learn things that would never have occurred to us, but, of course, the creativity of the human mind should remain more prominent." P22: "Using artificial intelligence while learning a language is like exploring a forest because the more you explore the forest, the more new things you learn and discover, and you begin to realize how many things you can benefit from."	

The fourth positive theme is ‘AI as a Reflective Learning Tool’ (f = 2) comprising the metaphors of ‘mirror’ and ‘scale/balance’ (See Table 7). Although its frequency is low compared to other themes, it is significant for highlighting a self-reflective dimension of students' GenAI usage. The ‘mirror’ metaphor points to a reflective function that helps students observe their own learning performance, shortcomings, or errors. The ‘scale/balance’ metaphor evokes concepts of evaluation, comparison, and the weighing of different aspects of a situation. When these metaphors are considered together, it becomes evident

that some students conceptualize GenAI as a means to evaluate their own learning processes and enhance self-awareness.

Table 7. AI as a Reflective Learning Tool (Theme 4)

Metaphors	Participants' Rationales	Example Expressions	Participant
mirror	identifying errors, recognizing strenghts and weaknesses,	P15: <i>"Using artificial intelligence while learning a language is like a mirror because it immediately shows you your mistakes and helps you recognize what you need to correct."</i>	
scale /balance	self-reflection	P33: <i>"Using artificial intelligence while learning a language is like a scale because which side outweighs the other depends on the individual."</i>	

The final and only negative theme is ‘AI as a Double-Edged Technology’ which refers to participants’ awareness of the potential limitations and risks associated with GenAI use in language learning (f = 3). Table 8 presents relevant metaphors and justifications. The metaphors of the ‘incompetent teacher,’ ‘packaged food,’ and ‘addiction’ demonstrate that students GenAI technology can create various problems depending on how it is used. For instance, the ‘incompetent teacher’ metaphor reveals that despite having an instructional role, it is not always viewed as a competent or reliable instructor. Similarly, the ‘packaged food’ metaphor reflects a critical perspective regarding the nature of the information GenAI provides which is ready-made, easily accessible, and consumable with minimal effort. Lastly, the ‘addiction’ metaphor takes this concern a step further, clearly articulating the fear that AI usage could escalate into excessive dependency and over-reliance.

Table 8. AI as a Double-Edged Technology (Theme 5)

Metaphors	Participants' Rationales	Example Expressions	Participant
incompetent teacher	providing suggestions	useful while	P3: "Using artificial intelligence while learning a language is like packaged food because you can buy it and consume it immediately, but it does not have the same taste as food prepared by hand. Similarly, AI-generated feedback cannot replace feedback from someone with relevant expertise and experience."
packaged food	sometimes producing inaccurate information		
addiction			P23: "Using artificial intelligence while learning a language is like addiction because if it is used improperly, it can negatively affect learning."

Conclusion and Implications

The results showed that students had a generally positive perception of GenAI for their English language learning. The majority of metaphors described it as an instructional resource that offers guidance and support, while some portrayed it as a learning companion, a space of creation, or a tool for reflection. At the same time, some students pointed out its dual nature, showing that they were aware that its advantages depend on how it is used. Overall, the metaphors suggest that GenAI is becoming an integral part of students' EFL learning experiences instead of being an external technological tool.

Pedagogically, these findings suggest that GenAI can be incorporated into EFL instruction as a complementary learning resource. Teachers may design activities that encourage students to engage with AI-generated feedback and content rather than simply accept them. Its potential for individualized practice and language support can also be used beyond the classroom. At the same time, students need opportunities to question, verify, and evaluate AI-generated content. Such an approach can help preserve students' active role in learning while making productive use of the support GenAI can offer.

Acknowledgement: This study was presented at the NEXUS 2026 International Conference on English Language Teaching & Translation, held in Cappadocia, Türkiye, on September 24–26, 2026.

Artificial Intelligence Statement: Figure 1 was prepared with the assistance of an AI tool, and the authors take full responsibility for its content and accuracy.

References

- Azap, S. (2025). The role of artificial intelligence in language teacher education: EFL learners' views on the concept of "ChatGPT" through metaphorical analysis. *Social Sciences & Humanities Open*, 11, 101394. <https://doi.org/10.1016/j.ssaho.2025.101394>
- Bayraktar-Çepni, S., & Erbay-Çetinkaya, Ş. (2026). Pre-service English teachers' tendency to integrate ChatGPT into exam routines: A metaphor analysis. *SN Social Sciences*, 6(6). <https://doi.org/10.1007/s43545-026-01500-x>
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.
- Derakhshan, A., & Ghiasvand, F. (2024). Is ChatGPT an evil or an angel for second language education and research? A phenomenographic study of research-active EFL teachers' perceptions. *International Journal of Applied Linguistics*, 34(4), 1246–1264. <https://doi.org/10.1111/ijal.12561>
- Eşki, M., & Eryılmaz, N. (2026). An exploration of Turkish EFL pre-service teachers' perceptions on the integration of generative AI in English language teacher education: Benefits, challenges, and future directions. *Asian-Pacific Journal of Second and Foreign Language Education*, 11(44). <https://doi.org/10.1186/s40862-026-00420-w>
- Hınız, G. (2026). A year of generative AI in English language teaching and learning – A case study. *Journal of Research on Technology in Education*, 58(2), 291–311. <https://doi.org/10.1080/15391523.2024.2404132>
- Karataş, F., Abedi, F. Y., Ozek Gunyel, F., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies*, 29, 19343–19366. <https://doi.org/10.1007/s10639-024-12574-6>
- Lo, C. K., Yu, P. L. H., Xu, S., Ng, D. T. K., & Jong, M. S.-Y. (2024). Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies. *Smart Learning Environments*, 11, Article 50. <https://doi.org/10.1186/s40561-024-00342-5>
- Orhan, A., & Çınar Yağcı, Ş. (2026). Using ChatGPT in EFL context: What do EFL teachers think about it? *International Journal of Technology in Education*. <https://doi.org/10.46328/ijte.5323>
- Selvi, B., Çoban, Z., & Demirkol Orak, S. (2025). Exploring preparatory students' metaphorical perceptions of ChatGPT. *Necmettin Erbakan Üniversitesi Ereğli Eğitim Fakültesi Dergisi*, 7(1), 116–137. <https://doi.org/10.51119/ereegf.2025.124>

- Seung, E., Park, S., & Jung, J. (2015). Methodological approaches and strategies for elicited metaphor-based research: A critical review. In W. Wan & G. Low (Eds.), *Elicited metaphor analysis in educational discourse* (pp. 39–64). John Benjamins Publishing Company.
- Teng, M. F. (2024). “ChatGPT is the companion, not enemies”: EFL learners’ perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence*, 7, 100270. <https://doi.org/10.1016/j.caeai.2024.100270>
- Vo, L.-H., & Huynh, N.-T. (2025). Vietnamese EFL teachers’ perspectives on ChatGPT: A conceptual metaphor analysis. *Arab World English Journal*, 16(1), 162–178. <https://doi.org/10.24093/awej/vol16no1.10>
- Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. *System*, 124, 103351. <https://doi.org/10.1016/j.system.2024.103351>
- Zirenko, M., Machura, I. A., Fabriz, S., Schulze-Vorberg, L., & Horz, H. (2025). AI and learning with AI: University students’ metaphorical conceptualizations. *Journal of Interactive Media in Education*, 2025(1), Article 13. <https://doi.org/10.5334/jime.994>

CHAPTER 6

**Science Teachers' Curriculum
Literacy Self-Efficacy Regarding
The 2024 Türkiye Century Maarif Model
Curriculum**

Büşra Özcan¹
Hasene Esra Yıldırım²

¹ Yüksek Lisans Öğrencisi, Balıkesir Üniversitesi,
<https://orcid.org/0009-0001-6557-4684>

² Doç. Dr., Balıkesir Üniversitesi, <https://orcid.org/0000-0002-9691-3730>

INTRODUCTION

The curriculum is a roadmap designed to ensure that teaching takes place in a deliberate and purposeful manner rather than haphazardly; rather than being a fixed and unchanging plan, it is a flexible framework that can be restructured to meet the needs of teachers and pupils within changing circumstances (TLA, 2023). The effective implementation of this framework depends to a large extent on teachers' ability to correctly interpret and understand the curriculum; this ability is referred to in the literature as 'curriculum literacy'.

Curriculum literacy is defined as teachers' ability to be knowledgeable about teaching programmes, to interpret them, to examine them from a critical perspective, and to implement them effectively in educational settings (Keskin & Korkmaz, 2021). Similarly, Aslan (2018) defines curriculum literacy as the ability to possess knowledge of the content within the curriculum, to understand this content correctly, to plan in accordance with the curriculum, and to implement the curriculum correctly. When these definitions are considered together, it becomes clear that curriculum literacy is not limited to merely knowing the curriculum; rather, it requires a multidimensional set of skills, such as interpreting the curriculum, critically evaluating it and adapting it for use in the classroom.

The skills that a curriculum-literate teacher should possess are discussed in detail in the literature (Erdem, 2023; Erdem & Yücel Toy, 2023). Accordingly, teachers are expected to understand the objectives/learning outcomes, content, teaching and learning process, and assessment and evaluation components that constitute the curriculum, as well as the relationships between these components; they are also expected to recognise the educational philosophy on which the curriculum is based and to assess the curriculum's consistency with this philosophy. In addition, they must be proficient in the principles of writing learning outcomes and their cognitive, affective and psychomotor dimensions; be able to evaluate content according to principles such as moving from the concrete to the abstract and from the simple to the complex; be able to plan activities by determining teaching methods and techniques appropriate to the learning outcomes; and be able to select and use appropriate assessment tools. Finally, the teacher's ability to adapt the official curriculum to pupils' readiness levels and classroom conditions by converting it into a written lesson plan, and to critically evaluate the effectiveness and strengths and weaknesses of the programme they implement, is also regarded as part of curriculum literacy. Given this multidimensional set of skills, determining teachers' levels of curriculum literacy with regard to current teaching programmes is important in terms of ensuring that the curriculum is reflected in the classroom environment as intended. Fen ve Teknoloji, Fen ve Tabiat Bilgisi, Fen Bilgisi, Fen Bilimleri)

The Historical Development of Science Curricula in Türkiye

Since the proclamation of the Republic, the curricula of this course, taught under different names (Science and Technology, Science and Nature, Science), have been updated in the years 1924, 1926, 1936, 1948, 1968, 1992, 2000, 2005, 2013, 2018, and 2024 (Arslan, 2000; Bakırcı et al., 2025; Dönmez & Gülen, 2023). In the curricula between 1924 and 1948, the aim was to provide students with basic and practical science knowledge, adopting an understanding where the teacher transmitted information and the student remained passive; while the 1948 curriculum continued this approach, it also included fostering the habit of observation (Arslan, 2000; Bakırcı et al., 2025; Demir, 2021). In 1968, the name of the course was changed to "Science and Nature," with the aim of developing scientific thinking and problem-solving skills in students, thereby producing scientifically literate individuals (Arslan, 2000; Dindar & Taneri, 2011).

In the curriculum drawn up in 1992, the name of the subject was changed to 'Science'. The primary aim of this programme is to nurture inquisitive and critical-thinking individuals; to this end, particular emphasis has been placed on developing scientific process skills and linking science lessons to everyday life (Ministry of National Education [MoNE], 1992; Dindar & Taneri, 2011). The programme drawn up in 2000, on the other hand, aimed to help pupils understand the relationship between science and technology and, in this respect, initiated the transition towards scientific literacy (MoNE, 2000; Dindar & Taneri, 2011).

A fundamental change in science education in Türkiye took place with the curriculum introduced in 2005 under the name 'Science and Technology'. With this programme, the behaviourist approach adopted in previous years was abandoned and replaced by a constructivist approach. This change signified a shift away from a rote-learning system towards a student-centred approach to education (Bakırcı et al., 2025; Şahin, 2009). The programme's vision was defined as nurturing individuals who are literate in science and technology (MoNE, 2005).

Following a programme change in 2013, the name of the course was updated once again to 'Science'. With the new curriculum, the aim is to nurture scientifically literate individuals who are inquisitive and critical thinkers, capable of using technology effectively, able to relate to engineering concepts, and possessing an interdisciplinary perspective (Bakırcı et al., 2025; MoNE, 2013). The 2018 Science Curriculum emphasised 21st-century life skills, engineering design skills and values education. In this curriculum, interdisciplinary interaction has been further strengthened under the influence of the STEM approach, and the curriculum's vision has once again been defined as 'nurturing scientifically literate individuals' (Bakırcı et al., 2025; Dönmez & Gülen, 2023; MoNE, 2018).

Looking broadly at the historical development of science curricula in Türkiye, it is clear that there has been a shift away from merely imparting knowledge towards equipping pupils with the skills to acquire knowledge through their own efforts. Throughout this process, efforts have been made to move students away from a passive role as recipients of knowledge and towards becoming active, inquisitive and inquisitive individuals. The ultimate aim has been defined as fostering individuals who possess scientific, digital and environmental literacy skills, rather than individuals who merely learn standardised scientific knowledge (Bakırcı et al., 2025; Dönmez & Gülen, 2023).

2024 Türkiye Century Maarif Model Science Curriculum

The 2024 Türkiye Century Maarif Model Science Curriculum is the result of a process of change and restructuring encompassing key areas such as approach, design model, objectives, content, the learning-teaching process and assessment (Bakırcı et al., 2025). The changes and restructuring in the 2024 science curriculum have given rise to the concept of multiple literacies, and pupils are expected to be environmentally, scientifically and digitally literate. Based on the principle of holistic development, the curriculum aims to develop pupils' cognitive, affective, mental and physical characteristics as a whole. The model aims to equip pupils with social-emotional learning skills, foster the adoption of an interdisciplinary perspective, and prioritise process-based assessment methods (MoNE, 2024).

Social-emotional learning skills, which have been systematically incorporated into the curriculum for the first time, encompass sub-dimensions such as self-awareness, self-management, social awareness, relationship skills and responsible decision-making, and have been integrated with subject-specific skills and literacy skills (MoNE, 2024; Demir et al., 2026). Within the framework of an interdisciplinary approach, the programme has moved away from the unit-based approach of previous programmes towards a thematic approach; the science curriculum has been structured around themes such as living things and life, matter and its nature, physical phenomena, and the Earth and the universe, thereby addressing the relationships between topics in a more holistic manner (MoNE, 2024). In line with a process-based assessment approach, the use of alternative assessment tools—such as a rubric, a checklist and a self-assessment form—which enable pupils to demonstrate their performance throughout the process, is recommended (MoNE, 2024). Within the framework of multiple literacies, the aim is for pupils to acquire a wide range of literacy types, including information, digital, data, visual, environmental and climate, health and financial literacy; these skills are presented in an integrated manner alongside subject-specific skills and social-emotional learning skills (MoNE, 2024).

In addition to these core orientations, the programme incorporates innovations across various dimensions. In line with the Sustainable Development Goals,

environmental awareness, the efficient use of resources and climate literacy have been included as key components of the programme (MoNE, 2024; Yüzbaşıoğlu, 2025). The programme aims not only to impart knowledge but also to foster various skills—such as conceptual skills, subject-specific skills and higher-order thinking skills—as well as values including justice, friendship, honesty, self-discipline, patience, respect, love, responsibility, patriotism and altruism (MoNE, 2024). In terms of content, there has been a marked simplification compared to previous programmes; information of an academic nature and that encouraging rote learning has been removed and deferred to higher education, thereby aiming to enable students to engage in deeper learning (MEB, 2024; Demir et al., 2026). Whilst science and engineering applications were previously largely limited to end-of-unit projects, in the 2024 programme they have been integrated with conceptual skills, social-emotional skills and literacy skills, and spread throughout the learning process as a whole (Demir et al., 2026). To enable pupils in heterogeneous classroom settings to realise their full potential, the curriculum incorporates ‘enrichment and support’ dimensions, thereby making it possible to simplify or enrich the content according to pupils’ needs (Demir et al., 2026). Finally, during the programme’s development, alignment with international assessments such as PISA and TIMSS, as well as contemporary educational approaches, was prioritised, with the aim of bringing science education in Türkiye up to a level competitive with international standards (MEB, 2024).

Previous Research on Levels of Self-Efficacy in Programme Literacy

A review of the literature in this field reveals that there has been an increasing number of studies on programme literacy in recent years; it is evident that these studies have predominantly been conducted using a quantitative approach and have involved different groups of participants, such as preservice teachers, teachers and school leaders.

Bolat (2017) developed the ‘Curriculum Literacy Scale’ and determined that the scale comprises two sub-dimensions: reading and writing. Erdem and Eǧmir (2018) and Aslan (2019) found that trainee teachers had high levels of curriculum literacy, but that they performed better in the ‘reading’ sub-dimension than in the ‘writing’ sub-dimension. It has been established that academic achievement and the field of study produce significant differences (Erdem & Eǧmir, 2018) and that this level predicts readiness for the teaching profession (Aygün, 2019). In terms of year group, Sural and Dedeali (2018) and Karaduman (2024) found that fourth-year teacher trainees’ levels were significantly higher than those of trainees in lower year groups; Karaduman (2024) also found that the ‘Science Teaching Programmes’ course increased this level and that fourth-year candidates’ levels were even higher than those of practising teachers.

Teachers’ competencies were generally found to be at a high (Nasırcı, 2022) or moderate (Kahramanoğlu, 2019; Saral, 2019) level; female teachers

outperformed their male counterparts (Kahramanoğlu, 2019; Saral, 2019), class teachers have higher levels than subject teachers (Keskin, 2020), and teachers with 21 years' or more of service have higher levels than other groups (Saral, 2019). It was also found that teachers who had received in-service training or participated in programme development processes had significantly higher perceptions of curriculum literacy (Erdamar, 2020; Keskin, 2020; Nasırcı, 2022). In line with this, Karaduman (2024) recommended the provision of in-service training.

Although the literature generally establishes that curriculum literacy is a fundamental competence for the teaching profession, no study has been found that directly addresses teachers' curriculum literacy competences in relation to the 2024 Science Curriculum. This indicates that the present study aims to fill this gap in the literature.

Previous Research on 2024 Türkiye Century Maarif Model Science Curriculum

In the literature on the 2024 Türkiye Century Maarif Model science curriculum, the following studies have been conducted: teacher perceptions (Ak & Köse, 2024; Bakırcı et al., 2025; Bilir, 2025), comparisons with the previous curriculum (Bilir, 2025; Demir et al., 2026; Kart, 2026; Özdemir & Özer, 2025; Torun & Karamustafaoğlu, 2025; Yener et al., 2025; Yılmaz & Çetin, 2025), taxonomic analysis (Özkan & Yaman, 2025), skills analysis (Cengiz, 2026; Harmanlı et al., 2025; Konanç & Bilir, 2025), textbook analysis (Pınar, 2025), learning objectives and student profiles (Yıldırım & Çalışkan, 2024), alignment between learning outcomes and textbooks (Günaltay & Kocaman, 2025), sustainability (Çelebi & Ekici, 2025; Yüzbaşıoğlu, 2025) and international comparison (Demirci et al., 2025; Mercan et al., 2026). In these studies, science teachers generally formed the sample (Ak & Köse, 2024; Bakırcı et al., 2025; Bilir, 2025), whilst the 2018 and 2024 curriculum texts were predominantly used as data collection tools (Bilir, 2025; Demir et al., 2026; Kart, 2026; Özdemir & Özer, 2025; Torun & Karamustafaoğlu, 2025; Yılmaz & Çetin, 2025) have been analysed directly; a semi-structured interview form (Ak & Köse, 2024; Bakırcı et al., 2025), a textbook (Pınar, 2025) and an assessment rubric (Günaltay & Kocaman, 2025) were utilised.

Studies indicate that the programme adopts a holistic, student-centred and skills-based approach rather than a knowledge-centred structure (Bilir, 2025; Kadirhan & Şat, 2025). Bakırcı et al. (2025) and Ak and Köse (2024) examined teachers' views on the programme's structure, innovations, the impact of PISA-TIMSS and the benefits of in-service training. Their findings revealed that, whilst teachers viewed simplification, the concept of 'learning outcomes' and an experiment/activity-focused approach positively, they also noted that process-

oriented assessment was challenging in large classes and that schools lacked adequate infrastructure (Ak & Köse, 2024; Bakırcı et al., 2025).

Comparative studies have revealed that the 2018 and 2024 programmes exhibit significant differences in terms of vision and content (Bilir, 2025; Demir et al., 2026; Özdemir & Özer, 2025; Torun & Karamustafaoğlu, 2025). Whilst the 2024 curriculum maintains the objective of scientific literacy, it has broadened its vision to include concepts such as sustainability, social-emotional skills, digital transformation and global citizenship; the number of learning outcomes has been reduced by 40 per cent to 182, the units have been restructured, and the topic of sustainable living has been added at the end of each unit (Bilir, 2025; Torun & Karamustafaoğlu, 2025). Whilst the curricula show similarities in terms of assessment, it has been observed that the assessment tools in the 2024 curriculum are more detailed (Yılmaz & Çetin, 2025). In international comparisons, it has been found that the Maarif Model is more structured than the Chinese curriculum, whilst being less tiered and featuring more detailed content than the Singaporean curriculum (Demirci et al., 2025; Mercan et al., 2026).

Taxonomic analyses have revealed that learning outcomes are concentrated primarily at the ‘conceptual knowledge’ and ‘analysis’ levels, whilst factual knowledge and recall skills are not given sufficient attention (Özkan & Yaman, 2025). Cengiz (2026) examined the distribution of higher-order thinking skills in learning outcomes, whilst Konanç and Bilir (2025) examined that of systems thinking skills; it was observed that conceptual skills were concentrated in Year 4 of primary school and Year 8 of lower secondary school (Harmanlı et al., 2025).

The studies generally focus on organising in-service training and workshops for teachers (Ak & Köse, 2024; Bilir, 2025; Çelebi & Ekici, 2025; Torun and Karamustafaoğlu, 2025; Yener et al., 2025; Yılmaz & Çetin, 2025), reducing class sizes (Torun & Karamustafaoğlu, 2025), the inclusion of STEM/engineering units (Demirci et al., 2025), the enrichment of learning outcomes (Özkan & Yaman, 2025), the wider adoption of process-oriented assessment tools (Kart, 2026), an increase in out-of-school learning environments (Yener et al., 2025) and the provision of sample assessment tools (Özdemir & Özer, 2025).

In conclusion, the literature emphasises that the 2024 Science Curriculum possesses a blended and holistic structure that addresses modern educational needs; however, it highlights the need to support teachers during the implementation phase in terms of material support, methodological knowledge and new assessment and evaluation techniques (Ak & Köse, 2024; Bakırcı et al., 2025; Torun & Karamustafaoğlu, 2025; Yüzbaşıoğlu, 2025).

Rationale for the Study

This study aims to determine the levels of curriculum literacy self-efficacy among science teachers regarding the 2024 Türkiye Century Maarif Model Science Curriculum, and to establish whether these levels vary according to various demographic variables. To achieve this aim, the following questions were addressed:

1. What is the level of curriculum literacy amongst science teachers?
2. Do science teachers' curriculum literacy levels differ significantly according to gender, years of professional experience, type of faculty graduated from, educational level, type of school currently working at, preferred method of obtaining a lesson plan, status of reviewing the current curriculum, and status of participating in curriculum-related training?

METHODOLOGY

In this study, the descriptive survey method—one of the quantitative research methods—was employed. Survey models are research approaches aimed at describing a situation that existed in the past or currently exists (Karasar, 2008).

Sample

The sample for this study consisted of 100 science teachers working in secondary schools under the Ministry of National Education during the 2025–2026 academic year. Criteria sampling, one of the purposeful sampling methods, was used to determine the sample; this method aims to work with a group that meets the criteria specified by the researcher (Yıldırım & Şimşek, 2016). The criterion for the study was that participants were teachers teaching in Years 5 and 6, where the 2024 Türkiye Century Maarif Model Science Curriculum was being implemented. Indeed, the programme was first implemented in Year 5 in the 2024–2025 academic year and in Year 6 in the 2025–2026 academic year. In this context, determining the levels of programme literacy self-efficacy among Year 5 and Year 6 science teachers—who are the implementers of the programme—is important for evaluating the programme's effectiveness and identifying potential needs. The frequency and percentage distribution of participants according to the variables is presented in Table 1.

Table 1. Frequency and percentage distribution of participants by variables

Variables		Gender		Total (f)	Total (%)
		Female (f)	Male (f)		
Years of professional experience	0-5	35	2	37	37
	6-10	16	1	17	17
	11-15	10	5	15	15
	16-20	11	2	13	13
	21 and over	14	4	18	18
Type of faculty graduated from	Faculty of Education	83	13	96	96
	Faculty of Arts and Sciences	3	1	4	4
Educational level	Undergraduate	53	10	63	63
	Graduate	33	4	37	37
Type of school currently working at	Public school	58	12	70	70
	Private school	28	2	30	30
Preferred method of obtaining a lesson plan	I use a ready-made plan	18	6	24	24
	I adapt a ready-made plan	44	5	49	49
	I prepare my own plan by reviewing the curriculum	24	3	27	27
Status of reviewing the current curriculum	Yes	76	14	90	90
	Partially	10	-	10	10
	No	-	-	-	-
Status of participating in curriculum-related training	Yes	50	12	62	62
	No	36	2	38	38
Total		86	14	100	100

When examining the data in Table 1, it is observed that 86% of the participants are female and 14% are male science teachers. In terms of years of professional service, 37% of the participants have 0-5 years, 17% have 6-10 years, 15% have 11-15 years, 13% have 16-20 years, and 18% have 21 years and above of seniority. 96% of the participants graduated from the Faculty of Education, while 4% graduated from the Faculty of Arts and Sciences; 63% hold a undergraduate degree, and 37% hold a graduate degree. In terms of the school where they work, 70% are employed at public schools, while 30% work at private schools. In the preferred type of plan, 24% use a ready-made plan, 49% adapt a ready-made plan, and 27% use their own plan by reviewing the program. while 10% reviewed it only partially, 62% attended a training on the current curriculum, while 38% did not.

Data Collection Tools

In this study, the ‘Curriculum Literacy Self-Efficacy Scale’ (CLSES), developed by Erdem and Yücel Toy (2023), was used to determine science teachers’ levels of self-efficacy regarding curriculum literacy in relation to the 2024 Türkiye Maarif Model Science Curriculum. In addition, a personal information form created by the researchers was used to determine the participants’ details (years of service in the profession, type of school from which they graduated, educational background, type of school where they are employed, preferred curriculum type, whether they had reviewed the current curriculum, and whether they had participated in training related to the curriculum).

The CLSES consists of 42 items and a single factor. Participants indicate their level of agreement with the scale items on a five-point Likert scale: ‘not at all’ (1), ‘somewhat’ (2), ‘moderately’ (3), ‘agree’ (4), and ‘strongly agree’ (5). No negative items have been reported in the scale. The highest possible total score on the scale is 210, whilst the lowest is 42. A score approaching 210 indicates a high level of self-efficacy in programme literacy, whilst a score approaching 42 indicates a low level of self-efficacy in programme literacy. The Cronbach’s alpha coefficients for all items on the scale were reported as .98 for the first study group, .99 for the second study group, and the scale as a whole was reported to be highly reliable (Erdem & Yücel Toy, 2023). Based on the data obtained from this study, Cronbach’s alpha coefficient for all items of the scale was found to be 0.990.

Data Analysis

The data obtained from the CLSES scale were analysed using IBM SPSS Statistics 27, and the score distributions were examined using descriptive statistics. To determine the appropriate analytical method (parametric or non-parametric), the skewness and kurtosis values for the participants’ scale scores were calculated; these values are presented in Table 2.

Table 2. Skewness and Kurtosis Values for Scores on the CLSES

Variables		Çarpıklık	SH	Basıklık	SH
CLSES	-	-0.485	0.241	-0.052	0.478
Years of professional experience	0-5	-0.326	0.388	-0.508	0.759
	6-10	-1.215	0.550	2.873	1.063
	11-15	-0.403	0.580	0.313	1.121
	16-20	-1.145	0.616	1.603	1.191
	21 and above	0.047	0.536	-0.568	1.038
Type of faculty graduated from	Faculty of Education	-0.528	0.246	0.040	0.488
	Faculty of Arts and Sciences	0.910	1.014	1.004	2.619
Educational level	Undergraduate	-0.385	0.302	-0.287	0.595
	Graduate	-0.723	0.388	0.955	0.759
Type of school currently working at	Public school	-0.495	0.287	-0.102	0.566
	Private school	-0.546	0.427	0.276	0.833
Preferred method of obtaining a lesson plan	I use a ready-made plan	-0.781	0.472	0.591	0.918
	I adapt a ready-made plan	-0.477	0.340	0.365	0.668
	I prepare my own plan by reviewing the curriculum	-0.428	0.448	-0.748	0.872
Status of reviewing the current curriculum	Yes	-0.569	0.254	0.140	0.503
	Partially	-0.066	0.687	0.167	1.334
	No	-	-	-	-
Status of participating in curriculum-related training	Yes	-0.487	0.304	-0.109	0.599
	No	-0.510	0.383	0.226	0.750

Note. CLSES: Curriculum Literacy Self-Efficacy Scale

The skewness and kurtosis coefficients were examined to determine whether the data met the assumption of a normal distribution. The skewness value for the total CLSES score was found to be −0.485, and the kurtosis value was −0.052. In the analyses conducted according to the independent variables, it was determined that nearly all of the skewness and kurtosis values fell within the ranges of -1.5

to +1.5 proposed by Tabachnick and Fidell (2014) and -2.0 to +2.0 proposed by George and Mallery (2021) (Table 2). Only the kurtosis value for the ‘6–10 years’ service group (2.873) fell outside both criteria; however, given the limited sample size of this group (n = 18) and the fact that all other groups met the assumption of normal distribution, it was concluded that the data generally followed a normal distribution.

Based on these results, it was decided to use parametric tests in the analysis of CLSES scores. An independent samples t-test (Büyüköztürk, 2020) was used to determine differences according to the variables of type of faculty graduated from, educational level, type of school currently working at, and status of participating in curriculum-related training, while a one-way analysis of variance (ANOVA) was used for the variables of years of professional experience, preferred method of obtaining a lesson plan, and status of reviewing the current curriculum. In cases where ANOVA revealed a significant difference, the homogeneity of variances was tested using Levene's test; since variances were found to be homogeneous across all groups, the post hoc Scheffe test was applied to determine the source of the difference (Büyüköztürk, 2020).

Ethical approval for this research was obtained from the Ethics Committee for Science and Engineering at Balıkesir University via an official letter dated 12 May 2026, reference number 19928322-050.04-675358. In this study, all rules specified in the “Guidelines on Scientific Research and Publication Ethics for Higher Education Institutions” have been adhered to. None of the actions listed under the heading “Actions Contrary to Scientific Research and Publication Ethics” in the second section of the Guidelines have been committed.

RESULTS

To address the first research question, descriptive statistics for the CLSES scores were calculated. The descriptive statistics for the CLSES scores are presented in Table 3.

Table 3. Descriptive Statistics for CLSES Scores

Ölçek	N	Minimum	Maksimum	$\bar{x}$	S
CLSES	100	66	210	154.65	33.972

Note. CLSES: Curriculum Literacy Self-Efficacy Scale, N: number of participants, $\bar{x}$: mean, S: standard deviation

An examination of Table 3 reveals that the average CLSES score is 154.65. The highest possible score on the CLSES is 210, whilst the lowest is 66. Given these values, the fact that the average CLSES score is above the midpoint of the scale indicates that the participants’ perceptions of their self-efficacy in programme literacy are at a high level. This finding suggests that the science teachers participating in the study hold positive perceptions regarding their knowledge and skills relating to the curriculum; they consider themselves

competent in the areas of understanding, interpreting, applying and evaluating, and their perceptions of curriculum literacy self-efficacy in these areas are at a high level.

To address the second research question, an independent samples t-test was conducted to determine whether CLSES scores differed according to the variables of gender, type of faculty graduated from, educational level, type of school currently working at, and status of participating in curriculum-related training. The results of the independent samples t-test are presented in Table 4.

Table 4. Results of the t-test for CLSES mean scores according to the relevant variables

Variables		N	$\bar{x}$	S	sd	t	p	Cohen d
Gender	Female	86	154.0116	34.12175	98	-0.464	.644	0.134
	Male	14	158.5714	34.01519				
Educational level	Lisans	63	149.4603	33.99280	98	-2.048	.044	0.419
	Lisansüstü	37	163.4865	32.50181				
Type of faculty graduated from	Faculty of Education	96	155.2708	34.02027	98	0.894	.373	0.456
	Faculty of Arts and Sciences	4	139.7500	33.56958				
Type of school currently working at	Public school	70	152.8143	33.30461	98	-0.824	.412	0.179
	Private school	30	158.9333	35.68824				
Status of participating in curriculum-related training	Yes	62	155.6129	34.33889	98	0.360	.719	0.074
	No	38	153.0789	33.76259				

Note. N: number of participants, $\bar{x}$: mean, S: standard deviation, sd: degree of freedom, p: statistical significance value, Cohen d: effect size coefficient

When Table 4 is examined, it can be seen that mean CLSES scores did not differ significantly according to gender [$t(98)=-0.464$, $p>.05$], type of faculty graduated from [$t(98)=0.894$, $p>.05$], type of school currently working at [$t(98)=-0.824$, $p>.05$], or status of participating in curriculum-related training [$t(98)=0.360$, $p>.05$]. Based on these findings, it can be interpreted that gender, type of faculty graduated from, type of school currently working at, and status of participating in curriculum-related training have no effect on science teachers' curriculum literacy self-efficacy levels.

Table 4 shows that mean CLSES scores differed significantly according to educational level. Teachers with a graduate degree ($\bar{x}=163.49$) had significantly higher curriculum literacy self-efficacy scores than those with an undergraduate degree ($\bar{x}=149.46$). According to the Cohen's d value ($d=0.419$), the observed difference indicates a small effect size (Cohen, 1988; Lenhard & Lenhard, 2022). Based on this finding, it can be interpreted that educational level has a small effect on teachers' curriculum literacy self-efficacy levels.

To address the second research question, descriptive statistics for CLSES were calculated according to the variables of years of professional experience,

preferred method of obtaining a lesson plan, and status of reviewing the current curriculum (Table 5).

Table 5. Descriptive Statistics for CLSES Scores

Variables	N	$\bar{x}$	S
Years of professional experience	37		
0-5		150.6216	35.70975
6-10	17	158.6471	33.16614
11-15	15	142.9333	32.70751
16-20	13	170.0769	37.20766
21 and over	18	157.7778	27.60979
Total	100	154.6500	33.97217
Preferred method of obtaining a lesson plan	24		
I use a ready-made plan		151.8333	34.41646
I adapt a ready-made plan	49	153.1837	32.48120
I prepare my own plan by reviewing the curriculum	27	159.8148	36.87613
Total	100	154.6500	33.97217
Status of reviewing the current curriculum	90		
Yes		156.8889	33.66272
Partially	10	134.5000	31.43335
No	-	-	-
Total	100	154.6500	33.9722

Note. N: number of participants, $\bar{x}$: mean, S: standard deviation

When Table 5 is examined, it can be seen that CLSES scores differed according to the variables of years of professional experience, preferred method of obtaining a lesson plan, and status of reviewing the current curriculum. An ANOVA was carried out on independent samples to test the statistical significance of the observed differences (Table 6).

Table 6. ANOVA Results for CLSES Average Scores by Variables

Variables	Source of Variance	Sum of Squares	sd	Mean of Squares	F	p	η^2	A significant difference
Years of professional experience	Between groups	6201.197	4	1550.299	1.363	0.253	0.054	-
	Within groups	108055.553	95	1137.427				
	Total	114256.750	99					
Preferred method of obtaining a lesson plan	Between groups	1015.996	2	507.998	0.435	0.648	0.009	-
	Within groups	113240.754	97	1167.430				
	Total	114256.750	99					
Status of reviewing the current curriculum	Between groups	4511.361	1	4511.361	4.029	0.047	0.039	Evet>Kısmen
	Within groups	109745.389	98	1119.851				
	Total	1142256.750	99					

Note. CLSES: Curriculum Literacy Self-Efficacy Scale, sd: degree of freedom, p; statistical significance value, η^2 : Eta-squared effect size

When the ANOVA results in Table 6 are examined, it can be seen that the mean total scores did not differ significantly according to years of professional experience [$F(4,95)=1.363$, $p>.05$] or preferred method of obtaining a lesson plan [$F(2,97)=0.435$, $p>.05$]. Based on these findings, it can be interpreted that years of professional experience and type of lesson plan preparation had no significant effect on participants' mean total scores.

In contrast, mean total scores differed significantly according to the status of reviewing the current curriculum, [$F(1,98)=4.029$, $p<.05$, $\eta^2=0.039$]. The effect size ($\eta^2=0.039$) corresponds to a small-to-medium effect (Lenhard & Lenhard, 2022). Examining the mean difference between the two groups, it was determined that the mean score of participants who reviewed the current curriculum ("Yes"; $\bar{x}=156.89$) was significantly higher than that of participants who partially reviewed it ("Partially"; $\bar{x}=134.50$).

CONCLUSION, DISCUSSION AND RECOMMENDATIONS

This study aimed to determine science teachers' curriculum literacy self-efficacy levels regarding the 2024 Türkiye Century Maarif Model science curriculum, and to reveal whether these levels differ according to years of professional experience, type of faculty graduated from, educational level, type of school currently working at, preferred method of obtaining a lesson plan, status of reviewing the current curriculum, and status of participating in curriculum-related training. In this section, the study's findings are discussed in comparison with findings from the relevant literature.

The first major finding of the study indicates that science teachers' curriculum literacy self-efficacy perceptions are at a high level. This finding suggests that

the participating teachers hold positive perceptions regarding their knowledge and skills related to the curriculum, and that they consider themselves competent in understanding, interpreting, implementing, and evaluating it. This result is thought to stem from the fact that the majority of participating teachers had reviewed the current curriculum, as well as from teachers being required to directly follow the curriculum due to their obligation to prepare lesson plans, design activities, and manage assessment-evaluation processes. The fact that more than half of the participants had attended training related to the curriculum may also have made an indirect contribution to this process. Indeed, according to Bandura's (1977) self-efficacy theory, individuals' perceptions of competence in a given domain are nourished by sources such as direct mastery experiences, vicarious learning, verbal persuasion, and physiological/emotional states. From this perspective, it can be said that the direct experience and knowledge accumulation teachers gained through reviewing the curriculum and participating in related training may have contributed to the increase in their curriculum literacy self-efficacy perceptions.

A review of the relevant literature reveals both similar and differing results. The majority of studies conducted with teachers (Adem, 2023; Aksoy & Şahin, 2025; Atlı et al., 2021; Coşkun Demirdağ & Taşgın, 2024; Güleş & Kılınç, 2024; Kuyubaşoğlu, 2019; Nasırcı, 2022; Sakallı & Altun, 2026; Saracaloğlu & Çetin, 2023; Tabak & Doğuş, 2025) and pre-service teachers (Aslan, 2019; Erdem & Eğmir, 2018; Karaduman, 2024; Sural & Dedeşali, 2018) found curriculum literacy proficiency to be at a high level. Kahramanoğlu (2019) and Saral (2019), on the other hand, identified a moderate level of proficiency. This discrepancy is thought to stem from the fact that these studies were conducted with teachers from different branches and in relation to different curricula, using different self-developed measurement instruments.

Regarding the gender variable, no significant difference was found in curriculum literacy self-efficacy levels according to gender in this study. This finding suggests that curriculum literacy is a domain developed through professional experience, individual effort, and professional activities rather than gender. A review of the relevant literature reveals studies with similar findings (Adem, 2023; Aslan, 2019; Atlı et al., 2021; Erdem & Eğmir, 2018; Tabak & Doğuş, 2025). Güleş and Kılınç (2024), in a study conducted with primary school teachers, also found no significant difference in curriculum literacy proficiency according to gender, attributing this to the fact that female and male teachers are mixed together in both pre-service and in-service training. In contrast, some studies have identified significant differences according to gender: Saracaloğlu and Çetin (2023), along with Kahramanoğlu (2019) and Saral (2019), found a difference in favor of female teachers. Coşkun Demirdağ and Taşgın (2024), on the other hand, found a result in favor of women in preschool teachers in terms

of the attitude sub-dimension and total score, while Erdamar (2020) identified a difference in favor of men.

Regarding the variable of type of faculty graduated from, no significant difference was found between graduates of faculties of education and faculties of arts and sciences in terms of curriculum literacy self-efficacy in this study. This result is thought to stem from the fact that 96% of the study group were graduates of faculties of education, and that teachers work in line with a common curriculum, similar professional standards, and similar implementation processes regardless of the type of faculty they graduated from. A review of the relevant literature shows that Saracaloğlu and Çetin (2023) also found that graduating from a faculty of education versus a faculty of arts and sciences did not create a difference in terms of curriculum literacy, a result that parallels the findings of the present study. Aslan (2018), while similarly finding no significant difference according to type of faculty graduated from, determined that the planning sub-dimension of the curriculum was performed better by teachers who graduated from faculties of education. Sakallı and Altun (2026), in a study conducted with secondary school mathematics teachers, found that graduates of faculties of education were more competent than graduates of faculties of arts and sciences in terms of the inquiry sub-dimension of curriculum competence, and explained this by the fact that the professional courses taken in faculties of education provide teachers with a different perspective.

Regarding the variable of type of school currently working at, no significant difference was found between teachers working in public schools and those working in private schools in terms of curriculum literacy self-efficacy in this study. This finding is thought to stem from the fact that teachers in both types of schools are required to implement the same national curriculum (The 2024 Türkiye Century Maarif Model). The literature includes studies with results similar to those of the present study (Keskin, 2020; Tabak & Doğuş, 2025). In contrast, Atlı et al. (2021) found that the curriculum literacy proficiency of physical education teachers working in private schools differed significantly from, and was higher than, that of teachers working in public schools. The researchers suggested that this difference may stem from private school teachers being more closely supervised by their schools and having more limited job security compared to public school teachers.

Regarding the variable of status of participating in curriculum-related training, no significant difference was found between teachers who had and had not participated in such training in terms of curriculum literacy self-efficacy levels in this study. This suggests that curriculum literacy may be a construct shaped through professional experience and continuous practice rather than through a single training activity. In the literature, while Keskin (2020) similarly found no significant difference, Atlı et al. (2021) determined that the curriculum literacy

perception levels of teachers who participated in in-service training were significantly higher than those who did not.

A significant difference was found in participants' curriculum literacy self-efficacy according to the educational level variable. Teachers with a graduate degree had significantly higher curriculum literacy self-efficacy scores than those with an undergraduate degree, although the observed difference had a small effect size. Similar findings are present in the literature (Aslan, 2018; Demir & Toraman, 2021; Coşkun Demirdağ & Taşgın, 2024; Güleş & Kılınç, 2024). Indeed, teachers with graduate education have been found to perform better than those with undergraduate degrees in recognizing the curriculum (Güleş & Kılınç, 2024) and in their competence to interpret the knowledge-skill sub-dimensions of the curriculum (Demir & Toraman, 2021). Taken together, these findings suggest that graduate education may be positively associated with teachers' curriculum literacy proficiency. In contrast, Keskin (2020) found no significant difference in curriculum literacy perception levels between teachers with undergraduate and graduate degrees.

No significant difference was found in participants' curriculum literacy self-efficacy according to years of professional experience. This suggests that curriculum literacy is not a competency acquired solely through professional experience. This may be because newly appointed teachers' efforts to learn everything about the teaching profession and avoid gaps in their knowledge may have offset differences in seniority; similarly, teachers' skills in understanding, interpreting, and implementing curricula may be related not so much to their years of service as to their participation in individual professional development activities, their efforts to keep up with current curricula, and the experience they gain through implementing the curriculum. Similar results are found in the literature (Aksoy & Şahin, 2025; Kahramanoğlu, 2019; Sakallı & Altun, 2026; Saracaloğlu & Çetin, 2023; Tabak & Doğuş, 2025). In contrast, Barut and Gündoğdu (2023) found that the curriculum literacy proficiency of teachers with 6–10 years of seniority was lower than that of teachers with 11–15 years and more than 25 years of seniority.

No significant difference was found in participants' curriculum literacy self-efficacy according to the preferred method of obtaining a lesson plan. This suggests that, regardless of whether teachers prepare their lesson plans by using a ready-made plan, adapting a ready-made plan, or developing their own plan by reviewing the curriculum, the common reference point for all of these approaches is the curriculum being implemented; indeed, whichever method is preferred, lesson plans are developed by taking into account the curriculum's content, learning outcomes, and teaching-learning and assessment-evaluation processes.

According to the examination of the current curriculum, it has been determined that there is a significant difference in the program literacy self-

efficacy of the participants. In the study, it was found that teachers who examined the current curriculum had higher levels of program literacy self-efficacy compared to teachers who partially examined the curriculum. This result may indicate that regularly reviewing the program is an important factor supporting program literacy. Atlı et al. (2021) revealed in their evaluation based on the importance given to the curriculum and the extent of its use that teachers who value the curriculum and frequently use it have a higher level of curriculum literacy. Keskin (2020), on the other hand, identified a significant difference between teachers who found the curriculum "important" and those who found it "very important," noting that the difference was in favour of the group that found it "very important." When these findings are evaluated together, it can be said that the importance given to the curriculum and the frequent use of the program can increase the level of program review and, consequently, the self-efficacy of program literacy.

In conclusion, this study revealed that science teachers' curriculum literacy self-efficacy perceptions were high; these perceptions did not differ according to gender, type of faculty graduated from, type of school currently working at, status of participating in curriculum-related training, years of professional experience, and preferred method of obtaining a lesson plan; however, they differed according to educational level and status of reviewing the current curriculum.

Based on the results of the research, the following recommendations have been developed:

1. Teachers should be encouraged to regularly review the curriculum, and their active participation in the curriculum review processes should be ensured.
2. Teachers can be directed towards postgraduate education to improve themselves professionally.
3. In-service training can be provided to teachers aimed at improving program literacy.
4. The research can be repeated in different provinces and with a larger sample.
5. By incorporating qualitative research methods, teachers' views and suggestions regarding program literacy can be examined in depth.

REFERENCES

- Adem, S. (2023). Branş öğretmenlerinin program okuryazarlık düzeylerinin incelenmesi. *Uluslararası Temel Eğitim Çalışmaları Dergisi*, 4(1), 14-20. <https://doi.org/10.59062/ijpes.1267978>
- Ak, B. S., & Köse, M. (2024). 2024 Fen bilimleri dersi öğretim programı hakkında öğretmen görüşlerinin incelenmesi. *Akademik Platform Eğitim ve Değişim Dergisi*, 7(2), 132-169. <https://doi.org/10.55150/apjec.1582677>
- Aksoy, A., & Şahin, Ç. (2025). Okul öncesi öğretmenlerinin program okuryazarlığı düzeylerinin belirli değişkenler açısından incelenmesi. *Sınıf Öğretmenliği Araştırmaları Dergisi (SÖAD)*, 5(2), 123-135. <https://soaddergi.com/wp-content/uploads/2025/12/123-135.pdf>
- Arslan, M. (2000). Cumhuriyet dönemi ilköğretim programları ve belli başlı özellikleri. *Milli Eğitim Dergisi*. Retrieved on July 25, 2021, https://dhgm.meb.gov.tr/yayimlar/dergiler/Milli_Egitim_Dergisi/144/arslan.htm
- Aslan, S. (2018). *Ortaokul öğretmenlerinin program okuryazarlık düzeyleri*. (Yayınlanmamış Yüksek Lisans Tezi), Hacettepe Üniversitesi, Ankara.
- Aslan, S. (2019). An analysis of prospective teachers' curriculum literacy levels in terms of reading and writing. *Universal Journal of Educational Research*, 7(4), 973-979. <https://doi.org/10.13189/ujer.2019.070408>
- Atlı, K., Kara, Ö., & Mirzeoğlu, A. D. (2021). Beden eğitimi öğretmenlerinin program okuryazarlık düzeylerine yönelik algılarının bazı değişkenlere göre incelenmesi. *Gazi Journal of Physical Education and Sport Sciences*, 26(2), 281-299. <https://izlik.org/JA73WT78EK>
- Aygün, H. E. (2019). The prediction of the teaching readiness level of pre-service teachers in terms of curriculum literacy. *International Journal of Curriculum and Instructional Studies*, 9(2), 203-220.
- Bakırcı, H., Yucakuş, D. N., & Karşı, G. (2025). Fen Bilimleri dersi öğretim programı hakkında fen bilimleri öğretmenlerinin görüşleri: Maarif Modeli örneği. *Uluslararası Anadolu Sosyal Bilimler Dergisi*, 9(4), 1049-1070. <https://doi.org/10.47525/ulasbid.1799269>
- Bandura, A. (1977). *Self-efficacy: Toward a unifying theory of behavioral change*. *Psychological Review*, 84(2), 191-215. <https://doi.org/10.1037/0033-295X.84.2.191>
- Barut, C., & Gündoğdu, K. (2023). Öğretmenlerin program okuryazarlık düzeyleri ve eleştirel düşünme eğilimleri arasındaki ilişki. *Bayburt Eğitim Fakültesi Dergisi*, 18(37), 198-217. <https://doi.org/10.35675/befdergi.1213781>

- Bilir, U. (2025). Türkiye’de 2018 ve 2024 yılları fen bilimleri dersi öğretim programlarının temel öğeler açısından karşılaştırılması. *Millî Eğitim*, 54(246), 793-836. <https://doi.org/10.37669/milliegitim.1532604>
- Bolat, Y. (2017). Eğitim programı okuryazarlığı kavramı ve eğitim programı okuryazarlığı ölçeği. *Turkish Studies*, 12(18), 121-138. <http://dx.doi.org/10.7827/TurkishStudies.12103>
- Büyüköztürk, Ş. (2020). *Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum*. Pegem Akademi.
- Cengiz, E. (2026). Türkiye Yüzyılı Maarif Modeli fen bilimleri dersi öğretim programlarının üst düzey düşünme becerileri bakımından incelenmesi. *Uşak Üniversitesi Eğitim Araştırmaları Dergisi*, 12(1), 115-139. <https://doi.org/10.29065/usakead.1759342>
- Coşgun Demirdağ, M., & Taşgın, A. (2024). Okul öncesi öğretmenlerinin program okuryazarlığı ve öğretim programı farkındalıkları arasındaki ilişki. *Uluslararası Türkçe Edebiyat Kültür Eğitim (TEKE) Dergisi*, 13(3), 1035-1052. <https://doi.org/10.7884/teke.1508003>.
- Çelebi, M. E., & Ekici, F. T. (2025). Türkiye Yüzyılı Maarif Modelinde fen bilimleri öğretim programında sürdürülebilirlik ve sürdürülebilir kalkınma hedefleri ile ilişkisi. *Denizli İl Millî Eğitim Müdürlüğü Bilim ve Eğitim Dergisi*, 1(1), 29-39. <https://izlik.org/JA83SS39GF>
- Demir, E. M., Aydın, S., & Şahin, Ü. D. (2026). Comparative analysis of the 2018 and 2024 science curricula: A study within the context of the Türkiye's Century Education Model. *Social Sciences*, 16(1), 1-8. https://www.jocress.com/eng/comparative-analysis-of-the-2018-and-2024-science-curricula-a-study-within-the-context-of-the-turkiyes-century-education-model_525
- Demir, E., & Toraman, Ç. (2021). Öğretmenlerin eğitim programı okuryazarlığı düzeyleri. *Trakya Eğitim Dergisi*, 11(3), 1516-1528. <https://doi.org/10.24315/tred.858813>
- Demir, G. T. (2021). Cumhuriyetin ilk yıllarında fen eğitimi: 1924-26 müfredatlarına göre süreli yayınların analizi. *Gaziantep University Journal of Social Sciences*, 20(1), 259-278. <https://doi.org/10.21547/jss.742856>
- Demirci, B., Bilge, Y., İnce, E. S., & Güven, İ. (2025). Türkiye Yüzyılı Maarif Modeli ve Çin fen bilimleri programlarının karşılaştırmalı olarak incelenmesi. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, 48, 20-43. <http://dx.doi.org/10.14582/zgefd.1729454>
- Dindar, H., & Taneri, A. (2011). MEB’in 1968, 1992, 2000 ve 2004 yıllarında geliştirdiği fen programlarının amaç, kavram ve etkinlik yönünden karşılaştırılması. *Kastamonu Eğitim Dergisi*, 19(2), 363-378. <https://izlik.org/JA27JE24UX>

- Dönmez, İ., & Gülen, S. (2023). Science curricula and science teachers training in Turkey: Past, present and future. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi*, 17(2), 939-970. <https://doi.org/10.17522/balikesirnef.1371492>
- Erdamar, F. S. (2020). *Sınıf Öğretmenlerinin program okuryazarlık algıları ve ilkökul yöneticilerinin öğretmenlerin program okuryazarlık becerisine yönelik algılarının ilerlemeci felsefe bağlamında analizi* [Yayınlanmamış Doktora Tezi]. Fırat Üniversitesi.
- Erdem, C., & Eǧmir, E. (2018). Öğretmen adaylarının eğitim programı okuryazarlığı düzeyleri. *Afyon Kocatepe Üniversitesi Sosyal Bilimler Dergisi*, 20(2), 123-138. <https://doi.org/10.32709/akusosbil.428727>
- Erdem, S. (2023). *Öğretmenlere yönelik program okuryazarlığı hizmet içi eğitim programının geliştirilmesi: Tasarım, uygulama ve değerlendirme* [Yayınlanmamış Doktora tezi]. Yıldız Teknik Üniversitesi.
- Erdem, S., & Yücel Toy, B. (2023). Program okuryazarlığı öz-yeterlik ölçeği (POKÖYÖ): Geçerlik ve güvenirlik çalışması. *Ulakbilge Sosyal Bilimler Dergisi*, 11(89), 991–1005. <https://doi.org/10.7816/ulakbilge-11-89-01>
- George, D., & Mallery, P. (2021). *IBM SPSS statistics 26 step by step: A simple guide and reference* (17th edition). Routledge. <https://doi.org/10.4324/9781003205333>
- Güleş, E., & Kılınç, H. H. (2024). İlkokul ve ortaokul öğretmenlerinin program okuryazarlıkları ve öğretim programına bağlılıklarının incelenmesi. *Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi*, 14(1), 183-202. <https://doi.org/10.30783/nevsosbilen.1413385>
- Günaltay, M. İ., & Kocaman, B. (2025). Maarif model middle school 5th grade science course curriculum learning outcomes and content relationship. *International Journal of Field Education*, 11(2), 75-87. <https://izlik.org/JA28AB54BF>
- Harmanlı, P., Özdemir, H. F., & Nayci, Ö. (2025). Öğretim programlarında kavramsal becerilerin sarmal yapısı: Türkiye Yüzyılı Maarif Modeli bağlamında bir değerlendirme. *Manisa Celal Bayar Üniversitesi Eğitim Fakültesi Dergisi*, 13(2), 769–793. <https://doi.org/10.52826/mcbuefd.1732512>
- Kadirhan., Z., & Şat, M. (2025). A review of studies on the Türkiye Century Maarif Model. *Primary Education*, 7(27), 84-105. <https://izlik.org/JA58BR76BM>
- Kahramanoğlu, R. (2019). Öğretmenlerin öğretim programı okuryazarlığına yönelik yeterlik düzeyleri üzerine bir inceleme. *Uluslararası Sosyal Araştırmalar Dergisi*, 12(65), 827-840. <http://dx.doi.org/10.17719/jisr.2019.3495>

- Karaduman, B. (2024). How curriculum literate are the implementers of the curriculum? *Educational Academic Research*, 54, 132-143. <https://doi.org/10.33418/education.1416962>
- Karasar, N. (2008). *Bilimsel araştırma yöntemi*. Nobel Yayın.
- Kart, M. (2026). Kazanımdan öğrenme çıktısına: 2018 ve 2024 fen bilimleri dersi öğretim programlarında öğrenme hedeflerinin yapısal dönüşümü. *Akademik Eğitim ve Sosyal Bilimler Dergisi (AJESS Journal)*, 3(2), 90-102. <https://doi.org/10.5281/zenodo.18721767>.
- Keskin, A. & Korkmaz, H. (2021). Öğretmenlerin öğretim programı okuryazarlığı algı ölçeği'nin geliştirilmesi. *Türk Eğitim Bilimleri Dergisi*, 19(2), 857-884. <https://doi.org/10.37217/tebd.917130>
- Keskin, A. (2020). *Öğretmenlerin öğretim programı okuryazarlık düzeylerine yönelik algılarının belirlenmesi* [Yayınlanmamış Doktora Tezi]. Hacettepe Üniversitesi.
- Konanç, S., & Bilir, V. (2025). Türkiye yüzyılı maarif modeli fen bilimleri dersi öğretim programının sistem düşüncesi becerileri açısından incelenmesi. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi*, 66, 4432-4454. <https://doi.org/10.53444/deubefd.1644771>
- Kuyubaşoğlu, R. M. (2019). Öğretmenlerin eğitim programı okuryazarlığı yeterliklerinin incelenmesi [Yayınlanmamış Yüksek Lisans Tezi]. Mersin Üniversitesi.
- Lenhard, W., & Lenhard, A. (2017). *Computation of effect sizes*. <https://doi.org/10.13140/RG.2.2.17823.92329>
- Mercan, B, Balkan Kıyıcı, F., & Beşoluk, Ş. (2026). Fen bilimleri dersi öğretim programlarının karşılaştırılması: Türkiye ve Singapur. *Bolu Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 26(1), 399-418. <https://doi.org/10.17240/aibuefd.2026.-1695007>
- Millî Eğitim Bakanlığı [MEB], (2000). *Tebliğler dergisi*, 63(2518), 1000-1105. Ankara: Millî Eğitim Bakanlığı.
- Millî Eğitim Bakanlığı [MEB], (2024). *Fen bilimleri dersi öğretim programı (3, 4, 5, 6, 7 ve 8. sınıflar)*. Millî Eğitim Bakanlığı Temel Eğitim Genel Müdürlüğü.
- Millî Eğitim Bakanlığı [MEB], (2018). *Fen bilimleri dersi öğretim programı (ilkokul ve ortaokul 3, 4, 5, 6, 7 ve 8. sınıflar)*. Millî Eğitim Bakanlığı Temel Eğitim Genel Müdürlüğü.
- Millî Eğitim Bakanlığı [MEB], (2013). *İlköğretim kurumları (ilkokullar ve ortaokullar) fen bilimleri dersi (3, 4, 5, 6, 7 ve 8. sınıflar) öğretim programı*. Ankara: Talim ve Terbiye Kurulu Başkanlığı.

- Millî Eğitim Bakanlığı [MEB]. (2005). *İlköğretim Fen ve Teknoloji Dersi (4-5. Sınıflar) Öğretim Programı*. Talim ve Terbiye Kurulu Başkanlığı.
- Millî Eğitim Bakanlığı [MEB]. (1992). *İlköğretim Fen Bilgisi Dersi Programı*. Talim ve Terbiye Kurulu Başkanlığı.
- Nasırcı, H. (2022). *Sınıf Öğretmenlerinin eğitim programı okuryazarlığı yeterliklerinin incelenmesi* [Doktora Tezi], Çukurova Üniversitesi.
- Özdemir, A., & Özer, B. (2025). 2018 fen bilimleri öğretim programı ile 2024 fen bilimleri öğretim programının karşılaştırılması. *Anadolu Türk Eğitim Dergisi*, 7(2), 281-297. <https://doi.org/10.29329/ated.2025.1388.7>
- Özkan, C., & Yaman, S. (2025). 2024 fen bilimleri dersi öğretim programı öğrenme çıktılarının yenilenmiş Bloom taksonomisine göre incelenmesi. *Turkish Journal of Primary Education (TUJPED)*, 10(1), 1-16. <https://izlik.org/JA75DK24GJ>
- Pınar, M. A. (2025). Türkiye Yüzyılı Maarif Modeli erdem-değer-eylem çerçevesi bağlamında 5. sınıf fen bilimleri ders kitabının incelenmesi. *International Journal of Düzce Educational Sciences*, 3(2), 92-116. <https://doi.org/10.62195/ijdes.1701554>
- Sakallı, Ç., & Altun, S. (2026). Ortaöğretim matematik öğretmenlerinin öğretim programı okuryazarlığı algı düzeyleri. *Education Sciences*, 21(1), 1-25. <http://dx.doi.org/10.12739/NWSA.2026.21.1.1C0709>
- Saracaloğlu, A. S., & Çetin, İ. G. (2023). Öğretmenlerin eğitim programı okuryazarlığı ve mesleki öz-yeterlik düzeyleri arasındaki ilişkinin incelenmesi. *Amasya Üniversitesi Eğitim Fakültesi Dergisi*, 12(1), 61-74. <https://doi.org/10.17539/amauefd.1243057>
- Saral, N. Ç. (2019). *Exploring curriculum literacy level of English Language teachers in Turkey* [Yayınlanmamış Yüksek Lisans Tezi]. Gazi Üniversitesi.
- Sural, S., & Dedebe, N. C. (2018). A study of curriculum literacy and information literacy levels of teacher candidates in department of social sciences education. *European Journal of Educational Research*, 7(2), 303-317. <https://doi.org/10.12973/eu-jer.7.2.303>
- Şahin, M. (2009). Cumhuriyetin kuruluşundan günümüze Türkiye’de hayat bilgisi dersi programlarının gelişimi. *The Journal of International Social Research*, 2(8), 402-410.
- Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th edition). Pearson Publication.
- Tabak, S., & Doğuş, F. (2025). Öğretmenlerin öğretim programı okuryazarlığı, Öğretim programına bağlılığı ve öğretmen özerkliği düzeyleri arasındaki ilişkinin analizi. *Uluslararası Türkçe Edebiyat Kültür Eğitim (TEKE) Dergisi*, 14(4), 1402-1431. <https://doi.org/10.7884/teke.1642131>

- Torun, B., & Karamustafaoğlu, O. (2025). 2024 Türkiye Yüzyılı Maarif Modeli ile 2018 fen bilimleri dersi öğretim programlarının temel öğeler bakımından karşılaştırılması. *Trakya Eğitim Dergisi*, 15(Özel Sayı), 346–388. <https://doi.org/10.24315/tred.1614676>
- Türk Dil Kurumu [TDK], (2023). *Genel açıklamalı sözlük*. TDK Yayınları.
- Yener, Y., Bektaş, M., & Araboğlu, Ü. (2025). İlkokul fen bilimleri dersi 2018-2024 programlarının karşılaştırılması ve 2024 programına genel bakış. *İnönü Üniversitesi Eğitim Fakültesi Dergisi*, 26(3), 2245-2277. <https://doi.org/10.17679/inuefd.1721688>
- Yıldırım, A. & Şimşek, H. (2016). *Sosyal bilimlerde nitel araştırma yöntemleri*. Seçkin.
- Yıldırım, Y., & Çalışkan, A. (2024). Evaluation of the Türkiye Century Education Model in terms of 21st century human profile. *Electronic Journal of Education Sciences*, 13(26), 205-221. <https://doi.org/10.55605/ejedus.1548121>
- Yılmaz, Ş., & Çetin, S. (2025). 2018 ve 2024 fen bilimleri dersi öğretim programlarının ölçme ve değerlendirme yaklaşımları açısından incelenmesi. *Eğitimde Kuram ve Uygulama*, 21(2), 71-88. <https://doi.org/10.17244/eku.1711764>
- Yüzbaşıoğlu, M. K. (2025). Sürdürülebilir yaşam için eğitim: 2024 yılı fen bilimleri dersi öğretim programının sürdürülebilir kalkınma açısından değerlendirmesi. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi*, 63, 948-966. <https://doi.org/10.53444/deubefd.1576551>

CHAPTER 7

**Mathematical Problem Solving in the
Age of Artificial Intelligence: A Proposed
Cognitive Responsibility-Based Model**

Sezer KÖSE Biber¹

Mahir Biber²

¹ Assoc. Prof.,

² Assoc. Prof.

INTRODUCTION

Mathematical problem solving is a central area of research and practice in mathematics education. It involves more than arriving at a correct answer: students must also make sense of the problem, develop appropriate solution paths, make mathematical decisions, and evaluate their results. In Schoenfeld's (1985) problem-solving framework, in addition to mathematical resources and heuristic strategies, control processes involving the monitoring and regulation of the solution process, as well as the beliefs of the individuals solving the problem, are considered important components. Current research in mathematics education likewise conceptualizes problem solving in terms of interrelated activities such as problem posing, representation, exploring different solution paths, formulating conjectures, mathematical reasoning, and justifying results (Santos-Trigo, 2024). Thus, although a correct answer is a necessary outcome of mathematical problem solving, it is not the sole indicator of a student's problem-solving competence.

The use of technology in mathematical problem solving did not begin with generative artificial intelligence (generative AI). Calculators have enabled numerical computation, computer algebra systems have supported symbolic manipulation, and dynamic mathematics environments have made it possible to represent mathematical objects in different ways and rapidly examine numerous cases. Petković and Herceg (2021) show that the role of computers in mathematics has expanded from numerical computation to symbolic computation, visualization, and computer-assisted proof, and that these tools not only accelerate operations but also enable the exploration of new patterns and relationships. Accordingly, the role of digital tools in problem solving is not limited to reducing computational load; they can also shape how students represent problems, which relationships they explore, and how they reason (Santos-Trigo, 2024). Generative AI systems have now become part of this trajectory of technological development.

Generative AI systems are noteworthy because they can play a role in a broader portion of the problem-solving process than earlier tools. These systems can interpret a mathematical problem through natural language and, in addition to producing an answer, can suggest solution strategies, explain solution steps, and generate alternative solution paths. The capacity of AI-supported mathematical assistants to process problems stated in natural language and provide step-by-step mathematical outputs extends the interaction between students and technological tools beyond that afforded by traditional computational tools (Nurhasanah et al., 2026).

The mathematical problem-solving capacity of generative AI systems is related not only to problem type and difficulty level but also to the criteria used to evaluate performance. In the MathOdyssey study, Fang et al. (2025) evaluated models on 387 expert-developed problems at the high school, university, and

olympiad levels and showed that the models could struggle, particularly with complex problems. The researchers further note that evaluations based solely on final-answer accuracy do not directly assess intermediate reasoning, creativity, or the quality of alternative solution paths. From an educational perspective, what is particularly noteworthy is that these systems can generate not only answers but also products—such as strategies, solution steps, and explanations—that have traditionally been regarded as indicators of a student's problem-solving process.

This development brings a fundamental distinction in mathematics education into sharper focus: accessing a solution is not the same as solving a problem. Students can now arrive at a solution without making sense of the problem themselves, developing an appropriate strategy, or constructing the solution steps. However, problem-solving performance demonstrated with AI support may not directly reflect a student's independent problem-solving competence. Salomon et al. (1991) distinguish between performance achieved through activities carried out with technology and the cognitive effects that remain with individuals when technology is no longer used. This distinction is also important in the context of generative AI. In a field experiment conducted in high school mathematics, Bastani et al. (2025) found that AI support improved students' performance on practice exercises; however, students who received standard AI support performed worse than the control group on assessments conducted without AI access. In contrast, students who used an AI version providing teacher-structured hints did not differ significantly from the control group in their performance on assessments without AI access. These findings indicate that the effects of AI on learning are related not only to its capacity to generate solutions but also to how the support provided to students is structured. Therefore, the findings neither support the conclusion that AI use necessarily undermines students' problem solving nor justify the assumption that higher performance achieved with AI support translates into independent learning.

In the age of generative AI, the central pedagogical question is not so much "Should AI be used?" as "Which mathematical thinking processes should students carry out themselves when AI can undertake certain parts of problem solving?" This chapter addresses this issue through the concept of **mathematical cognitive responsibility**. The concept refers to students retaining ownership of mathematical decisions related to sense-making, strategic decision-making, questioning, verification, justification, and reflection throughout the problem-solving process. The aim is not for students to perform every operation independently of technology, but for them to remain responsible for making mathematical decisions that are critical to learning while benefiting from AI support.

THE CHANGING NATURE OF MATHEMATICAL PROBLEM SOLVING

Mathematical problem solving goes beyond applying a familiar operation or an established algorithm; it requires students to organize their existing mathematical knowledge in a new situation and make decisions about how to proceed. Pólya's (1945) approach, commonly summarized as understanding the problem, devising a plan, carrying out the plan, and looking back to evaluate the solution, is one of the foundational frameworks demonstrating that problem solving involves more than performing mathematical operations. These dimensions continue to be used in contemporary empirical studies to assess problem-solving performance. For example, Noviyana et al. (2025) measured mathematical problem solving in terms of understanding the problem, developing a strategy, implementing the strategy, and evaluating the solution. Schoenfeld (1985), however, showed that problem solving cannot be explained solely by the application of particular stages or heuristic strategies; mathematical resources, heuristic strategies, control processes, and belief systems must be considered together. The control dimension is particularly important because it involves deciding which approach to use, monitoring the progress of a solution, and changing strategies when necessary. Thus, a substantial part of mathematical problem solving consists not of performing operations themselves but of making decisions about those operations.

The incorporation of technology into mathematical problem solving has changed how these operations and decisions are distributed between students and technological tools over time. However, the fact that technology performs an operation does not necessarily eliminate the student's mathematical role. Having a calculator perform a numerical computation does not remove the need for students to decide which calculation is required; having a computer algebra system perform symbolic manipulation does not eliminate the need to construct a mathematical model and interpret the result; and using a dynamic mathematics environment to display numerous cases does not make it unnecessary to justify an observed relationship mathematically. Table 1 compares the predominant functions of different technologies and the mathematical responsibilities that remain with students.

Table 1. Predominant functions of different technologies in mathematical problem solving and students' responsibilities

<i>Technology</i>	<i>Predominant function</i>	<i>Student responsibility</i>
<i>Calculator</i>	Performing numerical calculations	Identifying the appropriate operation, interpreting the result, and checking its correctness
<i>Computer algebra systems</i>	Performing symbolic operations and generating different representations	Selecting the mathematical model and operation, and making sense of the output
<i>Dynamic mathematics environments</i>	Supporting visualization, rapid experimentation, and the exploration of mathematical relationships	Interpreting relationships, formulating conjectures, and providing mathematical justification
<i>Generative AI</i>	Interpreting problems and generating strategies, solution steps, explanations, and alternatives	Making sense of solutions, evaluating strategies, questioning, verifying, and justifying

The distinctions presented in Table 1 do not constitute a rigid classification of technologies, as the functions of the same tool may vary depending on the learning goal and task design. The central point is that as the range of mathematical activities that technological tools can perform expands, the student's role in problem solving must also be reconsidered. A distinctive feature of generative AI in this development is its ability to produce not only computational outputs but also observable products of the problem-solving process. Students can therefore turn to technological tools not only for calculations but also with questions such as "How can I solve this problem?", "Is there another method?", or "Why does this result follow?"

The educational implications of this development are not determined solely by the technical capabilities of the tool. In the study by Noviyana et al. (2025), which integrated AI into an active learning approach, the AI-supported problem-based learning group was reported to outperform the traditional instruction group in understanding the problem, developing a strategy, implementing the strategy, and evaluating the solution. However, because AI support was integrated with problem-based learning, the observed differences cannot be attributed to the independent effect of AI. The study provides an example of how AI can be incorporated into a pedagogical structure that requires students to engage actively in problem solving. Accordingly, evaluating the educational role of AI in mathematical problem solving requires attention not only to the operations the system can perform but also to the learning goals for which these operations are used and the mathematical activities that students continue to undertake.

The transformation brought about by generative AI in mathematics education should not be interpreted as eliminating the need for students to engage in mathematical thinking. What has changed is the range of tasks that technology can undertake within the problem-solving process. As this range expands, the

question "Who performs the calculation?" must be accompanied by further questions: "Who makes sense of the problem? Who selects the strategy? Who determines whether the solution is mathematically valid?"

KEY CHALLENGES IN MATHEMATICAL PROBLEM SOLVING WITH GENERATIVE AI

The ability of generative AI to produce outputs at different points in the mathematical problem-solving process raises two fundamental pedagogical issues. The first concerns the extent to which cognitive operations that students would normally perform themselves can be delegated to AI. The second concerns the grounds on which an AI-generated solution can be considered mathematically reliable.

The first issue can be understood in relation to the concept of cognitive offloading. Risko and Gilbert (2016) define cognitive offloading as the use of external actions or resources to reduce the information-processing demands of a task. In this sense, delegating certain operations in mathematical activities to technology is not new and cannot, in itself, be regarded as detrimental to learning. When technology performs a lengthy calculation that is peripheral to the learning goal, students may be able to direct their attention to modeling the problem or interpreting its result. In the context of generative AI, the important change is the expanded scope of activities that can be offloaded. Students may delegate not only numerical or symbolic operations but also problem interpretation, solution strategy, and explanation to an external system. In such cases, students have access to a ready-made solution, but this does not mean that they have themselves engaged in the problem-solving processes through which the solution was produced. Students may follow the generated steps; however, this does not necessarily mean that they can develop an appropriate strategy or make mathematical decisions independently when confronted with the same problem on their own.

Nevertheless, an offloaded solution does not necessarily have to replace students' own thinking. An AI-generated solution can become an object of examination that requires students to identify errors, compare it with another solution, or verify it independently. In this case, rather than taking over the intended mathematical activity entirely, AI can provide a resource for students' questioning and evaluation processes. The learning goal largely determines which forms of cognitive offloading are appropriate. Delegating a calculation that falls outside the learning goal to AI may support learning, whereas obtaining a solution strategy directly from AI when the goal is to develop such a strategy may amount to offloading the very process students are expected to learn. Thus, the same AI function may support students in one task while replacing the thinking they are expected to undertake in another. The pedagogical appropriateness of cognitive offloading should therefore be evaluated not only in terms of which operations

are delegated to AI but also in terms of which mathematical decisions students assume responsibility for after receiving AI output.

The second fundamental issue concerns mathematical reliability. Although current AI models demonstrate considerable capacity for mathematical problem solving, this capacity is not without limitations. Sarıkaya and İnci Kuzu (2026) compared ChatGPT-4o's responses to 38 university-level probability problems at two points eight months apart and found a statistically significant improvement in the model's overall problem-solving performance when all problems were considered together. However, when easy, moderate, and difficult problem categories were examined separately, the differences between the two assessments were not statistically significant. The study also showed that the model could produce incorrect inferences and computational errors in some complex, multi-step problems. Because these findings concern the performance of a particular AI model on a specific set of problems, they cannot be directly generalized to all areas of mathematics or to other AI systems. Furthermore, evidence concerning a model's problem-solving performance should not be regarded as direct evidence of students' mathematical learning processes. Nevertheless, these findings highlight the importance of independently evaluating AI-generated mathematical solutions.

Another dimension of mathematical reliability concerns the extent to which a model remains robust when confronted with different formulations of the same underlying mathematical problem. In the PutnamGAP benchmark, Hao et al. (2025) observed substantial performance declines across 18 models when problems were modified by changing superficial linguistic elements or parameters while preserving the underlying mathematical problem. This finding indicates that a model's ability to arrive at a correct answer for a particular problem formulation does not guarantee that it will reliably maintain the same reasoning when presented with an equivalent formulation. Such benchmark findings do not constitute direct evidence of student learning; however, they provide an important rationale for not treating AI output as a mathematical authority and for preserving independent verification in instructional design.

The evaluation of AI output should not be reduced to checking the final answer. A mathematical solution may arrive at the correct result while containing an incorrect intermediate inference, employ correct operations but provide insufficient justification, or overlook a condition of the problem despite being presented convincingly. Accuracy and reliability have also been identified as major limitations of AI-supported mathematical assistants in recent research (Nurhasanah et al., 2026). At least three questions are therefore important when evaluating an AI-generated solution: Is the final answer correct? Are the solution steps mathematically valid? Is the justification sufficient?

It is important here to distinguish between a student's trust in AI and the mathematical reliability of AI output. A system's record of providing correct answers does not guarantee the correctness of a subsequent mathematical claim. Similarly, a detailed and convincing explanation does not constitute evidence of mathematical validity. Students therefore need to move from accepting a solution because "AI says it is correct" to accepting it because "I can justify its correctness on these mathematical grounds." At this point, the issues of cognitive offloading and reliability converge. When students delegate a mathematical operation or a solution process to AI, their responsibility for evaluating the quality of the resulting output may become even more important. Thus, in the age of generative AI, students do not necessarily have to produce every part of a solution independently from beginning to end. However, leaving processes such as making sense of the problem, evaluating appropriateness, identifying errors, verifying results, and developing justifications for accepting mathematical claims entirely to the authority of a technological tool may undermine the learning dimension of problem solving.

The following questions appear more useful for evaluating the use of generative AI:

- What mathematical work has AI undertaken?
- Is this work part of the learning goal of the current activity?
- What mathematical activity has the student performed after receiving AI output?
- Who makes the final mathematical decision about whether to accept the solution?

Taken together, these questions point to a central issue in mathematical problem solving in the age of generative AI: "How should the division of cognitive labor between students and AI be structured, and which mathematical responsibilities should remain with students?"

MATHEMATICAL COGNITIVE RESPONSIBILITY IN THE AGE OF AI

The ability of generative AI to produce solution paths, explanations, and justifications in mathematical problem solving makes it increasingly difficult to evaluate students' roles solely in terms of who produces the solution. Students may construct a solution entirely on their own, examine an AI-generated solution, choose between two alternatives, identify an error, or independently verify the validity of a solution. Conversely, students may interact extensively with AI while relying entirely on the system to determine how a problem should be interpreted, which strategy should be selected, and why a result is correct. Evaluating the educational role of AI in mathematical learning therefore requires

attention not only to the resulting product or the extent of students' technology use but also to which decisions concerning mathematical thinking students assume responsibility for. This brings the concept of **mathematical cognitive responsibility** to the forefront.

Mathematical cognitive responsibility refers to students taking ownership of mathematical decisions involving sense-making, strategic decision-making, questioning, verification, justification, and reflection throughout problem solving. In this sense, mathematical cognitive responsibility is not synonymous with AI literacy, academic integrity, or performing all operations independently. Rather than whether a technological tool is used, the focus is on ownership of decisions concerning mathematical meaning, strategic choices, and validity.

Although mathematical cognitive responsibility is related to existing theoretical approaches to mathematical problem solving and learning, it has a distinctive conceptual focus that differentiates it from these approaches. The control dimension in Schoenfeld's (1985) problem-solving framework encompasses metacognitive activities such as monitoring the solution process, making strategic decisions, and modifying an approach when necessary. Risko and Gilbert's (2016) account of cognitive offloading explains the transfer of cognitive operations to external resources. In the context of generative AI, Tankelevitch et al. (2024) emphasize that evaluating AI outputs and deciding when and how to use the system require metacognitive monitoring and control.

This conceptualization is also related to approaches in mathematics education concerning intellectual autonomy and norms of mathematical justification. Yackel and Cobb (1996) discuss students' ability to draw on their own intellectual resources when making mathematical decisions and judgments in relation to sociomathematical norms and intellectual autonomy developed within the classroom. In the context of generative AI, however, the external source of mathematical information encountered by students takes on a different character. Zhuang (2026) points out that AI-generated mathematical solutions may contain subtle errors despite appearing reliable and emphasizes that students should be positioned as active evaluators who critically examine and verify these solutions using mathematical criteria, rather than passively accepting them. Similarly, in her account of collective cognitive responsibility, Scardamalia (2002) emphasizes that strategic cognitive activities that guide learning are not solely the responsibility of teachers; students should also assume responsibility for these activities and contribute to the advancement of shared knowledge within the classroom community. The concept of mathematical cognitive responsibility proposed in this chapter aligns with these approaches in emphasizing students' active participation in mathematical decision-making and evaluation. Its distinctive focus, however, concerns how students assume responsibility for mathematical decisions essential to learning and the mathematical criteria they

use to accept AI outputs when generative AI can produce strategies, solutions, and justifications during problem solving. Within this framework, mathematical cognitive responsibility can be examined through the six interrelated dimensions presented in Table 2.

Table 2. Dimensions of mathematical cognitive responsibility in generative AI-mediated problem solving

<i>Dimension</i>	<i>Student's primary responsibility</i>	<i>Key risk when responsibility rests entirely with AI</i>
<i>Sense-Making</i>	Identifying what the problem asks, the given information, conditions, and mathematical relationships	Students do not construct the mathematical structure of the problem situation themselves
<i>Strategic Decision-Making</i>	Evaluating possible solution paths, selecting an appropriate strategy, and modifying it when necessary	Students do not make mathematical decisions about which solution path to follow
<i>Questioning</i>	Examining the assumptions, inferences, and relationships used in the solution	The generated solution is accepted without critical examination
<i>Verification</i>	Testing the validity of the result and solution steps using independent mathematical methods	AI becomes the authority for determining the correctness of its own output
<i>Justification</i>	Explaining why the solution or the mathematical relationship used is valid	A ready-made explanation replaces the student's own mathematical justification
<i>Reflection</i>	Evaluating one's own problem-solving process, decisions made, and AI support received	Students remain unaware of which cognitive operations they performed themselves and which were performed by AI

These dimensions should not be regarded as independent, linear stages of problem solving. How students make sense of a problem may influence their strategy selection; an issue identified during verification may lead them to reconsider the problem or revisit their strategy; and attempting to justify a solution may prompt them to question a relationship they initially accepted. Similarly, reflection is not an activity undertaken only after a solution has been completed. It is a metacognitive activity that helps students monitor their decisions and the support they use throughout the process. Mathematical cognitive responsibility therefore refers not so much to completing particular steps in a fixed sequence as to maintaining student responsibility for mathematical decisions throughout problem solving. Evaluating when and for what purpose to seek AI support, what type of assistance is needed, and how the support received influences one's own mathematical thinking is also part of this ongoing responsibility for monitoring and reflection.

It is important to distinguish cognitive responsibility from performing all operations independently. Mathematical cognitive responsibility does not mean that students cannot receive solutions, hints, or feedback from AI. For example,

students may develop their own solution, ask AI for an alternative solution path, compare the two methods, and decide which is more appropriate and why. In such cases, responsibility for mathematical evaluation may remain with students even though AI has contributed directly to solution generation. Similarly, an incorrect AI-generated solution can become a valuable problem-solving activity when students are required to identify the error and explain, using mathematical justification, why the solution is incorrect. Thus, cognitive responsibility is related less to who produces a mathematical solution than to who makes the mathematical decisions required to evaluate that solution.

This criterion also suggests that mathematical cognitive responsibility may vary depending on the task. For example, if the goal of an activity is to develop students' algebraic manipulation skills, having AI perform the algebraic operations may eliminate part of the intended learning activity. Conversely, if the goal is to compare different modeling approaches, having AI perform the same operations may allow students to focus their attention on the assumptions underlying the models and the interpretation of their results. Similarly, obtaining a solution method directly from AI may be inappropriate when students are expected to develop their own strategies. However, when the task involves evaluating the strengths and weaknesses of different strategies, AI-generated alternative solutions may serve as useful learning materials. The distribution of mathematical cognitive responsibility should therefore be guided primarily by the learning goal, with the technological tool as a secondary consideration.

This approach also requires distinguishing between students' apparent activity during AI interaction and their engagement in mathematical thinking. Students may write numerous prompts, revise AI responses, and engage in extended dialogue with the system. However, if they rely on AI for all mathematical decisions, their mathematical cognitive responsibility may remain low despite a high level of technological interaction. Conversely, students who use AI only once but thoroughly question the resulting solution, verify it independently, and provide mathematical justification may demonstrate greater mathematical cognitive responsibility. The goal, therefore, is not to increase or decrease the amount of interaction with AI, but to ensure that students retain their role as decision-makers concerning mathematical meaning and validity.

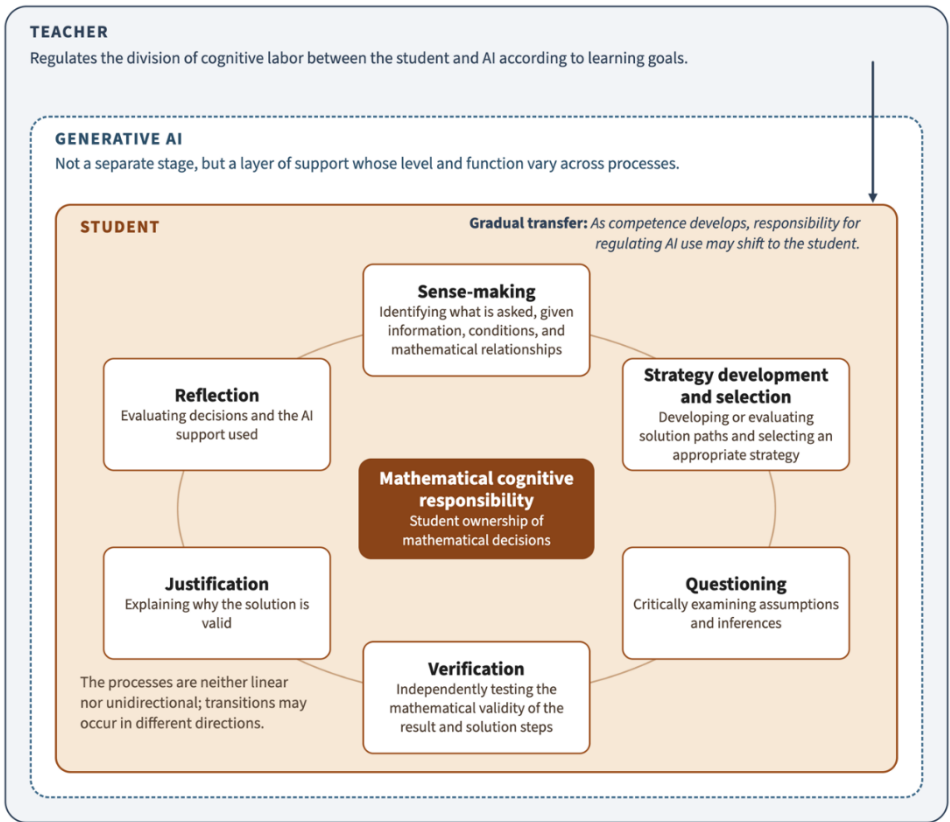
In the context of generative AI, students must also determine the criteria by which mathematical claims originating from another cognitive source should be accepted. Consequently, questioning, verification, and justification become particularly salient components of mathematical cognitive responsibility. The purpose of mathematical cognitive responsibility is not to distance students from technology, but to prevent them from becoming passive recipients of mathematical decisions while solving problems with technological support. Students may obtain certain calculations, representations, or even initial solution

suggestions from AI. However, to the extent required by the learning goal, they should be able to make sense of the problem, evaluate the appropriateness of the strategies used, verify the correctness of the solution, and explain why the mathematical result they accept is valid. From this perspective, the central pedagogical issue is not to establish a fixed division of tasks between students and AI, but to design learning activities that determine in which processes, and to what extent, cognitive responsibility should remain with students in accordance with the learning goal.

**COGNITIVE RESPONSIBILITY-BASED AI-SUPPORTED
PROBLEM-SOLVING MODEL**

To make the concept of mathematical cognitive responsibility operational for instructional purposes, it is necessary to explain how the teacher can structure the division of cognitive labor between students and generative AI throughout the problem-solving process. To this end, this chapter proposes a **Cognitive Responsibility-Based AI-Supported Problem-Solving Model**. Rather than constituting a new theory of mathematical problem solving, the model is conceived as a conceptual and pedagogical model that reorganizes established problem-solving processes in light of generative AI's capacity to produce solutions. Its central assumption is that the pedagogical value of AI use is related not to how extensively the system is used, but to who makes the mathematical decisions that are critical to learning. Accordingly, generative AI may serve various functions at different points in problem solving; however, responsibility for mathematical sense-making, selection, evaluation, and justification required by the learning goal should not be transferred entirely to AI. The model comprises six interrelated processes corresponding to the dimensions of cognitive responsibility presented in Table 2, as illustrated in Figure 1.

Figure 1. Processes of the Cognitive Responsibility-Based AI-Supported Problem-Solving Model



Like the dimensions presented in Table 2, these processes do not follow a linear or unidirectional sequence. Figure 1 provides a schematic representation of the processes and does not depict every possible transition between them. Within the model, interaction with AI does not constitute a separate stage of problem solving. Instead, generative AI is positioned as a **support layer** that can provide varying levels of assistance across any of the six processes. For example, AI may ask clarifying questions during sense-making, suggest alternative solution paths during strategy development, generate a solution for critical examination during questioning, provide an alternative representation or method during verification, help students identify gaps in their explanations during justification, and pose questions that encourage them to evaluate their own processes during reflection. Thus, the model conceptualizes AI use not as a fixed operation to be performed at a particular point, but as a flexible form of support whose level and function are regulated according to the learning goal.

Although the model does not treat strategy implementation as a separate dimension of cognitive responsibility, implementation remains part of the problem-solving process. The learning goal determines which operations students should perform themselves and which may be delegated to AI. If the learning goal is to apply a particular mathematical method, students should perform that application themselves. Conversely, if the learning goal focuses on a different aspect of mathematical thinking, certain operations may be delegated to AI as a form of support. Rather than defining implementation as a fixed task belonging exclusively to either students or AI, the model therefore emphasizes maintaining students' responsibility for mathematical sense-making, strategic decision-making, and evaluating validity during implementation, in accordance with the learning goal.

Within this structure, the teacher's role extends beyond simply permitting or restricting AI use. As the regulator of cognitive responsibility, the teacher makes instructional decisions about which mathematical processes students should perform themselves, where AI support may be used, and what additional cognitive operations students should undertake when working with AI output. In this respect, in addition to its six interrelated processes, the model involves three components with distinct roles: *(i) the student at the center of the problem-solving process, (ii) generative AI as a source of support across these processes, and (iii) the teacher, who structures the division of cognitive labor between students and AI according to learning goals.*

However, the intention is not for responsibility for regulating AI use to remain permanently with the teacher. In interactive learning environments, effective help seeking is associated with metacognitive and cognitive skills (Aleven et al., 2003). In human–AI systems, control can be shared in different ways among teachers, students, and technology (Molenaar, 2022). This means that decisions about when to seek AI support must also be addressed from an instructional perspective. Accordingly, the teacher may initially determine the purpose, timing, and limits of AI support. Students can then be encouraged to assess when they need particular types of support and justify their choices. As students gain experience and develop greater competence, they may be expected to regulate their AI use more independently. This progression is not a fixed developmental sequence that all students must follow in the same order, but an instructional arrangement that can be adapted to their learning needs. The teacher may provide more intensive guidance again when necessary. Thus, the teacher's regulatory role evolves to support students in progressively assuming more informed responsibility for deciding when, why, and how to use AI support. The roles of students, AI, and teachers within this framework are summarized in Table 3.

Table 3. Roles of students, AI, and teachers in the Cognitive Responsibility-Based AI-Supported Problem-Solving Model

<i>Process</i>	<i>Student's cognitive responsibility</i>	<i>Role of AI</i>	<i>Teacher's regulatory role</i>
<i>Sense-Making</i>	Structures what the problem asks, its conditions, and the mathematical relationships involved.	May explain concepts, rephrase the problem, or ask guiding questions.	Determines tasks that require students to analyze the problem themselves and establishes the limits of AI support.
<i>Strategy Development and Selection</i>	Develops or evaluates possible solution paths and selects an appropriate strategy.	May provide hints or suggest alternative strategies.	Regulates the boundary between providing a strategy directly and supporting students' strategic thinking.
<i>Questioning</i>	Critically examines the assumptions, steps, and inferences in a solution.	May provide solutions for examination, alternative explanations, or different approaches that can be deliberately compared.	Encourages students to question the correctness and underlying assumptions of solutions.
<i>Verification</i>	Independently tests the validity of the result and solution process using mathematical methods.	May suggest alternative solutions, representations, or methods for checking.	Designs tasks that prevent AI from becoming the sole source of verification for its own outputs.
<i>Justification</i>	Explains mathematically why a solution is valid.	May question students' explanations or help identify gaps in their justifications.	Makes students' mathematical justifications visible and evaluates the reasoning rather than focusing solely on the result.
<i>Reflection</i>	Evaluates their own problem-solving decisions, reasons for seeking AI support, and the effects of that support on problem solving.	May provide reflective questions or feedback.	Encourages students to evaluate their decisions and the support received; may gradually transfer responsibility for regulating AI use to students as their competence develops.

In the model, sense-making is not merely a formal step to be completed before solving begins; it involves students constructing mathematical meaning from the problem situation. Students distinguish between given information, unknowns, and conditions, and identify relationships among quantities or concepts. During this process, generative AI may simplify the wording of the problem or pose guiding questions. However, if the learning goal is to develop students' ability to structure a problem situation mathematically, having AI perform the entire sense-making process may reduce the cognitive activity that students are expected to undertake themselves.

Strategy development and selection involve a broader decision-making process than simply determining which operation to perform. Generative AI can suggest numerous alternative solution paths, but what is critical to the model is students' ability to evaluate these alternatives using mathematical criteria and explain why they prefer a particular approach. In the CREATIVEMATH benchmark, Ye et al. (2025) showed that some large language models could generate correct solution paths that differed from the reference solutions provided. However, they also demonstrated that correctness and originality are distinct characteristics that vary considerably across models. The researchers further reported that providing more reference solutions could improve correctness while reducing the tendency to generate novel solutions. These findings support the view that the novelty or distinctiveness of an AI-generated strategy is not, in itself, an indicator of mathematical quality. Students' responsibility is not merely to select the most distinctive option, but to evaluate a solution's validity, appropriateness for the problem, efficiency, and explanatory value. Thus, AI-generated strategies do not necessarily entail a loss of cognitive responsibility; generating alternatives can serve a supportive function as long as students retain responsibility for mathematically justifying their decisions.

In the model, questioning and verification serve related but distinct functions. Questioning involves critically examining the assumptions and inferences used in a solution, whereas verification refers to testing the mathematical validity of a solution using independent mathematical grounds. This distinction is particularly important in AI-mediated problem solving. Discussing the correctness of a solution by posing critical questions to AI does not mean that students have independently verified its mathematical validity. In a critical case study involving 12 university students, Zhuang (2026) observed that ChatGPT-3.5's agreement with students' criticisms could result in some correct solutions being changed into incorrect ones. This finding is important because it illustrates that engaging in a questioning dialogue with AI does not, in itself, constitute mathematical verification. Students should therefore avoid basing their judgments about the correctness of a solution solely on AI's confirmation and instead examine the relationships, operations, or inferences involved using independent mathematical grounds.

Justification requires students to go beyond identifying a correct answer or solution path and explain mathematically why the solution is valid. AI can generate a ready-made justification. However, the central concern of the model is not whether students can reproduce that justification, but whether they can develop their own mathematical explanations based on an understanding of the validity of the relationships involved. Accordingly, one potentially more appropriate role for AI in justification is to question students' explanations rather than produce complete justifications on their behalf. For this purpose, students

may be asked questions such as "Why is this step valid?", "What mathematical property supports this result?", or "Does this inference hold in all cases?"

Finally, reflection enables students to revisit not only the mathematical results they obtain but also the decisions they make throughout the problem-solving process and the AI support they use. Considering questions such as "Which mathematical decisions did I make myself?", "At what point and why did I turn to AI?", "What did I try before seeking AI support?", "Which AI suggestions did I accept or reject?", and "What would I do differently if I had to solve this problem again without AI?" may help prevent AI use from becoming an invisible support mechanism. This dimension is also intended to foster students' metacognitive awareness of when and how much AI support they need.

An important characteristic of the model is that it does not assign AI a fixed role. Depending on the task, the same system may function as a solution generator, a source of hints or alternative solutions, a questioning partner, or a feedback tool. None of these roles is inherently pedagogically beneficial or detrimental. For example, having AI generate a complete solution may replace students' own problem-solving activity when the goal is for them to solve the problem independently. However, when students are tasked with identifying errors in a generated solution, the same AI function may provide suitable learning material for questioning and verification. Sarıkaya and İnci Kuzu (2026) also identify the analysis of AI-generated solution errors as a pedagogical opportunity that can be used to support students' logical reasoning and problem-solving skills. In the proposed model, the learning goal is therefore the primary criterion for determining the role of AI. Three questions are recommended for consideration in each task:

1. What mathematical thinking is this activity intended to develop in students?
2. Is AI performing this thinking on students' behalf, or supporting them in doing it themselves?
3. After receiving AI output, what mathematical decision is the student expected to make, or what evidence is the student expected to produce?

These three questions provide a practical decision-making framework for teachers to regulate cognitive responsibility. For example, if the learning goal is to develop students' ability to apply an equation-solving algorithm, delegating all operations to AI may be inconsistent with that goal. If the goal is to select an appropriate mathematical model for a real-world situation, having AI perform certain calculations may allow students to focus more on the assumptions underlying the model and the interpretation of its results. In another activity,

when the goal is to identify mathematical errors, an incorrect or incomplete AI-generated solution may itself become learning material.

This approach also allows AI support to be adjusted according to students' levels. The study by Noviyana et al. (2025) provides an example of such instructional structuring. However, because the study involved a particular age group and an integrated intervention, its findings do not directly support the conclusion that the proposed model would function in the same way across different age groups and mathematical topics. The proposed model therefore requires teachers to regulate AI support according to students' prior knowledge, task difficulty, and the learning goal.

The proposed model does not seek to replace the problem-solving processes described in established frameworks such as those of Pólya and Schoenfeld. Sense-making, strategy development, monitoring, evaluation, and justification are already established components of the mathematics education literature. Rather, the model's contribution lies in making the preservation of students' mathematical cognitive responsibility an explicit dimension of problem-solving instructional design, given that some of these processes can now be performed on students' behalf by generative AI.

In this respect, the Cognitive Responsibility-Based AI-Supported Problem-Solving Model conceptualizes generative AI not as a single tool added to the problem-solving process or as a separate stage, but as a flexible form of support whose function is regulated according to learning goals within the division of cognitive labor between students and AI. Illustrating how this division of cognitive labor can be structured through concrete mathematical tasks, forms of interaction with AI, and assessment processes may help clarify the pedagogical potential of the model. Accordingly, the following section examines the model's implications for instruction and assessment.

IMPLICATIONS FOR INSTRUCTION AND ASSESSMENT

Implementing the Cognitive Responsibility-Based AI-Supported Problem-Solving Model in the classroom requires more than merely allowing students to use generative AI; it requires the joint design of mathematical tasks and interactions with AI in alignment with the learning goal. The central question, therefore, should not be "Can AI be used in this activity?" but "What mathematical thinking are students expected to engage in during this activity, and how should AI be integrated with that thinking?" The same AI function may support students' thinking in one task while replacing it in another. Consequently, instructional design should begin with the learning goal rather than the technology to be used.

In designing AI-supported mathematical tasks, students' roles in the solution process should be determined in advance. A task should require students not only

to produce an answer but also to make at least one mathematical decision and make the grounds for that decision explicit. Such a decision may concern how the problem is represented, which strategy is selected, whether an AI output is accepted, which verification method is used, or which of two solutions is more appropriate. In this way, as AI becomes capable of generating a wider range of outputs, students' roles can be defined not simply as "completing what AI cannot do," but as making decisions concerning mathematical meaning and validity.

The timing of students' interaction with AI may also be part of task design. Yoon et al. (2026) compared two approaches to mathematical proof construction among university students. In the first approach, students developed an independent proof draft before interacting with AI, whereas in the second, they first examined an AI-generated proof. In the former, AI was used primarily to review and revise existing drafts, whereas in the latter, it served as a resource guiding students' initial mathematical thinking. However, students in both groups drew on AI-generated proofs. Thus, working independently at the outset does not necessarily mean that mathematical cognitive responsibility is maintained. Students evaluated the initial independent-work approach positively in terms of its contribution to their thinking. Nevertheless, the findings do not establish that working independently before receiving AI support improves long-term conceptual learning. Therefore, the timing of AI support should be considered alongside how students make their initial mathematical thinking visible and subsequently evaluate AI output.

From this perspective, interaction with AI should not be reduced to the ability to "write good prompts." The prompts students give AI are important; however, the more fundamental pedagogical question concerns which mathematical thinking a prompt requires of students and which thinking it may displace. A request such as "Solve this problem and explain the steps" may delegate a substantial portion of problem solving to AI. By contrast, interactions such as "Ask me one guiding question to help me evaluate my solution," "Which step in this solution should I check particularly carefully?" or "Show me a strategy different from my own; I will decide which is more appropriate" may keep students' mathematical decision-making at the center of the process. Thus, the instructional goal is not to teach fixed prompting templates but to help students learn how to structure AI's role in ways that maintain their cognitive responsibility.

The teacher's role is critical at this point. The teacher regulates the boundary between the mathematical operations students need to perform themselves and the supporting operations that may be delegated to AI, according to learning goals. Teachers also design tasks that require students to question AI-generated solutions, verify them independently when necessary, and justify their

mathematical decisions. Table 4 summarizes five key decisions for considering these relationships together in instructional design.

Table 4. Instructional design decisions in the Cognitive Responsibility-Based AI-Supported Problem-Solving Model

Design decision	Key question	Expected student cognitive responsibility	Potential role of AI
Identifying the Learning Goal	What mathematical thinking are students expected to develop by the end of this activity?	Engaging in the targeted cognitive activity.	Facilitating operations that fall outside the learning goal or support its achievement.
Timing and Regulation of AI Support	Will students interact with AI before or after developing their own mathematical thinking, and who will make this decision?	Making the targeted mathematical thinking visible and justifying decisions to seek AI support in relation to the learning goal and their own needs.	Providing hints, feedback, alternative approaches, or materials for examination.
Working with AI Output	What will students do with the output provided by AI?	Comparing, questioning, verifying, or justifying.	Generating solutions, explanations, or alternatives for examination.
Generating Process Evidence	How can students' mathematical thinking be made visible?	Articulating the mathematical grounds for their choices and judgments.	Providing reflective questions or feedback.
Determining AI's Role in Assessment and Feedback	For what purpose will AI analyze students' solutions: preliminary diagnosis and feedback, or final assessment decisions?	Making their solution processes and critical mathematical decisions visible, and evaluating feedback using mathematical evidence.	Providing stage-level preliminary diagnosis, identifying potential errors, and offering guiding feedback without becoming the sole authority for mathematical correctness.

Different classroom activities can be designed based on these decisions. Two examples are presented below to illustrate how the model can be implemented.

In the "**Questioning and Verifying an AI-Generated Solution**" activity, students may be presented with a mathematical problem accompanied by an AI-generated solution. Rather than immediately classifying the solution as correct or incorrect, students are first asked to identify the strategy used, examine critical inferences mathematically, and test the result without relying on AI confirmation. If the solution is incorrect, students are expected to identify and correct the step at which the error first occurs. If the solution is correct, they are expected to explain why it is mathematically valid. The teacher reviews the AI-generated output in advance to ensure that the mathematical relationships to be examined are aligned with the learning goal. In this way, the AI-generated solution becomes

a learning resource that requires students not merely to arrive at an answer but to justify their mathematical decisions to accept or reject it.

In the **"Problem Solving with Guiding AI Support"** activity, students initially make sense of the problem independently and develop their first solution attempt. Rather than providing a complete solution, AI is intended to offer limited support that helps students make progress. Students may request a hint, a clarifying question, or feedback on a potential issue in their own solution. However, after receiving this support, they remain responsible for deciding what changes to make and for providing the mathematical justification for those changes. The teacher may initially determine the scope of assistance and, as students' competence develops, gradually transfer responsibility for deciding when and what type of support is needed to the students. This activity resembles the AI intervention used by Bastani et al. (2025), in which students received teacher-designed hints. However, that study did not demonstrate a statistically significant advantage of structured AI support over the control condition in assessment performance without AI access. The extent to which guiding AI support contributes to independent problem solving must therefore be evaluated separately.

Illustrating the Model Through an Example Mathematical Problem

To illustrate how the Cognitive Responsibility-Based AI-Supported Problem-Solving Model can be applied to a mathematical task, consider an example involving a comparison of costs between two printing centers. The primary learning goals are for students to model linear cost relationships, evaluate different solution strategies, and justify their mathematical decisions.

Poster Printing for a School Exhibition

A school plans to print posters for an upcoming exhibition. The school administration has received the following price quotations from two printing centers.

Printing center	Fixed setup fee	Cost per poster
Printing Center A	80 TL	6 TL
Printing Center B	20 TL	9 TL

Problem: Which printing center is more economical depending on the number of posters the school needs to print? Determine the number of posters for which both printing centers charge the same total amount and justify your decision mathematically.

Sense-Making and Strategic Decision-Making: Students first distinguish between fixed and variable costs in the problem. By letting x denote the number

of posters to be printed (a positive integer), students can express the total costs as $A(x) = 80 + 6x$ and $B(x) = 20 + 9x$. They may then formulate the equation $80 + 6x = 20 + 9x$ to determine the number of posters for which the costs are equal. Students may choose to formulate an equation to identify the break-even point, construct a table to compare costs for different numbers of posters, or draw a graph to examine how the costs change. Even if they select equation formulation as an appropriate strategy for determining the break-even point, they need to recognize that an additional comparison of costs is necessary to explain which printing center is more economical under different conditions. During this process, AI may provide guiding questions that encourage students to consider how to represent the costs rather than supplying ready-made equations.

Questioning and Verification: After developing their initial solution, students may ask AI for an alternative evaluation. To illustrate how the model operates, suppose AI generates an explanation stating that both printing centers charge the same amount for 20 posters, but Printing Center A is more economical when fewer than 20 posters are printed. Although this explanation correctly identifies the break-even point, it contains an error in the cost comparison. Students first distinguish between the two claims in the explanation: they recognize that the break-even point is consistent with their own model, but that the comparison regarding which printing center is more economical requires further examination. Rather than asking AI to confirm this comparison, they independently verify it by calculating the costs for different numbers of posters. For 10 posters, the total cost is 140 TL at Printing Center A and 110 TL at Printing Center B. For 30 posters, the corresponding costs are 260 TL and 290 TL, respectively. These calculations reveal that the cost comparison in the illustrative AI-generated explanation is incorrect. Students then develop a general mathematical justification that establishes which printing center is more economical for all relevant numbers of posters.

Justification: Students are expected to go beyond calculating costs for particular numbers of posters and explain, using a general mathematical relationship, which printing center is more economical under different conditions. The difference between the two costs can be expressed as $A(x) - B(x) = 60 - 3x$. Because this difference is positive when $x < 20$ and negative when $x > 20$, Printing Center B is more economical for fewer than 20 posters, whereas Printing Center A is more economical for more than 20 posters. When $x = 20$, the costs are equal. Thus, students base their decisions on a mathematical justification that holds for all relevant numbers of posters rather than on a conjecture derived from a few examples.

Reflection: At the end of the problem-solving process, students compare their initial mathematical model with the explanation provided by AI. They evaluate which AI suggestions they accepted, which claims they rejected and why, and

how independent verification influenced their decisions. The teacher may examine not only students' final answers but also the mathematical models they constructed, the verification methods they selected, and the mathematical justifications they provided.

This example illustrates how generative AI can assume different roles during mathematical problem solving. AI may initially provide guiding questions and subsequently present a solution or explanation for students to question. However, the pedagogical value of these roles is related less to the outputs generated by AI than to the mathematical activities students undertake in response to those outputs. In this example, cognitive responsibility is therefore demonstrated not by students producing every component of the solution themselves, but by their continued ownership of mathematical decisions when making sense of relationships, testing claims, and justifying results.

Implications of the Model for Assessment

This instructional approach also has implications for assessment. In an environment where generative AI can produce correct answers, solution steps, and explanations on students' behalf, examining only the final product may provide increasingly limited information about the mathematical thinking students have engaged in. However, this does not diminish the importance of correct answers or mathematical correctness. Rather, assessment should consider mathematical correctness alongside evidence of the decisions made throughout the problem-solving process.

Generative AI may also play a supportive role in assessment, although the boundaries of this role need to be carefully designed. Jin et al. (2024) developed a system intended to diagnose students' problem-solving processes in elementary-level mathematical word problems through four stages: Parse, Extract, Retrieve, and Combine (PERC). This preliminary conference paper identified the potential of stage-level AI diagnosis to provide scalable feedback. However, a mathematics education expert emphasized the need to align the rubrics, particularly those for the Extract, Retrieve, and Combine stages, more closely with the target grade level and curriculum. Zhang et al. (2026) noted that identifying the first incorrect step in a solution can be more difficult than solving the problem directly. Reliably evaluating each step of a solution requires tracking the problem-solving process step by step and accurately locating where the first error occurs. Accordingly, AI may be used for preliminary diagnosis and feedback in formative assessment. However, basing final judgments about students' mathematical thinking solely on AI-generated assessments would be inconsistent with the cognitive responsibility approach proposed in this chapter.

In this context, four types of evidence may be particularly useful, whether assessment is conducted by teachers or supported by AI-generated feedback. The first is **sense-making and strategy evidence**, which reveals how students make

sense of a problem and why they select a particular solution path. The second is **verification evidence**, which demonstrates the independent mathematical method students use to test a result or an important inference, rather than merely stating that they have checked it. The third is **justification evidence**, through which students explain in their own mathematical terms why a solution is valid. The fourth is **interaction, questioning, and reflection evidence**, which shows what support students received from AI, which aspects of that support they accepted or rejected, and how they might have proceeded without AI.

This approach does not require all students to submit extensive AI conversation records or detailed process journals. A more manageable alternative is to request brief process evidence documenting a few critical decisions concerning strategy selection, verification, and the evaluation of AI suggestions alongside the final product. However, because students may also prepare these explanations with AI assistance, such evidence alone may not reliably demonstrate independent problem-solving competence. Therefore, when aligned with the learning goal, assessment may also include a brief **transfer task** completed without AI access. Students may be asked to solve a new problem involving the same mathematical relationship as the original task but with a different context or numerical values, and to justify their critical mathematical decisions. For example, after completing the poster-printing problem, students could compare two services with different fixed and unit costs. This would allow problem-solving performance demonstrated with AI support to be examined alongside performance demonstrated independently. Nevertheless, a single transfer task does not provide conclusive evidence of long-term learning. Stronger conclusions about the retention of learning require assessments conducted at different points in time.

Assessment criteria may likewise be kept limited and clearly defined. While mathematical correctness remains a fundamental criterion, additional criteria may include the appropriateness of the strategy, the quality of verification, mathematical justification, and critical evaluation of AI output. Not all dimensions need to be assessed with equal weight in every activity. The criteria should be selected in alignment with the learning goal. For example, verification and justification may be emphasized in an error-analysis activity, whereas strategic decision-making and the evaluation of alternatives may play a more central role in an open-ended problem.

This approach does not imply that AI use should be assessed as a separate criterion. The ability to write numerous effective prompts is not, in itself, an indicator of mathematical competence. What should be assessed is the mathematical thinking students engage in, whether through interaction with AI or independently of it. Therefore, the ability to "use AI effectively" should

become an assessment criterion only when it is directly related to the mathematical learning goal.

Integrating generative AI into mathematics instruction requires a more comprehensive approach to instructional design than simply adding an AI tool to existing mathematical problems. The learning goal, the role of AI, the timing of interaction, and the process evidence expected from students should be considered together. The teacher's primary responsibility is not to create a learning environment in which technology is used as extensively as possible, but to align the support AI can provide with the learning goal while maintaining students' mathematical cognitive responsibility.

CONCLUSION

The ability of generative AI to rapidly produce solutions, explanations, and alternative solution paths for mathematical problems necessitates reconsidering some fundamental assumptions about the teaching of mathematical problem solving. This transformation does not diminish the importance of mathematical problem solving; rather, it makes more apparent that problem solving involves more than simply arriving at a correct answer. Students' ready access to a solution does not necessarily mean that they have made sense of the problem, can select an appropriate strategy, evaluate the validity of solution steps, or justify the result obtained. Consequently, the distinction between accessing a solution and engaging in mathematical problem solving has become increasingly important in the age of generative AI.

Nevertheless, viewing generative AI solely as a technology that replaces students' thinking would also be incomplete. The educational effects of AI depend on the nature of the support provided and how that support is structured within the learning process. In a field experiment conducted in high school mathematics, standard AI support improved students' performance on practice exercises. However, on an assessment conducted without AI access, students who had received standard AI support performed worse than students in the control group. In contrast, students who used an AI version providing teacher-structured hints did not differ significantly from the control group on the same assessment (Bastani et al., 2025). Similarly, the intervention examined by Noviyana et al. (2025), in which AI was integrated with problem-based learning, provides an example of an instructional arrangement that may support students' active participation in problem solving. However, the independent effect of AI could not be distinguished from that of problem-based learning in this study. These findings indicate that the educational significance of AI lies not merely in whether it is used or whether it generates correct solutions, but in the mathematical functions it performs and the thinking processes students continue to engage in during problem solving.

To address this issue, this chapter has proposed the concept of **mathematical cognitive responsibility**. The concept does not require students to perform every mathematical operation independently of technology. Rather, it refers to students retaining ownership of mathematical decisions related to sense-making, strategic decision-making, questioning, verification, justification, and reflection throughout the problem-solving process. From this perspective, the generation of a solution by AI does not necessarily imply a loss of cognitive responsibility, just as extensive interaction with AI does not, in itself, demonstrate that cognitive responsibility has been maintained. The central criterion is who makes the decisions concerning mathematical meaning and validity.

Building on this conceptualization, the proposed **Cognitive Responsibility-Based AI-Supported Problem-Solving Model** treats sense-making, strategy development and selection, questioning, verification, justification, and reflection as interrelated processes. Within the model, generative AI is positioned not as a separate or mandatory stage, but as a support layer that can provide varying levels of assistance at different points in problem solving. The model's contribution lies in making the allocation of cognitive responsibility within established problem-solving processes an explicit dimension of instructional design. The teacher's role is to regulate the scope of this support according to learning goals and structure the division of cognitive labor between students and AI. Thus, the purpose of AI is not to perform as many operations as possible on students' behalf, but to support the mathematical thinking that students are expected to undertake themselves.

This approach also calls for a shift in instruction and assessment. Although mathematical correctness remains important, the assessment of students' problem-solving competence should not rely solely on the final answer. The reasons students select particular strategies, the criteria they use to accept or reject AI output, how they verify solutions, and how they justify their results also provide important evidence of learning. Accordingly, tasks in AI-supported mathematics instruction should be designed not only to elicit answers but also to make explicit the mathematical grounds on which students base their decisions.

Together, the concept of mathematical cognitive responsibility and the accompanying model offer a conceptual perspective on mathematical problem solving with generative AI. They should not be regarded as an independently validated theory or a fixed instructional method. Empirical research is needed to examine how the model functions across different student levels, mathematical topics, and forms of AI use. In particular, the timing and level of AI support, the gradual transfer to students of responsibility for regulating that support, and the relationships between these variables and students' performance in questioning, verification, and justification represent important areas for future research. Furthermore, examining the extent to which problem-solving performance demonstrated with AI support transfers to comparable mathematical tasks

completed without AI, and whether these gains are sustained over time, would contribute to evaluating the model in terms of learning.

In conclusion, the central issue in mathematical problem solving in the age of generative AI is not whether technology can produce an answer to a problem. Rather, it is which mathematical thinking processes students will continue to assume responsibility for in an environment where solution generation can increasingly be shared with technology. The focus of mathematics instruction is broadening beyond answer production to include making sense of solutions, questioning them, verifying them, and justifying them. The value of generative AI in mathematics education should therefore be evaluated in terms of the extent to which it supports students' cognitive responsibility and makes that responsibility visible.

REFERENCES

- Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help seeking and help design in interactive learning environments. *Review of Educational Research*, 73(3), 277–320. <https://doi.org/10.3102/00346543073003277>
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Fang, M., Wan, X., Lu, F., Xing, F., & Zou, K. (2025). MathOdyssey: Benchmarking mathematical problem-solving skills in large language models using Odyssey Math data. *Scientific Data*, 12, Article 1392. <https://doi.org/10.1038/s41597-025-05283-3>
- Hao, Y., Wan, X., & Zhai, C. (2025). An investigation of robustness of LLMs in mathematical reasoning: Benchmarking with mathematically-equivalent transformation of advanced mathematical problems [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2508.08833>
- Jin, H., Kim, Y., Park, Y. S., Tilekbay, B., Son, J., & Kim, J. (2024). Using large language models to diagnose math problem-solving skills at scale. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (pp. 471–475). Association for Computing Machinery. <https://doi.org/10.1145/3657604.3664697>
- Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *European Journal of Education*, 57(4), 632–645. <https://doi.org/10.1111/ejed.12527>
- Noviyana, H., Rahmawati, F., Kirana, A. R., & Tanod, M. J. (2025). Enhancing elementary students' mathematical problem-solving skills through AI-assisted problem-based learning. *Journal of Integrated Elementary Education*, 5(2), 254–268. <https://doi.org/10.21580/jieed.v5i2.27576>
- Nurhasanah, F., Usodo, B., Kuswardi, Y., Chrisnawati, H. E., & Sutopo. (2026). AI-powered mathematical assistants: Redefining problem-solving paradigms. In *2026 2nd International Conference on Sustainable Computing and Integrated Communication in Changing Landscape of AI (ICSCAI)* (pp. 1366–1373). IEEE. <https://doi.org/10.1109/ICSCAI68849.2026.11649329>
- Petković, I., & Herceg, Đ. (2021). Computer tools for solving mathematical problems: A review. *Facta Universitatis, Series: Mathematics and Informatics*, 36(1), 205–236. <https://doi.org/10.22190/FUMI201203017P>
- Pólya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.
- Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>

- Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher*, 20(3), 2–9. <https://doi.org/10.3102/0013189X020003002>
- Santos-Trigo, M. (2024). Problem solving in mathematics education: Tracing its foundations and current research-practice trends. *ZDM – Mathematics Education*, 56(2), 211–222. <https://doi.org/10.1007/s11858-024-01578-8>
- Sarıkaya, K., & İnci Kuzu, Ç. (2026). The evolution of artificial intelligence in mathematical problem-solving processes: A mathematics education perspective. *Sage Open*, 16(1), 1–13. <https://doi.org/10.1177/21582440261426456>
- Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp. 67–98). Open Court.
- Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press. <https://doi.org/10.1016/C2013-0-05012-8>
- Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 680, pp. 1–24). Association for Computing Machinery. <https://doi.org/10.1145/3613904.3642902>
- Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, 27(4), 458–477. <https://doi.org/10.5951/jresmetheduc.27.4.0458>
- Ye, J., Gu, J., Zhao, X., Yin, W., & Wang, G. (2025). Assessing the creativity of LLMs in proposing novel solutions to mathematical problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24), 25687–25696. <https://doi.org/10.1609/aaai.v39i24.34760>
- Yoon, H., Lee, Y., Lee, K., & Kwon, O. N. (2026). Before or after? Beyond timing of students' interaction with generative artificial intelligence in proving mathematical statements. *ZDM – Mathematics Education*, 58(5), 945–962. <https://doi.org/10.1007/s11858-026-01800-9>
- Zhang, L., Fu, Y., & Jin, X. (2026). Is mathematical problem-solving expertise in large language models associated with assessment performance? In C. Stephanidis, G. Margetis, S. Ntoa, M. Antona, & G. Salvendy (Eds.), *HCI International 2026 Posters: Proceedings, Part VI* (Communications in Computer and Information Science, Vol. 3052, pp. 202–212). Springer. https://doi.org/10.1007/978-3-032-30836-8_23
- Zhuang, Y. (2026). Empowering students to critically validate AI-generated mathematical solutions through the rational questioning approach.

Educational Studies in Mathematics. Advance online publication.
<https://doi.org/10.1007/s10649-026-10528-y>