

A COMPARISON OF SCIENCE TEACHERS AND CHATGPT IN IDENTIFYING STUDENTS' INCORRECT IDEAS ABOUT HEAT AND TEMPERATURE AND PROVIDING FEEDBACK

TUFAN İNALTEKİN^{1*}, TOLGA SAKA²

¹PHD, ASSOCIATE PROFESSOR, DEPARTMENT OF MATHEMATICS & SCIENCE,
EDUCATION, DEDE KORKUT FACULTY OF EDUCATION, KAFKAS UNIVERSITY, TÜRKİYE
inaltekintufan@gmail.com,

<https://orcid.org/0000-0002-3843-7393>

*CORRESPONDING AUTHOR

²PHD, ASSISTANT PROFESSOR, DEPARTMENT OF MATHEMATICS & SCIENCE,
EDUCATION, DEDE KORKUT FACULTY OF EDUCATION, KAFKAS UNIVERSITY, TÜRKİYE.
tsaka61@gmail.com,

<https://orcid.org/0000-0002-0042-0836>

Abstract

This study aimed to compare the competencies of science teachers and ChatGPT in identifying students' incorrect ideas and generating feedback ideas regarding the topics of "Heat and Temperature". In this context, it was also revealed, in terms of professional seniority and educational level, how science teachers compare to ChatGPT in generating these ideas. This study employed a comparative case study design, which is one of the qualitative research approaches. The sample comprised 28 science teachers and the artificial intelligence program ChatGPT. The data were collected using the "Student Error Identification and Feedback Form". The ideas generated regarding student errors were descriptively analyzed based on their explanation levels (adequate, partially adequate, and inadequate), while the feedback ideas were analyzed descriptively using predetermined codes. The findings revealed that approximately half of the teachers were able to identify student errors at an adequate level, whereas ChatGPT tended to provide superficial or incomplete analyses when detecting mistakes in student responses. In terms of professional seniority, teachers in the activeness stage (7–18 years of experience) with a master's degree were the most productive group both in identifying student errors and in providing feedback, and they exhibited a level of feedback productivity comparable to that of ChatGPT. It was observed that novice teachers and those with only a bachelor's degree demonstrated substantial shortcomings in the productivity of both identifying student errors and generating feedback ideas. This group was found to lag significantly behind ChatGPT in terms of productivity. These findings suggest that the combination of educational attainment and professional experience plays a critical role in accurately identifying students' thinking and producing high-quality feedback, and that ChatGPT, in particular, could serve as a complementary tool that more strongly supports learning for novice teachers in generating feedback ideas.

Keywords: Artificial Intelligence, ChatGPT, Science Teacher, Heat and Temperature Questions, Feedback

INTRODUCTION

In science courses, it is of great importance for students to learn fundamental physics topics accurately, as understanding physical phenomena enables a better comprehension of nature and the universe. Among these topics, "Heat and Temperature" holds a significant place in the science curriculum, beginning in the early years of middle school and continuing throughout students' entire educational journey. Research indicates that "Heat and Temperature" is among the topics with which students experience the greatest learning difficulties (Salame et al., 2025; Soeharto et al., 2019). Therefore, the depth of professional knowledge that science teachers possess when teaching heat and temperature becomes particularly significant (İnaltekin & Akcay, 2021). Teachers' knowledge of understanding and assessing students is defined as one of the fundamental dimensions within teacher professional knowledge models (Carlson et al., 2019; Gess-Newsome et al., 2019; Magnusson et al., 1999; Park & Oliver, 2008). Moreover, these dimensions are also regarded as core components that influence one another within the broader framework of teacher professional knowledge (Barendsen & Henze, 2019). In this context,

feedback, situated within the assessment knowledge dimension of science teachers, serves as a powerful pedagogical tool that guides students' ways of thinking, helps them recognize their mistakes, and promotes meaningful learning (Black & Wiliam, 2018; Brown, 2018). Recognized as one of the most effective instructional strategies for supporting student learning (Hattie & Timperley, 2007), feedback plays a critical role in deepening students' understanding and ensuring accurate knowledge acquisition, particularly in conceptually dense subjects such as science (Deverel-Rico & Furtak, 2025). Topics such as "Heat and Temperature" which students often find difficult to learn, are among the important subjects that require accurate and timely feedback (Andini et al., 2024). In such physics topics, high-quality feedback directly influences students' understanding (Wancham & Tangdhanakanond, 2022). Therefore, it is considered that teachers' ability to understand their students' incorrect ideas regarding "heat and temperature" and to provide high-quality feedback in response to these incorrect ideas has a direct impact on the quality of instruction in this topic. Supporting this view, Miller et al. (2013) describe feedback as a fundamental element of formative assessment, as it contributes to correcting learning errors and reinforcing achievement. In science courses, effective feedback should be clear, constructive, and tailored, fostering conceptual development and self-regulation in a way that supports student learning (Güney & Feyzioğlu, 2023; Irons & Elkington, 2021; Limbere et al., 2022). Wancham and Tangdhanakanond (2022) emphasize that effective feedback not only enhances students' achievement but also improves their problem-solving skills. Thus, in science education, high-quality feedback enables students to achieve deeper learning. However, it is known that some science teachers have significant deficiencies in accurately identifying student understanding and providing effective feedback (Sadler & Sonnert, 2016). Various factors, including limited time, large class sizes, knowledge, pedagogical competence, educational background, and professional experience, may contribute to this situation (Hartelt et al., 2022). In recent years, the increasing integration of artificial intelligence (AI) technologies into education has led teachers to question their professional knowledge and skills to an even greater extent (Almasri, 2024; Al Darayseh, 2023; Park et al., 2023). The potential and impact of AI technologies in the context of high-quality science education have reached remarkably notable levels (Huang, 2024; Jia et al., 2024). It has been emphasized that by leveraging AI, educators can move away from traditional, one-size-fits-all approaches to education and instead provide students with personalized and interactive learning experiences (Almasri, 2024). One of the most common and effective applications in this field, ChatGPT offers information that can influence science education for students at all levels (Cooper, 2023). Today, ChatGPT presents the image of a mechanized human in various aspects of science education, including delivering scientific information, problem-solving, planning, and generating original ideas (Zhai, 2023). Discussions regarding the potential societal impacts of AI have recently attracted considerable attention among educators, as debates over the relationship between AI outputs and substantial job losses in education continue to grow (Pavlik, 2023). In this context, due to its highly efficient ability to "understand" natural language, ChatGPT has been described as the large-scale data model most closely resembling human intelligence (Binz & Schulz, 2023). Therefore, it is increasingly asserted that, within the next decade, AI will replace teachers just as it is expected to replace many other professions (Bill Gates, 2025). Research aiming to substantiate this claim has recently intensified in the field of education. In this context, the number of studies focusing on ChatGPT within the domain of science education has been steadily increasing (Ng et al., 2024; Park & Martin, 2024; Valeri et al., 2025). These studies primarily aim to reveal the pedagogical effects of ChatGPT. Furthermore, in recent years, AI-based systems have attracted attention for their potential to provide students with instant and personalized feedback (Hooda et al., 2022; Wongvorachan et al., 2022; Xu et al., 2023). One such system, ChatGPT, is capable of generating responses that are explanatory, corrective, and directive in nature to students' answers (Ba et al., 2025; Sain et al., 2024; Wang et al., 2025). Considering ChatGPT's apparent uses and the concerns surrounding its role in education, evaluating its outcomes in the context of science education is of significant importance (Cooper, 2023). A review of the literature reveals that previous research examining the potential of ChatGPT in generating instructional texts on science topics is insufficient (Lee & Zhai, 2024). As AI applications increasingly influence science education, understanding their effectiveness in comparison with human-centered knowledge and skills has become highly important. In particular, the extent to which AI-based tools such as ChatGPT align with or differ from teachers, both in identifying students' incorrect ideas and in providing feedback from a content and pedagogical perspective, has not yet been sufficiently explored. Furthermore, how teachers and AI systems like ChatGPT evaluate the same student response, and the nature of the feedback they provide, have not been adequately investigated. This situation creates an uncertainty in science education research. In this context, the present study aims to reveal how teachers and an AI application, ChatGPT, approach understanding student responses, to evaluate the pedagogical effectiveness of tools such as ChatGPT in providing feedback, and to establish a foundation for future teacher–AI-supported hybrid learning environments. Furthermore, this study is expected to fill an important gap by providing scientific data on the in-class use of AI tools and by contributing to the optimization of human–technology collaboration in science education. Additionally, by comparing science teachers' professional knowledge in understanding and assessing students with AI-generated information, this study seeks to make significant contributions to both science education practitioners and AI developers. Within the scope of the study, the ability of two different sources (teachers and ChatGPT) to

identify errors in students' answers to questions on the topic of heat and temperature, as well as the feedback they provided to students, was compared from a pedagogical perspective. In addition, during this comparison process, teachers' professional seniority and educational background were taken into account to examine their existing strengths and weaknesses in greater depth relative to ChatGPT. In this way, the idea-generation potential of both teachers and AI-supported systems such as ChatGPT was evaluated in the context of science education with respect to different variables.

Student Assessment and Feedback on the Topics of Heat and Temperature

The concepts of heat and temperature form the basis for understanding many physical and chemical processes in thermodynamics and thermochemistry (Elkababi et al., 2020). Heat and temperature are fundamental physics topics that often pose significant learning challenges for students (Eric et al., 2021). The difficulties students encounter in understanding these concepts can negatively affect their ability to analyze and solve real-life problems (Yulianti et al., 2025). Many students struggle to answer questions posed by teachers on these topics because they are unable to directly observe the phenomena involved (Başer, 2006). However, Kotsis et al. (2023) report that everyday experiences are a particularly important factor in students' perception of thermal phenomena. The researchers found that students' perceptions of the concept of heat remain highly resistant throughout the educational process from primary school to university. In this process, providing students with effective feedback can help them accurately construct their knowledge on the topic and correct their misconceptions (Andini et al., 2024). One of the most important characteristics of effective feedback is the teacher's ability to accurately analyze student understanding and, based on this analysis, to present strategies that correct learning errors (Brookhart, 2017). In feedback related to the topics of heat and temperature, it is important for the teacher to encourage the student to question their own ideas (Åhman & Jeppsson, 2020). Therefore, student assessments and effective feedback on the topics of heat and temperature not only serve to measure learning levels but also function as tools to guide and enhance learning (Jewaru et al., 2022).

Assessment in Science Education and ChatGPT

AI plays an important role in the rapidly increasing digitalization of society. Its ability to process large amounts of data and provide meaningful insights is leading to increasingly significant transformations in many areas of daily life (Cooper, 2023; Yang, 2022). As the variety of fields in which AI-based learning applications are used expands, the opportunities offered to improve the education system and keep pace with developments also increase (AI Darayseh, 2023). In recent years, there has been a growing interest in the integration of AI applications into science education, and it has been emphasized that these applications have the potential to significantly transform the ways in which science is taught and learned (Almasri, 2024; Cai et al., 2025). Jia et al. (2024), in their study reviewing trends and practices in the use of AI in science education over the past decade, demonstrate that AI is having an increasingly significant impact on science education. Similarly, Almasri (2024) reveals that AI-supported tools have been extensively integrated into science education for various pedagogical purposes, such as improving the learning environment, preparing examinations, assessing students' work, and predicting their academic performance. In particular, recent emphasis has been placed on the use of AI applications in the design and refinement of assessment tools and performance tasks, highlighting their role in determining the student's next step in the learning process (Jin, 2019). This is because AI applications in science education provide students with opportunities for interactive guidance to enhance the learning process. Students can also benefit from instant feedback and adaptive learning pathways, thereby addressing any misconceptions or gaps in their understanding of scientific phenomena (Mavroudi et al., 2018). The integration of AI and ChatGPT into assessment has led to a remarkable transformation in science education (Alabadi et al., 2023). AI enables science educators to effectively monitor students' progress, allowing for targeted interventions and the provision of necessary support. Moreover, it makes science education more engaging and accessible for students with diverse learning needs (Almasri, 2024). There is a rather limited body of research on the use of AI and ChatGPT in student assessment within science education. Among these studies, Alabidi et al. (2023) highlight that ChatGPT's personalized feedback and adaptive assessment features facilitate students' active engagement in original scientific research, use of critical thinking skills, and participation in problem-solving activities. Zhai and Nehm (2023) emphasize the significant contributions of AI in providing formative feedback to students, assisting teachers in assessment practices, and supporting instructional decision-making. Wu et al. (2025) note that AI (ChatGPT) has substantial potential to transform education in large classroom settings by promoting active learning and self-assessment. The study by Lee and Zhai (2024) with preservice science teachers reveals that ChatGPT has a significant impact on assessment and feedback. Taani and Alabidi (2024) emphasize the positive effects of ChatGPT on science teachers. In their study, teachers drew attention to ChatGPT's effectiveness in providing instant feedback to students. They also stated that ChatGPT facilitates teachers' ability to identify and address students' weaknesses by analyzing student responses and interactions. In addition, it was highlighted that ChatGPT makes important contributions to delivering targeted interventions, personalized guidance, and individualized instruction.

Research Problem and Purpose

Teachers' understanding of students' responses to topics such as "Heat and Temperature," where students experience significant learning difficulties, and the feedback they provide play a crucial role in shaping the instructional process for these concepts. However, the quality of such insights is significantly influenced by various teacher-related factors (Demirci & Şahin, 2016). In particular, teacher feedback on students' responses can often be quite limited today due to constraints of time and competence (Wan & Chen, 2024). At this point, it is noted that ChatGPT, an AI-supported program, has the potential to rapidly and effectively support students' learning processes. However, there is concern that ChatGPT may not reach the level of teacher feedback, particularly in terms of contextual understanding of errors and pedagogical consistency. Moreover, it remains unclear to what extent students can trust AI-generated feedback. This gap stems primarily from the lack of concrete data on effective AI feedback that supports student development, particularly in the context of science education. Therefore, the extent to which the pedagogical quality of AI-generated insights is as effective as that of teacher-generated insights has not yet been sufficiently investigated. Moreover, this situation presents itself as an uncertainty in technology-assisted instructional practices. In line with this gap, comparing the extent to which science teachers and ChatGPT can identify inaccuracies in the same student responses, as well as the feedback they provide on these responses, serves an important purpose for the future design of teacher-AI interactive learning environments. The study can provide data for hybrid learning models by revealing the strengths and weaknesses of both human-based and AI-based sources regarding incorrect ideas and feedback ideas in student responses on the topics of heat and temperature. In this context, the present study aims to compare science teachers and ChatGPT, an AI-supported system, in terms of their ability to detect incorrect ideas in student responses to questions on "Heat and Temperature" at the middle school level and the feedback they offer on these incorrect ideas. Within the scope of the study, the inaccuracies in students' responses, as well as the content and pedagogical quality of the feedback provided by both sources (teachers and ChatGPT), were evaluated. In this way, the intellectual similarities and differences between traditional teacher feedback and AI-supported feedback, as well as how they can complement each other, were revealed. Accordingly, the guiding research question of the study was determined as follows: "In middle school science courses, based on students' responses to questions on the topics of 'Heat and Temperature,' what are the levels at which teachers and ChatGPT identify inaccuracies in these responses, and what similarities and differences exist between them in terms of the feedback that should be provided?"

The sub-questions to be answered in the study are as follows:

- What are the levels at which science teachers and ChatGPT identify errors in students' responses?
- How do science teachers' levels of identifying errors in students' responses vary according to them professional seniority and educational background, and how do they perform compared to ChatGPT?
- How do the feedback provided by science teachers on students' responses vary according to them professional seniority and educational background, and how do they perform compared to ChatGPT?

METHOD

Research Design

This study was conducted using comparative case study design, one of the qualitative research methods (Mello, 2021). Within this design, the study aimed to compare, both pedagogically and in terms of content, the extent to which student errors in the same student responses were identified by two different sources (teachers and ChatGPT) and the feedback they provided.

Participants and Data Source

A total of 28 science teachers working at the middle school level, including 15 males and 13 females, participated in the study. Of these teachers, 2 were in the "commitment" stage (novice, ≤ 3 years), 8 were in the "career stability" stage (mid-level, 4–6 years), 14 were in the "activeness" stage (experienced, 7–18 years), and 4 were in the "stagnation" stage (highly experienced but with low interaction, 19–30 years). In addition, 16 of the participating teachers held a bachelor's degree, 10 held a master's degree, and 2 held a doctoral degree. As the comparative source, the standard (non-Pro) version of ChatGPT was used among artificial intelligence applications.

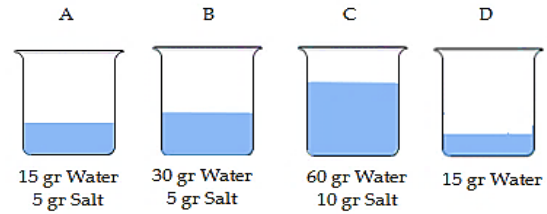
Data Collection

In the study, two questions designed to reveal student incorrect ideas about the topics of "Heat and Temperature," which had been prepared in a previous study by one of the researchers, were used to collect data from both teachers and ChatGPT. From the responses collected from 8th-grade students who had studied the topic of heat and temperature, one response for each question was used within the scope of this research. Based on these questions and the student responses, a second form, the "Student Error Identification and Feedback Form", was prepared. In this form, teachers were asked to identify the errors in the student's response and to write the feedback that should be provided based on the student's answers (see Table 1). These forms were distributed to each participating teacher by the researchers with detailed explanations. Teachers were given approximately 60 minutes to complete

the forms. During data collection, some teachers completed the forms in the researchers' office, while others gathered at their own institutions at convenient times to do so. In addition, the same questions and student responses were entered into ChatGPT, and a "prompt" was written regarding the ideas it was asked to generate (see Table 1). The reason for focusing on two questions on the topic of heat and temperature in this study was to verify the quality of the responses provided by both teachers and ChatGPT.

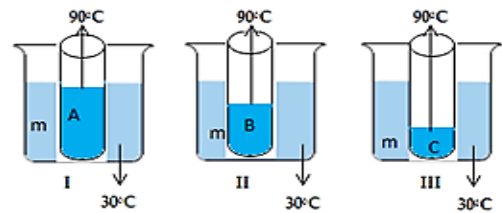
Table 1. Data Collection Forms Administered to Teachers and ChatGPT Prompts

Q1. Containers with the same properties contain water and salt; their amounts are indicated below the containers. Let us imagine that, on a cold winter day, we place the water in these containers on the windowsill. In what order do you think the water in these containers would freeze? Explain.



Student's response: "When we observe the water and salt in these containers, the water in container D freezes first. Then, the water in containers A and B freeze at the same time. Finally, the water in container C freezes. The reason for this is that salt lowers the freezing point of water. Therefore, when we look at the amount of salt in the containers, since container D has no salt, the water in this container freezes first. Then, since containers A and B each have an equal amount of 5 g of salt, they freeze at the same temperature. Lastly, the water in container C, which has the highest amount of salt added, 10 g, freezes."

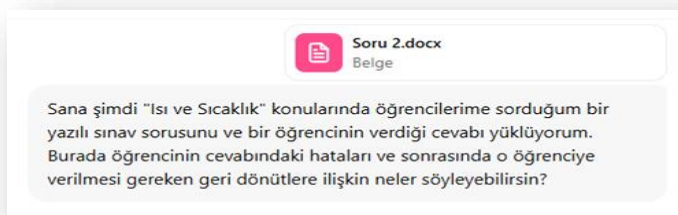
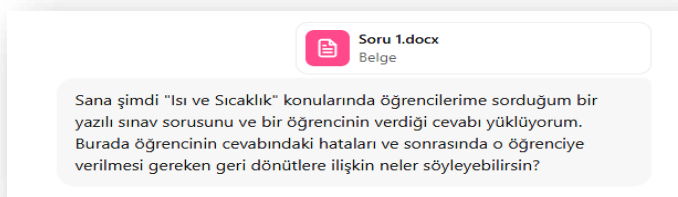
Q2. The initial temperature of identical tubes A, B, and C is 90 °C, and their masses are 400 g, 300 g, and 200 g, respectively. These tubes are immersed in containers I, II, and III, each containing an equal amount of water at a temperature of 30 °C. When the heat exchange between the tubes and the water in the containers is complete, what do you think will be the temperatures of the water in containers I, II, and III? Explain.



Student's response: "Here, the temperature of each tube is equal and 90 °C. It is also known that the masses of the containers are equal. Since the mass of the liquid in tube A is the greatest, at 400 g, the temperature change in container I will be the greatest; then, the change in container II, into which tube B is immersed, will be slightly less, and the smallest temperature change will occur in container III, into which the smallest-mass tube C is immersed."

By examining the above responses from a student to whom you asked these questions, write your thoughts regarding the student's answers and your feedback appropriate to these responses.

Prompts entered into ChatGPT:



The prompt entered for both questions:

"I have now provided you with a written exam question on the topics of 'Heat and Temperature' that I asked my students, along with a response given by a student. What can you say here regarding the errors in the student's answer and the feedback that should be given to the student in relation to this answer?"

Data Analysis

The data obtained in this study were analyzed using the descriptive analysis method. In this context, the researchers first conducted a literature review to identify the types of feedback provided for student responses (McMillan, 2014; Panadero & Lipnevich, 2022). Eight types of feedback common to the reviewed sources were identified (see Table 2), and the researchers evaluated the feedback given by both the teachers and ChatGPT in line with this classification. For data analysis, six sample responses related to student errors and feedback generated by the science teachers and ChatGPT were first evaluated by a committee consisting of a science education expert and two researchers, using a double-blind protocol. During the evaluation process, the descriptions of both the experts and the researchers were compared, and discussions were held on evaluations where inconsistencies were identified. These evaluations yielded a consistency rate of over 90%. Subsequently, the two researchers separately reviewed all the remaining response forms in the same manner, and as a result of this review, an agreement rate of over 90% was achieved both in identifying student errors and in classifying the feedback. Differences in evaluations were reviewed, and complete consensus was reached among the researchers on all responses. In this way, the error identification and feedback suggestions of both the teachers and ChatGPT were analyzed descriptively. These analyses were presented in the findings in the form of frequencies, percentages, and sample excerpts.

Table 2. Feedback Classification (McMillan, 2014)

Types of Feedback		Description
Verification (Informing the students that the answer is correct)	Correct Answer	Without providing additional information, the student is only informed whether their answer is correct or incorrect.
	Try Again	It is stated that the students' answer is incorrect, and they are given an opportunity to try again.
	Marking errors	The students' errors are marked and indicated, but the correct answer is not given, and the student is not told what needs to be done to improve their performance.
Detailing (Include why the answer is wrong and give additional time for relearning)	Abstraction of the attribute	The fundamental features of the learned topic are separated from other subtopics and explained by the teacher.
	Answer-based feedback	Based on the student's answer and performance, the teacher explains why the answer is correct or incorrect.
	Providing hints	Hints are provided by the teacher to the student in order for the correct answer to be found by the student.
	Error Analysis	Misunderstood topics are identified, and error analysis is conducted together with the students.
	Guided feedback	Providing strategies on how the student can demonstrate the required performance without giving the correct answer, including the marking of errors.
Goal Based Feedback		It is the provision of information about the extent to which students have or have not achieved a predetermined goal.
Support Based Feedback		Teachers interact with students to determine their levels and provide the necessary support for them to learn each step by breaking down the activities they are expected to participate in into smaller steps. Teachers offer hints, strategies, and new materials as supportive elements to the students and help them acquire learning skills step by step. The focus is placed on the skills that the student can accomplish with assistance rather than the performance they can demonstrate independently.
Person Based Feedback		Here, the teacher compares the student's performance and the expected performance with their previous performance. The teacher will help increase the student's self-belief regarding their ability to learn the topics. In addition, they will assist the student in understanding what needs to be done in the next step. The teacher should show the student which of their behaviors and efforts played a role in reaching the correct answer.

Feedback that is based on established standards

It is the act of comparing the students' performance with predetermined standards and providing feedback accordingly. The teacher focuses on the student's level of learning based on the established criteria. Here, rather than making a comparison with the student's previous performance, the emphasis is placed on the student's achievements and shortcomings according to the predetermined criteria and performance indicators.

To determine the levels at which both teachers and ChatGPT generated ideas regarding the errors in the student's response, the researchers also developed an "Error Identification Level Rubric" (Table 3). This rubric comprises three levels for identifying errors in a student's response: adequate, partially adequate, and inadequate.

Table 3. Classification of Teachers' Idea Generation Levels Regarding Students' Thinking

Level	Q1 Scientifically Correct Approach to Student Error	Q2 Scientifically Correct Approach to Student Error
Adequate: The response fully conveys the errors in the student's answer or provides a detailed elaboration of the student's errors in many aspects.	The student could not think that the liquid in container D would freeze first, then B, then C, and lastly the mixture in container A. The student does not know that the amount of substance affects the freezing time, not the freezing point. Therefore, since the salt ratios in containers B and C are the same, the student did not understand that the freezing points would also decrease equally. However, the student did not think that container B would freeze earlier because it has less amount of substance. In addition, the student could not determine that the salt ratio in container A is the highest. Therefore, the student did not predict that this salt ratio represents the lowest freezing point and that the mixture in container A would freeze last. The student correctly interprets that the water in container D will freeze first and that freezing will start there because there is no salt in this container. The student claims that after container D, the mixtures in containers A and B will freeze at the same time, and finally, the mixture in container C will freeze. This interpretation is based on the amount of salt in the containers. In fact, the student knows that salt lowers the freezing point of water. However, by focusing on the fact that there is an equal amount of salt in containers A and B, the student has shown a misunderstanding. Moreover, although the student considered that the salt amount in container C is high, this misunderstanding emerged because the student did not compare the water amount in this container and therefore the mixture ratio with containers A and B.	The student could not think that when the heat exchange is completed, the equilibrium temperature would be the highest in container I, followed by II, and finally III. The student knows that the temperature of each tube is 90 °C and that the amount (m) and temperature (30 °C) of the water in the containers are equal. The student could not think that the mass of tube A is higher here and therefore the amount of heat that can be transferred would also be greater. For this reason, the student could not think that the temperature increase of the water in container I would be greater than the temperature increase in container II, and that the increase in container II would be greater than that in container III. The student thinks here that the mass of the liquid in tube A is 400 g, and therefore the least temperature change will occur in container I, in which tube A is immersed, followed by slightly less change in container II, in which tube B is immersed, and the greatest temperature change will occur in container III, in which tube C with the smallest mass is immersed. The student could not think here that a tube with a greater mass would contain more heat energy and therefore the water in container I, in which tube A is immersed, would heat up more.
Partially Adequate: The response provides limited insight into the student's errors or does not contribute sufficiently to the identification of the student's errors.		
Inadequate: The response offers no accurate insight into the errors in the students' ideas or contains no appropriate ideas regarding the students' errors.		

Findings

The comparative analysis of science teachers' and ChatGPT's levels of identifying errors in students' responses to questions on "heat and temperature" and their ideas regarding the feedback to be provided is presented in Table 4.

Table 4. Descriptive Analysis Results of Science Teachers' and ChatGPT's Levels of Identifying Student Errors and Feedback Ideas

		Feedback Source		Teacher n= 28		ChatGPT	
		Learning Status Determination Question		Q1	Q2	Q1	Q2
		Descriptive Analysis		n		Identification	
Codes of Feedback Ideas	Level of Identifying Probable Student Errors	Adequate		15	13		
		Partially Adequate		12	13	✓	✓
		Inadequate		1	2		
		Correct–Incorrect Answer		4	5		
		Try Again		9	8		
		Marking Errors		4	5	✓	✓
		Abstraction of the Attribute		10	7	✓	✓
		Answer-Based Feedback		14	12	✓	✓
		Providing hints		7	5		
		Error Analysis		14	9	✓	✓
		Guided Feedback		4	4	✓	✓
		Goal Based Feedback		8	6	✓	✓
		Support-based Feedback		10	8		

Table 4 presents the levels at which both teachers and ChatGPT identified errors in students' responses and their feedback ideas. According to these findings, it was determined that approximately 54% of the teachers (n = 28, Q1: 15) in the first question and 46% (n = 28, Q2: 13) in the second question identified errors at an adequate level. In addition, it was determined that approximately 43% of the teachers in Q1 and 46% in Q2 identified errors related to student incorrect ideas at a partially adequate level. The rates of teachers who identified errors at an inadequate level were found to be quite low in both questions. In contrast, it came to the forefront that ChatGPT identified errors at a "partially adequate" level in both questions. This shows that although ChatGPT defined the errors, it provided superficial or incomplete analyses in identifying errors in students' responses to these two questions on heat and temperature. When the analysis results regarding the feedback provided by the teachers were examined, it was found that "answer-based feedback", "error analysis", "support-based feedback", and "abstraction of the attribute" were the most prominent. ChatGPT, on the other hand, was found to provide especially "marking errors", "abstraction of the attribute", "answer-based", "error analysis", "guided", and "goal-based" feedback, which overlapped with the feedback ideas emphasized by the teachers. However, it was determined that ChatGPT did not provide any "try again", "correct–incorrect answer", "providing hints", or "support-based" feedback. While ChatGPT was strong in providing explanatory, analytical, and guiding feedback, it was found to be weak in offering feedback that encourages students to try again, provides direct support, or advances the learning process with hints. ChatGPT showed similarities with teachers in certain types of feedback (particularly explanatory and analytical) but was found to be lacking in some pedagogical aspects (such as encouraging retry). While ChatGPT tended to provide more conceptual, guiding, and analytical feedback, teachers were observed to provide much more diverse, student-centered, and encouraging feedback. In particular, feedback such as "try again", which directly contributes to the learning process, was found to be used only by teachers. It can be said that ChatGPT is strong in explanatory and guiding feedback but limited in terms of pedagogical diversity.

Table 5 presents the distribution of teachers' feedback according to professional seniority.

Table 5. Descriptive Analysis Results of the Comparison of Science Teachers' Levels of Identifying Student Errors and Feedback Ideas, According to Professional Seniority, with ChatGPT

Learning Status Determination Question		Q1				Q2			
Teacher Professional Career Stage Codes		Commitment	Career stability	Activeness	Stagnation	Commitment	Career stability	Activeness	Stagnation
Descriptive Analysis		n				n			
Level of Identifying Probable Student Errors	Adequate	-	5	8	2	-	4	6	3
	Partially Adequate	1	3	6	2	-	4	8	1
	Inadequate	1	-	-	-	2	-	-	-
Codes of Feedback Ideas	Correct–Incorrect Answer	1	1	2	-	1	2	2	-
	Try Again	-	3	4	2	2	4	2	1
	Marking Errors	-	1	2	1	-	1	2	2
	Abstraction of the Attribute	1	5	4	1	-	3	3	1
	Answer-Based Feedback	1	6	5	2	1	4	5	2
	Providing hints	-	3	3	1	-	2	2	1
	Error Analysis	1	5	5	3	1	2	4	2
	Guided Feedback	-	1	2	1	-	1	3	1
	Goal Based Feedback	-	3	5	-	-	2	4	-
	Support-based Feedback	-	4	4	2	-	2	4	2

Note. Professional career stages: ≤ 3 years – Commitment; 4–6 years – Career stability; 7–18 years – Activeness; 19–30 years – Stagnation; > 30 years – Withdrawal, based on Richter et al.'s (2011) classification of teachers' professional career stages.

The findings regarding ChatGPT are identical to those presented in Table 1 and, therefore, are omitted here to prevent redundancy.

Table 5 presents the types of feedback provided by teachers and their levels of identifying student errors according to career stages (Commitment, Career stability, Activeness, Stagnation). According to these findings, teachers in the “Activeness” stage identified errors at the “adequate” level most frequently for both questions. This suggests that analytical skills improve alongside professional maturity. Among teachers in the “Commitment” stage, “inadequate” error identification was notable (Q1: 1, Q2: 2). Similarly, teachers in this stage demonstrated significant shortcomings in providing feedback. In contrast, teachers in the “Activeness” stage were found to employ nearly all types of feedback. Teachers in the “Stagnation” stage, however, provided more limited feedback and, in some areas, none at all (e.g., no “goal-based” feedback was given). These findings indicate that teachers in the activeness stage tend to both identify student errors more clearly and provide a greater variety of feedback types, whereas novice and stagnated teachers tend to offer more superficial or incomplete feedback. As professional experience increases, teachers' ability to analyze student responses and provide high-quality feedback becomes more prominent. According to these results, teachers in the “Activeness” stage were found to provide the greatest variety of feedback. Teachers in the “Commitment” stage constituted the weakest group in both error detection and feedback provision (with no contributions in certain areas). These results indicate that experienced teachers use feedback types in a richer and more student-centered manner, whereas novice teachers can be considered a group more open to support. Active teachers were observed to be more effective than ChatGPT in both pedagogical and constructive feedback. ChatGPT employs many types of feedback and stands out particularly in pedagogically strong areas such as “attribute abstraction”, “error analysis”, and “goal-based guidance.”

Table 6 presents the analyses of teachers' levels of identifying errors in student responses and their feedback ideas for these responses according to their educational levels.

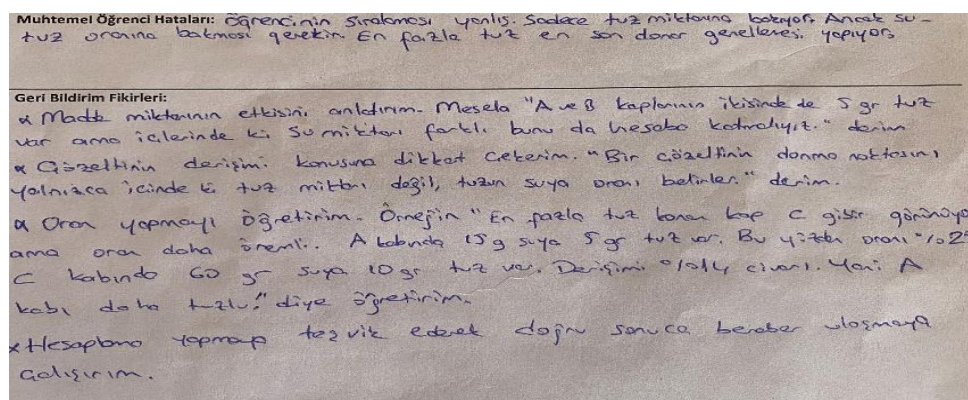
Table 6. Descriptive Analysis Results of the Comparison of Science Teachers' Levels of Identifying Student Errors and Feedback Ideas, According to Educational Level, with ChatGPT

Learning Status Determination Question		Q1			Q2		
Teacher Education Level		Bachelor	Master	Doctorate	Bachelor	Master	Doctorate
Descriptive Analysis		n			n		
Level of Identifying Probable Student Errors	Adequate	8	5	2	8	4	2
	Partially Adequate	7	5	-	6	6	-
	Inadequate	1	-	-	2	-	-
Codes of Feedback Ideas	Correct-Incorrect Answer	2	2	-	3	2	-
	Try Again	4	4	1	3	4	1
	Marking Errors	1	1	-	1	1	-
	Abstraction of the Attribute	3	5	2	2	4	1
	Answer-Based Feedback	6	6	2	4	6	2
	Providing hints	2	4	1	1	3	1
	Error Analysis	6	7	1	2	6	1
	Guided Feedback	2	2	-	2	2	-
	Goal Based Feedback	3	4	1	2	4	-
	Support-based Feedback	4	4	2	3	3	2

Table 6 presents the analyses of teachers' levels of identifying errors in students' responses and their feedback ideas, categorized according to their educational background. The findings indicate that teachers with graduate-level education (master's and doctoral) education were in a stronger position in error detection. Among teachers with a bachelor's degree, "partially adequate" or "inadequate" levels were more prominent, with some remaining superficial in their analyses. Moreover, teachers with master's and doctoral degrees were found to provide feedback particularly in the forms of "attribute abstraction," "answer-based," and "goal-based" feedback, whereas teachers with a bachelor's degree tended to focus more on direct and practice oriented feedback types such as "correct-incorrect", "try again", and "providing hints". Based on these findings, it can be concluded that as the level of education increases, the analytical depth and qualitative differentiation of feedback also increase, and that graduate-level education degree holders are especially more effective in conceptual abstraction and directive analyses. Furthermore, as the level of education increases, teachers' ability to provide conceptual and goal-based feedback was found to improve. The findings revealed that teachers with higher levels of education employed a richer variety of feedback types in a more student-centered manner. Graduate-level education degree holders emerged as the strongest group in feedback types with pedagogical depth (error analysis, goal-based, attribute abstraction). Teachers with a bachelor's degree tended to provide more direct and superficial feedback. Nevertheless, it can be stated that ChatGPT provided feedback that was more consistent, in-depth, and pedagogically effective than that of all educational-level groups. Examples of the findings of science teachers and ChatGPT regarding the marking errors in student responses and the feedback ideas they proposed to support student understanding are presented in Table 7.

Table 7. Science Teachers' and ChatGPT's Identifications of Errors in Student Responses and Feedback Ideas

Teachers' Ideas Based on the Student's Response to Q1



(The information in the image above is provided below.)

Student Error: The student's ranking is incorrect. Because the student only looks at the amount of salt. However, the student should look at the water-salt ratio. The student makes the generalization that the container with the most salt freezes last.

Feedback Ideas:

* I would explain the effect of the amount of substance. For example, I would say, "Both containers A and B have 5 g of salt, but the amount of water in them is different, so we should take that into account."

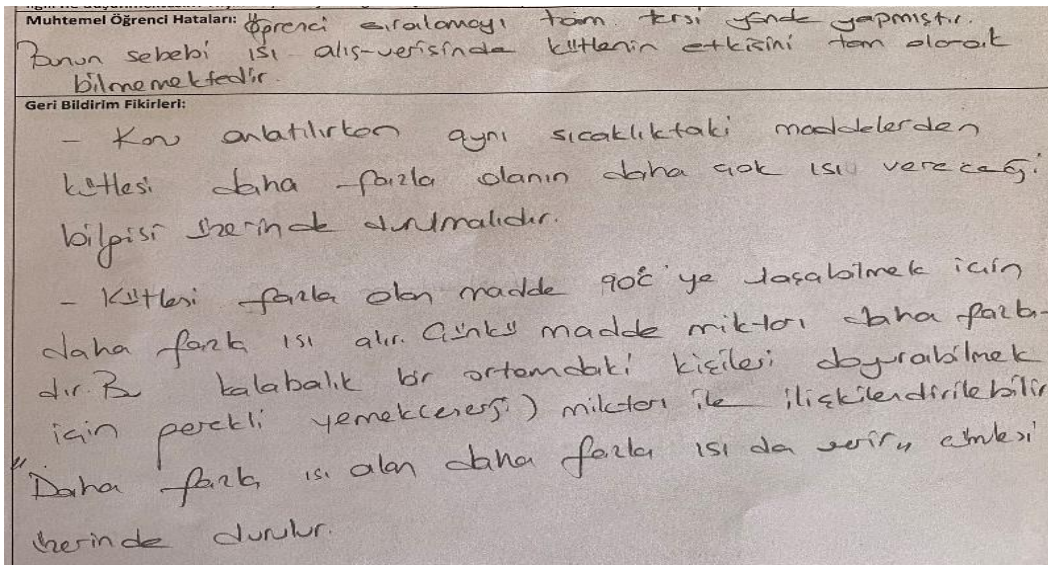
* I would draw attention to the topic of solution concentration. I would say, "A solution's freezing point is determined not only by the amount of salt it contains but also by the ratio of salt to water."

* I would teach how to calculate ratios. For example, I would give the idea, "It seems that container C has the most salt. However, the ratio is more important. In container A, there are 5 g of salt in 15 g of water, so the ratio is 25%. In container C, there are 10 g of salt in 60 g of water, so the concentration is around 14%. Therefore, container A is saltier."

* I would encourage the student to make the calculation and then try to reach the correct result together.

(Teacher21, Master's Degree, Activeness)

Teachers' Ideas Based on the Student's Response to Q2



(The information in the image above is provided below.)

Student Error: The student made the ranking in the exact opposite order. The reason for this is that the student does not fully know the effect of mass in heat exchange.

Feedback Ideas: -While explaining the topic, emphasis should be placed on the fact that, among substances at the same temperature, the one with the greater mass will release more heat.

- A substance with greater mass requires more heat to reach 90 °C. Because the amount of substance is greater. The idea that the one which absorbs more heat also releases more heat should be emphasized.

(Teacher13, Bachelor, Commitment)

ChatGPT's Ideas Based on the Student's Response to Q1

ChatGPT

Yeni sohbet
Sohbetleri ara
Kitaplık
Sora
GPT'ler

Sohbetler
Öğrenci cevabı değerlendirme
Laboratuvar sonuçları analizi

Öğrencinin verdiği yanıt genel olarak doğru bir mantığa dayansa da, hem bilimsel açıklama hem de ifade bakımından bazı eksiklik ve hatalar içeriyor. Aşağıda hem hataları hem de bu öğrenciye verilebilecek yapıcı geri bildirimleri bulabilirsiniz:

Öğrencinin Cevabındaki Hatalar ve Eksiklikler

- Bilimsel Kavramların Kullanımı Eksik:**
 - Öğrenci, tuzun suyun donma noktasını düşürdüğünü doğru bir şekilde belirtmiş. Ancak "donma noktası düşmesi" terimi açıkça kullanılmamış. Bu terim, öğrencinin kavram bilgisini göstermek için önemlidir.
- Sıralama Açıklaması Yüzeysel:**
 - Sıralama yapılmış ancak her kap için nicel ya da nitel bir analiz eksik. Örneğin; neden 5 gr tuzlu sular aynı anda donar, bu net değil. Öğrenci bunun için sıcaklık, tuzluluk oranı ve donma noktası ilişkisini daha açık şekilde anlatmalıydı.
- Varsayımlar Belirtilmemiş:**
 - Dış ortam sıcaklığı, kap hacmi gibi tüm diğer değişkenlerin sabit olduğu varsayılmış ama bu açıkça belirtilmemiş. Bilimsel düşünme açısından bu tür varsayımları belirtmek önemlidir.

(The complete output of ChatGPT identifying the errors in a student's response to Question 1 is presented below.)

Although the student's response is generally based on correct reasoning, it contains some shortcomings and errors in both scientific explanation and expression. Below, you can find both the errors and the constructive feedback that can be given to this student.

Errors and Shortcomings in the Student's Response

1. Lack of Use of Scientific Concepts:

- The student correctly stated that salt lowers the freezing point of water. However, the term "freezing point depression" was not explicitly used. This term is important for demonstrating the student's conceptual knowledge.

2. Ranking Explanation is Superficial:

- A ranking was made, but a quantitative or qualitative analysis for each bowl is missing, for example, why 5 g saltwater freezes at the same time is not clear. The student should have explained the relationship between temperature, salinity ratio, and freezing point more clearly.

3. Assumptions Not Stated:

- It was assumed that all other variables, such as ambient temperature and bowl volume, were constant, but this was not explicitly stated. From a scientific thinking perspective, it is important to state such assumptions.

ChatGPT

Bilimsel Geri Bildirimler

1. Donma Noktası Alçalması Kavramının Açıklanması

Geri bildirim:

"Cevabında tuzun suyun donma sıcaklığını düşürdüğünü belirtmişsin, bu doğru bir gözlem. Ancak bu olayın adı 'donma noktası alçalması' (ya da 'donma noktası düşmesi') olarak bilinir. Bu kavram, koligatif özellikler başlığı altında ele alınır ve yalnızca çözünen maddenin cinsine değil, çözeltideki tanecik sayısına da bağlıdır. Bu nedenle sadece 'tuz miktarı fazla = daha geç donar' demek yerine, çözeltinin molalite değeriyle donma noktasının ilişkilendirildiğini ifade etmen daha bilimsel olurdu."

2. Koligatif Özelliklerle İlişki Kurma

Geri bildirim:

"Donma noktası alçalması, çözücüye eklenen uçucu olmayan bir maddenin (örneğin NaCl gibi bir tuz) çözeltinin donma noktasını saf çözücüye göre düşürmesi anlamına gelir. Bu, koligatif özelliklerden biridir ve molalite ile doğru orantılıdır. Açıklamanda bu ilişkiye değinmiş olman kavramsal anlayışını derinleştirdi."

(Complete of ChatGPT's corrective feedback on a student's incorrect idea is presented below.)

Scientific Feedback

1. Explanation of the Concept of Freezing Point Depression

"In your answer, you stated that salt lowers the freezing point of water, which is a correct observation. However, this phenomenon is known as 'freezing point depression.' This concept is discussed under the topic of colligative properties and depends not only on the type of solute but also on the number of particles in the solution. Therefore, instead of simply saying 'more salt = freezes later,' it would be more scientific to state that the freezing point is related to the molality of the solution."

2. Establishing a Connection with Colligative Properties

"Freezing point depression refers to the lowering of a solution's freezing point compared to that of the pure solvent when a non-volatile solute (such as a salt like NaCl) is added to the solvent. This is one of the colligative properties and is directly proportional to molality. Referring to this relationship in your explanation would have deepened your conceptual understanding."

3. Lack of Molecular-Level Explanation

"When salt dissolves, it dissociates into its ions, and these ions interfere with the ability of water molecules to form a regular crystalline structure. As a result, the freezing of water occurs at a lower temperature. If you add such a molecular-level explanation to your response, you will be providing not only an observation but also a scientific justification."

4. Lack of Quantitative Data Use for Ranking

"In making the ranking, you directly considered the amounts of salt. However, scientifically, the change in a solution's freezing point can be expressed with the formula $\Delta T_f = K_f \cdot m$, where m represents molality and K_f is the freezing point depression constant of the solvent. Including this formula in your explanation would allow you to provide an explanation based not only on observation but also on scientific calculation."

5. Experimental Design and Control Variables

"It is good that you stated the bowls have the same properties; however, for a more systematic scientific approach, you should have explicitly stated all the control variables (ambient temperature, duration of light exposure, bowl volume, amount of water, etc.). This is important for validity and reliability when designing an experimental scenario."

6. Lack of Causal Connection Strength in Explanations

"In your answer, you said that since the amounts of salt in containers A and B are the same, they freeze at the same time. This may be correct, but why exactly does this result occur? To strengthen your answer, you should also consider that the ion concentration in the solution lowers the freezing point by the same amount and how this can be demonstrated experimentally."

Responses obtained from ChatGPT based on the prompt entered on 27.07.2025.

ChatGPT's Ideas Based on the Student's Response to Q2

ChatGPT

Partial
screenshot of
ChatGPT's
identification
of the
student's
errors

Öğrencinin Yanıtındaki Bilimsel Hatalar ve Kavramsal Eksiklikler

1. Sıcaklık ve Isı Kavramları Karıştırılmış

- Öğrenci, sıcaklık değişimiyle ilgili yorum yaparken "sıcaklık değişimi en fazla olur" gibi ifadeler kullanmış ancak bunun hangi yönde olduğu (artan mı, azalan mı) ve neye göre fazla olduğu belirsiz kalmıştır.
- Hata: Öğrenci ısı ile sıcaklığı birbirine karıştırmış. Isı bir enerji türüdür ve kütleye bağlıdır, ancak sıcaklık bir maddenin ortalama kinetik enerjisidir ve kütleye doğrudan ilişkili değildir.

2. Isı Aktarımı Mekanizması Yetersiz Açıklanmış

- Tüplerin sıcaklığı 90°C, suyun sıcaklığı 30°C olduğuna göre ısı, tüplerden suya geçer.
 - Öğrenci, bu enerji aktarımını anlamış gibi görünse de ısı miktarı hesaplamaya yönelik hiçbir ifade veya formül kullanmamış.
- Oysa bu tür bir soruda temel fiziksel ilişki şudur:

$$Q = m \cdot c \cdot \Delta T$$

Bu formül hem tüplerin verdiği ısı miktarını hem de suyun aldığı ısıyı nicel olarak değerlendirmemize olanak sağlar.

(The complete output of ChatGPT identifying the errors in a student's response to Question 2 is presented below.)

Scientific Errors and Conceptual Gaps in the Student's Response

1. Confusion Between the Concepts of Temperature and Heat

- While commenting on temperature change, the student used expressions such as "the temperature change would be the greatest," but it remained unclear in which direction this change occurs (increase or decrease) and according to what it is considered greater.
- Error: The student confused heat with temperature. Heat is a form of energy and depends on mass, whereas temperature is the average kinetic energy of a substance and is not directly related to mass.

2. Heat Transfer Mechanism Poorly Explained

- Since the tubes are at 90 °C and the water is at 30 °C, heat flows from the tubes to the water.
- Although the student seems to have understood this energy transfer, no statement related to calculating the amount of heat was provided.

However, in this type of question, the fundamental physical relationship is:

$$Q = m c \Delta T.$$

This formula allows us to quantitatively evaluate both the amount of heat given by the tubes and the heat absorbed by the water.

3. Equilibrium Temperature Concept Overlooked

- The question asks about "the final temperatures in the containers," which refers to the equilibrium temperatures after heat exchange. Instead of predicting the equilibrium temperatures, the student commented only on the magnitude of the "temperature change," and in this respect, did not fully answer the question.

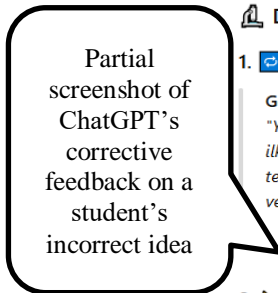
4. Superficial Response and Lack of Quantitative Approach

- No numerical estimation, comparison, or ratio calculation was made. The mass data provided in the question was not used appropriately (for example, a quantitative analysis could have been done on how much heat the 400 g tube would release).

5. Neglect of the Specific Heat Concept

- The answer was based solely on mass; however, from a scientific perspective, when calculating the amount of energy, specific heat capacity (c) and temperature difference (ΔT) are also taken into account ($Q = m c \Delta T$). It appears that the student either did not know or did not apply this equation.

ChatGPT ▾



Daha Bilimsel ve Derinlemesine Geri Bildirimler

1. Isı Transferi Mekanizmasına Bilimsel Yaklaşım Eksikliği

Geri Bildirim:

"Yanıtında ısı alışverişi sürecinden bahsetmişsin ancak bu süreç, enerji korunumuna dayalı bir termodinamik ilkedir. Yani, tüplerin verdiği toplam ısı miktarı, suyun aldığı toplam ısıya eşit olmalıdır. Açıklamanda bu temel fiziksel yasa eksik kalmış. Isı alışverişi denklemleriyle desteklemen bilimsel açıklamanın doğruluğunu ve derinliğini artırdı."

2. $Q = m \cdot c \cdot \Delta T$ Denkleminde Yararlanmamış Olman Eksiklikler

Geri Bildirim:

"Sistemdeki her tüp eşit sıcaklık farkına ($\Delta T = 90^\circ\text{C} - 30^\circ\text{C} = 60^\circ\text{C}$) sahip olduğuna göre, her birinin suya verdiği ısı yalnızca kütlelerine bağlı olur (öz ısı ve sıcaklık farkı sabit). Bu durumda, A tüpü $400\text{ g} \times c \times 60^\circ\text{C}$ kadar ısı verirken, C tüpü yalnızca $200\text{ g} \times c \times 60^\circ\text{C}$ kadar verir. Bu mantıkla suyun son sıcaklıklarının $I > II > III$ şeklinde sıralanacağını nicel olarak temellendirmeliydin."

(Complete of ChatGPT's corrective feedback on a student's incorrect idea is presented below.)

Scientific Feedback

1. Lack of Scientific Approach to the Heat Transfer Mechanism

"In your answer, you mentioned the heat exchange process; however, this process is a thermodynamic principle based on the conservation of energy. In other words, the total amount of heat released by the tubes must equal the total amount of heat absorbed by the water. This fundamental physical law was missing in your explanation. Supporting your answer with heat exchange equations would have increased the accuracy and depth of your scientific explanation."

2. Failure to Utilize the $Q = m \cdot c \cdot \Delta T$ Equation is a Shortcoming

"Since each tube in the system has the same temperature difference ($\Delta T = 90^\circ\text{C} - 30^\circ\text{C} = 60^\circ\text{C}$), the heat each one gives to the water depends only on its mass (specific heat and temperature difference are constant). In this case, while tube A gives $400\text{ g} \times c \times 60^\circ\text{C}$ of heat, tube C gives only $200\text{ g} \times c \times 60^\circ\text{C}$. With this reasoning, you should have quantitatively justified that the final temperatures of the water would be ordered as $I > II > III$."

3. Equilibrium Temperature Concept Overlooked

"Since the question asks for the final temperatures in the containers, your answer is expected to explain the process of reaching equilibrium temperature. Equilibrium temperature is reached when the temperatures of the substances giving and receiving heat become equal. As different amounts of heat transfer occur in each container, this difference is reflected in the final temperatures as well. Therefore, you should have focused not only on 'temperature change' but also on the 'final temperature.'"

4. Inadequacy in Establishing Relationships Between Concept

"Not expressing the relationship between the amount of heat (Q), mass (m), temperature change (ΔT), and specific heat (c) using the equation $Q = m \cdot c \cdot \Delta T$ has limited your ability to establish conceptual connections. Since the term 'identical tubes' is used in the question, c can be considered constant. In this case, it is possible to compare heat quantities solely based on mass differences. Establishing this connection would have demonstrated scientific reasoning."

5. Lack of Scientific Modeling

"You could have constructed a simple heat balance model when solving this question. For example, you could have approximately calculated the equilibrium temperatures by equating the energy required to heat the water in the containers with the energy transferred from the tubes. Creating such quantitative models is important for developing your physical reasoning skills."

6. Failure to Explicitly State Assumptions Weakens the Scientific Approach

"In your answer, you did not specify basic assumptions such as heat insulation, whether the system is closed, or whether there is energy exchange with the environment. In scientific problem-solving, analyses without clearly stated assumptions remain incomplete. For example, you should have explicitly stated assumptions such as 'the containers are insulated' or 'the water quantities are equal and in identical containers'."

7. Causal Connections Are Needed Instead of Superficial Commentary

"Instead of superficial statements such as 'the temperature change is the greatest,' establishing cause-and-effect connections like 'because the tube releases more heat energy, this energy increases the kinetic energy of the water and raises its temperature' would have reflected your scientific thinking skills."

Responses obtained from ChatGPT based on the prompt entered on 27.07.2025.

DISCUSSION

This study was conducted to compare the abilities of science teachers and ChatGPT in analyzing students' responses to heat and temperature questions and in providing feedback on these topics. The findings, in summary, offer important insights into the idea-generation capabilities of both teachers and ChatGPT in this regard. Within the scope of the study, it was found that approximately half of the science teachers were inadequate in detecting errors in students' responses to questions on the topics of heat and temperature. This finding has been consistent with the results of previous research (Inaltekin & Akcay, 2021). Additionally, the study revealed that ChatGPT also produced superficial or incomplete analyses when identifying student errors. These results indicate that both many science teachers and ChatGPT have notable limitations in detecting and deeply analyzing student errors, which may constitute a barrier to conducting an effective instructional process and providing high-quality feedback. This finding is also consistent with previous research emphasizing that ChatGPT's responses should be

carefully evaluated by both instructors and students, particularly in terms of scientific accuracy (Dahlkemper et al., 2023). Conversely, another study by Bewersdorff et al. (2023) has demonstrated that there are varying levels of accuracy in error detection between AI systems and human evaluators. They particularly highlight that AI systems are capable of accurately identifying many fundamental student errors (Bewersdorff et al., 2023).

It was observed that teachers provided a more diverse, student-centered, and pedagogically rich structure in terms of feedback types. Moreover, it was concluded that feedback types that directly promote learning, such as encouraging students to try again and providing hints, were employed exclusively by teachers. ChatGPT, on the other hand, was found to demonstrate strong performance particularly in explanatory, analytical, and directive feedback. However, it was observed that ChatGPT did not employ feedback strategies with pedagogical purposes, such as repetition, emphasizing correct-incorrect, and providing hints. This situation indicates that ChatGPT remains passive in certain important components of the feedback process (Yoon et al., 2023). These results revealed that science teachers and ChatGPT possess different strengths in their feedback processes (Seßler et al., 2025). Therefore, integrating teachers' pedagogical approaches with ChatGPT's knowledge-based analytical skills can be considered to hold significant potential for supporting students' learning processes in conceptually critical topics such as heat and temperature. This finding is also consistent with previous research suggesting that ChatGPT summarizes students' performance more fluently and coherently than human instructors and can provide more detailed feedback accordingly (Dai et al., 2023).

Another finding revealed that teachers in the "activeness" stage of professional seniority (7–18 years) and those holding a master's degree demonstrated the highest productivity in both identifying student errors and providing feedback. In particular, teachers with a master's degree and the activeness stage showed a level of feedback productivity comparable to that of ChatGPT. This finding serves as evidence that experience enhances pedagogical analysis and interaction skills (Kini & Podolsky, 2016). Conversely, it was found that novice teachers with only a bachelor's degree had significant shortcomings in both error detection and feedback productivity, performing at a lower level compared to ChatGPT. This result indicates that pedagogical awareness is a dynamic process that develops in relation to educational attainment and professional experience. It particularly underscores the direct impact on the depth of pedagogical analysis and the ability to interact with students. Solak (2024) states that human teachers are more effective in providing feedback due to their understanding, personalized guidance, and emotional intelligence, whereas AI-generated feedback is perceived as more detailed and comprehensive. This highlights the importance of balancing efficiency with the human touch in the feedback process. Supporting this result, the study by Yu and Xie (2025) has shown that ChatGPT's feedback is comprehensive, accurate, and performs qualitatively and quantitatively better than that of teachers. The researchers emphasize that ChatGPT feedback should be supported as an additional resource alongside teacher feedback in classroom settings (Demir & Çetin, 2025). It was found that teachers with graduate-level education were particularly strong in providing "conceptual abstraction," "analytical analysis," and "goal-oriented" feedback. In contrast, teachers with only a bachelor's degree tended to focus more on practical and direct feedback types, such as "correct-incorrect," "hints," and "try again". This finding reveals that educational attainment is a determining factor not only in terms of knowledge but also in the diversity of pedagogical feedback. Therefore, it can be stated that teachers with higher levels of education are able to support students' conceptual development at a more advanced level. The differences observed in feedback types according to educational level highlight the variety of factors that shape teachers' pedagogical repertoires (Gan et al., 2021). The fact that teachers with graduate-level education were strong in "conceptual abstraction," "analytical analysis," and "goal-oriented" feedback, whereas those with only a bachelor's degree tended to focus more on practice-oriented strategies such as "correct-incorrect," "hints," and "try again," has shown that educational attainment shapes the quality of feedback types. This finding reveals that teachers' educational level is a determining factor not only in content knowledge but also in conceptual and analytical thinking skills.

CONCLUSION

This study highlights important findings regarding the abilities of teachers and ChatGPT to understand student ideas and provide feedback on heat and temperature, which are topics in the science curriculum that are particularly challenging for students. One of these findings revealed that science teachers' skills in error detection and feedback generation vary depending on their professional experience and educational level. Both experienced teachers and those with graduate-level education demonstrated performance similar to that of ChatGPT in identifying student errors and providing feedback that was both of higher quality and more strongly supportive of learning. In contrast, novice teachers with only a bachelor's degree were found to have notable deficiencies in both error detection and feedback generation. ChatGPT can serve as a pedagogical complement, particularly for novice science teachers; however, it needs to be supported by teacher experience and expertise to ensure pedagogical depth. For beginning teachers, tools such as ChatGPT can be useful for providing analytical examples and offering response-based analyses, yet they may be insufficient for fostering pedagogical diversity. This study revealed that hybrid models based on a comparative analysis of teachers and ChatGPT (human–AI collaboration) have the potential to

strengthen processes of identifying student errors and providing feedback in science education.

Limitations

The study's sample size and its restriction to a single subject area (heat and temperature) may limit the generalizability of the findings.

Implications

Based on the findings of this research, ChatGPT can be used as a pedagogical support tool to help novice teachers identify student errors and develop feedback strategies. In this regard, AI-assisted assessment applications could be integrated into both pre-service and in-service teacher training programs. The strengths of experienced and highly educated science teachers in providing feedback could be utilized in the development of AI systems. Future research could explore different science topics, larger groups of teachers, and various AI models. In addition, professional development programs and AI algorithm enhancements are needed to improve the abilities of both teachers and AI-based systems to diagnose and interpret student incorrect ideas. In this context, it is recommended that AI-based systems be developed to enhance their capacity to provide feedback aligned with pedagogical strategies, while teachers' abilities to use technology effectively should also be strengthened.

Ethics Approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Conflict of Interest: The authors declare no competing interests.

REFERENCES

1. Åhman, N., & Jeppsson, F. (2020). Teachers' and pupils' scientific dialogue in learning about invisible thermal phenomena. *International Journal of Science Education*, 42(18), 3116–3133. <https://doi.org/10.1080/09500693.2020.1852334>
2. Alabidi, S., AlArabi, K. S., Alsalhi, N. R., & al Mansoori, M. (2023). The dawn of ChatGPT: Transformation in science assessment. *Eurasian Journal of Educational Research*, 106(106), 321-337. <https://doi.org/10.29333/ejedus/11400>
3. Al Darayseh, A. (2023). Acceptance of artificial intelligence in teaching science: Science teachers' perspective. *Computers and Education: Artificial Intelligence*, 4, 100132. <https://doi.org/10.1016/j.caeai.2023.100132>
4. Almasri, F. (2024). Exploring the impact of artificial intelligence in teaching and learning of science: A systematic review of empirical research. *Research in Science Education*, 54(5), 977–997. <https://doi.org/10.1007/s11165-024-10176-3>
5. Andini, R., Khusairi, S., & Khusaini, K. (2024, May). Development of moodle-based formative feedback of temperature and heat to overcome learning difficulties for high school students. In *AIP Conference Proceedings* (Vol. 3106, No. 1, p. 060002). AIP Publishing LLC.
6. Ba, S., Zhan, Y., Huang, L., & Lu, G. (2025). Investigating the impact of ChatGPT-assisted feedback on the dynamics and outcomes of online inquiry-based discussion. *British Journal of Educational Technology*. <https://doi.org/10.1111/bjet.13605>
7. Barendsen, E., & Henze, I. (2019). Relating teacher PCK and teacher practice using classroom observation. *Research in Science Education*, 49, 1141–1175. <https://doi.org/10.1007/s11165-017-9637-z>
8. Başer, M. (2006). Fostering conceptual change by cognitive conflict based instruction on students' understanding of heat and temperature concepts. *Eurasia journal of mathematics, science and technology education*, 2(2), 96-114. <https://doi.org/10.12973/ejmste/75458>
9. Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence*, 5, 100177. <https://doi.org/10.1016/j.caeai.2023.100177>
10. Bill Gates predicts AI will replace doctors, teachers within 10 years - and claims humans won't be needed 'for most things'. (*New York Post*). (2025, March 27). Retrieved from: https://people.com/bill-gates-ai-will-replace-doctors-teachers-in-next-10-years-11705615?utm_source=chatgpt.com
11. Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences of the United States of America*, 120(6). <https://doi.org/10.1073/pnas.2218523120>
12. Black, P., & Wiliam, D. (2018). *Classroom assessment and pedagogy. Assessment in Education: Principles, Policy & Practice* (pp. 1–25). <https://doi.org/10.1080/0969594X.2018.1441807>
13. Brookhart, S. M. (2008). *How to give effective feedback to your students*. Association for Supervision and Curriculum Development.
14. Brown, G. T. L. (2018). *Assessment of student achievement*. Routledge.
15. Cai, Y., Deng, Q., Lv, T., Zhang, W., & Zhou, Y. (2025). Impact of GPT on the Academic Ecosystem. *Science & Education*, 34(2), 913-931. <https://doi.org/10.1007/s11191-024-00561-9>

16. Carlson, J., Daehler, K. R., Alonzo, A. C., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Kam, K., Chan, H., Cooper, R., Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Nilsson, P., ... Wilson, C. D. (2019). The refined consensus model of pedagogical content knowledge in science education. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 77–94). Springer. https://doi.org/10.1007/978-981-13-5898-2_2
17. Cooper, G. (2023). Examining science education in chatgpt: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*, 32(3), 444–452. <https://doi.org/10.1007/s10956-023-10039-y>
18. Dahlkemper, M. N., Lahme, S. Z., & Klein, P. (2023). How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT. *Physical Review Physics Education Research*, 19(1), 010142(1)- 010142(25). <https://doi.org/10.1103/physrevphyseducres.19.010142>
19. Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y.-S., Gasevic, D., & Chen, G. (2023). Can Large Language models provide feedback to students? A case study on ChatGPT. <https://doi.org/10.35542/osf.io/hcgzj>
20. Demirci, C., & Şahin, E. (2014). Fen ve Teknoloji Öğretmenlerinin Öğrencilerin Isı ve Sıcaklık Konusundaki Kavram Yanılgıları Hakkındaki Görüşleri. *Eğitim Ve İnsani Bilimler Dergisi: Teori Ve Uygulama*, 5(9), 67-76.
21. Demir, S. Ö., & Çetin, M. (2025). Bir Üretken Yapay Zekâ Aracı Olarak ChatGPT'nin Yabancı Dil Öğretiminde Kullanılmasına İlişkin Görüşler. *Dil Eğitimi ve Araştırmaları Dergisi*, 11(1), 207-234. <https://doi.org/10.31464/jlere.1501596>
22. Deverel-Rico, C., & Furtak, E. M. (2025). How Can Features of Classroom Activity Systems Center Students' Ideas and Experiences in Formative Assessment? A Study of Sensemaking and Answer-Making Frames. *Educational Assessment*, 30(2), 91-114. <https://doi.org/10.1080/10627197.2025.2504081>
23. Elkababi, I., Atıbi, A., Radıd, M., Belaaouad, S., & Tayane, S. (2020). Assessment of learners' learning about temperature and heat concepts. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(2), 956-962. <https://doi.org/10.30534/ijatcse/2020/08922020>
24. Eric A., Daniel N., Richard A., Frederick A. (2021). Diagnostic assessment of students' misconceptions about heat and temperature through the use of two-tier test instrument. *British Journal of Education, Learning and Development Psychology*, 4(1), 90-104. <https://doi.org/10.52589/BJELDPO22B5EPK>
25. Gan, Z., An, Z., & Liu, F. (2021). Teacher feedback practices, student feedback motivation, and feedback behavior: How are they associated with learning outcomes? *Frontiers in Psychology*, 12, 697045. <https://doi.org/10.3389/fpsyg.2021.697045>
26. Gess-Newsome, J., Taylor, J. A., Carlson, J., Gardner, A. L., Wilson, C. D., & Stuhlsatz, M. A. (2019). Teacher pedagogical content knowledge, practice, and student achievement. *International Journal of Science Education*, 41(7), 944–963. <https://doi.org/10.1080/09500693.2016.1265158>
27. Güney, S., & Feyzioğlu, B. (2023). Types of feedback in inquiry science learning classroom in secondary education. *Psycho-Educational Research Reviews*, 12(2), 390-408. https://doi.org/10.52963/PERR_Biruni_V12.N2.03
28. Hartelt, T., Martens, H., & Minkley, N. (2022). Teachers' ability to diagnose and deal with alternative student conceptions of evolution. *Science Education*, 106(3), 706-738. <https://doi.org/10.1002/sce.21705>
29. Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112. <https://doi.org/10.3102/003465430298487>
30. Hooda, M., Rana, C., Dahiya, O., Rizwan, A., & Hossain, M. S. (2022). Artificial intelligence for assessment and feedback to enhance student success in higher education. *Mathematical Problems in Engineering*, 2022(1), 5215722. <https://doi.org/10.1155/2022/5215722>
31. Huang, X. (2024). Application of artificial intelligence APP in quality evaluation of primary school science education. *Educational Studies*, 50(6), 1215-1235. <https://doi.org/10.1080/03055698.2022.2066462>
32. Inaltekin, T. & Akcay, H. (2021). Examination the knowledge of student understanding of pre-service science teachers on heat and temperature. *International Journal of Research in Education and Science*, 7(2), 445-478. <https://doi.org/10.46328/ijres.1805>
33. Irons, A., & Elkington, S. (2021). *Enhancing learning through formative assessment and feedback*. Routledge.
34. Jewaru, A. A. L., Parno, Nasikhudin, Kusairi, S., & Latifah, E. (2022, November). The profile of students' science process skill on thermodynamics topic through authentic PBL with STEM approach and formative assessment. In *AIP Conference Proceedings* (Vol. 2659, No. 1, p. 120011). AIP Publishing LLC.
35. Jia, F., Sun, D., & Looi, C. K. (2024). Artificial intelligence in science education (2013–2023): Research trends in ten years. *Journal of Science Education and Technology*, 33(1), 94–117. <https://doi.org/10.1007/s10956-023-10077-6>
36. Jin, L. (2019). Investigation on potential application of artificial intelligence in preschool children's

- education. *Journal of Physics: Conference Series*, 1288, Article 012072. <https://doi.org/10.1088/1742-6596/1288/1/012072>
37. Kini, T., & Podolsky, A. (2016). *Does Teaching Experience Increase Teacher Effectiveness? A Review of the Research*. Palo Alto: Learning Policy Institute. Retrieved from: https://learningpolicyinstitute.org/sites/default/files/productfiles/Teaching_Experience_Report_June_2016.pdf
38. Kotsis, K. T., Stylos, G., Houssou, P. ., & Kamaratos, M. . (2023). Students' perceptions of the heat and temperature concepts: a comparative study between primary, secondary, and university levels. *European Journal of Education and Pedagogy*, 4(1), 136–144. <https://doi.org/10.24018/ejedu.2023.4.1.577>
39. Lee, G., & Zhai, X. (2024). Using ChatGPT for science learning: A study on pre-service teachers' lesson planning. *IEEE Transactions on Learning Technologies*, 17, 1643–1660. <https://doi.org/10.1109/TLT.2024.3401457>.
40. Limbere, A. M., Munakata, M., Klein, E. J., & Taylor, M. (2022). Exploring the tensions science teachers navigate as they enact their visions for science teaching: What their feedback can tell us. *International Journal of Science Education*, 44(12), 1897–1915. <https://doi.org/10.1080/09500693.2022.2105413>
41. Magnusson, S., Krajcik, L., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 95–132). Kluwer.
42. Mavroudi, A., Giannakos, M., & Krogstie, J. (2018). Supporting adaptive learning pathways through the use of learning analytics: Developments, challenges and future opportunities. *Interactive Learning Environments*, 26(2), 206–220. <https://doi.org/10.1080/10494820.2017.1292531>
43. McMillan, J. (2014). *Classroom assessment: Principles and practice for effective standards-based instruction* (6th ed.). New York: Pearson.
44. Mello, P.A. (2021). *Qualitative comparative analysis: an introduction to research design and application*. Georgetown University Press.
45. Miller, M. D., Linn, R., & Gronlund, N. (2013). *Measurement and assessment in teaching* (11th ed.). Pearson.
46. Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. *British Journal of Educational Technology*, 55(4), 1328–1353. <https://doi.org/10.1111/bjet.13454>
47. Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, 35, 100416. <https://doi.org/10.1016/j.edurev.2021.100416>
48. Park, S., & Oliver, J. S. (2008). Revisiting the conceptualisation of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Research in Science Education*, 38(3), 261–284. <https://doi.org/10.1007/s11165-007-9049-6>
49. Park, H. K., & Martin, S. N. (2024). Exploring the role of ChatGPT in science education for Asia-Pacific and beyond: a systematic review. *Asia-Pacific Science Education*, 10(2), 233–264. <https://doi.org/10.1163/23641177-bja10079>
50. Park, J., Teo, T. W., Teo, A., Chang, J., Huang, J. S., & Koo, S. (2023). Integrating artificial intelligence into science lessons: Teachers' experiences and views. *International Journal of STEM Education*, 10(1), 61. <https://doi.org/10.1186/s40594-023-00454-3>
51. Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, 107769582211495. <https://doi.org/10.1177/10776958221149577>
52. Richter, D., Kunter, M., Klusmann, U., Lüdtke, O., & Baumert, J. (2011). Professional development across the teaching career: Teachers' uptake of formal and informal learning opportunities. *Teaching and Teacher Education*, 27(1), 116–126. <https://doi.org/10.1016/j.tate.2010.07.008>
53. Sadler, P. M., & Sonnert, G. (2016). Understanding misconceptions: Teaching and learning in middle school physical science. *American Educator*, 40(1), 26–32.
54. Sain, Z. H., Vasudevan, A., Thelma, C. C., & Asfahani, A. (2024). Harnessing ChatGPT for effective assessment and feedback in education. *Journal of Computer Science and Informatics Engineering*, 3(2), 74–82. <https://doi.org/10.55537/cosie.v3i2.856>
55. Salame, I. I., Fadipe, O., & Akter, S. (2025). Examining difficulties, challenges, and alternative conceptions students exhibit while learning about heat and temperature concepts. *Interdisciplinary Journal of Environmental and Science Education*, 21(2), e2510. <https://doi.org/10.29333/ijese/15997>
56. Seßler, K., Bewersdorff, A., Nerdel, C., & Kasneci, E. (2025). Towards adaptive feedback with AI: Comparing the feedback quality of LLMs and teachers on experimentation protocols. *arXiv preprint arXiv:2502.12842*, 1–27. <https://doi.org/10.48550/arXiv.2502.12842>
57. Solak, E. (2024). Examining writing feedback dynamics from CHATGPT AI and human educators: A comparative study. *Παδαγωγικά*, 96(7), 955–970.

58. Taani, O., & Alabidi, S. (2024). ChatGPT in education: Benefits and challenges of ChatGPT for mathematics and science teaching practices. *International Journal of Mathematical Education in Science and Technology*, 1–30. <https://doi.org/10.1080/0020739X.2024.2357341>
59. Valeri, F., Nilsson, P., & Cederqvist, A. M. (2025). Exploring students' experience of ChatGPT in STEM education. *Computers and Education: Artificial Intelligence*, 8, 100360. <https://doi.org/10.1016/j.caeai.2024.100360>.
60. Wan, T., & Chen, Z. (2024). Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review Physics Education Research*, 20(1), 010152. <https://doi.org/10.1103/PhysRevPhysEducRes.20.010152>
61. Wancham, K., & Tangdhanakanond, K. (2022). Effects of feedback types and opportunities to change answers on achievement and ability to solve physics problems. *Research in Science Education*, 52(2), 427-444. <https://doi.org/10.1007/s11165-020-09956-4>
62. Wang, W. S., Lin, C. J., Lee, H. Y., Huang, Y. M., & Wu, T. T. (2025). Enhancing self-regulated learning and higher-order thinking skills in virtual reality: the impact of ChatGPT-integrated feedback aids. *Education and Information Technologies*, 1-27. <https://doi.org/10.1007/s10639-025-13557-x>
63. Wongvorachan, T., Lai, K. W., Bulut, O., Tsai, Y.-S., & Chen, G. (2022). Artificial intelligence: transforming the future of feedback in education. *Journal of Applied Testing Technology*, 23, 95–116.
64. Wu, T. T., Lee, H. Y., Chen, P. H., Lin, C. J., & Huang, Y. M. (2025). Integrating peer assessment cycle into ChatGPT for STEM education: A randomised controlled trial on knowledge, skills, and attitudes enhancement. *Journal of Computer Assisted Learning*, 41(1), e13085. <https://doi.org/10.1111/jcal.13085>
65. Yang, W. (2022). Artificial intelligence education for young children: Why, what, and how in curriculum design and implementation. *Computers and Education: Artificial Intelligence*, 3, 100061. <https://doi.org/10.1016/j.caeai.2022.100061>
66. Yoon, S. Y., Miszogladd, E., & Pierce, L. R. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. *arXiv preprint arXiv:2310.06505*. <https://doi.org/10.48550/arXiv.2310.06505>
67. Yu, H., & Xie, Q. (2025). Generative AI vs. teachers: Feedback quality, feedback uptake, and revision. *Language Teaching Research Quarterly*, 47, 113-137. <https://doi.org/10.32038/ltrq.2025.47.07>
68. Yulianti, E., Abdul Rahman, N. F., Suwono, H., & Phang, F. A. (2025). Transdisciplinary STEAM learning in improving students' conceptual understanding of heat and temperature. *Research in Science & Technological Education*, 1–21. <https://doi.org/10.1080/02635143.2025.2452542>
69. Xu, W., Meng, J., Raja, S. K. S., Priya, M. P., & Kiruthiga Devi, M. (2023). Artificial intelligence in constructing personalized and accurate feedback systems for students. *International Journal of Modeling, Simulation, and Scientific Computing*, 14(01), 2341001. <https://doi.org/10.1142/S1793962323410015>
70. Zhai, X. (2023). ChatGPT for next generation science learning. *XRDS: Crossroads, the ACM Magazine for Students*, 29(3), 42–46. <https://doi.org/10.1145/3589649>
71. Zhai, X., & Nehm, R. H. (2023). AI and formative assessment: The train has left the station. *Journal of Research in Science Teaching*, 60(6), 1390-1398. <https://doi.org/10.1002/tea.21885>