HUMAN-ROBOT INTERACTION

Observing a robot peer's failures facilitates students' classroom learning

Liuqing Chen[1]*†, Yu Cai[1]†, Yuyang Fang[1], Ziqi Yang[2], Duowei Xia[1], Jiaxiang You[3], Shuhong Xiao[1], Yaxuan Song[1], Lingwei Zhan[1], Juanjuan Chen[4], Lingyun Sun[1]*

According to productive failure (PF) theory, experiencing failure during problem-solving can enhance students' knowledge acquisition in subsequent instruction. However, challenging students with problems beyond their current capabilities may strain their skills, prior knowledge, and emotional well-being. To address this, we designed a social robot–assisted teaching activity in which students observed a robot's unsuccessful problem-solving attempts, offering a PF-like preparatory effect without requiring direct failure. We conducted two classroom-based studies in a middle school setting to evaluate the method's effectiveness. In study 1 ($N = 135$), we compared three instructional methods—observing robot failure (RF), individual problem-solving failure, and direct instruction—in an eighth-grade mathematics lesson. Students in the RF condition showed the greatest gains in conceptual understanding and reported lower social pressure, although no significant differences were found in procedural knowledge or knowledge transfer. Follow-up study 2 ($N = 110$) further validated the method's effectiveness in supporting knowledge acquisition after a 2-week robot-involved adaptation phase, when the novelty effect had largely subsided. Students confirmed their perception of the robot as a peer, and they offered positive evaluations of its intelligence and neutral views of its anthropomorphism. Our findings suggest that observing the robot's failure has a comparable, or even greater, effect on knowledge acquisition than experiencing failure firsthand. These results underscore the value of social robots as peers in science, technology, engineering, and mathematics education and highlight the potential of integrating robotics with evidence-based teaching strategies to enhance learning outcomes.

INTRODUCTION

Educational studies emphasize the benefits of experiencing impasses and failures in the classroom (*1*), which can stimulate deep thinking and improve learning outcomes, often exceeding the effects of direct instructions from teachers (*1*–*3*). A prominent instructional strategy grounded in this principle is productive failure (PF) (*2*). Within the PF framework, students are encouraged to attempt solving problems before receiving formal instruction on specific knowledge (*4*, *5*). Drawing on their prior knowledge, students generate multiple suboptimal solutions and experience initial failure. This problem-solving phase prepares them for subsequent instruction, fostering deeper conceptual understanding (*6*, *7*) and facilitating knowledge transfer (*5*, *6*). However, implementing PF in classrooms presents substantial challenges and risks. The effectiveness of PF depends heavily on the variety and quality of solution attempts generated by individual students during the problem-solving phase (*3*, *8*). This reliance places substantial demands on students' prior knowledge and domain proficiency (*9*), potentially leading to inequalities in learning outcomes. Furthermore, the deliberate lack of support during the problem-solving phase can increase frustration, anxiety, and cognitive load, potentially leading to negative self-assessments and unfavorable perceptions of the learning experience (*10*, *11*).

Extensive research has explored the use of socially assistive robotics in education, where robots provide cognitive or emotional support through social interaction (*12*–*15*), serving as peers, teachers, or evaluators (*14*, *16*–*19*). Compared with robots in teaching roles, robot peers are particularly effective in supporting lower-performing students and are often perceived as more friendly and sociable (*17*, *20*). Most robot peers are designed as knowledgeable companions who guide the learning process (*16*, *21*, *22*), whereas some are programmed as novice learners with abilities comparable to those of students (*16*, *20*, *23*). When functioning as novice learners, robots often adopt a learning-by-teaching model (*20*, *24*), which enhances students' confidence, task commitment, and motivation and fosters learning outcomes in areas such as reading and writing (*24*). Notably, Alemi *et al.* showed that a robot peer who frequently mispronounced words helped reduce students' anxiety in language learning, but the study did not investigate its effect on learning outcomes (*23*, *25*). Nevertheless, whether observing a robot's failure, particularly during problem-solving, can positively influence knowledge acquisition in a manner similar to firsthand failure remains underexplored.

Despite growing expectations that social robots can enhance classroom interactions by asking and answering questions on behalf of students, providing guidance, and fostering competition and collaboration (*26*), empirical evidence of their influence on learning outcomes during regular classroom instruction remains scarce. A few studies situated in school or classroom environments have reported benefits such as increased engagement and motivation (*22*, *27*), stronger peer relationships and interest in scientific topics (*21*), and improved language proficiency (*28*). However, most of these studies were carried out during noninstructional periods (such as during class breaks) (*21*, *28*) or specially designed activities that departed from typical lessons (such as drama-based activities) (*22*, *27*, *29*). Moreover, research in science, technology, engineering, and mathematics (STEM) education, which involves more complex cognitive tasks (*14*, *30*), is particularly rare (*24*, *25*, *30*).

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China. [2]Donald Bren School of Information and Computer Sciences, University of California, Irvine, Irvine, USA. [3]School of Software Technology, Zhejiang University, Ningbo, China. [4]College of Education, Zhejiang University, Hangzhou, China.
*Corresponding author. Email: chenlq@zju.edu.cn (L.C.); sunly@zju.edu.cn (L.S.)
†These authors contributed equally to this work.

To address these challenges, we introduced the NAO social robot into classroom learning, positioning it as a peer to students. Instead of directly engaging in problem-solving and experiencing failure themselves, students observed the robot's problem-solving attempts and failures before receiving formal instruction [referred to as robot failure (RF) here]. In this approach, social robots were expected to demonstrate the proposal of diverse, high-quality suboptimal solutions and the experience of failure, aligning with the core mechanisms of PF theory: activating relevant prior knowledge and enhancing awareness of knowledge gaps (*4*). Additionally, we anticipated that using a robot peer would help shield students from directly experiencing failure, peer pressure, and performance anxiety (*23*).

Here, we conducted two empirical studies in a middle school mathematics classroom to evaluate the effectiveness of the RF method. Study 1 ($N = 135$) compared RF with direct instruction and standard PF (*5*, *31*) in a lesson on standard deviation, examining effects on knowledge acquisition, classroom experience, and underlying mechanisms. Study 2 ($N = 110$) included a 2-week adaptation phase and a lesson on derivatives to determine whether the benefits of RF persisted beyond the novelty effect—the tendency for initial psychological responses to social robots to differ from long-term patterns (*32*, *33*)—and to investigate students' perceptions of the robot's social role, problem-solving performance, and human-robot interaction over time. Our results show that the RF method enhances conceptual knowledge acquisition and knowledge transfer, with effects lasting beyond the initial novelty phase and across different instructional topics. Compared with traditional PF, RF reduces classroom pressure but still leads to lower perceived competence and mental effort. These findings highlight the potential of integrating social robots as peers with evidence-based instructional strategies to improve learning outcomes while minimizing emotional costs, an important factor for sustainable classroom implementation. This study makes two key contributions to educational robotics: It validates the transferability of PF, traditionally centered on interpersonal interactions, to human-robot interaction contexts and provides one of the few empirical studies demonstrating the positive effects of social robots on STEM knowledge acquisition in real classroom settings.

RESULTS

Using robots to perform problem-solving processes

We used four NAO robots as peers to participate in preinstructional problem-solving activities, enabling students to prepare for instruction by observing the robots' problem-solving processes. The robots' interaction procedure for the problem-solving activity was designed on the basis of the PF framework (*31*). Each robot, accompanied by a monitor, was positioned next to a group of six to eight students.

The procedure began with the teacher introducing a mathematical problem related to the intended instructional content (standard deviation in study 1, derivatives in study 2), deliberately set beyond the students' current knowledge. The teacher then invited the robot peers to attempt to solve the problem, asking them to generate as many solutions as possible and present their thought processes to the group of students (*10*).

In response, the robot peers were programmed to present six suboptimal solutions to the problem, tailored to students' prior knowledge (see the Supplementary Materials). These solutions were derived from previous research (*7*, *34*), the pretest, and consultations with middle school teachers to ensure their representativeness. Each solution was presented in 1 to 2 min, following a structured format: The robot introduced the solution's initial intention, explained the calculation steps, presented the results, and reflected on its limitations. This narrative structure aimed to emphasize the key characteristics of the suboptimal solutions (*10*), enhance awareness of knowledge gaps, and foster a focus on deeper patterns (*4*, *31*), which are central mechanisms of PF theory.

The robots used voice and body movements for anthropomorphic demonstrations, with the calculation steps displayed on the monitor screen (see Figs. 1 and 2B and Movie 1; all images in the main text have been edited to display translated English text; the original Chinese versions are available in the Supplementary Materials). In addition to presenting suboptimal solutions, the robots simulated common social behaviors, such as indicating thought (Fig. 2A), reengaging attention (Fig. 2C), expressing confusion and frustration (Fig. 2D), and inspiring thinking (for example, "What other ideas do you have?"). The robots were designed to autonomously organize these behaviors within a predefined narrative structure (Fig. 3B). We predesigned multiple variations for each type of behavior, incorporating subtle differences in speech and actions. This enabled the robot to randomly select from these variations, ensuring the interactions remained dynamic and natural. The experimenter retained the ability to adjust the robot's behavior in real time on the basis of student interactions, ensuring contextually appropriate engagement in the dynamic classroom environment. The robots then spent 15 min presenting the solutions. Afterward, the teacher reviewed and critically evaluated the robots' solutions and then introduced the definitions and calculation methods of the intended instructional content.

Study 1: Field test of PF with robots

Study 1 primarily investigates whether observing the problem-solving process and failures of a robot peer can enhance students' learning outcomes and experiences and whether the underlying mechanisms align with those of PF theory. In the middle of the fall semester, we invited eighth-grade students from six classes at a regular public middle school in Southeast China to participate in a 45-min math lesson on standard deviation, a topic commonly explored in PF research (*5*, *7*, *10*). The six classes were randomly assigned to one of

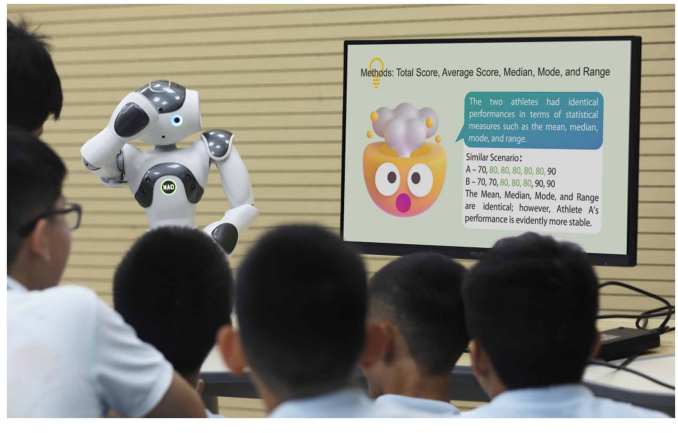

Fig. 1. Students observing the robot peer's problem-solving attempts before instruction. The teacher presented the introductory problem to the class and invited the robot to solve it. The robot presented six predetermined unsuccessful solutions, articulating its thought process, steps, and reflections aloud. Students left their seats to observe the robot peer's problem-solving process.

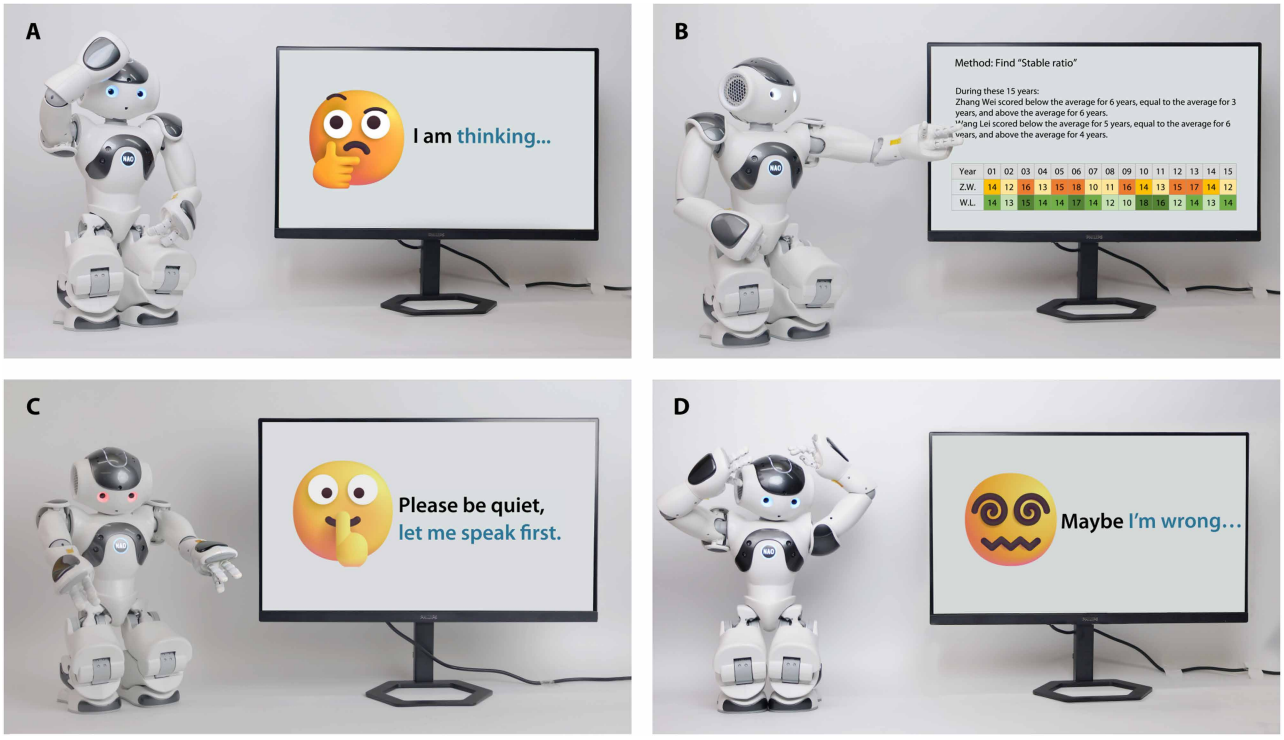

Fig. 2. Robot behaviors in problem-solving and social interactions. (**A**) Thinking, typically before proposing a suboptimal solution. (**B**) Problem-solving steps, with the robot pointing to specific locations on the display screen. (**C**) Mild annoyance and requesting a quiet audience to refocus students' attention. (**D**) Disappointment and confusion, usually when reflecting on the shortcomings of a suboptimal solution.

Movie 1. Overview of the RF method and key findings.

three experimental conditions, with two classes in each condition (Fig. 3A): RF; standard PF; and direct instruction (DI), which served as the control condition. In total, the study collected valid data from 135 students (72 males and 63 females; mean age of 13.6 years, SD of 0.3 years; DI, $n = 44$; PF, $n = 45$; RF, $n = 46$). The gender distribution in these conditions was roughly balanced, and there were no significant age differences across the conditions.

We used a mixed-methods approach to assess students' learning outcomes and emotional experiences. Before the lesson, students completed a pretest measuring prior knowledge, including concepts such as mean, median, mode, and standard deviation (*7*). During the 45-min lesson, students completed questionnaires reporting their engagement, mental effort (*7*, *8*), awareness of knowledge gaps (*11*), perceived competence (*10*), and emotional responses (*35*). At the end of the day, students took a 30-min math quiz to evaluate their knowledge acquisition on standard deviation. This quiz evaluated students' understanding of standard deviation concepts (conceptual knowledge), their ability to perform calculation procedures (procedural knowledge), and their capacity to apply this knowledge in new contexts (knowledge transfer) (*10*). Additionally, students wrote down their subjective impressions of the instructional method on the posttest answer sheet.

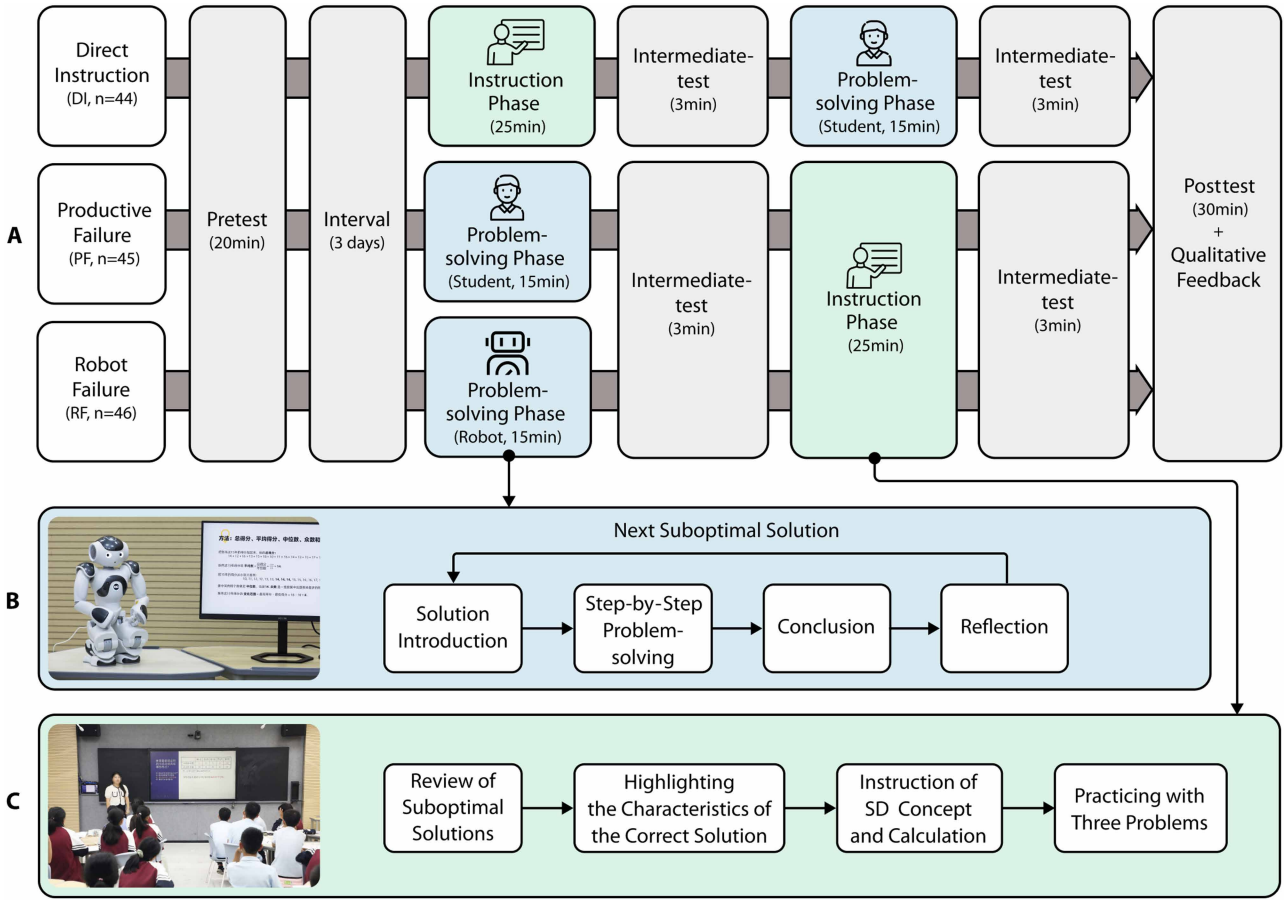

Fig. 3. Experimental procedures of study 1. (**A**) The specific procedures for each condition, consisting of a pretest, problem-solving phase, intermediate test, instruction phase, and posttest. (**B**) The narrative structure used by the robot peer to introduce each suboptimal solution. (**C**) The teacher's explanation steps for standard deviation during the instruction phase.

Study 2: Validation of novelty effect and social perception

Qualitative feedback from study 1 revealed students' mixed attitudes toward the robot peer, prompting further investigation into students' perceptions of the robot's abilities and social roles (for instance, whether it was viewed as a peer or demonstration tool). Additionally, a key question in social robot-assisted instruction is whether the positive effect on learning outcomes persists after the novelty effect subsides and whether this effect can be generalized across various instructional content.

To address these questions, we conducted study 2, a 15-day classroom experiment, 5 months later at the same school. Students from the four classes in the DI and RF conditions from study 1 were reinvited to participate in study 2 and were assigned to their original conditions, with two classes in each condition. In total, study 2 collected valid data from 110 students (61 males and 49 females; mean age of 14.0 years, SD of 0.3 years; DI, $n = 53$; RF, $n = 57$). Compared with study 1 (DI, $n = 44$; RF, $n = 46$), more students from these original DI and RF classes provided complete data. Given that study 2 no longer focused on the mechanisms influencing learning outcomes, the PF condition was not included.

The study began with a 2-week adaptation phase, during which the robot peer participated daily in mathematics lessons and was occasionally called upon by the teacher to answer questions and present problem-solving steps (Fig. 4). This phase allowed students to become familiar with the robot's abilities and social presence, gradually diminishing the novelty effect. After the adaptation phase, the instructional experiment was conducted, with content focused on derivatives. The instructional procedures for the RF and DI conditions were identical to those in study 1. Pre- and posttests on derivatives were administered 3 school days before and on the evening of the instructional experiment, respectively. After the instructional experiment, students completed questionnaires on their perceptions of the robot's social roles, robot-peer likeness, intelligence, and anthropomorphism. Additionally, written qualitative feedback was collected to capture students' attitudes toward the robot and how their perceptions evolved over time.

Analysis of study 1

Knowledge acquisition

We conducted an analysis of covariance (ANCOVA) with prior knowledge, engagement (*6*–*8*), and students' birth month, common individual difference factors that might influence students' learning outcomes, as covariates to examine the effects of experimental conditions on students' knowledge acquisition (Fig. 5, A to C). A significant effect of experimental condition was found for both conceptual knowledge ($F_{2,129} = 3.447$, $P = 0.035$, effect size partial $\eta^2 = 0.051$) and

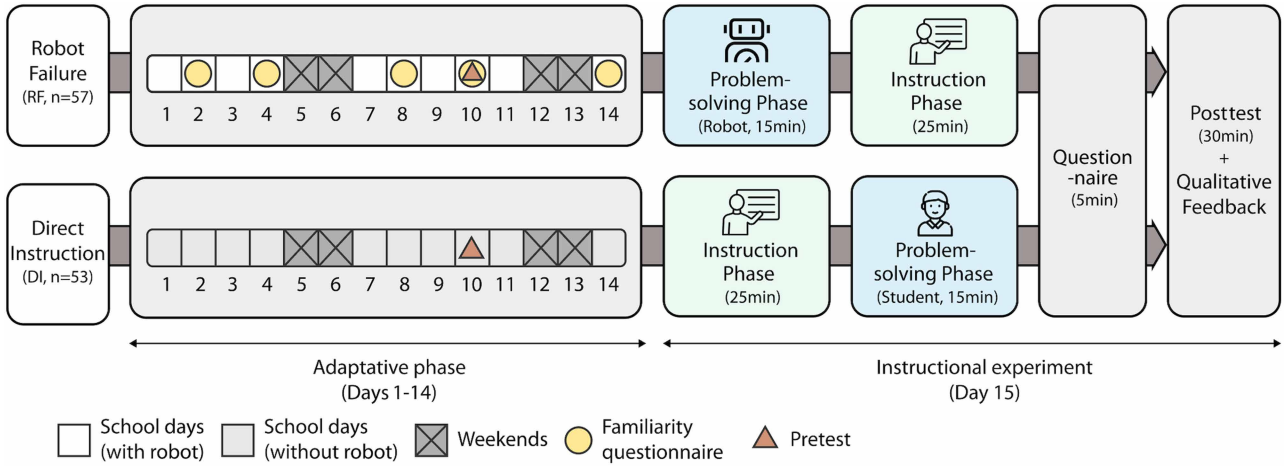

Fig. 4. Experimental procedure of study 2. The procedure included a 2-week adaptation phase, a pretest, and an instructional experiment on day 15. The instructional experiment comprised a problem-solving phase, an instruction phase, postlesson questionnaires, and a posttest.

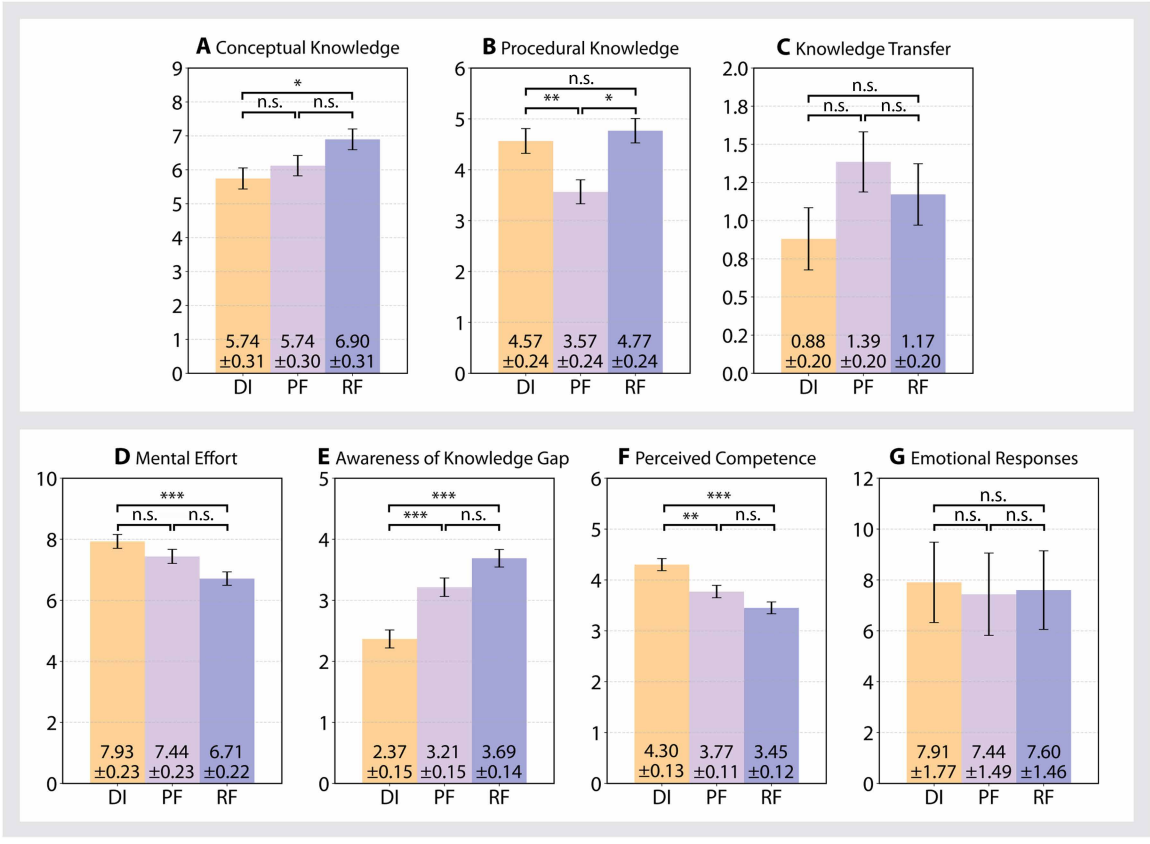

Fig. 5. Quantitative results of study 1 (*N* = 135). Students' postlesson knowledge acquisition was divided into three components: (**A**) conceptual knowledge, (**B**) procedural knowledge, and (**C**) knowledge transfer. Students' self-reported learning experiences included (**D**) mental effort, (**E**) awareness of knowledge gaps, (**F**) perceived competence, and (**G**) emotional responses toward their own experiences (in PF and DI) or the robot peer (in RF) after the problem-solving phase. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; n.s., not significant. Bar heights represent EMMs from ANCOVA or MANOVA, with error bars showing model-derived SEs based on valid participant data after exclusion (DI, $n = 44$; PF, $n = 45$; RF, $n = 46$).

procedural knowledge ($F_{2,129} = 7.471$, $P = 0.001$, partial $\eta^2 = 0.104$), but not for knowledge transfer. Students in the RF condition demonstrated significantly greater conceptual knowledge than those in the DI condition ($P = 0.034$). For procedural knowledge, students in the PF condition scored significantly lower than those in both the RF ($P = 0.002$) and DI conditions ($P = 0.011$).

Among the covariates, only prior knowledge showed a significant effect on conceptual knowledge ($F_{2,129} = 19.422$, $P < 0.001$, partial $\eta^2 = 0.131$). No other covariates or the interaction between covariates and experimental conditions showed significant effects on the three knowledge acquisition variables.

Classroom performance and experiences

A multivariate analysis of variance (MANOVA) showed a significant difference in mental effort among the three instructional conditions ($F_{2,126} = 7.568$, $P = 0.001$, partial $\eta^2 = 0.107$). Students in the RF condition reported significantly lower mental effort than those in the DI condition ($P < 0.001$; Fig. 5D). Instructional condition also significantly affected students' awareness of knowledge gaps, a key mechanism in PF theory ($F_{2,126} = 21.098$, $P < 0.001$, partial $\eta^2 = 0.251$), with both the RF and PF conditions leading to significantly higher awareness than the DI condition ($P < 0.001$; Fig. 5E). A similar pattern was observed for perceived competence ($F_{2,126} = 13.339$, $P < 0.001$, partial $\eta^2 = 0.175$), with students in the RF and PF conditions reporting lower perceived competence than those in the DI condition (RF versus DI, $P < 0.001$; PF versus DI, $P = 0.007$; Fig. 5F). No significant differences were found in students' emotional responses toward themselves or the robot peer after the problem-solving phase (Fig. 5G). We also recorded the number of suboptimal solutions generated by students in the PF condition during the problem-solving phase, with an average of 1.32 suboptimal solutions per student (SD = 0.83).

Qualitative feedback

Students' qualitative feedback in study 1 is organized into six themes (for example, theme S1-3 refers to the third theme identified in the study 1 feedback). Themes S1-1, S1-2, and S1-3 (Fig. 6A) highlight differences in experiences shared across the three instructional conditions. Theme S1-4 (Fig. 6B) reflects students' preferences for the different instructional methods. Themes S1-5 and S1-6 focus on students' perceptions of the robot peers, which are explored further in study 2. Participant quotes are identified by group designation and within-group ID (for example, RF-7 or DI-12). Further details on the codebook and frequency of each code are provided in tables S2 and S4.

Classroom engagement and interaction (theme S1-1). Across all conditions, students reported on their active engagement in the classroom, with approximately one-third explicitly reflecting on their participation. Feelings of pressure and anxiety, often stemming from

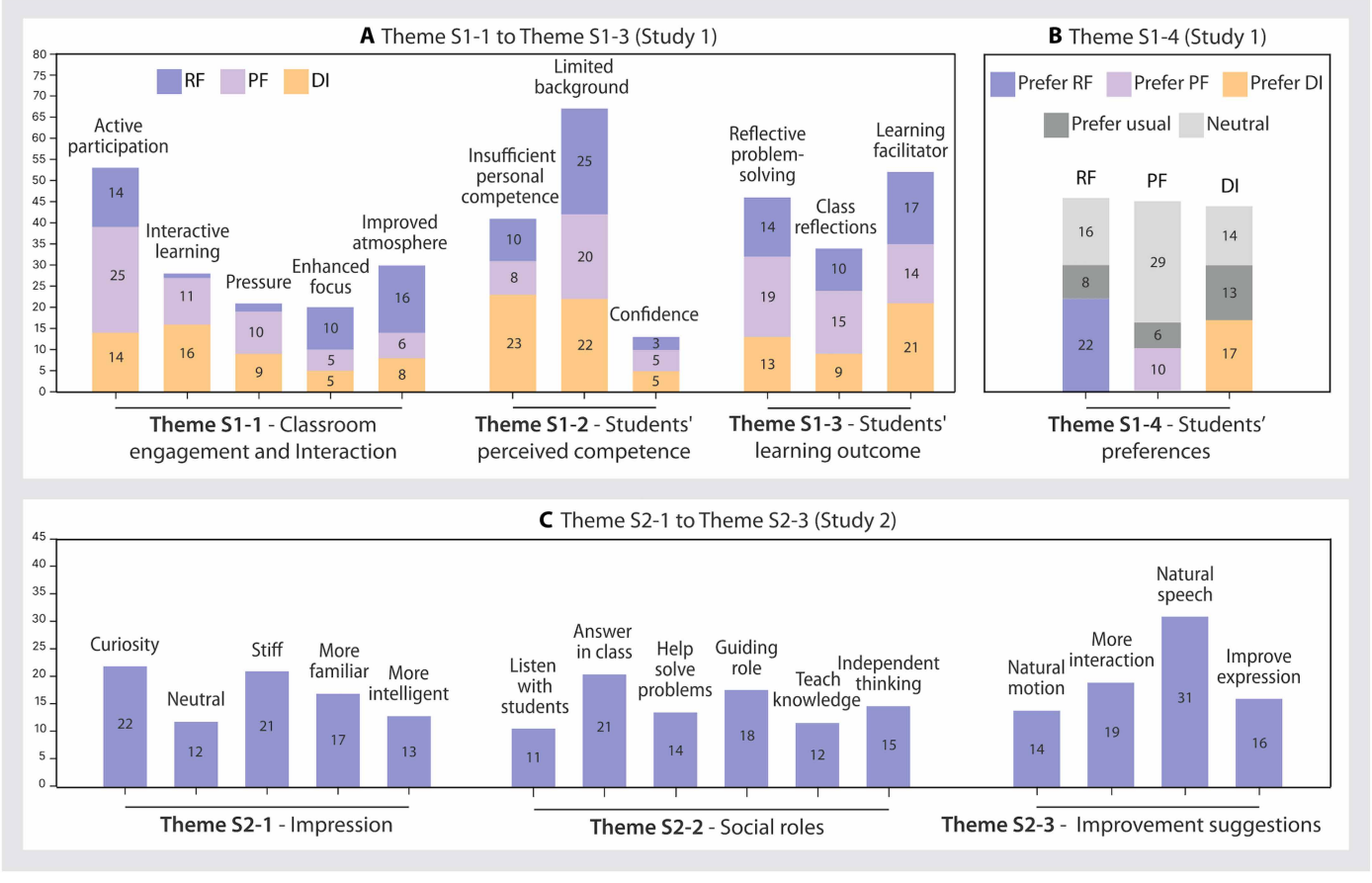

Fig. 6. Thematic analysis of qualitative feedback from study 1 and study 2, organized by themes and codes. (**A**) Three of the six themes from study 1 (themes S1-1, S1-2, and S1-3), comparing students' classroom experiences across different instructional methods. (**B**) Student preferences for instructional methods (theme S1-4). (**C**) Three themes from study 2 (themes S2-1, S2-2, and S2-3), describing students' attitudes toward the robot. The numbers in the figure represent the number of participants who mentioned each code under the respective experimental condition.

peer comparison or self-evaluation, were more frequently reported in the PF condition by 10 students. For instance, PF-18 wrote, "Because everyone else already had ideas and were writing, I didn't know how to do it and felt a little awkward." Similar concerns were mentioned by only two students in the RF condition. In contrast, 16 students in the RF condition noted that the robot helped reduce stress and tension, with many expressing relief at not having to solve the problem themselves.

Self-evaluation and learning outcomes (themes S1-2 and S1-3). In each condition, 13 to 19 students reflected on their problem-solving approaches. For example, PF-10 noted, "My method is more convenient compared to the method in class, but it's not accurate." In the RF condition, 17 students mentioned that the robot's solutions inspired their thinking: "His thinking indeed went from shallow to deep, and it was very enlightening" (RF-16). Additionally, students reflected on their limitations, such as challenges with mathematical skills (10 students in RF and 8 in PF) and gaps in background knowledge (25 in RF and 20 in PF).

Students also compared the experimental instructional methods with their usual classroom experiences (theme S1-4). In the RF condition, 22 students preferred the experimental method over their usual activities (eight students). In the PF and DI conditions, the ratios were 10 to 6 and 17 to 13, respectively. Students in the RF condition commented, "(Prefer) having a robot involved in the teaching because it is more interesting" (RF-03), and "My thinking keeps updating and evolving along with the robot's… (This) helps me realize my knowledge gaps in time, making the classroom rhythm very 'pleasant' (for me)" (RF-14).

In addition to the above themes, students expressed curiosity, novelty, and mixed attitudes toward the robot's capabilities (themes S1-5 and S1-6). Although 29 students felt the robot performed well, 13 others noted that the robot's problem-solving approach was simplistic and limited, expressing impatience and frustration with its repeated failures and self-critical responses. For instance, RF-08 commented, "The robot classmate is a bit silly. It doesn't need to think about the content that's obviously unreasonable." These observations prompted further exploration of students' perceptions of the robot peers, which became the focus of study 2.

Analysis of study 2

Perception changes regarding the robot during the adaptation phase

During the 2-week adaptation phase, students' familiarity with the robot peer initially increased and then gradually stabilized (Fig. 7D) ($F_{4,314} = 4.841$, $P = 0.001$, partial $\eta^2 = 0.058$). Post hoc tests revealed that only the familiarity score from the first measurement (day 2) differed significantly from subsequent measurements. Additionally, teachers reported that between day 2 and day 4, students from various classes began initiating more diverse interactions with the robot, such as requesting assistance when answering questions. These observations suggest, both directly and indirectly, that students quickly became familiar with the robot's capabilities.

The students' qualitative feedback reflected changes in their perceptions of the robot (theme S2-1 in Fig. 6C). Twenty-two students reported an initial impression of curiosity or interest, whereas 12 students expressed unfamiliarity or indifference. After the adaptation phase, students noted positive shifts in their perceptions of the robot's familiarity and intelligence, with 17 and 13 students mentioning these aspects, respectively. RF-20 explained that "As time went by, its appearance made me more familiar with it, and I already got used to its presence." RF-06 remarked that "As the class went deeper, I found that its interaction ability was way better than I expected—it could also explain questions in real time based on what students and teachers asked."

Knowledge acquisition

An ANCOVA controlling for prior knowledge, classroom engagement, and students' birth month revealed that students in the RF condition scored significantly higher than those in the DI condition on both conceptual knowledge ($F_{1,105} = 4.329$, $P = 0.040$, partial $\eta^2 = 0.040$) and knowledge transfer ($F_{1,105} = 19.421$, $P < 0.001$, partial $\eta^2 = 0.156$) in the posttest. No significant difference was found between the conditions for procedural knowledge (Fig. 7, A to C).

As a covariate, prior knowledge had a significant positive effect on conceptual knowledge ($F_{1,105} = 31.033$, $P < 0.001$, partial $\eta^2 = 0.228$), procedural knowledge ($F_{1,105} = 27.509$, $P < 0.001$, partial $\eta^2 = 0.208$), and knowledge transfer ($F_{1,105} = 7.325$, $P = 0.008$, partial $\eta^2 = 0.065$). These results indicate that students with stronger prior knowledge generally achieved better learning outcomes. Neither engagement nor birth month significantly affected these knowledge acquisition variables. Consistent with study 1, no significant interactions were found between experimental condition and any of the covariates, including prior knowledge, birth month, or engagement, on any of the three knowledge acquisition variables.

Students' perceptions of the robot's social roles and performance

We used a paired comparison method (*36*) to ask students which social role they viewed the robot as in the classroom—teacher, peer or classmate, demonstration tool, or care receiver (Fig. 7E) (*16*, *37*). The results confirmed that students predominantly perceived the robot as a peer or classmate, although some still regarded it as a more dynamic demonstration tool. Furthermore, students gave positive evaluations of whether the robot's behavior resembled that of their peers, significantly surpassing the midpoint ($t = 4.482$, $P < 0.001$, Cohen's $d = 0.594$) (Fig. 7F). Specifically, 25 of the 57 students agreed that the robot resembled their classmate, and 37 students believed that the robot's responses were generated autonomously.

Students rated the robot as highly intelligent (Fig. 7F), with perceived intelligence scores significantly exceeding the scale's midpoint ($t = 7.920$, $P < 0.001$, Cohen's $d = 1.040$). In contrast, students' perceptions of the robot's anthropomorphism were largely neutral (Fig. 7F; $t = -0.212$, $P = 0.833$).

In the qualitative feedback, half of the students (29 of 57) expressed a tendency to perceive the robot as a peer or classmate, primarily attributing this to its student-like behavior (theme S2-2 in Fig. 6C). RF-32 said, "Once or twice, I saw that after the teacher asked a question, it even raised its hand to show it wanted to answer, just like us who were listening carefully in class"; RF-09 said, "It was like a learner just like us, having a thinking process similar to ours." Students also shared their expectations for improvement, including the robot's physical movement, interaction abilities, language naturalness, and preference for slower-paced communication to allow more time for reflection (theme S2-3).

DISCUSSION

Preparatory effects of RF for instruction

Our study found that the RF condition significantly enhanced students' knowledge acquisition compared with the DI condition. In

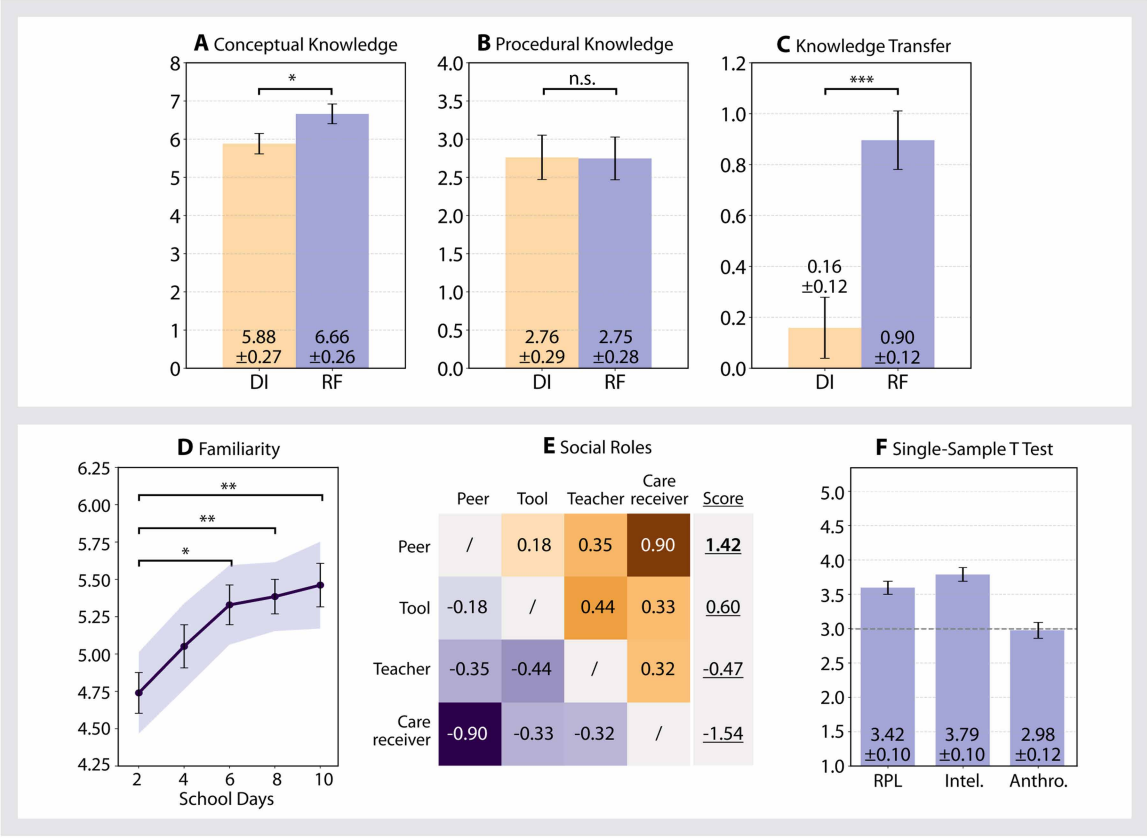

Fig. 7. Quantitative results of study 2 (*N* = 110). Students' knowledge acquisition in (**A**) conceptual knowledge, (**B**) procedural knowledge, and (**C**) knowledge transfer ability. (**D**) The change in students' familiarity with the robot peer throughout the adaptation phase. Additionally, students reported their perceptions of (**E**) the robot's social roles within the classroom context and evaluations of the robot's (**F**) robot-peer likeness (RPL), intelligence (Intel.), and anthropomorphism (Anthro.). $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; n.s., not significant. Bar heights represent EMMs from ANCOVA or means from *t* tests, with error bars showing model-derived SEs (DI, $n = 53$; RF, $n = 57$; after exclusions).

study 1, students in the RF condition showed significant gains in conceptual knowledge, whereas study 2 revealed improvements in both conceptual knowledge and knowledge transfer. These findings are consistent with previous PF research, which similarly reported significant gains in conceptual knowledge, occasional improvements in knowledge transfer, and minimal changes in procedural knowledge (*5*, *10*, *38*). In study 2, the positive effect of the RF method on knowledge acquisition was sustained after students' familiarity with the robot peer increased and stabilized during the adaptation phase. This further supports the idea that the positive preparatory effects of the RF method may not be attributable to the social robot's novelty effect but rather to its potential long-term effectiveness. Furthermore, studies 1 and 2 collectively demonstrate the generalizability of the RF method to different instructional content, highlighting its potential for broader application in classroom settings.

Key factors in PF theory, such as awareness of knowledge gaps, engagement, and prior knowledge, were assessed to determine whether the mechanisms underlying the RF method's preparatory effect align with the PF framework (*6*, *7*, *10*, *11*). In study 1, both the PF and RF conditions increased students' awareness of knowledge gaps, a key mechanism through which PF helps students identify the deeper features of the correct problem-solving approach, ultimately fostering positive learning outcomes (*4*, *31*). Consistent with previous PF literature, both study 1 and study 2 found no interaction effect between instructional methods and prior knowledge on knowledge acquisition, aside from the main effect of prior knowledge itself (*6*). This supports the idea that, similar to traditional PF methods, our approach produces positive effects across students with varying levels of prior knowledge (*4*, *6*). This finding also aligns with another key mechanism in PF theory: the activation of prior knowledge through preinstructional problem-solving (*4*). These results suggest that the mechanisms driving knowledge acquisition in both the PF and RF conditions may not differ substantially.

In study 1, we observed a stronger preparatory effect in the RF condition compared with PF. This advantage may be attributed to the robot peer's superior problem-solving and presentation performance compared with that of human students. Previous research has shown that both the quantity and quality of suboptimal solutions are key factors in determining whether PF can produce positive effects (*3*, *8*). In study 1, PF students generated an average of 1.32 solutions, notably fewer than the three or four solutions typically reported in prior PF studies (*8*, *34*) and the six solutions presented in the RF condition, possibly explaining the weaker learning outcomes in the PF condition despite following a similar trend (*7*, *34*). By providing a broader range of strategies and a more structured presentation (*39*), the RF method more effectively activates prior knowledge and enhances students' understanding of the key features of correct solutions, particularly benefiting low-performing

students and time-constrained instructional scenarios. Further support, although indirect, comes from study 2, where the RF method produced even greater improvements in knowledge transfer when addressing more advanced and abstract concepts such as derivatives, by preventing students from stagnating during problem-solving.

Another explanation for RF's advantage lies in its observational perspective, allowing students to focus on critically evaluating solutions without spending time recovering from pressure or confusion (theme S1-1). This advantage also relates to vicarious failure, an instructional method within the PF framework (*11*), where students observe their peers' unsuccessful problem-solving attempts. Although vicarious failure typically supports conceptual knowledge acquisition, it is generally less effective than PF (*8*) because students often struggle to fully grasp peers' cognitive processes, such as intentions, conclusions, and reflections (*10*, *11*), aspects a robot peer effectively conveys.

Students' perceptions of the RF method and the robot peer

One motivation for our study was to use robots to replace human students in preinstructional problem-solving, thereby reducing students' performance anxiety and social pressure (*23*). In study 1, students in the RF condition reported experiencing less pressure and fewer negative emotions (theme S1-1). In the PF condition, social comparisons often arose from classmates' problem-solving discussions and achievements, leading to peer pressure. These comparisons were largely reduced when the problem-solver was a robot, allowing students to easily separate the robot's failure from their own abilities or talents. However, students in the RF condition still reported decreases in perceived competence and mental effort, like those in the PF condition. The decrease in perceived competence was closely related to the students' increased awareness of knowledge gaps, whereas the reduction in mental effort was largely due to the robot taking over the problem-solving process, thus lessening the students' cognitive effort and struggles.

Another issue is whether students perceive the robot as a true peer. Study 1 found no significant differences in emotional responses between the PF and RF conditions, which may indicate some level of human-robot empathy. In study 2, according to the robot-peer likeness questionnaire and paired comparison of social roles, students were more likely to perceive the robot as a peer or classmate rather than as a demonstration tool or teacher. This perception was primarily shaped by the robot's interactions with the teacher and students (theme S2-1). Qualitative feedback revealed that interactions such as "being called on by the teacher to answer questions like us" and "sitting with us" in the teacher-student-robot dynamic, along with the robot's hesitation and overreflection during problem-solving, led students to view the robot as an equal classmate. Students' positive perception of the robot-peer likeness did not coincide with a positive evaluation of its anthropomorphism. The neutral perception of anthropomorphism may be attributed to our use of text-to-speech–generated voice materials, rather than the prerecorded human voices commonly used in previous studies (*40*) (theme S2-3). This restrained approach to anthropomorphism aimed to prevent distractions from the robot's social elements, which could negatively influence learning outcomes (*17*), and to avoid students perceiving the robot as overly preprogrammed, potentially viewing it as a teacher or authority figure (*41*).

We observed an evolution in students' perceptions of the robot peer over time, with the most notable change occurring in their evaluation of the robot's intelligence. In study 1, students often criticized the robot for repeatedly presenting incorrect solutions, even labeling it as unintelligent or "stupid." These negative judgments may reflect students' perceptions of their human peers' mathematical performances and the social pressure such judgments may create, given that the robot's performance matched or even exceeded that of real students. Unlike with human peers, students appeared to feel no obligation to suppress such negative comments or offer social-emotional support to the robot, as noted in previous studies (*20*). From a PF perspective, this critical viewpoint and the freedom of expression could potentially promote critical thinking. However, a concern remains regarding whether such negative judgments would persist and extend to human peers. Throughout study 2, students' perceptions of the robot's abilities shifted positively. They rated its intelligence more favorably, with 13 students noting that they gradually recognized its intelligence (theme S2-1). Although students continued to point out the robot's mistakes when it failed, this feedback was generally nonaggressive. This suggests that, after the adaptation phase, students recognized the robot's problem-solving abilities and restored a more friendly human-robot social interaction.

Reflections on design and implementation of educational social robots

Although our study makes an important contribution to the application of robots in classroom education, several limitations should be noted. Social robots in educational settings can produce novelty effects that last from hours to months (*28*, *42*, *43*), temporarily boosting engagement and learning outcomes (*32*, *33*). To evaluate whether the positive effects of the RF method persist after the novelty effect has subsided, we implemented a 2-week adaptation phase, with one 45-min session per school day, consistent with prior social robot studies and some long-term research (*28*, *44*). Although students' self-reported familiarity, qualitative feedback, and behavior generally suggest that the novelty effect has substantially diminished (*45*), certain measures, such as eye-tracking and interaction frequency (*45*, *46*), remain difficult to collect or lack sufficient statistical value in large-scale classroom settings.

In our study, students were seated in random groups to facilitate classroom management, in line with previous PF research (*2*, *6*, *8*, *47*). Given that student discussions were not prohibited (mainly in the PF condition) (*2*, *6*), there may have been a subtle correlation between students' performance and their seating groups. However, because of the voluntary participation principle, fewer than two-thirds of the students in each group provided complete data for the final analysis, limiting our ability to model the effect of seating group through hierarchical or nested analysis. Additionally, previous PF research suggests that seating groups and related discussions typically have minimal effect on learning outcomes, with individual differences in prior knowledge being far more influential (*2*). Thus, we treated all students as independent data points when comparing instructional methods.

Autonomy is another crucial aspect of applying robots in classroom education (*48*). In the RF condition, our robot was designed to autonomously perform preset actions within a predefined framework (Fig. 3B). Researchers supervised the robot and reordered the interactions when student feedback conflicted with the robot's planned sequence. This represents a medium level of autonomy (*49*) and led most students to believe the robot "came up with the solutions on its own," according to the robot-peer likeness questionnaire. However,

a greater challenge is whether robots should have autonomy over teaching content (*50*), given that suboptimal solutions in the RF method must be carefully designed to fit middle school teaching scaffolding and avoid introducing unintended alternatives (*31*). For example, teachers typically avoid introducing mean absolute deviation when teaching standard deviation, because justifying the preference for standard deviation exceeds the scope of middle school knowledge. Highly autonomous robots may struggle to adapt to the varying prior knowledge levels of middle school students, potentially diverting the students from the intended instructional progression. Therefore, educators should remain closely involved in the design of robot- or artificial intelligence (AI)–generated materials to ensure proper control over the learning process.

The preparation workload for teachers is a key consideration in the practical implementation of the RF method. Teachers can easily compile the suboptimal solutions for the robot on the basis of mistakes made by previous same-grade students. They then need to prepare the robot's 10- to 15-min problem-solving slides and text-to-speech–generated audio files. This process is efficient when existing course materials are available. No additional programming of the robot is required. Instead, teachers only need to arrange preset actions (Fig. 2) by creating a timestamped file specifying when to switch slides and trigger specific actions. Once teachers are familiar with the robot's interaction program, preparation can be completed within 1 to 2 days. Although the workload may still be demanding for an individual teacher, it would be manageable for a middle school teaching team.

Despite these limitations, our study offers strong evidence that, when integrated with appropriate instructional strategies, social robots can provide sustainable and broadly effective benefits for both learning outcomes and classroom experiences. By implementing the RF method in real-world middle school lessons, we demonstrate the potential of social robots to serve as effective classroom peers and to assume roles beyond those of a conventional companion, mentor, or role model. Future research could further investigate adaptive and autonomous capabilities, enabling previously unidentified forms of interaction that might otherwise be constrained by social norms or traditional classroom structures. Looking ahead, we envision educational social robots not merely as novel media for content delivery but as socially and behaviorally capable partners who engage in shared experiences of challenge, reflection, encouragement, and growth alongside students and teachers.

MATERIALS AND METHODS

Objectives and experiment design of study 1

Study 1 explored whether observing a social robot's unsuccessful problem-solving process before instruction positively influenced students' mathematics knowledge acquisition. It compared this approach with actively engaging in problem-solving and experiencing failure, focusing on both the effects and underlying mechanisms. Additionally, the study examined how this method influenced students' classroom experiences, such as pressure and perceived competence.

Eighth-grade students from six classes at a public middle school in Southeast China participated in a mathematics lesson on the concept of standard deviation. The six classes were randomly assigned to three instructional conditions, with two classes in each condition. In the PF condition, students first attempted to solve an introductory problem related to data stability (problem-solving phase) before receiving teacher instruction on standard deviation (instruction phase) (*31*). In the RF condition, students observed a robot peer attempting to solve the problem before the instruction phase. In the DI condition, the control group, students received instruction on standard deviation before attempting the problem (*6*, *7*). The pretest, intermediate questionnaires, and posttest were administered before, during, and after the lesson to collect data on students' prior knowledge, classroom experience, and learning outcomes (see the Supplementary Materials) (*7*, *8*, *10*).

Cases were excluded on the basis of the following criteria: prior knowledge of the instructional content, as indicated by the pretest, and submission of blank or irregular responses, such as selecting the same option for all items, which suggested disengagement from the surveys and quizzes. These exclusion criteria were applied consistently in study 2. Further details on case exclusions are provided in the Supplementary Materials. The final sample ($N = 135$) provided sufficient power to detect a medium effect size ($f = 0.27$, $1 - \beta = 0.8$, as estimated using G*Power).

Problem-solving phase

During the 15-min problem-solving phase, students were tasked with a classic introductory problem related to standard deviation: determining which of two athletes had a more stable performance (*7*). In the PF condition, the teacher encouraged students to propose as many reasonable solutions as possible to address the problem (Fig. 8C) (*7*, *10*). Given that students had not yet received instruction on standard deviation, this problem was likely to lead to experiences of failure during problem-solving. Students were allowed to discuss their ideas with each other (*7*, *31*). At the same time, the teacher moved around the classroom, periodically announcing typical solutions proposed by students or recording them on the blackboard to inspire further ideas.

In the RF condition, a group of six to eight students closely observed a robot peer's problem-solving process (Fig. 8, A and B). The teacher pretended to read the same problem-solving requirements to the robot. The robot paused as if thinking, then proceeded to outline six classic suboptimal solutions to standard deviation (see the Supplementary Materials). After each solution, the robot paused for reflection. If a student attempted to interject during the explanation, then the robot expressed a preference to finish its explanation first. When a student suggested an alternative solution substantially different from the robot's planned sequence, the researcher intervened backstage to adjust the order of the robot's solutions, ensuring contextual alignment. Last, the robot asked students if they had better solutions before returning control of the class to the teacher. In the DI condition (Fig. 8D), problem-solving occurred after teacher instruction on standard deviation, with the same problem presented as a postinstruction exercise (*7*).

Instruction phase

During the 25-min instruction phase, the teacher formally introduced the concept and calculation of standard deviation (Fig. 3C). In line with previous literature (*7*, *31*), this process included several key steps: First, the teacher reviewed the limitations of suboptimal solutions and summarized the characteristics of a correct method for describing data stability. Next, the teacher introduced the concept of standard deviation and provided a step-by-step guide for computing it (*7*). Last, the teacher explained the problem-solving process for standard deviation using two practice problems and invited students to solve a third problem independently.

In the DI condition, because students had not engaged in prior problem-solving, the lesson began with a simplified situational

Fig. 8. Student participation in classroom activities under different conditions. (**A**) The positions of the robot, monitor screen, students, and researcher in the RF condition. (**B**) In the RF condition, a group of six to eight students gathered around the robot peer to observe its problem-solving process. (**C**) In the PF condition, students were asked to solve the introduced problem independently but were allowed to communicate. (**D**) In the DI condition, the teacher directly introduced the concept of standard deviation and problem-solving steps to the students.

introduction. The teacher quickly used students' familiar mathematical concepts (mean and mode) to address the introductory problem, guiding students to recognize that their prior knowledge was insufficient to resolve the issue of data stability.

Objectives and experiment design of study 2

Study 2 was conducted 5 months after study 1 at the same middle school. It had two main objectives: to assess whether the RF method's positive effect on knowledge acquisition persisted after the novelty effect had subsided and to explore students' perceptions of the robot's social role, perceived intelligence, and anthropomorphism. The same four classes from study 1's RF and DI conditions were invited to participate. The study consisted of a 2-week adaptation phase, followed by an instructional experiment on day 15. The final sample ($N = 110$) provided sufficient power to detect a medium effect size ($d = 0.54$, $1 - \beta = 0.8$).

Adaptation phase

During the 2-week adaptation phase (10 school days), the robot peer participated daily in 45-min mathematics lessons with students in the RF condition. The teaching content followed the school's regular curriculum, covering topics such as quadratic functions, parallelograms, and symmetry.

Throughout the lessons, the teacher occasionally selected the robot peer to answer questions during roll call, with an average of three instances per lesson. This ensured that each student in the RF condition ideally experienced 30 interactions with the robot peer. The teacher informed the research team in advance of the questions to be asked, allowing the researchers to prepare the robot's responses accordingly. These responses occasionally included incorrect answers or indications of uncertainty, which accounted for approximately one-quarter of all responses. This approach helped students adapt to the possibility of the robot providing failed solutions. We specifically instructed the teacher not to ask the robot to perform tasks outside the scope of a peer's role, such as summarizing answers from various students.

Following principles summarized in previous research (*42*, *45*), we collected multiple types of evidence suggesting a decline in the novelty effect over the course of the study. Students completed a familiarity questionnaire every other day after the robot-involved lessons. In addition, teachers were asked to observe when students began initiating interactions beyond the designed scripts, indicating habituation to the robot's functions (*45*). Qualitative feedback collected after the instructional experiment further reflected changes in students' perceptions of the robot (*45*, *51*).

Instructional experiment

On day 15, students in both the RF and DI conditions participated in the instructional experiment, which focused on derivatives. The structure of the experiment mirrored that of study 1 (see Fig. 4),

with the introductory problem and suboptimal solutions provided in the Supplementary Materials. Students' prior knowledge of derivatives was assessed through a 20-min pretest conducted 3 days before the instructional experiment. Questionnaires regarding the robot's social role, robot-peer likeness, perceived intelligence (*52*), and anthropomorphism (*53*) were administered immediately after the experiment. A 30-min posttest was given later that evening, along with written qualitative feedback.

Data collection

All data collection procedures were approved by the Institutional Review Board at Zhejiang University [approval no. (2024)095] and the administrative committee of the participating middle school. To maintain a single-blind design, only the school's administrative committee was informed in advance about the experimental conditions and objectives. Before the experiment, teachers informed their classes that students were invited to participate in a series of lessons aimed at evaluating different instructional methods, emphasizing that participation in the lessons and related surveys was entirely voluntary. Informed consent was obtained from all participating students and their parents for all cases in this study. For study 1 and the instructional experiment in study 2, teachers reassured students that the involved instructional content would be formally covered later in the academic year, ensuring that participation would not affect their academic progress.

In study 1, 227 students from six classes were invited to participate. Of these, 144 students completed all experimental procedures, and 135 valid samples were retained after excluding participants who did not meet the criteria (described in the "Objectives and experiment design of study 1" section). In study 2, 150 students from the same RF and DI classes as in study 1 were initially invited, and 124 students completed all experimental procedures, resulting in 110 valid samples. Among these 110 students, 74 had fully participated in study 1.

In both study 1 and study 2, we assessed students' prior knowledge and knowledge acquisition using pretests and posttests. The 20-min pretest was administered during an after-school study session 3 days before the experiment. In study 1, the pretest included questions on means, medians, modes, and standard deviation. In study 2, the pretest covered equations, linear functions, quadratic functions, and derivatives. On the evening after the experiment, students completed the posttest independently under supervision, within a 30-min time limit. The posttest assessed three dimensions of knowledge acquisition: procedural knowledge, conceptual knowledge, and knowledge transfer (*6*, *8*). The test content and scoring criteria were developed on the basis of prior research and consultations with the middle school teaching team.

The scales measuring classroom engagement, mental effort (*7*, *8*), awareness of knowledge gaps (*11*), perceived competence (*10*), and emotional responses (*35*) in study 1 and those measuring perceived intelligence (*52*) and anthropomorphism (*53*) in study 2 were adapted from previously validated measures. Additionally, we designed scales to evaluate students' familiarity with the robot and robot-peer likeness and used paired comparisons (*36*) to assess the robot's social role in study 2. All scales used Likert-type items with response options ranging from five to nine points.

We collected qualitative feedback by asking students to write ~200 words on their subjective impressions of specific topics. Study 1 focused on students' evaluations and comparisons of their classroom experiences under different instructional methods. Study 2 explored students' perceptions of the robot peer's abilities and social attributes. All quiz papers, scales, qualitative feedback surveys, and related materials mentioned in this section are provided in the Supplementary Materials.

Statistical analysis

All pretest and posttest answer sheets from studies 1 and 2 were independently scored by two researchers on the basis of predefined criteria developed from previous studies and consultations with the school's mathematics teachers (see the Supplementary Materials). The final score was determined by averaging the two researchers' scores. The scoring demonstrated high interrater reliability, with intraclass correlation coefficients of 0.972 for study 1 and 0.978 for study 2.

Students' qualitative feedback was recorded and anonymized by researchers in a digital spreadsheet. The coding process followed a constructivist grounded theory approach (*54*). Two researchers independently conducted an initial round of open coding to describe the content of each reflection. Given the large sample size, random sampling was used to select 10 reflections from each experimental condition (*55*). Once saturation was reached (typically seven to nine reflections per group in both studies), the research team collaboratively reviewed and refined the codes, ultimately producing a codebook with several themes and codes. This codebook was then applied to the remaining students' reflections. Detailed descriptions of the themes and codebook are provided in the Supplementary Materials. Interrater reliabilities in both studies exceeded the reliability threshold of $\kappa > 0.80$ ($\kappa = 0.831$ for study 1 and $\kappa = 0.827$ for study 2).

Each scale score was calculated as the average of its item scores. The scales measuring engagement ($\alpha = 0.932$), awareness of knowledge gaps ($\alpha = 0.722$), perceived competence ($\alpha = 0.876$), familiarity ($\alpha = 0.876$), robotic-peer likeness ($\alpha = 0.724$), perceived intelligence ($\alpha = 0.815$), and anthropomorphism ($\alpha = 0.812$) all demonstrated acceptable reliability (Cronbach's $\alpha > 0.7$). Internal consistency was not calculated for the mental effort, emotional response, and paired comparison questions regarding the robot's social roles, because these scales do not consist of multiple items measuring the same construct.

To examine the effects of experimental conditions on knowledge acquisition, we conducted an ANCOVA, controlling for prior knowledge, engagement, and students' birth month as covariates. Other outcome variables were analyzed using MANOVA or *t* tests. Reported means, estimated marginal means (EMMs), SDs, and model-derived standard errors (SEs) were calculated for each experimental condition after excluding data on the basis of predefined criteria. In addition to the model-based estimates from ANCOVA and MANOVA, descriptive statistics calculated directly from the raw data are provided in tables S1 and S3. All statistical analyses were performed at a significance level of 0.05, using two-tailed tests, with Bonferroni correction applied for post hoc analyses. All assumptions required for these analyses were met. Any missing values identified during specific analyses were temporarily excluded.

Supplementary Materials

The PDF file includes:

Methods
Figs. S1 to S4
Tables S1 to S4

Other Supplementary Material for this manuscript includes the following:
Data files S1 to S4
MDAR Reproducibility Checklist

REFERENCES AND NOTES

1. K. VanLehn, S. Siler, C. Murray, T. Yamauchi, W. B. Baggett, Why do only some events cause learning during human tutoring? *Cogn. Instr.* **21**, 209–249 (2003).
2. M. Kapur, Productive failure. *Cogn. Instr.* **26**, 379–424 (2008).
3. M. Kapur, J. Saba, I. Roll, Prior math achievement and inventive production predict learning from productive failure. *npj Sci. Learn.* **8**, 15 (2023).
4. K. Loibl, I. Roll, N. Rummel, Towards a theory of when and how problem solving followed by instruction supports learning. *Educ. Psychol. Rev.* **29**, 693–715 (2017).
5. T. Sinha, M. Kapur, When problem solving followed by instruction works: Evidence for productive failure. *Rev. Educ. Res.* **91**, 761–798 (2021).
6. M. Kapur, Productive failure in learning the concept of variance. *Instr. Sci.* **40**, 651–672 (2012).
7. M. Kapur, Productive failure in learning math. *Cognit. Sci.* **38**, 1008–1022 (2014).
8. M. Kapur, Comparing learning from productive failure and vicarious failure. *J. Learn. Sci.* **23**, 651–677 (2014).
9. C. Mazziotti, K. Loibl, N. Rummel, "Collaborative or individual learning within productive failure: Does the social form of learning make a difference?" in *Exploring the Material Conditions of Learning: The Computer Supported Collaborative Learning (CSCL) Conference 2015* (ISLS, 2015).
10. C. Hartmann, T. van Gog, N. Rummel, Productive versus vicarious failure: Do students need to fail themselves in order to learn? *Appl. Cogn. Psychol.* **36**, 1219–1233 (2022).
11. C. Hartmann, T. van Gog, N. Rummel, Preparatory effects of problem solving versus studying examples prior to instruction. *Instr. Sci.* **49**, 1–21 (2021).
12. D. Feil-Seifer, M. J. Mataric, "Defining socially assistive robotics" in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005* (IEEE, 2005), pp. 465–468.
13. N. Zhang, J. Xu, X. Zhang, Y. Wang, Social robots supporting children's learning and development: Bibliometric and visual analysis. *Educ. Inf. Technol.* **29**, 12115–12142 (2024).
14. I. Papadopoulos, R. Lazzarino, S. Miah, T. Weaver, B. Thomas, C. Koulouglioti, A systematic review of the literature regarding socially assistive robots in pre-tertiary education. *Comput. Educ.* **155**, 103924 (2020).
15. T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, F. Tanaka, Social robots for education: A review. *Sci. Robot.* **3**, eaat5954 (2018).
16. V. Rosanda, A. Istenic Starcic, "The robot in the classroom: A review of a robot role" in *Emerging Technologies for Education: 4th International Symposium, SETE 2019, Held in Conjunction with ICWL 2019, Magdeburg, Germany, September 23–25, 2019, Revised Selected Papers* (Springer-Verlag, 2019), pp. 347–357.
17. J. Kennedy, P. Baxter, T. Belpaeme, "The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning" in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2015), pp. 67–74.
18. J. M. Kory-Westlund, C. Breazeal, A long-term study of young children's rapport, social emulation, and language learning with a peer-like robot playmate in preschool. *Front. Robot. AI* **6**, 81 (2019).
19. S. Ekström, L. Pareto, The dual role of humanoid robots in education: As didactic tools and social actors. *Educ. Inf. Technol.* **27**, 12609–12644 (2022).
20. S. Ekström, L. Pareto, S. Ljungblad, Teaching in a collaborative mathematic learning activity with and without a social robot. *Educ. Inf. Technol.* **30**, 1301–1328 (2024).
21. M. Shiomi, T. Kanda, I. Howley, K. Hayashi, N. Hagita, Can a social robot stimulate science curiosity in classrooms? *Int. J. Soc. Robot.* **7**, 641–652 (2015).
22. V. Hakim, S.-H. Yang, M. Liyanawatta, J.-H. Wang, G.-D. Chen, Robots in situated learning classrooms with immediate feedback mechanisms to improve students' learning performance. *Comput. Educ.* **182**, 104483 (2022).
23. M. Alemi, A. Meghdari, M. Ghazisaedy, The impact of social robotics on l2 learners' anxiety and attitude in English vocabulary acquisition. *Int. J. Soc. Robotics* **7**, 523–535 (2015).
24. F. Jamet, O. Masson, B. Jacquet, J.-L. Stilgenbauer, J. Baratgin, Learning by teaching with humanoid robot: A new powerful experimental tool to improve children's learning ability. *J. Robot.* **2018**, 4578762 (2018).
25. B. Zhong, L. Xia, A systematic review on exploring the potential of educational robotics in mathematics education. *Int. J. Sci. Math. Educ.* **18**, 79–101 (2020).
26. J. Ceha, E. Law, D. Kulić, P. Y. Oudeyer, D. Roy, Identifying functions and behaviours of social robots for in-class learning activities: Teachers' perspective. *Int J. Soc. Robotics* **14**, 747–761 (2022).
27. V. G. A. Hakim, S.-H. Yang, T.-H. Tsai, W.-H. Lo, J.-H. Wang, T.-C. Hsu, G.-D. Chen, "Interactive robot as classroom learning host to enhance audience participation in digital learning theater" in *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)* (IEEE, 2020), pp. 95–97.
28. T. Kanda, T. Hirano, D. Eaton, H. Ishiguro, Interactive robots as social partners and peer tutors for children: A field trial. *Hum. Comput. Interact.* **19**, 61–84 (2004).
29. F. Tanaka, S. Matsuzoe, Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *J. Hum. Robot Interact.* **1**, 78–95 (2012).
30. K. Wang, G. Y. Sang, L. Z. Huang, S. H. Li, J. W. Guo, The effectiveness of educational robots in improving learning outcomes: A meta-analysis. *Sustainability* **15**, 4637 (2023).
31. M. Kapur, K. Bielaczyc, Designing for productive failure. *J. Learn. Sci.* **21**, 45–83 (2012).
32. J. Sung, H. I. Christensen, R. E. Grinter, "Robots in the wild: Understanding long-term use" in *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (ACM/IEEE, 2009), pp. 45–52.
33. R. Kühne, J. Peter, C. de Jong, A. Barco, How does children's anthropomorphism of a social robot develop over time? A six-wave panel study. *Int. J. Soc. Robotics* **16**, 1665–1679 (2024).
34. C. Hartmann, T. van Gog, N. Rummel, Do examples of failure effectively prepare students for learning from subsequent instruction? *Appl. Cogn. Psychol.* **34**, 879–889 (2020).
35. S. H. Seo, D. Geiskkovitch, M. Nakane, C. King, J. E. Young, "Poor thing! Would you feel sorry for a simulated robot? A comparison of empathy toward a physical and a simulated robot" in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2015), pp. 125–132.
36. H. A. David, *The Method of Paired Comparisons* (Hafner Publishing Company, 1963).
37. M. Ghosh, F. Tanaka, "The impact of different competence levels of care-receiving robot on children," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2011), pp. 2409–2415.
38. J. B. K. Yap, S. S. H. Wong, Deliberately making and correcting errors in mathematical problem-solving practice improves procedural transfer to more complex problems. *J. Educ. Psychol.* **116**, 1112–1128 (2024).
39. L. Schalk, R. Schumacher, A. Barth, E. Stern, When problem-solving followed by instruction is superior to the traditional tell-and-practice sequence. *J. Educ. Psychol.* **110**, 596–610 (2018).
40. S. Ko, J. Barnes, J. Dong, C. H. Park, A. Howard, M. Jeon, The effects of robot voices and appearances on users' emotion recognition and subjective perception. *Int. J. Human. Robot.* **20**, 2350001 (2023).
41. A.-L. Vollmer, R. Read, D. Trippas, T. Belpaeme, Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Sci. Robot.* **3**, eaat7111 (2018).
42. R. Van Den Berghe, J. Verhagen, O. Oudgenoeg-Paz, S. Van Der Ven, P. Leseman, Social robots for language learning: A review. *Rev. Educ. Res.* **89**, 259–295 (2019).
43. Z. J. You, C. Y. Shen, C. W. Chang, B. J. Liu, G. D. Chen, "A robot as a teaching assistant in an English class" in *Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06)* (IEEE, 2006), pp. 87–91.
44. J. H. Han, M. H. Jo, V. Jones, J. H. Jo, Comparative study on the educational use of home robots for children. *J. Inf. Process. Syst.* **4**, 159–168 (2008).
45. I. Leite, C. Martinho, A. Paiva, Social robots for long-term interaction: A survey. *Int. J. Soc. Robot.* **5**, 291–308 (2013).
46. G. Laban, A. Kappas, V. Morrison, E. S. Cross, Building long-term human–robot relationships: Examining disclosure, perception and well-being across time. *Int. J. Soc. Robot.* **16**, 1–27 (2024).
47. S. S. H. Wong, Deliberate erring improves far transfer of learning more than errorless elaboration and spotting and correcting others' errors. *Educ. Psychol. Rev.* **35**, 16 (2023).
48. H. Woo, T. Pham-Shouse, Y. Xiong, The use of social robots in classrooms: A review of field-based studies. *Rev. Educ. Res.* **33**, 100388 (2021).
49. T. Salter, F. Michaud, H. Larouche, How wild is wild? A taxonomy to characterize the 'wildness' of child-robot interaction. *Int. J. Soc. Robot.* **2**, 405–415 (2010).
50. A. Gunturu, Y. Wen, N. Zhang, J. Thundathil, R. H. Kazi, R. Suzuki, "Augmented physics: Creating interactive and embedded physics simulations from static textbook diagrams" in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (ACM, 2024), pp. 1–12.
51. J. Sung, H. I. Christensen, R. E. Grinter, "Robots in the wild: Understanding long-term use" in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (ACM, 2009), pp. 45–52.
52. C. Bartneck, T. Kanda, O. Mubin, A. Al Mahmud, "The perception of animacy and intelligence based on a robot's embodiment," in *2007 7th IEEE-RAS International Conference on Humanoid Robots* (IEEE, 2007), pp. 300–305.
53. C. Bartneck, D. Kulić, E. Croft, S. Zoghbi, Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **1**, 71–81 (2009).
54. J. S. Olson, W. A. Kellogg, Eds., *Ways of Knowing in HCI*. (Springer, 2014).
55. L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan, K. Hoagwood, Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Adm. Policy Ment. Health* **42**, 533–544 (2015).

Acknowledgments: We thank Wenhai Middle School in Hangzhou for supporting the implementation of this experiment. Special appreciation is extended to teachers S. Yan, C. Liu,

M. Zhao, C. Xia, and C. Xu for professional insights on revising the robot's suboptimal solutions, classroom procedures, and quiz standards. We also wish to acknowledge F. Gao (Zhejiang University), Y. Du (University of Southern California), and Z. Shi (University of Southern California) for valuable feedback during the early stages of manuscript preparation. **Funding:** This work was supported by the National Key R&D Program of China (2022YFB3303304) and the National Natural Science Foundation of China (62207023). **Author contributions:** Conceptualization: L.C. and Y.C. Methodology: L.C., Y.C., and Z.Y. Software: Y.F., D.X., and Z.Y. Investigation: Y.F., Y.C., J.Y., S.X., D.X., L.Z., and L.C. Formal analysis: Y.C., Z.Y., J.Y., Y.F., S.X., D.X., and L.Z. Visualization: Y.S. and D.X. Funding acquisition: L.C. and L.S. Project administration: L.C. Supervision: L.C. Writing—original draft: Y.C., Y.F., and Z.Y. Writing—review and editing: Y.C., Y.F., L.C., S.X., L.S., and J.C. Z. Yang completed contributions to software and methodology during a summer internship at Zhejiang University before enrolling at the University of California, Irvine, where she made the remaining contributions to the study. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data and code are available in the main text, the Supplementary Materials, or via Zenodo (https://doi.org/10.5281/zenodo.16756851). There are no restrictions on the availability of data or materials after publication.

Submitted 18 November 2024
Accepted 12 August 2025
Published 10 September 2025
10.1126/scirobotics.adu5257

Science Robotics

Observing a robot peer's failures facilitates students' classroom learning

Liuqing Chen, Yu Cai, Yuyang Fang, Ziqi Yang, Duowei Xia, Jiaxiang You, Shuhong Xiao, Yaxuan Song, Lingwei Zhan, Juanjuan Chen, and Lingyun Sun

Sci. Robot. **10** (106), eadu5257. DOI: 10.1126/scirobotics.adu5257

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.