

Peers4U: Designing Learner-Centered Virtual Peers for Children’s Multimodal Mathematics Learning

ZHI ZHENG, Department of Computer Science and Technology, Tsinghua University, China

BINGLIN LIU, Department of Computer Science and Technology, Tsinghua University, China

WEIHAO CHEN, Department of Computer Science and Technology, Tsinghua University, China

LAN LUO, The Hong Kong University of Science and Technology (Guangzhou), China

YU MEI, Department of Computer Science and Technology, Tsinghua University, China

JIE CAI, College of Software, Nankai University, China

CHUN YU, Department of Computer Science and Technology, Tsinghua University, China

YUANCHUN SHI, Department of Computer Science and Technology, Tsinghua University, China

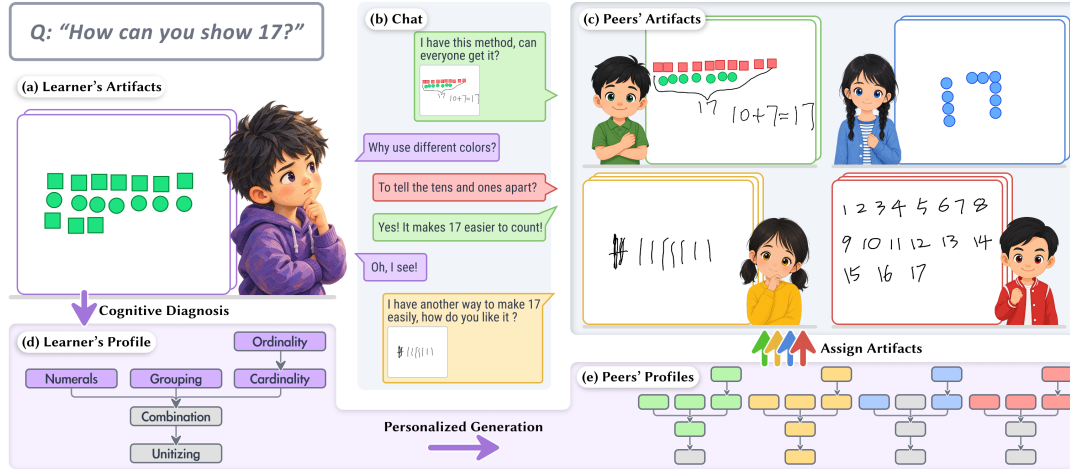


Fig. 1. Overview of *Peers4U* through an example task to represent 17. After authoring solution artifacts (a), the learner joins a personalized group of four virtual peers with diverse artifacts (c) and engages in discussion (b). The lower part of the figure shows the backend pipeline: the system analyzes the learner’s initial solutions to derive a cognitive profile (d), generates heterogeneous peer cognitive profiles tailored to the learner (e), and assigns artifacts accordingly.

In early mathematics, children learn abstract concepts through concrete representations, such as manipulating objects, drawing, and speaking. Learning with peers can enrich this process by exposing children to diverse strategies and enabling them to exchange ideas through interaction. We realized this potential with virtual peers in *Peers4U*, a multimodal learning system to enhance children’s mathematical conceptual understanding. Virtual peers driven by LLMs are personalized to the learner’s cognitive level and assigned multimodal solutions with inspectable processes. During learning, peers engage in mixed-initiative interactions by responding, initiating topics, and advancing discussions along a pedagogical progression. A user study with 22 children and two experts showed

Authors’ Contact Information: Zhi Zheng, Department of Computer Science and Technology, Tsinghua University, Beijing, China; Binglin Liu, Department of Computer Science and Technology, Tsinghua University, Beijing, China; Weihao Chen, Department of Computer Science and Technology, Tsinghua University, Beijing, China; Lan Luo, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China; Yu Mei, Department of Computer Science and Technology, Tsinghua University, Beijing, China; Jie Cai, College of Software, Nankai University, Tianjin, China; Chun Yu, Department of Computer Science and Technology, Tsinghua University, Beijing, China; Yuanchun Shi, Department of Computer Science and Technology, Tsinghua University, Beijing, China.

our learner-centered topic orchestration significantly outperformed a persona-centered baseline, and children spontaneously authored new artifacts to demonstrate inspired ideas or share their own strategies, producing significantly more solutions when grouped with more Leading Peers. We contribute design insights for multimodal, learner-centered virtual peer systems.

CCS Concepts: • **Applied computing** → **Interactive learning environments**; *Collaborative learning*; *Computer-assisted instruction*; • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: Virtual Peers, Peer Learning, Multimodal Learning, Mathematics Education, Children, Intelligent Tutoring Systems

1 Introduction

In early mathematics, it is crucial for children to ground their understanding of abstract concepts and operations in concrete experiences [5, 64]. For example, children can conceptualize place value by bundling ten individual objects into a group, drawing corresponding quantities, and connecting these actions to written numerals. Such multimodal learning through multiple representations, including manipulating, drawing, writing, and speaking [11, 12, 52, 53], helps children develop deeper conceptual understanding and representational fluency [25, 62].

Learning with peers offers natural opportunities to enrich this process [18, 96]. Peers often bring diverse solution strategies across different representations [113], exposing one another to unfamiliar ways of thinking and sparking curiosity [23]. Through interaction, they can further explain their reasoning, compare alternative approaches, exchange ideas, and challenge or build on one another’s understanding [27, 85].

The emergence of Large Language Models (LLMs) has opened new possibilities for introducing virtual peers into Intelligent Tutoring Systems (ITSs). However, many recent virtual-peer systems target older learners or other learning domains [56, 102, 106], with peer interactions primarily centered on text- and voice-based conversation [97, 104, 112]. This conversational paradigm fits poorly with the visual and manipulable needs of early mathematics learning [53, 63]. Furthermore, little is known about how young children would participate and learn in an open virtual-peer environment.

To bridge this gap, we explore how to design a multimodal virtual peer-learning system for early mathematics. Specifically, we investigate: *How can virtual peers be designed to support children’s multimodal mathematics learning, and how do children participate in the resulting peer-learning interactions?*

We first conducted a formative study consisting of exploratory probe sessions with 8 children and interviews with 5 experienced teachers. We identified four interrelated design challenges: (1) children’s mathematical peer communication needs to be grounded in shareable multimodal solution artifacts; (2) peers need to be sufficiently different from the learner to provide new perspectives while remaining understandable; (3) children vary substantially in how they prefer to participate; and (4) open peer interaction alone does not naturally produce a coherent learning progression. These findings informed four corresponding design goals (DGs).

Guided by these goals, we designed and developed *Peers4U*, a multimodal virtual-peer ITS for early mathematics learning, as shown in Figure 1. In *Peers4U*, after independently completing multimodal solutions to an open-ended problem, the learner is presented with a personalized learning group of 4 peers based on the learner’s demonstrated mathematical understanding. Each peer possesses a diverse set of solution artifacts whose multimodal construction processes can be inspected. The interaction follows a mixed-initiative design [36]: the learner can engage with peers to discuss specific artifacts and author and share new solutions at any time, while virtual peers can likewise respond, proactively initiate topics, and advance discussions along a pedagogical trajectory.

To characterize children’s peer-learning experiences and evaluate the system design, we conducted an exploratory user study with 22 children across different peer-composition conditions and a post-hoc evaluation with 2 experts. *Peers4U*’s learner-centered topic orchestration received high pedagogical appropriateness ratings (4.69/5), significantly outperforming a persona-centered baseline. Furthermore, children spontaneously authored new artifacts to teach peers, demonstrate newly inspired ideas, or improve and share their own strategies. Children produced significantly more additional solutions when learning with a Leading Group, while those in a Mixed Group showed significantly greater willingness to learn with the same peers again. We also identified how different types of peer messages were associated with children’s likelihood of responding. We further discuss broader implications for designing multimodal virtual peers that children can both learn from and teach.

Our contributions are threefold:

- We identify core challenges of introducing virtual peers into children’s mathematics learning and derive four DGs spanning modality mediation, peer heterogeneity, flexible initiative, and progressive orchestration.
- We present the design and implementation of *Peers4U*, instantiating an artifact-mediated multi-peer learning paradigm that combines mixed-initiative interaction with learner-centered pedagogical orchestration.
- We provide empirical findings revealing children’s interaction patterns in multimodal virtual peer learning, and outline transferable design implications for virtual peer learning systems.

2 Related Work

2.1 Theoretical Foundations: Multiple Representations and Peer Learning in Mathematics

Mathematics learning is fundamentally a process of cognitive development. Rooted in Embodied Cognition, children construct abstract concepts through physical manipulation and visual perception [6, 64]. Theories such as Bruner’s Enactive-Iconic-Symbolic (EIS) framework [11, 12] and Lesh’s translation model [53] emphasize the necessity of practicing and translating across multiple representations—including concrete manipulatives, pictorial drawings, spoken language, and written symbols [63, 66]. Engaging in these translational activities allows children to develop Representational Fluency and achieve a deep conceptual understanding [25, 62]. Consequently, multimodal interactive learning systems inherently align with this pedagogical need, which forms the foundational rationale of our system.

Social learning environments naturally complement this need for multi-representational learning [107], where peers bring diverse strategies and varying representational choices [23]. Drawing on Social Learning Theory [3], children can effectively acquire new strategies by observing and imitating peer behaviors. Furthermore, exposure to such heterogeneous peer strategies often triggers Socio-Cognitive Conflict, compelling children to reflect on their own reasoning and ultimately facilitating deeper knowledge internalization [18, 69].

Consequently, peer learning paradigms are widely adopted in pedagogical practice [16, 100, 101]. Specifically, Reciprocal Peer Tutoring—where children alternate between tutor and tutee roles—yields mutual learning gains through preparation and explanation [22, 24, 99], while cooperative learning enhances academic achievement and strengthens social support [37, 41]. Recently, advancements in Large Language Models (LLMs) have made virtual peers a reality, providing a unique opportunity to offer highly controlled and tailored peer interactions. Building on this, our research explores how to effectively design such virtual peers to facilitate multimodal mathematics learning for children.

2.2 Multimodal Interactive Systems for Mathematics Learning

Grounded in the theories outlined above, mathematical understanding hinges on multiple representations. Existing HCI research seeks to move beyond abstract symbols by externalizing mathematical relationships into directly perceptible, manipulable objects to scaffold learning.

One line of research combines physical environments with digital feedback to contextualize the mathematical learning experience. For instance, Tangible Tens [70] and Bricksmart [57] leverage tabletop block play to visually represent quantitative relationships and spatial structures; Balance Board Math [93] and MOVES-NL [1, 17] map embodied action onto symbolic forms such as graphs and number lines. Extended Reality (XR) environments, such as ARMath [42], Mathland [44], and MaR-T [9], further augment mathematical inquiry with immersive visualizations. While grounding math learning in embodied cognition provides strong intuition, these systems remain bound by the inherent properties of physical artifacts, constraining their applicability to particular mathematical topics.

Beyond physical interfaces, digital systems can cover broader learning scenarios. They can support learners' understanding of multiple interconnected representations [74, 81], using on-screen virtual manipulatives to bridge concreteness and abstraction [61, 82]. Specifically, Fractions Tutor dynamically integrates graphical models with number lines [73, 74], and MathWorlds links graphs, animations, and equations [78]. Yet, these systems are predominantly designed for solo learning and lack built-in peer-interaction mechanisms, missing opportunities for the collision of different perspectives. While OpenCD further captures speech alongside these modalities, it is designed for diagnostic assessment rather than interactive learning [113].

To overcome the limitations of solo learning, we aim to introduce virtual peers embedded within a multimodal interactive environment, fostering rich cognitive friction and learning opportunities.

2.3 Virtual Peers in Learning Environments

Since their emergence, LLMs have been widely adopted for role-playing to generate anthropomorphic dialogue and behavior [68, 88]. Building on this, educational research employs LLMs to simulate students [40, 65], laying the groundwork for virtual peers and complex social learning dynamics.

The HCI community has long focused on the design of virtual peers [13, 72, 87]. Applications span from collaborative child-agent co-reading [102] to systems like Your Thoughtful Opponent and ArgueMate, which induce cognitive conflict through disagreement and debate [45, 104]. Research has also explored robots as physical companions [8, 35, 67, 79]. A separate line of work has investigated teachable agents—where students learn by teaching AI—to enhance learning outcomes, foster metacognitive skills, and boost student engagement [39, 59, 77, 80, 110].

In more social learning settings, virtual peers can participate alongside human learners [60]. For instance, PeerGPT joins children's small-group collaboration, while Saya takes part in classroom activities as an "additional student" [56, 97]. Beyond introducing a single peer, some studies employ multiple virtual peers to create richer social learning environments, such as MAIC [111], ClassMeta [58], among others [60, 94, 106, 112]. However, these virtual peers primarily communicate through conversations (text/voice), which constrains their use in children's mathematics learning, where interaction often involves richer forms of engagement. Thus, how young children interact with and learn alongside virtual peers, and how such interactions can be facilitated in early mathematics, remain underexplored.

3 Formative Study

To inform the design of our system, we conducted a two-part formative study consisting of exploratory probe sessions with children and semi-structured interviews with elementary mathematics teachers. We synthesized evidence from both parts into four design goals (DGs).

3.1 Study Rationale and Design

We specifically focused on the early stages of mathematical and social learning—spanning late kindergarten to early primary school. Mathematics at this stage is strongly grounded in concrete and visual representations. Rather than relying solely on verbal explanations, children at this stage commonly express and develop mathematical ideas through manipulating objects, drawing, writing, and speaking. This setting therefore provides a particularly relevant context for examining how virtual peers might participate in and support multimodal mathematical learning.

Our formative study combined two complementary perspectives. First, we used an exploratory probe to observe how children attended to, interpreted, and interacted with peers' multimodal mathematical solutions. Second, we interviewed experienced teachers to contextualize these behaviors within classroom practice and to identify pedagogical opportunities, risks, and desirable forms of peer interaction.

3.2 Part 1: Child Probe Sessions

Probe. We developed an exploratory probe that intentionally kept the social behaviors of virtual peers relatively simple while providing a rich and open multimodal solution space. The probe was intended to elicit how children and teachers expected peers to participate in such a space and what forms of interaction they found useful or problematic.

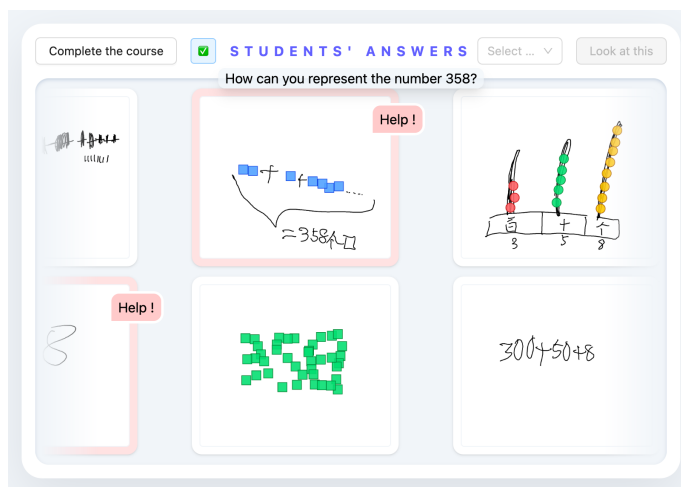


Fig. 2. User interface of the probe system. Each card presents a multimodal artifact; children could replay its construction process and chat with the associated virtual peer.

Children first selected a mathematics task and created one or more solutions using multiple modalities, including manipulating virtual objects, drawing, writing, and verbal explanation, using a multimodal authoring workspace described in prior work [113]. After submitting their own solution, children were presented with a gallery of peer

solutions, as shown in Figure 2. They could replay the multimodal process of each solution and chat with the associated LLM-driven virtual peer via voice input and text-to-speech (TTS) output. Artifacts included flawed solutions where peers actively asked for help.

The peer-solution repository was constructed from an existing dataset described in OpenCD [113], with its reuse approved under the applicable IRB protocol. All data were de-identified prior to use, and audio responses additionally underwent voice conversion to reduce speaker identifiability. The mathematics tasks are listed in Appendix A.1; details of the peer-solution presentation order and the LLM configuration of the virtual peers are provided in Appendix A.2.

Participants. We recruited 8 children (4 boys and 4 girls; age $M = 6.7$ years, $SD = 0.61$, range = 5y11m–7y5m) spanning the final year of kindergarten through first grade (4 K, 4 G1). Recruitment was conducted via personal accounts on RedNote for in-person sessions in a Chinese city. Parents registered their children voluntarily, and we filtered excess applicants solely to balance gender and grade. Following IRB approval and parental written consent, children participated voluntarily and were informed that they could pause or withdraw at any time. They received certificates as gifts upon completion. Participant details are provided in Appendix A.3.

Procedure. Each child participated in an approximately 60-minute, in-person session using the probe on a 13-inch iPad with an Apple Pencil. The probe included nine mathematics tasks across three categories (Appendix A.1). Children could freely choose tasks, complete one or more solutions, browse peer solutions, inspect their multimodal construction processes, and interact with the corresponding virtual peers. We did not explicitly tell children beforehand that the peers were AI-driven; children could nevertheless infer this from occasional imperfections in the interactions.

The sessions followed a semi-structured observation protocol: as children navigated the system, the researcher occasionally probed their willingness to engage in specific interactions, prompted them to articulate their feelings and thoughts, and observed their spontaneous behaviors. Children could work on multiple tasks during the session, and the researcher occasionally varied the peer-solution presentation configuration to elicit user reactions across different social contexts (Appendix A.3).

3.3 Part 2: Teacher Interviews

Participants. We recruited 5 experienced elementary school mathematics teachers from different schools through personal networks. They had 7–17 years of teaching experience ($M = 11.2$, $SD = 4.15$), including 3–17 years of experience teaching Grades 1–2 ($M = 7.2$, $SD = 5.67$). Participants received 200 RMB (approx. \$30 USD) for the 1.5-hour interview. See Appendix A.4 for detailed teacher profiles.

Procedure. We conducted semi-structured interviews centered on three themes: (1) the role and dynamics of peer interaction in real-world early math classrooms; (2) their professional critique and design suggestions after the demonstration of our technology probe; and (3) their interpretations and pedagogical attributions of the specific child behaviors we observed in Part 1.

3.4 Data Collection and Analysis

Throughout the study, we collected screen recordings, video footage, and raw interaction logs from the children’s sessions, alongside audio recordings from the teacher interviews. All audio was transcribed verbatim.

We conducted a team-based codebook thematic analysis [10]. The child-session and teacher-interview data were first analyzed separately. For the child-session data, screen recordings and video footage were synchronized and analyzed together, using segments containing a coherent expression or behavior as analytic units. Two researchers independently open-coded a subset of each dataset (three child sessions and two teacher transcripts) and developed codebooks. One

researcher coded the remaining data, while the second reviewed the coded segments. Within each dataset, related codes were grouped into candidate themes and reviewed against the full data.

Finally, the researchers compared and integrated themes across the two datasets, connecting children’s observed behaviors with teachers’ pedagogical interpretations.

3.5 Findings and Design Goals

Overall, the probe successfully elicited rich interactions. Children selected an average of 4.0 ± 1.2 tasks, initiated chats around 5.5 ± 4.1 peer artifacts, and sent 19.6 ± 16.5 messages. Through our observations of children’s interactions with the probe and interviews with expert teachers, we identified four core challenges and derived corresponding Design Goals (DGs) for integrating virtual peers into multimodal early math learning.

Finding 1: Peer interaction around mathematics extends beyond language to multimodal artifacts. A key challenge is that text or speech alone could not accommodate how children wanted to communicate mathematical ideas. In our probe sessions, children often sought to share or demonstrate their own multimodal solutions, making solution artifacts part of the “language” through which they observed, questioned, explained, and responded to one another.

The dynamic solution process drew particular attention, and children frequently directed their questions to specific actions within it (S2, S4, S5, S8). For example, S4 noticed the sequence of a peer’s actions, asking, “Why did you drag 8 squares first, and then 8 circles?” Teachers similarly emphasized the importance of making mathematical thinking visible (T1, T4). As T4 explained, young children often need to “see it while listening” in order to follow an explanation.

Solution artifacts also became a natural medium for interacting with peers. When attempting to respond or explain, some children wanted to directly share or demonstrate their own solutions rather than rely on verbal explanation (S1, S2, S5). Peer solutions could also prompt further creation: S4 and S7 voluntarily proposed returning to the canvas to create another solution after viewing others’ approaches. Children sometimes even used their solutions for self-presentation, proudly telling peers how many methods they had created (S4, S6).

⇒ **DG1: Ground peer interaction in shareable multimodal mathematical artifacts.** Peer interactions should revolve around mathematical solutions rather than purely conversational text. Artifacts should serve not only as initial assessments but also as an ongoing medium of communication. The system should allow children to revisit, share, and discuss artifacts during interaction, as well as create new solutions.

Finding 2: Useful peers must be different enough to inspire, yet close enough to understand. A second challenge lies in calibrating peer heterogeneity. Across all eight children, overly simple or similar solutions were often quickly skipped. Conversely, unfamiliar methods sparked curiosity, with 7 children expressing a desire to communicate upon seeing a novel solution. However, when the cognitive gap was too large, children sometimes misinterpreted the methods or incorrectly judged them as wrong (S2, S4, S5). Teachers therefore noted that peer materials must provide an “intermediate ladder”, guiding children from understandable methods toward unfamiliar ones (T2, T4).

We also observed the unique pedagogical value of peers presenting less developed or flawed solutions. Six children showed a strong willingness to help when peers explicitly asked for assistance, and re-explained their own thoughts while helping (S1–S6). S1 eagerly claimed, “Of course I want to help, I’ll be the little teacher!” Some children even transferred strategies they had just learned from another peer (S1, S3). Teachers noted that explaining to others can force children to reorganize their reasoning and expose gaps in their own understanding (T2, T3).

Consequently, different peer levels serve distinct purposes: more advanced peers inspire, similar peers provide relatability, and struggling peers offer opportunities for correction and confidence-building (T2–T5).

⇒ **DG2: Construct heterogeneous peers relative to the learner.** The system should dynamically assign a group of peers with diverse mathematical approaches and cognitive levels, tailored to the user’s initial understanding. This heterogeneity must balance inspiration with approachability, deliberately providing diverse opportunities for observation, explanation and learning-by-teaching [24, 27, 99].

Finding 3: Children’s participation styles vary substantially, requiring flexible interaction initiative. A third challenge is the substantial variation in how children prefer to participate in peer learning. A rigid interaction paradigm—whether primarily system-led or learner-led—is unlikely to work well for all children.

Interaction logs revealed a participation divide. On average, some children sent over 10 messages per task (S3, S6), while others sent no more than 1 message per task (S7, S8). S7 even explicitly rejected a peer’s request for help and later explained, “I don’t want to talk to them anymore because I want to go to the next problem faster.” T5 pointed out that while peer presence encourages some children to express themselves, it can make others more silent, thus for introverted learners, “listening while other peers ask questions” was also an effective and necessary form of participation.

⇒ **DG3: Support flexible interaction initiative and participation roles.**

The system’s conversational agency must be flexible, as reflected in prior learning systems that combine learner- and agent-initiated interaction [106, 111]. Virtual peers should be able to naturally respond to user inquiries, but also proactively initiate topics, move the discussion forward, or gracefully conclude them when the user is passive. This allows children with different participation styles to fluidly shift among the roles of leader, responder, or observer.

Finding 4: Open Peer Interaction Lacks Focus Without Progressive Pedagogical Orchestration. Finally, young children may lack the metacognitive capacity to independently navigate multiple peers, representations, and mathematical approaches. An open peer environment does not by itself produce an effective learning trajectory. As T3 emphasized, “Learning is not just about interaction... it must have clear pedagogical objectives.”

Teachers outlined a consistent progression for productive peer learning. First, children must form their own initial ideas before being exposed to peers (T1, T5). Subsequently, peer materials should be introduced progressively—starting from relatable, easier methods to increasingly diverse representations (T1, T2, T4). Furthermore, T1 noted that after exploring divergent methods, there must be a convergent process where children are guided to compare differences, analyze commonalities, and extract mathematical essences.

This suggests a tension between *social openness* and *pedagogical structure*: peer interaction should feel exploratory, while its underlying progression still needs to be organized toward specific learning objectives.

⇒ **DG4: Orchestrate peer interaction through cognitively progressive learning topics.** The system must act as an “invisible director”, dynamically orchestrating learning topics around the learner’s current understanding. The discussion of solution methods should follow the developmental trajectory of mathematical cognition [30, 89], and the types of discussion topics should deepen progressively according to cognitive levels [50].

4 System Design

In this section, we present the design of *Peers4U* based on the four DGs derived from our formative study. Together, they shaped a holistic interaction paradigm for *Peers4U*.

4.1 System Overview

For each task, the learner first independently constructs their own solutions. This serves multiple purposes: it ensures independent mathematical thinking, provides materials for subsequent peer interaction, and allows the system to estimate the learner’s current mathematical understanding.

Following this, the learner enters a virtual “classroom”, as shown in Figure 3, where the system has tailored a personalized learning group (DG2). Mirroring real-world small-group dynamics, this group consists of four virtual peers, each equipped with multiple multimodal solution artifacts. Learners can review their peers’ problem-solving processes, engage in group or private chats, directly discuss specific artifacts, and create and share new artifacts at any time (DG1). Furthermore, peers can proactively initiate learning topics to engage the user and move the discussion forward (DG3). These topics are diverse, unfold iteratively, and are dynamically orchestrated based on the learner’s performance, progressing from shallow to deep following pedagogical strategies (DG4).

Ultimately, these mechanisms aim to foster deep participation, effective peer collisions, and an orderly deepening of conceptual understanding. Below, we detail the system’s design.

4.2 Multimodal Artifact Creation and Peer Interaction (DG1)

DG1 requires the interface to facilitate natural, multimodal interactions centered around mathematical artifacts.

To support children’s mathematical expression and learning through multiple representations, *Peers4U* allows learners to create solutions using a variety of modalities, including virtual manipulatives [61, 82], drawing, writing, and speaking, as shown in Figure 3b. We adopted the multimodal workspace introduced in OpenCD [113] for this purpose.

After completing their initial solutions, learners enter the “Learning Group” page (Figure 3a). The four peers and their associated solutions are presented together for an overview. Learners can open the chat panel to participate in group discussions, or inspect a peer’s solutions, replay their complete multimodal construction process, and communicate privately with that peer (Figure 3d).

Beyond text or voice chat, students can return to the workspace at any point to author new content, and send their entire multimodal creation process to peers. Naturally, students can also forward other peers’ solutions into the chat for discussion, a capability shared by the virtual peers as well.

4.3 Personalized Peer Construction (DG2)

DG2 requires the system to dynamically assemble an appropriate peer group for each learner. Specifically, the system needs to estimate the learner’s current mathematical understanding, determine heterogeneous but appropriate cognitive profiles for multiple peers, and assign each peer a set of solutions consistent with that profile.

4.3.1 Artifact Analysis and Cognitive Mapping. A learner’s cognitive level can be estimated from their initial solution artifacts. The system must comprehend the full multimodal problem-solving process and map it to task-specific mathematical competencies. Because *Peers4U* uses the same mathematical tasks and multimodal workspace as OpenCD [113], we directly reused its knowledge bases, which contain typical solution behaviors and task-specific cognitive graphs organized as directed acyclic graphs (DAGs), and adapted its VLM-based artifact-analysis pipeline:

Solution Artifact Comprehension Agent: Translates the multimodal problem-solving process into a concise text description, enabling downstream interaction agents to efficiently handle a large volume of artifacts during conversations.

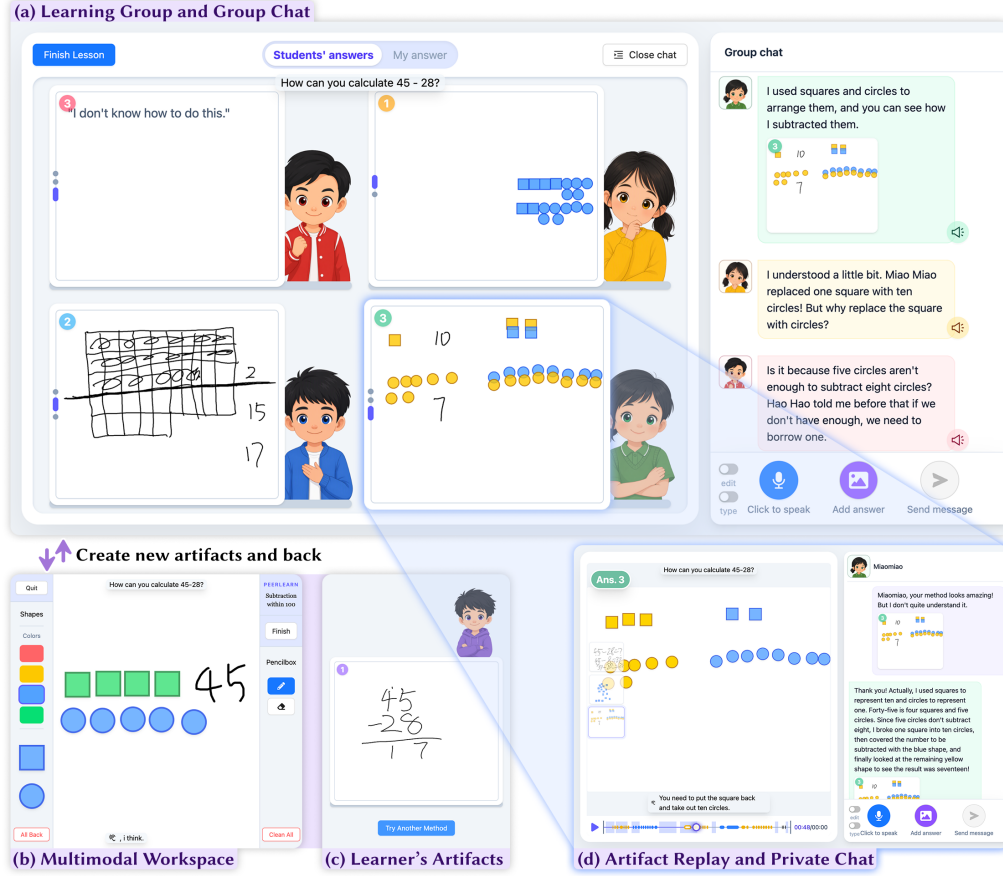


Fig. 3. User interface and interaction flow of *Peers4U*. Learners can browse all peers' artifacts and chat with the group (a). They can also create new artifacts in the multimodal workspace (b), revisit their own artifacts (c), or replay a peer's multimodal solution process and discuss with them in a private chat (d).

Behavior Detection and Cognitive Mapping Agent: Identifies task-relevant behavioral evidence from the artifact and maps it to corresponding nodes in the cognitive graph. This indexes the artifact at both behavioral and cognitive levels, establishing a structural correspondence between the solutions and cognitive states.

Consequently, a student's initial solution artifacts are mapped as a set of evidence on the cognitive graph, as shown in Figure 4a, forming a task-specific cognitive profile that supports subsequent peer construction.

Furthermore, all artifacts within the system's artifact repository are preprocessed using the same pipeline, making them ready for peer allocation in the subsequent step.

4.3.2 Assigning Artifacts to Peers. Based on the learner's cognitive profile and initial solutions, *Peers4U* constructs multiple peers with heterogeneous cognitive levels and assigns each peer a diverse set of appropriate artifacts. Leveraging the aforementioned analysis preprocessing, this construction can be performed using explicit rules rather than LLM generation, making the process more efficient and controllable.

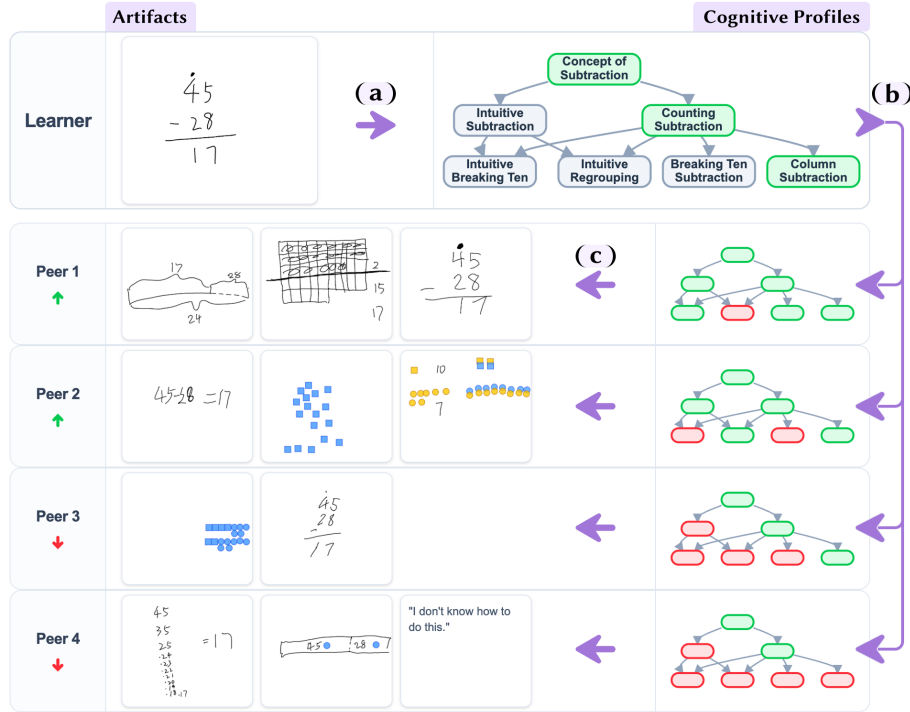


Fig. 4. Example of personalized peer construction. The system maps the learner’s artifacts to a task-specific cognitive profile (a), constructs heterogeneous profiles for four peers (b), and assigns each peer a diverse portfolio of artifacts consistent with their profile (c). Within the profiles, green and red nodes denote mastered and unmastered competencies, respectively, while gray nodes denote competencies not demonstrated. Green upward and red downward arrows on the left indicate Leading and Struggling Peers, respectively.

Following DG2, we distinguish two broad peer roles relative to the learner: *Leading Peers* and *Struggling Peers*. Leading Peers are assigned valuable strategies that the learner has not yet demonstrated, providing opportunities to encounter new perspectives. Struggling Peers contain more incomplete or flawed solutions, creating opportunities for correction and learning-by-teaching. Peers within the same role can also differ in their specific cognitive profiles. The system supports flexible combinations of these roles; the configurations used in our user study are described in § 5.2.

Once the peers’ cognitive profiles are determined (Figure 4b), the system assigns solutions that both support these personas and provide useful materials for interaction (Figure 4c). Motivated by prior work showing both the benefits and cognitive demands of learning from multiple solution methods [31, 75], we bounded each group’s solution pool around 10 artifacts, with at least 2 per peer. Using the preprocessing results described above, the assignment procedure prioritizes solutions according to their cognitive fit with the target peer, similarity to existing solutions and modality coverage. This produces peer portfolios that are consistent with the intended cognitive profiles while remaining diverse in strategy and form. The construction and assignment rules, together with examples across tasks and group compositions, are provided in Appendix B.2.

4.4 Learning Topic Initiation and Progression (DG3, DG4)

Rather than allowing each virtual peer to independently pursue a persona-driven agenda, *Peers4U* coordinates peer behaviors around the learner’s learning progression to achieve learner-centered interaction. To fulfill this, we design the interaction mechanism from two dimensions: adaptive conversational agency and progressive topic orchestration.

4.4.1 Adaptive Conversational Agency via Implicit Interaction. DG3 requires the coordination to remain flexible: peers should respond when the learner takes initiative, while also being able to proactively drive the discussion when appropriate, consistent with mixed-initiative interaction principles [36]. However, rigid system interventions can easily interrupt the child’s active learning process. We operationalize this through **implicit interaction**: using the user’s operational silence as an implicit contextual cue to dynamically modulate conversational agency [15, 86].

The system continuously monitors the learner’s activity stream and triggers generation through a conversation control flow shown in Figure 5. When the learner sends a chat message, the system directly invokes the *Reactive Agent* for a response. For other activities, the system differentiates the user’s state: if the user is engaged in high-initiative tasks such as recording speech, constructing an artifact, or replaying a solution process, the system classifies the state as *Engaged* and remains silent to avoid interruption. Conversely, after the learner remains *Idle* for a period, the system treats this as an opportunity for a lightweight proactive intervention. It invokes the *Topic Agent* to initiate a new topic or the *Follow-up Agent* to advance the current discussion. Both the Reactive and Follow-up Agents can close the current topic when the topic’s goal has been reached or has expired. Implementation details are provided in Appendix B.3.

Through this mechanism, children who prefer to lead can do so without interruption, while those who participate less actively can still receive timely invitations to engage, fostering an elastic and natural group chat rhythm.

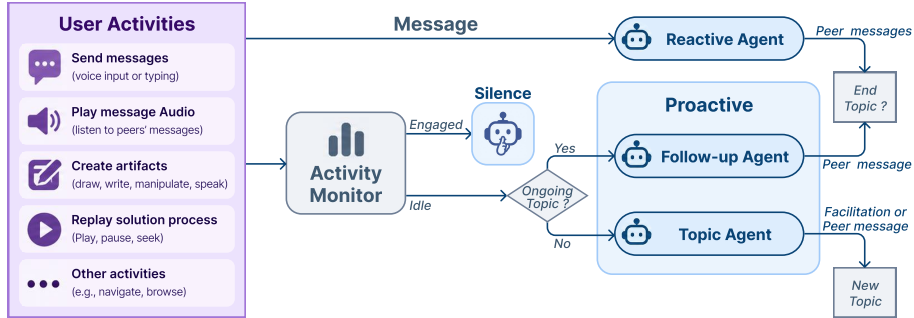


Fig. 5. Message-generation flow of *Peers4U*. A learner message invokes the Reactive Agent directly. Other activities are monitored to estimate the learner’s state. If the learner is engaged, the system remains silent; if idle, it invokes either the Follow-up Agent or the Topic Agent. Both the Reactive and Follow-up Agents can conclude the current topic.

4.4.2 Orchestrating Progressive Learning Topics. To ensure an orderly deepening of discussions (DG4), we designed various topic types incorporating suggestions from our formative study teachers, roughly corresponding to different levels of Bloom’s Taxonomy [50]. Because some advanced topics require pedagogical orchestration rather than a peer’s perspective, we introduced *Facilitation Messages*. These act similarly to a teacher’s guiding prompts or group announcements, injected into the chat to stimulate peer-wide discussion. The topic types are detailed in Table 1.

The Topic Agent leverages these topic types to dynamically structure the interaction over time. Its prompt encodes the intended pedagogical progression, enabling the agent to select the next topic based on the learner’s demonstrated

Table 1. Taxonomy and types of learning topics.

Taxonomy	Topic Type	Description
Understanding	Sharing and Verification (<i>Peer Message</i>)	A peer shares their artifact in the group and asks if others understand, introducing new strategies and perspectives.
	Clarification Request (<i>Peer Message</i>)	A peer privately asks about the learner’s artifact and invites the learner to explain its meaning or reasoning, encouraging them to externalize their understanding.
Applying	Help Request (<i>Peer Message</i>)	A peer privately seeks help with a flawed artifact, asking the learner to identify errors or demonstrate fixes, thereby creating a chance for learning-by-teaching.
Analyzing	Finding Commonalities (<i>Facilitation Message</i>)	Invites the group to analyze commonalities between two artifacts, converging from surface forms to mathematical essence.
	Comparing Differences (<i>Facilitation Message</i>)	Invites the group to identify differences between two distinct artifacts and evaluate advantages, establishing connections between alternatives.
Evaluating	Overall Evaluation (<i>Facilitation Message</i>)	Invites the group to review all solutions and evaluate the strengths and limitations, and encourages creating new artifacts, fostering knowledge internalization.

understanding, peers’ cognitive configurations, and prior chat history. Guided by this logic, earlier interactions primarily draw on lower-demand peer messages, which provide opportunities to practice mathematical explanation and probe the learner’s understanding. As interaction progresses, deeper analytical topics are introduced to connect multiple solutions and reveal relationships among mathematical strategies. Finally, an evaluation topic is deployed as a concluding discussion to scaffold convergence and bring the learning sequence to a coherent close.

5 User Study

We conducted a user study to answer the following research questions:

RQ1: How do children engage in and experience multimodal peer learning with *Peers4U*?

RQ2: How does peer composition influence children’s engagement and experience?

RQ3: How pedagogically appropriate are the learning topics generated from children’s actual interactions?

We conducted an exploratory, in-person, between-subjects study with children to address RQ1 and RQ2, and a post-hoc evaluation with two expert teachers to answer RQ3.

5.1 Children Participants

We initially recruited 24 children, while two withdrew, and their incomplete records were excluded. The final sample comprised 22 children (11 girls and 11 boys; age $M = 7.00$ years, $SD = 0.67$, range = 6y0m–8y2m, denoted as U1–U22). The study took place during the summer break; 11 children had completed their final kindergarten year and 11 had completed Grade 1. Participant details are provided in Appendix C.1.

Recruitment was conducted via personal accounts on RedNote for in-person sessions in a Chinese city, with parents registering their children voluntarily. Because learning ahead of the school curriculum is common in this context, parents reported their child’s actual mathematics progress. We excluded applicants who had already completed the Grade 2 curriculum to minimize ceiling effects, and filtered excess applicants solely to balance gender and grade. Following IRB approval and parental written consent, children participated voluntarily and could pause or withdraw at any time. Each child received a small toy and a certificate as gifts upon completion.

5.2 Study Setup

We employed a between-subjects design comparing two configurations ($n = 11$ each): **Condition A (Mixed Group)** included two Leading and two Struggling Peers, providing a balanced cognitive distribution; **Condition B (Leading Group)** included four Leading Peers, maximizing exposure to inspiring strategies. In both conditions, peer cognitive profiles and solution artifacts were personalized based on the learner’s initial solutions. Peer appearance, screen position, and voice were randomized for each participant, with two male and two female peer characters per group.

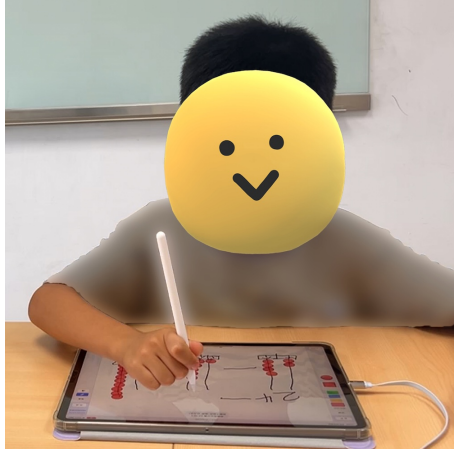


Fig. 6. In-person user-study setup. A child used *Peers4U* on a 13-inch iPad Pro with an Apple Pencil.

All sessions used a 13-inch iPad Pro with an Apple Pencil, as shown in Figure 6. We selected *Subtraction with Regrouping* as the learning task and prepared three task tiers corresponding approximately to entry into Grade 1, the first semester, and the second semester. Each tier paired a focal learning task with a different but comparable post-test task (Appendix C.2). The artifact repository for these tasks was assembled from the OpenCD dataset [113], our formative study, and researcher-authored supplements, with details provided in Appendix C.2; all artifacts were preprocessed using the pipeline described in § 4.3.1.

5.3 Procedure

Each child first completed a guided tutorial covering the multimodal workspace, solution replay, and chats. Then the researcher suggested a task tier based on the child’s grade and parent-reported progress; the child could confirm or adjust this choice. Participants were then assigned to one of the two conditions to balance task tier and gender.

Children then completed one full learning session in *Peers4U*. Their independently authored initial solutions served as the pre-test and as the input for personalizing the four peers. Subsequently, they explored peer solutions and interacted with the peers. The system explicitly notified the user when planned learning topics were completed, though children were free to continue interacting if they wished. Children could also choose to end the lesson or decline any remaining procedures at any time. After the lesson, children completed a questionnaire, which provided a short break from problem solving and then authored solutions to the paired post-test task. A brief structured interview concluded the session.

The researcher remained present during the formal task, but assistance was strictly limited to reading text for children who needed it, correcting speech-recognition transcripts, and answering basic interface queries.

5.4 Data Collection and Analysis

We collected camera and screen recordings, along with system logs capturing the complete pre- and post-test solution processes, additional solutions created during learning, chat messages, and interaction events. The post-study questionnaire captured perceptions of individual peers, social pressure, and comparisons with regular classes. Most ratings used a five-point pictorial Smileyometer [33]. The interview probed perceived learning, preferred solution strategies, experiences of giving and seeking help, and liked or disliked aspects of the experience. The complete instruments are provided in Appendix C.3.

We combined descriptive analyses of logs and questionnaires with structured coding of messages, multimodal artifacts, and interviews. Messages and artifacts were coded using researcher-defined schemes, and structured interview responses were analyzed question by question using qualitative content analysis. Between-condition outcomes were compared using two-sided exact Mann–Whitney U tests, while message-level response comparisons used user-stratified exact permutation tests. Coding procedures and statistical details are provided in Appendices C.4 and D.1.

5.5 Post-hoc Expert Evaluation

To assess the pedagogical quality of system-initiated topics, we conducted an expert evaluation examining (1) whether topic messages were pedagogically appropriate, (2) whether our learner-centered orchestration produced more appropriate topics than a persona-centered agent, and (3) why topics were sometimes inappropriate.

Evaluation Dataset. We evaluated all 157 topic message points from U1–U22. Each paired the learner-centered message delivered during the user study with a persona-centered counterpart generated from the same preceding interaction context. The latter initiated topics from a peer’s own cognitive state and interaction history, representing a vanilla agent-driven approach to peer participation. Further details are provided in Appendix C.5.1.

Expert Participants. We recruited two expert mathematics teachers from different primary schools. E1 had 28 years of teaching experience, including 15 years teaching Grades 1–2; E2 had 29 years, including 5 years teaching Grades 1–2.

Procedure. Using a custom replay interface (Figure 12; Appendix C.5.2), the experts independently and blindly rated the paired topics. After calibration, they reviewed the remaining sessions independently, and reconciled rating differences exceeding one point, while briefly explaining each inappropriate judgment to the researchers (Appendix C.5.3). During independent review, the experts agreed exactly on 65.0% of topics and differed by at most one point on 97.0%; the linearly weighted Cohen’s κ was 0.466. Appendix C.5.3 provides a distribution-sensitive interpretation.

Measures. Experts rated the overall pedagogical appropriateness of each topic—considering both its instructional value and its fit with the learner and ongoing interaction—from 1 (*very inappropriate*) to 5 (*very appropriate*). A rating of 3 or below required a reason, provided through predefined tags or written feedback. After reconciliation, the two ratings were averaged to obtain the final message-level appropriateness score. For the reasons for inappropriateness, we retained all distinct reasons identified by either expert, interpreting them with reference to the experts’ tags, comments, and consensus-meeting explanations.

6 Findings

Our findings are organized around the three research questions. We first characterize how children engage in and experience multimodal peer learning with *Peers4U* (RQ1), then examine differences associated with peer composition (RQ2), and finally evaluate the pedagogical appropriateness of learning topics generated from children’s interactions (RQ3). For readability, we report only p values for statistical comparisons in the main text with full details in Appendix D.1.

6.1 Children’s Engagement and Experience in Multimodal Peer Learning (RQ1)

Children engaged with *Peers4U* in varied but predominantly learning-focused ways, and they generally reported positive experiences with the system; more than half also showed observable positive changes in their post-test artifacts. Their engagement extended beyond chat, as peer interactions elicited some children to create additional artifacts during learning. Conversational participation varied substantially across children and message types, with direct invitations and adaptive follow-ups associated with higher response rates.

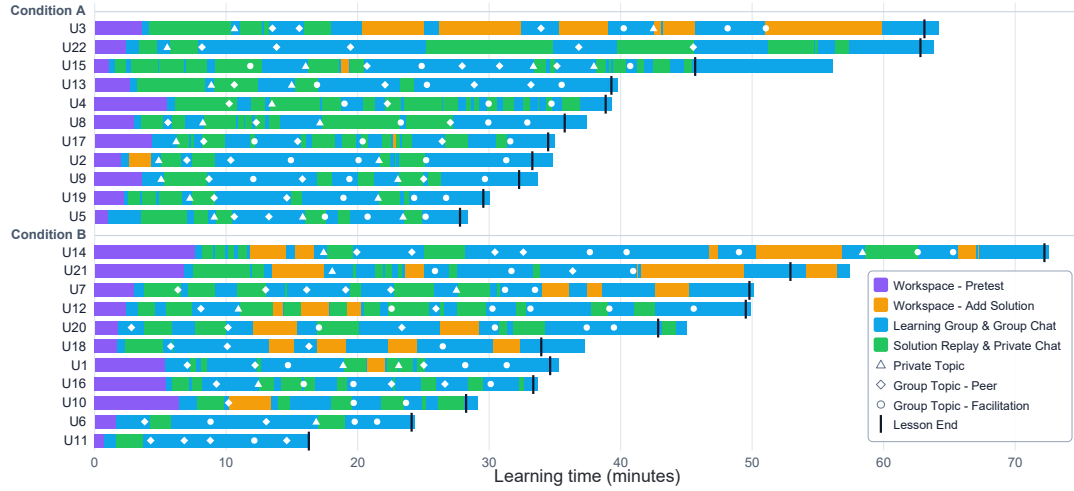


Fig. 7. Interaction timelines for all 22 users, grouped by condition and ordered by total learning time. Colored segments show time spent in different views. White symbols mark different kinds of topics; vertical ticks mark lesson completion, after which users could continue interacting.

6.1.1 Overall Engagement and Experience. All 22 children participated in the learning session, with timelines shown in Figure 7. One child (U11) chose to end it early, while the remaining 21 completed all topics orchestrated by the system. Sessions lasted 41.5 ± 14.3 minutes on average (range: 16.4–72.6), covering 7.1 ± 1.9 learning topics (range: 3–11). Children sent 17.5 ± 16.1 messages per session on average (range: 0–79). **Most messages remained focused on peer learning:** 73.8% concerned mathematical learning content and 14.0% concerned coordinating peer interactions, whereas only 12.2% were off-task casual conversation.

Twenty children (except U11 and U22) subsequently completed the post-test, all producing correct answers. **Notably, 11 children (55%) showed observable positive changes compared to their pre-test artifacts**, including newly adopted calculation strategies (4 children), new visual representations (8 children), or improved correctness (2 children); these categories were non-exclusive. U6, for example, attempted to represent subtraction with manipulatives in the pre-test but did not successfully show regrouping, while in the post-test, he correctly represented it using a strategy previously shown by a peer, later explaining that he was “inspired by others”.

Overall, children exhibited high acceptance of *Peers4U*. Compared with their usual mathematics learning experiences, 17 rated *Peers4U* as more or somewhat more interesting. After finishing the session, some children spontaneously expressed strong interest in future use, asking if they could join this as an extracurricular class or download the application at home, or responding “Great!” when they mistakenly thought another learning session was about to begin.

(U3, U19, U22). U12 expressed a desire to play with the peers in person and, after being told that they were virtual, immediately asked her mother to buy the entire setup.

6.1.2 Peer Interaction Elicited Further Artifact Creation. Peer interaction frequently elicited children to return to mathematical creation rather than remaining within the chat interface. During peer interaction, 11 children voluntarily created 29 additional artifacts ($M = 2.6 \pm 1.5$ among these children, range: 1–5).

These additional artifacts were often closely tied to ongoing peer interaction. We observed three primary contextual triggers: **(1) Peer sharing:** three children (U3, U20, U21) created and immediately shared artifacts right after seeing a peer share one in the group chat; **(2) Topic-prompted evaluation:** three (U7, U14, U21) authored new artifacts in response to the final *Overall Evaluation* topic, which invited them to express what they considered a better approach; and **(3) Peer help-seeking:** three (U1, U12, U15) created artifacts to demonstrate a method while responding to a peer’s request for help. For example, in response to repeated inquiries and scaffolding from a struggling peer, U12 consecutively authored three distinct artifacts, progressing from an equation to a visual demonstration and then to a step-by-step computational diagram. Social comparison could also motivate creation; as U3 remarked before entering the workspace, “I feel like I’m falling behind everyone—I’ve only done two!”

The content of these new artifacts also reflected ongoing exposure to peers: 8 additional artifacts closely resembled a peer solution the child had previously viewed; 2 followed suggestions raised in chat; and 6 improved upon the child’s own earlier work. Thus, peer artifacts functioned not only as objects for discussion, but also as resources for subsequent mathematical production.

6.1.3 Participation Varied with Message Framing. Children’s response rates varied substantially with conversational framing: explicit invitations and progressively structured follow-ups lowered the barrier to participation.

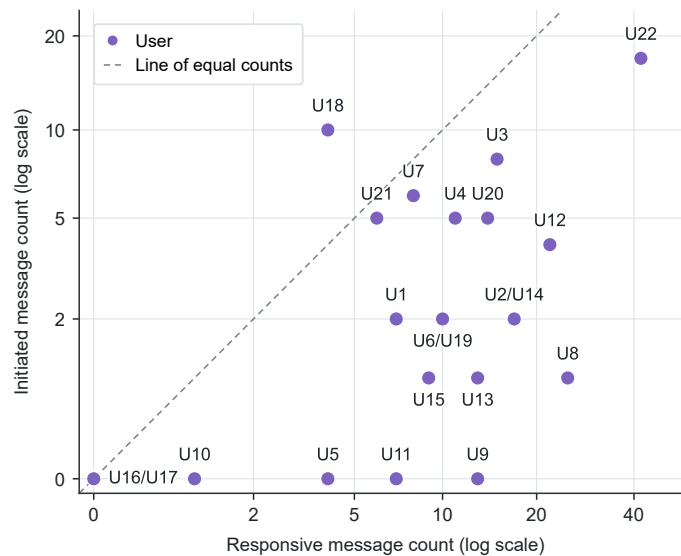


Fig. 8. Learning-related message counts for each child, separated into user-initiated and responsive messages. Both axes use logarithmic scales, and the dashed diagonal denotes equal counts. Most children fall below the diagonal, indicating more responsive than user-initiated messages.

Overall, children exhibited diverse participation styles, ranging from passive observation to largely responsive, and to highly active, self-initiated engagement. As shown in Figure 8, most children sent more responsive messages than self-initiated ones (with U18 being an exception). Children who responded rarely also produced almost no self-initiated messages, whereas initiative only began to emerge among highly responsive children.

Table 2. Learner response rates by message framing. Open Q, Closed Q, Peer, All, and Learner refer to group chat; Private refers to private-chat questions.

<i>Message form</i>	Non-Q	Q	Open Q	Closed Q
	52/275 (18.9%)	187/442 (42.3%)	48/174 (27.6%)	74/183 (40.4%)
<i>Question target</i>	Peer	All	Learner	Private
	0/19 (0.0%)	80/277 (28.9%)	42/61 (68.9%)	65/85 (76.5%)

Across all 717 response-eligible incoming messages from peers and system facilitations, the overall response rate was 33.3% and varied substantially with message characteristics (Table 2). Questions were more likely to receive a response than non-question messages ($p < .001$). Response rates also varied markedly by target: within group chats, questions directed to another peer received no user replies, while questions broadcast to the whole group or directed at the learner received progressively more responses. Private-chat questions reached the highest response rate. Additionally, closed-ended questions in group chats were more likely to receive answers than open-ended questions ($p = .011$).

Moreover, **adaptive follow-ups recovered participation** in many topics that did not initially receive a response. As shown in Figure 9, among 106 eligible group topics, 29.2% received an initial response. For those that did not, the system could issue up to two peer follow-ups; an additional 17.9% received a response to the first follow-up and another 21.7% to the second, resulting in a cumulative participation rate of 68.9%. Across these three turns, the proportion of closed-ended questions increased sharply (10.5% $\rightarrow$ 56.9% $\rightarrow$ 79.2%), as did explicit targeting of the learner (0.0% $\rightarrow$ 2.8% $\rightarrow$ 20.8%). Together, this indicates that dynamically adjusting agency and progressively scaffolding questions from open-ended inquiry to targeted, low-barrier verification can re-engage children.

6.2 Engagement and Experience across Peer Compositions (RQ2)

Children’s additional artifact creation, peer interaction, and perceptions of the group differed across the two peer compositions.

6.2.1 Leading Groups Elicited More Additional Artifact Creation. Children in Condition B (Leading Group) created significantly more additional artifacts than those in Condition A (Mixed Group) ($p = .037$). Descriptively, 8 children in Condition B created 22 additional artifacts, compared with 3 children and 7 artifacts in Condition A. This pattern is consistent with greater exposure to Leading Peers providing children with more material or motivation for further solution creation.

Interestingly, despite being surrounded exclusively by Leading Peers in Condition B, children did not perceive their peers as more mathematically capable, nor did they report greater peer pressure or perceived mathematical inferiority relative to their peers (Appendix D.1). Interview responses suggested that children inferred peer ability from varied and sometimes ambiguous cues, including the neatness and speed of peers’ artifacts and how they interacted. For example, U12 interpreted continued discussion as a sign of lower ability: “I could tell right away—they just kept discussing it.” We further discuss this pattern in § 7.2.

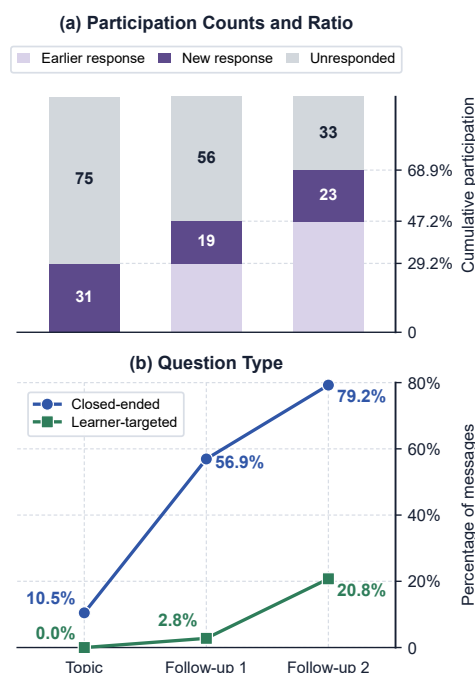


Fig. 9. Participation across 106 group topics and their adaptive follow-ups. (a) Of the 106 topic messages, 31 received a direct response (29.2%); the first and second follow-ups received 19 and 23 additional responses, increasing cumulative participation to 47.2% and 68.9%, respectively. (b) Across successive turns, messages increasingly used closed-ended questions and targeted the learner. These two message characteristics are not mutually exclusive.

6.2.2 Mixed Groups Afforded More Help-oriented Interaction. Condition A, by design, included Struggling Peers and therefore afforded more opportunities for private help-oriented interaction, including *Help Requests* and *Clarification Requests*. Accordingly, the system initiated significantly more private topics in Condition A ($p = .006$), and these sessions also contained significantly more private-chat messages ($p = .029$). Children in Condition A were furthermore significantly more willing to learn with the same peers again ($p = .015$). U15, for example, said she liked a particular peer because that peer had “talked to me first.”

Seven children explicitly described enjoying teaching or helping peers, suggesting that help-seeking Struggling Peers could provide children with a valued contributory role in the group. Teaching could also prompt reflection on their own work. As U8 explained, “At first, I didn’t notice that I was wrong. I only realized it when I was teaching him.”

Together, these results suggest a complementary pattern across peer compositions: Leading Groups were associated with more continued solution creation, whereas Mixed Groups afforded richer one-to-one interaction and were associated with stronger willingness to continue learning with the same peers.

6.3 Pedagogical Appropriateness of Generated Learning Topics (RQ3)

Finally, an expert evaluation assessed the pedagogical appropriateness of topic-initiating messages from *Peers4U*’s learner-centered orchestration and a persona-centered baseline.

Learner-centered topic messages achieved a high overall appropriateness score 4.69 ± 0.58 , with 91.7% regarded as appropriate (i.e., both experts rated ≥ 4). Overall, learner-centered messages were rated significantly more appropriate than their persona-centered alternatives ($p = .0013$; Figure 10a).

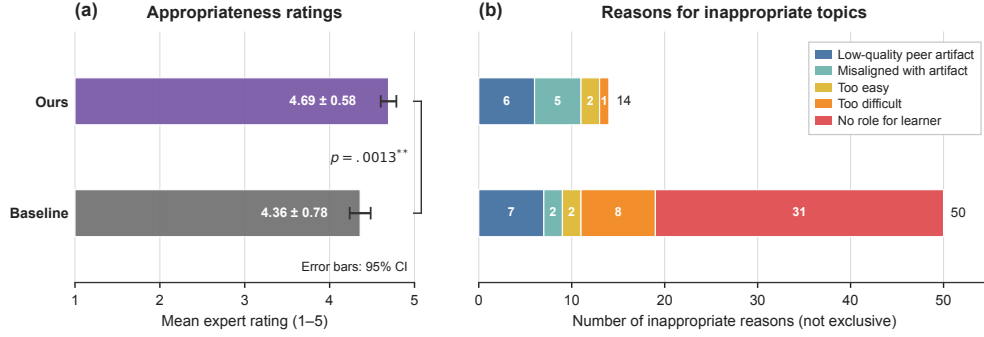


Fig. 10. Post-hoc expert evaluation of learner-centered topic orchestration (Ours) and the persona-centered baseline across 157 topic rounds. (a) Mean expert ratings of pedagogical appropriateness on a five-point scale. (b) Counts of reasons coded by the experts for inappropriate topics; multiple reasons could be assigned to the same topic.

Figure 10b summarizes the non-exclusive reasons for inappropriateness. Among the 13 “inappropriate” learner-centered topics (i.e., at least one rating < 4), the most common reasons concerned the underlying mathematical content: 6 were grounded in low-quality peer artifacts which experts considered inadequate as learning examples, and 5 contained misalignments with underlying artifacts, potentially misleading the learner. These cases primarily reflected limitations in solution quality or in the system’s interpretation of multimodal mathematical artifacts.

In contrast, 43 persona-centered topics were rated as inappropriate, yielding 50 reason counts. Their dominant reason was interactional: 72.1% (31/43) gave the learner little role in the topic, often because one virtual peer initiated a discussion primarily with another peer. Another common reason was excessive difficulty: 8 messages were considered too difficult for the learner because the topic was generated from the initiating peer’s own perspective without accounting for the learner’s current understanding. Thus, persona-centered topics were inappropriate predominantly because of learner involvement and cognitive alignment, whereas *Peers4U*’s inappropriate topics more often reflected content quality and artifact interpretation.

Our analysis also identified valuable forms of persona-centered interaction. Experts explained that, when appropriately phrased, peer-to-peer discussions (e.g., one peer commenting on or asking another for clarification) could still capture the child’s interest and support observational participation even if the learner did not respond directly.

7 Discussion

Drawing on our findings, we discuss two central questions for designing virtual peers in children’s mathematics learning: how peer interaction takes shape, and how it becomes productive for learning. Beyond these, we discuss how artifact-mediated peer learning may extend to broader educational contexts.

7.1 Supporting Peer Interaction through Multimodal Artifacts and Flexible Participation

For young children, both the medium of peer interaction and the available modes of participation shape how accessible and meaningful that interaction becomes. Our findings show how multimodal solution artifacts can ground interaction in observable mathematical thinking, while flexible participation can accommodate children who prefer to initiate, respond, or observe. Children’s emotional experiences also emerged as an important design consideration.

7.1.1 Multimodal Solution Artifacts as the Medium of Peer Interaction. Existing virtual peers are often designed first as “conversational agents”, treating mathematical content merely as information exchanged through dialogue [60, 112]. Our findings suggest that for early mathematics learning, this relationship may be reversed: concrete mathematical solutions can themselves become the primary medium of interaction, with language revolving around these observable, authorable artifacts. Indeed, the act of authoring an artifact is, in itself, a form of communication [21, 34].

In *Peers4U*, children did not merely observe peer solutions; they directed their inquiries and explanations at specific actions, symbols, and strategies embodied in those solutions. More importantly, exposure to peer methods frequently elicited children to imitate, adapt, or completely author new solutions, and then bring these new artifacts back into the interaction. Consequently, mathematical ideas can flow across artifacts through a continuous cycle of observation, explanation, imitation, modification, and sharing [49, 91]. Supporting multimodal mathematical peer learning is thus not simply a matter of adding whiteboards, images, or videos to a chat interface; it requires elevating mathematical artifacts to become the primary shared objects of social interaction.

This perspective also opens a broader design space. Children need not always start from a blank canvas. Future systems could allow them to directly annotate, remix, or build upon a peer’s existing solution, thus making comparison and strategy transfer easier to enact. Likewise, virtual peers need not be restricted to statically pre-assigned solutions. As interaction unfolds, peers could also “author” new artifacts, either through real-time AI generation or dynamic retrieval from a solution repository. Such capabilities could strengthen the *symmetry of action* between learners and peers [20], while further encouraging children to engage in mathematical construction and practice.

We therefore argue that the “peer-like” quality of mathematical virtual peers need not rely only on conversational styles, personas, or anthropomorphic avatars [87]. It can also stem from a symmetry of mathematical agency: both children and peers can author, share, explain, and respond to one another’s solutions. Such *artifact-mediated peer interaction* turns mathematical representations from static learning resources into evolving objects of social interaction, grounding peer learning more directly in concrete mathematical thinking.

7.1.2 Flexible Participation across Initiating, Responding, and Observing. Our findings highlight multiple ways for children to participate in peer learning. Some frequently initiated questions and shared ideas, others primarily responded to explicit invitations, while still others preferred to observe. The critical design question, therefore, is whether the interaction lets children engage in ways that fit their preferences.

First, virtual peers can adjust the form of interaction to alter the threshold for participation. Our results showed that specific invitations directed toward the learner were more likely to receive responses than open, generalized group questions. When initial topics went unanswered, progressively direct and specific follow-ups brought some previously silent children into the discussion. Effective proactive peer behavior therefore depends on modulating the directness and openness of questions to provide accessible entry points. At the same time, observational participation can itself be a meaningful form of engagement [3, 71]. For children who are reluctant to speak, observing questions and explanations exchanged among other peers may also support vicarious learning [95].

Second, the timing of system-initiated interaction matters. While proactive peer messages can stimulate participation, they can also compete for children’s attention and interrupt ongoing mathematical activity [105]. For this reason, *Peers4U* employs a strategy of implicit interaction [15, 86]: the system remains silent when the learner is engaged in active learning behaviors, and gradually takes greater initiative only after a sustained period of operational silence. This design also highlights a broader challenge: because operational silence is inherently ambiguous, learner-centered initiative should be grounded in continuous awareness of the learner’s ongoing activity [55], rather than treating silence uniformly as disengagement.

This suggests that an ideal peer-learning system should personalize interaction through joint models of learners’ mathematical cognition and social interaction styles [7, 54]. Systems could estimate a learner’s initiative tendency, response habits, and topic preferences from prior interactions [83]. For relatively passive children, a system might allow longer periods of observation, avoid repeated prompting, and offer clearer, lower-threshold invitations; for highly active children, it might reduce interventions and preserve more space for self-expression. In this way, different learners could first engage through modes that feel comfortable; over time, the system could gently nudge them toward more diverse forms of participation, helping them take fuller advantage of varied reciprocal learning opportunities, as we will discuss later.

Together, these findings position *flexible participation* as a central design principle. Learner-centered virtual peers should adapt the threshold and timing of their interventions to each child’s ongoing activity and interaction style.

7.1.3 Children’s Emotional Experiences in Virtual Peer Learning. Because young children are still developing emotion regulation [84, 90], virtual peer learning can introduce complex emotional dynamics alongside pedagogical opportunities. In our study, U9 frequently exhaled deeply, showed tension, and responded only passively; he relaxed only after the lesson ended, spending an extended time carefully inspecting the peers’ artifacts. U11 communicated normally at first, but grew increasingly impatient as peer responses and audio continued to appear, eventually covering his ears, stating “I don’t want to listen anymore,” and asking to stop. U15 initially engaged actively in explaining and creating solutions, expressing delight when a peer used the same method. However, when that peer later praised a different approach, U15 sent a blaming message, became angry, and stopped responding across several consecutive topics. She rejoined the chat only after telling her peers that she was angry because “they all said things that did not support her”.

These cases highlight the need for peer-learning systems to respond appropriately to children’s potential emotional reactions. First, systems should enhance affect-awareness for young users through in-system interactions [2] and external signals such as tone of voice, facial expressions, gaze, and body movements [43, 108], provided privacy is preserved [4]. Second, systems should adapt the timing, frequency, and content of messages—for example, by enabling peers to affirm children’s emotions [28], perform relational repairs [19], or acknowledge the value in children’s solutions. Together, affect-aware sensing and interactional adaptation can help young children remain safe and engaged in virtual peer learning.

7.2 Turning Multi-Peer Interaction into Productive Mathematics Learning

Creating rich opportunities for peer interaction does not guarantee that learning will automatically occur. Our research highlights that translating interaction into learning depends on at least three interconnected factors: whether the system creates reciprocal learning opportunities, whether children can recognize the mathematical value of peers’ methods, and whether interactions are orchestrated into a cognitively progressive learning process.

7.2.1 Enabling Reciprocal Learning through Heterogeneous Peers. The value of heterogeneous peers extends beyond presenting solutions at different levels; it lies in providing learners with cognitively complementary peers, thereby creating reciprocal learning opportunities [22, 27, 47]. Leading Peers can introduce novel strategies that push the learner’s boundaries, whereas Struggling Peers can provide opportunities to explain, correct, and help. Crucially, these roles need not be rigidly fixed. A Struggling Peer may still possess a basic strategy that the learner has not used, while a Leading Peer may ask the learner for clarification or assistance on a particular issue.

Virtual peers can therefore support a dynamic form of *reciprocal learning*. The direction of learning—“who learns from whom”—can shift fluidly across different methods and topics. This differs from a typical teachable-agent paradigm, where the child is always cast as the tutor [14, 39]. Here, the contributory role emerges locally and temporarily within a holistic peer relationship of mutual learning, preserving the benefits of learning-by-teaching without permanently positioning participants as stronger or weaker.

This flexibility represents a practical advantage of virtual peers. Physical peer groupings are constrained by children’s existing abilities and social relationships, whereas virtual groups can be tailored to an individual learner and adjust their roles as interaction unfolds. Heterogeneity can thus become a dynamic pedagogical resource within an enduring reciprocal relationship.

7.2.2 Making Virtual Peers’ Expertise Recognizable to Children. Designing virtual peers with greater mathematical expertise does not ensure that children will recognize it. Although every peer in Condition B was designed as a Leading Peer, children did not perceive these peers as having higher ability. Their judgments of peer competence may not be based solely on the mathematical value of the artifacts, but also on social signals such as help-seeking behaviors, question frequency, and conversational style. For instance, some children interpreted peers’ repeated pedagogical questions as a sign of ignorance, concluding that they themselves were the more knowledgeable learners.

This gap matters pedagogically: a novel method can influence learning only when the learner recognizes its value. This may be particularly challenging for younger children, who may rely on their own familiar methods and lack the metacognitive skills to actively compare and adopt new strategies [38, 92]. Systems should therefore construct cognitively useful peers while also making their expertise and learning value perceptible and interpretable to the child.

One possible direction is to leverage richer social and pedagogical signals to help learners identify content worth attending to [29, 51, 98]. This could involve having other peers explicitly endorse a valuable method, introducing lightweight facilitator guidance, or designing clearer expertise cues. Future designs could also integrate teacher-led instruction organically with peer learning, allowing teachers or facilitators to provide scaffolding at critical moments. Yet a strong authoritative presence may undermine the reciprocal and equal character of peer interaction. Balancing horizontal *peer equality* with vertical *pedagogical authority* remains a crucial challenge for future system design.

7.2.3 Orchestrating Peer Interaction along a Cognitive Progression. Pedagogical orchestration remains necessary for moving peer interaction toward deeper mathematical learning. Young children may not spontaneously move from understanding a method toward analyzing, comparing, and transferring it, nor identify common mathematical structures across multiple representations—all of which requires appropriate guidance [46, 75, 76]. An underlying cognitive progression can guide open peer interaction from casual social exchange toward deeper cognitive activity.

This also explains the limitations of allowing virtual peers to initiate topics independently based on their own personas. As our expert evaluation demonstrated, persona-centered topic generation can produce naturalistic, socially plausible peer-to-peer messages, yet such messages frequently fail to provide a clear participatory role for the learner or align with the learner’s current level of understanding.

Multi-peer learning systems therefore need to maintain both *social openness* and *pedagogical structure*. The interaction can feel open, equal, and emergent, while the system orchestrates topics along a cognitive trajectory, moving from observation and comprehension toward application, comparison, evaluation, and eventually, creation and transfer. Such orchestration can evolve from predefined sequences of learning activities [48, 103] into a dynamic process that adapts to the learner’s evolving understanding. This more adaptive form can preserve the openness of peer interaction while guiding children toward deeper conceptual understanding and strategy transfer.

7.3 Extending Artifact-Mediated Peer Learning to Broader Educational Contexts

Although this study focuses on children’s mathematics, artifact-mediated peer learning may extend to broader educational contexts. In more advanced mathematics, learners still construct, interpret, and connect representations to understand abstract concepts [26]. Science learning likewise relies on dynamic representations that make abstract relationships observable. For example, *Augmented Physics* turns static textbook diagrams into interactive simulations of topics such as optics, circuits, and kinematics [32]. Such representations could become shared artifacts that learners and virtual peers manipulate and discuss together.

Scaling this paradigm, however, requires artifacts that are both credible and cognitively diverse. During development, we found that directly generating multimodal problem-solving processes with LLMs often produced overly polished and mature artifacts, missing the irregularities, partial errors, and subtle diversity of human reasoning. Similar limitations have been reported in LLM-based student simulation [109]. We therefore assembled *Peers4U*’s artifact repository from authentic child-produced artifacts and researcher-authored supplements. This approach was manageable for early mathematics, but constructing comparable repositories may become a bottleneck in domains where authentic artifacts are more costly to obtain or produce.

8 Limitations and Future Work

First, the generalizability of our findings is constrained by the small sample of 22 children aged 6–8, recruited from one Chinese city, and by our focus on one mathematics topic. The single, researcher-present laboratory session may also amplify novelty effects and cannot reveal how participation, peer relationships, or interest evolve over time. Longitudinal deployments across populations, topics, and everyday learning settings are needed.

Second, our immediate pre–post task and small between-subjects conditions provide limited evidence of learning effectiveness. Without a learning control condition or delayed measures, we cannot establish durable learning gains or transfer. The persona-centered baseline was evaluated offline by two experts and never entered children’s interactions; an online comparison is required to understand its effects on participation and learning.

Finally, the pedagogical value of peer profiles and artifact assignments produced by our VLM analysis and hand-crafted rules remains unvalidated, while expert-identified low-quality artifacts call for stronger curation. We also held group size and total artifacts constant at four peers and approximately ten artifacts, leaving their effects on social presence and cognitive load unknown. Future work should validate and adapt peer construction, group size, and artifact exposure to support productive learning.

References

- [1] Jacqueline Anton, Giulia Cosentino, Mirko Gelsomini, Kshitij Sharma, Michail N. Giannakos, and Dor Abrahamson. 2025. Mathematics MOVES Me—Digital Solutions for Co-ordinating Enactive and Symbolic Resources: The Case of Positive and Negative Integer Arithmetic. *Digital Experiences in Mathematics Education* 11, 2 (Aug. 2025), 262–275. doi:10.1007/s40751-024-00158-5

- [2] Ivon Arroyo, Beverly Park Woolf, Winslow Burelson, Kasia Muldner, Dovan Rai, and Minghui Tai. 2014. A Multimedia Adaptive Tutoring System for Mathematics That Addresses Cognition, Metacognition and Affect. *International Journal of Artificial Intelligence in Education* 24, 4 (Dec. 2014), 387–426. doi:10.1007/s40593-014-0023-y
- [3] Albert Bandura. 1977. *Social Learning Theory*. Prentice-Hall, Oxford, England. viii, 247 pages.
- [4] Allison Macey Banzon, Jonathan Beever, and Michelle Taub. 2024. Facial Expression Recognition in Classrooms: Ethical Considerations and Proposed Guidelines for Affect Detection in Educational Settings. *IEEE Transactions on Affective Computing* 15, 1 (Jan. 2024), 93–104. doi:10.1109/TAFFC.2023.3275624
- [5] Arthur J Baroody. 2017. The Use of Concrete Experiences in Early Childhood Mathematics Instruction. In *Advances in Child Development and Behavior*. Vol. 53. Elsevier, 43–94.
- [6] Lawrence W. Barsalou. 2008. Grounded Cognition. *Annual Review of Psychology* 59, 1 (Jan. 2008), 617–645. doi:10.1146/annurev.psych.59.103006.093639
- [7] Paul Baxter, Emily Ashurst, Robin Read, James Kennedy, and Tony Belpaeme. 2017. Robot education peers in a situated primary school study: Personalisation promotes child learning. *PLOS ONE* 12, 5 (May 2017), e0178126. doi:10.1371/journal.pone.0178126
- [8] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social Robots for Education: A Review. *Science Robotics* 3, 21 (Aug. 2018), eaat5954. doi:10.1126/scirobotics.aat5954
- [9] Ceylan Beşevli, Elif Salman, Tilbe Goksun, Hakan Urey, and Oğuzhan Özcan. 2019. MaR-T: Designing a Projection-Based Mixed Reality System for Nonsymbolic Math Development of Preschoolers: Guided by Theories of Cognition and Learning. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children (IDC '19)*. Association for Computing Machinery, New York, NY, USA, 280–292. doi:10.1145/3311927.3323147
- [10] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp0630a
- [11] Jerome S. Bruner. 1964. The Course of Cognitive Growth. *American Psychologist* 19, 1 (Jan. 1964), 1–15. doi:10.1037/h0044160
- [12] Jerome Seymour Bruner. 1974. *Toward a Theory of Instruction*. Harvard university press.
- [13] J. Cassell, M. Ananny, A. Basu, T. Bickmore, P. Chong, D. Mellis, K. Ryokai, J. Smith, H. Vilhjálmsson, and H. Yan. 2000. Shared Reality: Physical Collaboration with a Virtual Peer. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems (CHI EA '00)*. Association for Computing Machinery, New York, NY, USA, 259–260. doi:10.1145/633292.633443
- [14] Catherine C. Chase, Doris B. Chin, Marily A. Oppezzo, and Daniel L. Schwartz. 2009. Teachable Agents and the Protégé Effect: Increasing the Effort Towards Learning. *Journal of Science Education and Technology* 18, 4 (Aug. 2009), 334–352. doi:10.1007/s10956-009-9180-4
- [15] Longfei Chen, Shenghan Gao, Shiwei Wang, Ken Lin, Yun Wang, Yan Lu, and Quan Li. 2026. When Systems Take Initiative: A Design Framework for Adaptive, Mixed-initiative Database Querying. In *Proceedings of the 2026 Designing Interactive Systems Conference (DIS '26)*. Association for Computing Machinery, New York, NY, USA, 1170–1187. doi:10.1145/3800645.3812906
- [16] Marie A. Cooper. 1999. Classroom Choices From a Cognitive Perspective on Peer Learning. In *Cognitive Perspectives on Peer Learning*. Routledge.
- [17] Giulia Cosentino, Jacqueline Anton, Kshitij Sharma, Mirko Gelsomini, Michail Giannakos, and Dor Abrahamson. 2025. Exploring Children's Embodied Interactions through Digitally Facilitated Enactment: A Case Study When Math Education MOVES. *Int. J. Hum.-Comput. Stud.* 201, C (June 2025). doi:10.1016/j.ijhcs.2025.103509
- [18] William Damon. 1984. Peer Education: The Untapped Potential. *Journal of Applied Developmental Psychology* 5, 4 (Oct. 1984), 331–343. doi:10.1016/0193-3973(84)90006-6
- [19] Aayushi Dangol, Meghna Gupta, Daeun Yoo, Robert Wolfe, Jason Yip, Franziska Roesner, and Julie A. Kientz. 2026. Toys That Listen, Talk, and Play: Understanding Children's Sensemaking and Interactions with AI Toys. In *Proceedings of the 25th Annual ACM Interaction Design and Children Conference (IDC '26)*. Association for Computing Machinery, New York, NY, USA, 818–834. doi:10.1145/3773077.3806142
- [20] P. Dillenbourg. 1999. What Do You Mean by Collaborative Learning.
- [21] Pierre Dillenbourg and David Traum. 2006. Sharing Solutions: Persistence and Grounding in Multimodal Collaborative Problem Solving. *Journal of the Learning Sciences* 15, 1 (Jan. 2006), 121–151. doi:10.1207/s15327809jls1501_9
- [22] David Duran and Charles Monereo. 2005. Styles and Sequences of Cooperative Interaction in Fixed and Reciprocal Peer Tutoring. *Learning and Instruction* 15, 3 (June 2005), 179–199. doi:10.1016/j.learninstruc.2005.04.002
- [23] Kelley Durkin, Jon R. Star, and Bethany Rittle-Johnson. 2017. Using Comparison of Multiple Strategies in the Mathematics Classroom: Lessons Learned and next Steps. *ZDM* 49, 4 (Aug. 2017), 585–597. doi:10.1007/s11858-017-0853-9
- [24] John Fantuzzo and Marika Ginsburg-Block. 1998. Reciprocal Peer Tutoring: Developing and Testing Effective Peer Collaborations for Elementary School Students. In *Peer-Assisted Learning*. Routledge.
- [25] Nicole L. Fonger. 2019. Meaningfulness in Representational Fluency: An Analytic Lens for Students' Creations, Interpretations, and Connections. *The Journal of Mathematical Behavior* 54 (June 2019), 100678. doi:10.1016/j.jmathb.2018.10.003
- [26] Nicole L. Fonger. 2019. Meaningfulness in Representational Fluency: An Analytic Lens for Students' Creations, Interpretations, and Connections. *The Journal of Mathematical Behavior* 54 (June 2019), 100678. doi:10.1016/j.jmathb.2018.10.003
- [27] Ellice A. Forman and Courtney B. Cazden. 1998. Exploring Vygotskian Perspectives in Education: The Cognitive Value of Peer Interaction *. In *Learning Relationships in the Classroom*. Routledge.

- [28] Elin Marie Frivold Kostøl and David Lansing Cameron. 2021. Teachers' Responses to Children in Emotional Distress: A Study of Co-Regulation in the First Year of Primary School in Norway. *Education 3-13* 49, 7 (Oct. 2021), 821–831. doi:10.1080/03004279.2020.1800062
- [29] Maria Fusaro and Paul L. Harris. 2008. Children Assess Informant Reliability Using Bystanders' Non-Verbal Cues. *Developmental Science* 11, 5 (Sept. 2008), 771–777. doi:10.1111/j.1467-7687.2008.00728.x
- [30] Karen C. Fuson, Diana Wearne, James C. Hiebert, Hanlie G. Murray, Pieter G. Human, Alwyn I. Olivier, Thomas P. Carpenter, and Elizabeth Fennema. 1997. Children's Conceptual Structures for Multidigit Numbers and Methods of Multidigit Addition and Subtraction. *Journal for Research in Mathematics Education* 28, 2 (March 1997), 130. doi:10.2307/749759
- [31] Cornelia S. Große and Alexander Renkl. 2006. Effects of Multiple Solution Methods in Mathematics Learning. *Learning and Instruction* 16, 2 (April 2006), 122–138. doi:10.1016/j.learninstruc.2006.02.001
- [32] Aditya Gunturu, Yi Wen, Nandi Zhang, Jarin Thundathil, Rubaiat Habib Kazi, and Ryo Suzuki. 2024. 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, Pittsburgh PA USA, 1–12. doi:10.1145/3654777.3676392
- [33] Lynne Hall, Colette Hume, and Sarah Tazzyman. 2016. Five Degrees of Happiness: Effective Smiley Face Likert Scales for Evaluating with Children. In *Proceedings of the The 15th International Conference on Interaction Design and Children (IDC '16)*. Association for Computing Machinery, New York, NY, USA, 311–321. doi:10.1145/2930674.2930719
- [34] S. Hennessy. 2011. The Role of Digital Artefacts on the Interactive Whiteboard in Supporting Classroom Dialogue: Dialogue and IWB Artefacts. *Journal of Computer Assisted Learning* 27, 6 (Dec. 2011), 463–489. doi:10.1111/j.1365-2729.2011.00416.x
- [35] Hui-Ru Ho, Bengisu Cagiltay, Nathan Thomas White, Edward M. Hubbard, and Bilge Mutlu. 2021. RoboMath: Designing a Learning Companion Robot to Support Children's Numerical Skills. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference (IDC '21)*. Association for Computing Machinery, New York, NY, USA, 283–293. doi:10.1145/3459990.3460709
- [36] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. doi:10.1145/302979.303030
- [37] Christine Howe. 2009. *Peer Groups and Children's Development*. John Wiley & Sons.
- [38] Junichi Ishida. 2002. Students' Evaluation of Their Strategies When They Find Several Solution Methods. *The Journal of Mathematical Behavior* 21, 1 (Jan. 2002), 49–56. doi:10.1016/S0732-3123(02)00102-5
- [39] Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. 2024. Teach AI How to Code: Using Large Language Models as Teachable Agents for Programming Education. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–28. doi:10.1145/3613904.3642349
- [40] Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. 2025. TeachTune: Reviewing Pedagogical Agents Against Diverse Student Profiles with Simulated Students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–28. doi:10.1145/3706598.3714054
- [41] David W. Johnson and Roger T. Johnson. 2009. An Educational Psychology Success Story: Social Interdependence Theory and Cooperative Learning. *Educational Researcher* 38, 5 (2009), 365–379. doi:10.3102/0013189X09339057
- [42] Seokbin Kang, Ekta Shokeen, Virginia L. Byrne, Leyla Norooz, Elizabeth Bonsignore, Caro Williams-Pierce, and Jon E. Froehlich. 2020. ARMath: Augmenting Everyday Life with Math Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. doi:10.1145/3313831.3376252
- [43] Ashish Kapoor and Rosalind W. Picard. 2005. Multimodal Affect Recognition in Learning Environments. In *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA '05)*. Association for Computing Machinery, New York, NY, USA, 677–682. doi:10.1145/1101149.1101300
- [44] Mina Khan, Fernando Trujano, Ashris Choudhury, and Pattie Maes. 2018. Mathland: Playful Mathematical Learning in Mixed Reality. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3170427.3186499
- [45] Kyuwon Kim, Jaeryeong Hwang, Younseo Lee, Jeanhee Lee, Sung-Eun Kim, and Hyo-Jeong So. 2025. Your Thoughtful Opponent: Embracing Cognitive Conflict with Peer Agent. In *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '25)*. Association for Computing Machinery, New York, NY, USA, 1–3. doi:10.1145/3746058.3758410
- [46] Alison King. 1997. ASK to THINK-TEL WHY: A Model of Transactive Peer Tutoring for Scaffolding Higher Level Complex Learning. *Educational Psychologist* 32, 4 (Sept. 1997), 221–235. doi:10.1207/s15326985ep3204_3
- [47] Cornelia Kneser and Rolf Plöetzner. 2001. Collaboration on the Basis of Complementary Domain Knowledge: Observed Dialogue Structures and Their Relation to Learning Success. *Learning and Instruction* 11, 1 (Feb. 2001), 53–83. doi:10.1016/S0959-4752(00)00015-3
- [48] Ingo Kollar, Frank Fischer, and Friedrich W. Hesse. 2006. Collaboration Scripts – A Conceptual Analysis. *Educational Psychology Review* 18, 2 (June 2006), 159–185. doi:10.1007/s10648-006-9007-2
- [49] Timothy D. Koschmann. 1996. *CSCL, Theory and Practice of an Emerging Paradigm*. Routledge.
- [50] David R. Krathwohl. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice* 41, 4 (2002), 212–218. jstor:1477405
- [51] Jonathan D. Lane and Paul L. Harris. 2015. The Roles of Intuition and Informants' Expertise in Children's Epistemic Trust. *Child development* 86, 3 (May 2015), 919–926. doi:10.1111/cdev.12324

- [52] Yew Hoong Leong, Weng Kin Ho, and Lu Pien Cheng. 2015. Concrete-Pictorial-Abstract: Surveying Its Origins and Charting Its Future. *The Mathematics Educator* 16, 1 (2015), 1–18.
- [53] Richard Lesh, Tom Post, and Merlyn Behr. 1987. Representations and Translations among Representations in Mathematics Learning and Problem Solving. In *Problems of Representation in the Teaching and Learning of Mathematics*. Routledge.
- [54] Mike E. U. Ligthart, Mark A. Neerincx, and Koen V. Hindriks. 2022. Getting Acquainted: First Steps for Child-Robot Relationship Formation. *Frontiers in Robotics and AI* 9 (Sept. 2022). doi:10.3389/frobt.2022.853665
- [55] Ailin Liu, Yesmine Karoui, Fiona Draxler, Frauke Kreuter, and Francesco Chiossi. 2026. Sensing What Surveys Miss: Understanding and Personalizing Proactive LLM Support by User Modeling. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3772318.3791191
- [56] Jiawen Liu, Yuanyuan Yao, Pengcheng An, and Qi Wang. 2024. PeerGPT: Probing the Roles of LLM-based Peer Agents as Team Moderators and Participants in Children's Collaborative Learning. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3613905.3651008
- [57] Yujia Liu, Siyu Zha, Yuewen Zhang, Yanjin Wang, Yangming Zhang, Qi Xin, Lun Yiu Nie, Chao Zhang, and Yingqing Xu. 2025. BrickSmart: Leveraging Generative AI to Support Children's Spatial Language Learning in Family Block Play. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–19. doi:10.1145/3706598.3714212
- [58] Ziyi Liu, Zhengzhe Zhu, Lijun Zhu, Enze Jiang, Xiyun Hu, Kylie A Peppler, and Karthik Ramani. 2024. ClassMeta: Designing Interactive Virtual Classmate to Promote VR Classroom Participation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3613904.3642947
- [59] Bailing Lyu, Chenglu Li, Hai Li, and Rui Guo. 2026. Designing AI Teachable Agents with Personality: Supporting Student Emotions in Mathematics Learning. In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference (LAK '26)*. Association for Computing Machinery, New York, NY, USA, 868–875. doi:10.1145/3785022.3785078
- [60] Wenhao Lyu, Yimeng Wang, Murong Yue, Yifan Sun, Jennifer Suh, Meredith Kier, Ziyu Yao, and Yixuan Zhang. 2026. Designing AI Peers for Collaborative Mathematical Problem Solving with Middle School Students: A Participatory Design Study. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3772318.3791138
- [61] Patricia S. Moyer, Johnna J. Bolyard, and Mark A. Spikell. 2002. What Are Virtual Manipulatives? *Teaching Children Mathematics* 8, 6 (2002), 372–77.
- [62] Mitchell J Nathan, Ana C Stephens, DK Masarik, Martha W Alibali, and Kenneth R Koedinger. 2002. Representational Fluency in Middle School: A Classroom Study. In *Proceedings of the Twenty-Fourth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, Vol. 1. ERIC Clearinghouse for Science, Mathematics and Environmental Education ..., 462–472.
- [63] Oi-Lam Ng. 2016. The Interplay between Language, Gestures, Dragging and Diagrams in Bilingual Learners' Mathematical Communications. *Educational Studies in Mathematics* 91, 3 (March 2016), 307–326. doi:10.1007/s10649-015-9652-9
- [64] Erin R. Ottmar, Candace Walkington, Dor Abrahamson, Mitchell J. Nathan, Avery Harrison, and Carmen Smith. 2019. *Embodied Mathematical Imagination and Cognition (EMIC) Working Group*. Technical Report. North American Chapter of the International Group for the Psychology of Mathematics Education.
- [65] Sitong Pan, Robin Schmucker, Bernardo Garcia Bulle Bueno, Salome Aguilar Llanes, Fernanda Albo Alarcón, Hangxiao Zhu, Adam Teo, and Meng Xia. 2025. TutorUp: What If Your Students Were Simulated? Training Tutors to Address Engagement Challenges in Online Learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3706598.3713589
- [66] Stephen J. Pape and Mourat A. Tchoshanov. 2001. The Role of Representation(s) in Developing Mathematical Understanding. *Theory Into Practice* 40, 2 (2001), 118–127. jstor:1477273
- [67] Hae Won Park, Rinat Rosenberg-Kima, Maor Rosenberg, Goren Gordon, and Cynthia Breazeal. 2017. Growing Growth Mindset with a Social Robot Peer. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*. Association for Computing Machinery, New York, NY, USA, 137–145. doi:10.1145/2909824.3020213
- [68] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3586183.3606763
- [69] Jean Piaget. 2015. *The Grasp of Consciousness (Psychology Revivals): Action and Concept in the Young Child*. Psychology Press.
- [70] Taciana Pontual Falcão, Tanja Dackermann, Madeleine Schüler, Christine Ulrich, Andre Klemke, and Korbinian Moeller. 2018. Tangible Tens: Evaluating a Training of Basic Numerical Competencies with an Interactive Tabletop. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3174125
- [71] Yingxiao Qian, Yi-Chun Hong, and Michelene Chi. 2024. Learning from Watching Dialog and Monolog Videos in Online STEM Courses. *International Journal of STEM Education* 11, 1 (Sept. 2024), 49. doi:10.1186/s40594-024-00505-3
- [72] Emilee Rader, Margaret Echelbarger, and Justine Cassell. 2011. Brick by Brick: Iterating Interventions to Bridge the Achievement Gap with Virtual Peers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2971–2974. doi:10.1145/1978942.1979382

- [73] Martina A. Rau, Vincent Aleven, and Nikol Rummel. 2009. Intelligent Tutoring Systems with Multiple Representations and Self-Explanation Prompts Support Learning of Fractions. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*. IOS Press, NLD, 441–448.
- [74] M. A. Rau, V. Aleven, N. Rummel, and Z. Pardos. 2014. How Should Intelligent Tutoring Systems Sequence Multiple Graphical Representations of Fractions? A Multi-Methods Study. *International Journal of Artificial Intelligence in Education* 24, 2 (June 2014), 125–161. doi:10.1007/s40593-013-0011-7
- [75] Bethany Rittle-Johnson. 2007. Does Comparing Solution Methods Facilitate Conceptual and Procedural Knowledge? An Experimental Study on Learning to Solve Equations. *Journal of Educational Psychology* 99 (Aug. 2007), 561–574. doi:10.1037/0022-0663.99.3.561
- [76] Bethany Rittle-Johnson, Abbey M. Loehr, and Kelley Durkin. 2017. Promoting Self-Explanation to Improve Mathematics Learning: A Meta-Analysis and Instructional Design Principles. *ZDM: The International Journal on Mathematics Education* 49, 4 (Aug. 2017), 599–611. doi:10.1007/s11858-017-0834-z
- [77] Kantwon Rogers, Michael Davis, Mallesh Maharana, Pete Etheredge, and Sonia Chernova. 2025. Playing Dumb to Get Smart: Creating and Evaluating an LLM-based Teachable Agent within University Computer Science Classes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3706598.3713644
- [78] Jeremy Roschelle, James J. Kaput, and Walter Stroup. 2000. Simcalc: Accelerating Students' Engagement With the Mathematics of Change. In *Innovations in Science and Mathematics Education*. Routledge.
- [79] Edger P. Rutatola, Elina C. Ntahomvukye, Koen Stroeken, and Tony Belpaeme. 2026. Robot Tutors or Peers? Evaluating Math Learning and Conformity with LLM-Powered Robots in Tanzanian Primary Schools. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26)*. Association for Computing Machinery, New York, NY, USA, 237–245. doi:10.1145/3757279.3785594
- [80] Nihar Sabnis and Tomohiro Nagashima. 2024. Empowering Learners: Chatbot-Mediated 'Learning-by-Teaching'. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3613905.3650754
- [81] Nazmus Saquib, Rubaiat Habib Kazi, Li-yi Wei, Gloria Mark, and Deb Roy. 2021. Constructing Embodied Algebra by Sketching. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. doi:10.1145/3411764.3445460
- [82] Julie Sarama and Douglas H. Clements. 2016. Physical and Virtual Manipulatives: What Is "Concrete"? In *International Perspectives on Teaching and Learning Mathematics with Virtual Manipulatives*, Patricia S. Moyer-Packenham (Ed.). Springer International Publishing, Cham, 71–93. doi:10.1007/978-3-319-32718-1_4
- [83] Silvia Schiaffino and Analia Amandi. 2004. User – Interface Agent Interaction: Personalization Issues. *International Journal of Human-Computer Studies* 60, 1 (Jan. 2004), 129–148. doi:10.1016/j.ijhcs.2003.09.003
- [84] Juliane Schlesier, Ingo Roden, and Barbara Moschner. 2019. Emotion Regulation in Primary School Children: A Systematic Review. *Children and Youth Services Review* 100 (May 2019), 239–257. doi:10.1016/j.childyouth.2019.02.044
- [85] Daniel L. Schwartz. 1995. The Emergence of Abstract Representations in Dyad Problem Solving. *Journal of the Learning Sciences* 4, 3 (July 1995), 321–354. doi:10.1207/s15327809jls0403_3
- [86] Barış Serim and Giulio Jacucci. 2019. Explicating "Implicit Interaction": An Examination of the Concept and Challenges for Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3290605.3300647
- [87] Narek Shamamyan, Man Su, and Tomohiro Nagashima. 2025. Co-Designing Trustworthy Peer Agents with Middle-School Students. In *Proceedings of the 24th Interaction Design and Children (IDC '25)*. Association for Computing Machinery, New York, NY, USA, 535–544. doi:10.1145/3713043.3728847
- [88] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role Play with Large Language Models. *Nature* 623, 7987 (Nov. 2023), 493–498. doi:10.1038/s41586-023-06647-8
- [89] Robert S Siegler. 1984. Strategy Choices in Addition and Subtraction: How Do Children Know What to Do? *Origins of cognitive skills* (1984).
- [90] Judith Rebecca Silkenbeumer, Eva-Maria Schiller, and Joscha Kärtner. 2018. Co- and Self-Regulation of Emotions in the Preschool Setting. *Early Childhood Research Quarterly* 44 (July 2018), 72–81. doi:10.1016/j.ecresq.2018.02.014
- [91] G. Stahl, S. Ludvigsen, N. Law, and U. Cress. 2014. CSCL Artifacts. *International Journal of Computer-Supported Collaborative Learning* 9, 3 (Sept. 2014), 237–245. doi:10.1007/s11412-014-9200-0
- [92] Elsbeth Stern. 1992. Spontaneous Use of Conceptual Mathematical Knowledge in Elementary School Children. *Contemporary Educational Psychology* 17, 3 (July 1992), 266–277. doi:10.1016/0361-476X(92)90065-7
- [93] Sofia Tancredi. 2024. Balance Board Math: Exploring the Sense of Balance as a Basis for Functions and Graphing and Number Line Concepts. *Digital Experiences in Mathematics Education* 10, 2 (Aug. 2024), 202–227. doi:10.1007/s40751-024-00140-1
- [94] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2022. Authoring Virtual Peer Interactions for Lecture Videos. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3491101.3519835
- [95] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2023. Scripted Vicarious Dialogues: Educational Video Augmentation Method for Increasing Isolated Students' Engagement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–25. doi:10.1145/3544548.3581153

- [96] Harriet R. Tenenbaum, Naomi E. Winstone, Patrick J. Leman, and Rachel E. Avery. 2020. How Effective Is Peer Interaction in Facilitating Learning? A Meta-Analysis. *Journal of Educational Psychology* 112, 7 (Oct. 2020), 1303. doi:10.1037/edu0000436
- [97] Satomi Tokida, Koki Usui, Godai Tanaka, Shin Osuga, Masanori Kayano, Yuta Itoh, and Yoshio Ishiguro. 2026. Design and Evaluation of a Photorealistic AI Virtual Peer in Elementary Collaborative Classroom. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3772318.3791450
- [98] Yu Tong, Fuxing Wang, and Judith Danovitch. 2020. The Role of Epistemic and Social Characteristics in Children's Selective Trust: Three Meta-Analyses. *Developmental Science* 23, 2 (2020), e12895. doi:10.1111/desc.12895
- [99] K. J. Topping. 1996. The Effectiveness of Peer Tutoring in Further and Higher Education: A Typology and Review of the Literature. *Higher Education* 32, 3 (Oct. 1996), 321–345. doi:10.1007/BF00138870
- [100] Keith J. Topping. 2005. Trends in Peer Learning. *Educational Psychology* 25, 6 (Dec. 2005), 631–645. doi:10.1080/01443410500345172
- [101] Eva Vass and Karen Littleton. 2010. Peer Collaboration and Learning in the Classroom. Leeds, 105–136.
- [102] Olga Viberg, Patrik Larsson, Elias Hedlin, and Olov Engwall. 2026. Co-Designing AI-mediated Paired Reading for Children in Swedish Classrooms. In *Proceedings of the 25th Annual ACM Interaction Design and Children Conference (IDC '26)*. Association for Computing Machinery, New York, NY, USA, 1266–1272. doi:10.1145/3773077.3812154
- [103] Freydis Vogel, Christof Wecker, Ingo Kollar, and Frank Fischer. 2017. Socio-Cognitive Scaffolding with Computer-Supported Collaboration Scripts: A Meta-Analysis. *Educational Psychology Review* 29, 3 (Sept. 2017), 477–511. doi:10.1007/s10648-016-9361-7
- [104] Chenyang Wang, Filippo Forte, Léane Wettstein, Pierre Dillenbourg, and Thiméo Wambsganss. 2026. ArgueMate: Designing an Arguing Agent with Maximised Disagreement to Support Student Peer-Argumentation Exercise. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3772363.3799300
- [105] Duo Wang, Wenxuan Song, Jianwei Ni, Qingxiao Zheng, Yuhang Zhou, Kyrian Liang, Mike Yao, and Caroline G. L. Cao. 2026. Should the AI Speak First? Evaluating Proactive vs. Reactive Facilitation in Mixed-Reality Medical Training. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3772318.3791910
- [106] Tianjia Wang, Tong Wu, Huayi Liu, Chris Brown, and Yan Chen. 2025. Generative Co-Learners: Enhancing Cognitive and Social Presence of Students in Asynchronous Learning with Generative AI. *Proc. ACM Hum.-Comput. Interact.* 9, 1 (Jan. 2025), GROUP19:1–GROUP19:24. doi:10.1145/3701198
- [107] Tobin White and Roy Pea. 2011. Distributed by Design: On the Promises and Pitfalls of Collaborative Learning with Multiple Representations. *Journal of the Learning Sciences* 20, 3 (July 2011), 489–547. doi:10.1080/10508406.2010.542700
- [108] Beverly Woolf, Winslow Burleson, Ivon Arroyo, Toby Dragon, David Cooper, and Rosalind Picard. 2009. Affect-Aware Tutors: Recognising and Responding to Student Affect. *International Journal of Learning Technology* 4, 3/4 (2009), 129. doi:10.1504/IJLT.2009.028804
- [109] Tao Wu, Jingyuan Chen, Wang Lin, Mengze Li, Yumeng Zhu, Ang Li, Kun Kuang, and Fei Wu. 2025. Embracing Imperfection: Simulating Students with Diverse Cognitive Levels Using LLM-based Agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 9887–9908. doi:10.18653/v1/2025.acl-long.488
- [110] Wanli Xing, Yukyeong Song, Chenglu Li, Zifeng Liu, Wangda Zhu, and Hyunju Oh. 2025. Development of a Generative AI-powered Teachable Agent for Middle School Mathematics Learning: A Design-Based Research Study. *British Journal of Educational Technology* 56, 5 (2025), 2043–2077. doi:10.1111/bjet.13586
- [111] Ji-Fan Yu, Daniel Zhang-Li, Zhe-Yuan Zhang, Yu-Cheng Wang, Hao-Xuan Li, Joy Jia Yin Lim, Zhan-Xin Hao, Shang-Qing Tu, Lu Zhang, Xu-Sheng Dai, Jian-Xiao Jiang, Shen Yang, Fei Qin, Ze-Kun Li, Bing-Lin Liu, Xin Cong, Bin Xu, Lei Hou, Man-Li Li, Juan-Zi Li, Hui-Qin Liu, Yu Zhang, Zhi-Yuan Liu, and Mao-Song Sun. 2026. From MOOC to MAIC: Reimagine Online Teaching and Learning Through LLM-Driven Agents. *Journal of Computer Science and Technology* 41, 1 (April 2026), 394–414. doi:10.1007/s11390-025-6000-0
- [112] Murong Yue, Wenhan Lyu, Jennifer Suh, Yixuan Zhang, and Ziyu Yao. 2025. MathVC: An LLM-Simulated Multi-Character Virtual Classroom for Mathematics Education. arXiv:2404.06711 [cs.CL] doi:10.48550/arXiv.2404.06711
- [113] Zhi Zheng, Chun Yu, Weihao Chen, Minzheng Song, Binglin Liu, Jianyang Liu, Shiyi Wang, Xutong Wang, Jie Cai, and Yuanchun Shi. 2026. OpenCD: Empowering Diagnosis of Children's Mathematical Cognition through Open-ended Multimodal Tasks. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, Barcelona Spain, 1–31. doi:10.1145/3772318.3790318

A Formative Study

A.1 Mathematics Content and Tasks

The formative study covered three mathematics topics, each at three progressively harder tiers, resulting in nine tasks (Table 3).

Table 3. Mathematics Content and Tasks Used in the Formative Study.

Content	Task
Number Representation	
Numbers within 20	“How can you represent the number 17?”
Numbers within 100	“How can you represent the number 46?”
Numbers within 1,000	“How can you represent the number 358?”
Addition with Regrouping	
Addition within 20	“How can you calculate $6 + 8$?”
Addition within 30	“How can you calculate $15 + 7$?”
Addition within 100	“How can you calculate $27 + 14$?”
Subtraction with Regrouping	
Subtraction within 20	“How can you calculate $13 - 7$?”
Subtraction within 30	“How can you calculate $24 - 15$?”
Subtraction within 100	“How can you calculate $45 - 28$?”

A.2 Probe Implementation

Peer-solution configurations. To expose participants to different forms of peer heterogeneity, we implemented three configurations for presenting peer solutions: (1) a developmental configuration that followed the developmental progression of the underlying mathematical strategies; (2) a diversity configuration that maximized variation among the presented artifacts; and (3) a researcher-selected configuration that enabled targeted presentation of unusual, incomplete, or flawed solutions. In the developmental configuration, the probe mapped the child’s submitted solutions to a task-specific cognitive graph adapted from OpenCD [113], prioritized artifacts near the child’s next developmental steps, and updated its estimate as artifacts were viewed. The diversity configuration represented solutions by their mathematical behavior features and selected artifacts dissimilar to both the child’s solutions and previously presented artifacts. These two configurations initially displayed four artifacts and added two when the researcher requested more. In the researcher-selected configuration, flawed solutions were marked as needing help and opened with a short, pre-scripted request for assistance before the conversation began.

Virtual-peer interaction. The probe used one artifact-grounded Chat Agent based on Gemini 3 Flash Preview (temperature 0.7). For each turn, it received the associated task, the peer artifact’s complete multimodal construction process and analysis summary, and the conversation history, and produced a concise reply from a peer persona aligned with the mathematical understanding reflected in that artifact. Children’s speech was transcribed with Volcengine streaming ASR, and peer replies were streamed as speech using ElevenLabs’ eleven_v3 model and a childlike voice. To further de-identify the original children’s narration during artifact replay, its pitch was shifted.

A.3 Child Participants

Table 4 presents the demographic details of the child participants.

Task and configuration exposure. Table 5 summarizes which configurations each child encountered for each task. We varied configurations to elicit reactions to contrasting peer sets within this exploratory probe.

Table 4. Demographics of Child Participants in the Formative Study.

ID	Grade	Gender	Age	ID	Grade	Gender	Age
S1	K	Male	5y11m	S5	1	Male	7y5m
S2	1	Male	7y4m	S6	1	Female	7y2m
S3	K	Female	6y5m	S7	K	Male	5y11m
S4	1	Female	6y11m	S8	K	Female	6y4m

Table 5. Peer-solution configurations encountered by each child and task. D = developmental; V = diversity-maximizing; R = researcher-selected; – = task not attempted.

ID	Number Repr.			Addition			Subtraction		
	17	46	358	6 + 8	15 + 7	27 + 14	13 – 7	24 – 15	45 – 28
S1	R	D	R	V	R	–	–	V	–
S2	R	–	R	–	–	D	–	–	–
S3	V	R	–	–	R	D	–	D	–
S4	–	V	R	–	–	D	–	–	D
S5	–	V	R	–	V	–	–	–	V
S6	–	V	R	–	–	–	–	–	–
S7	R	V	D	–	V	–	–	–	–
S8	D	R	–	R	–	–	V	–	–

A.4 Teacher Participants

The professional backgrounds of the teacher participants are detailed in Table 6.

Table 6. Profiles of Teacher Participants in the Formative Study.

ID	Teaching Exp. (years)	Grades 1–2 Exp. (years)	Gender	School Type
T1	9	4	Male	Public school
T2	9	3	Male	Tutoring institution
T3	14	7	Female	Private school
T4	17	17	Female	Public school
T5	7	5	Male	Private school

B System Design

B.1 Artifact Analysis Agents

Artifact preprocessing used the two VLM agents described in the main text. The Solution Artifact Comprehension Agent received the complete multimodal construction process and produced a concise description of the solution. The Behavior Detection and Cognitive Mapping Agent compared that process with the task-specific behavior rubrics and mapped the detected mathematical behaviors to competencies in the cognitive graph. When a behavior fell outside the predefined rubrics, this analysis included an additional outlier-diagnosis call to identify the relevant competencies. These outputs supported peer-profile construction and artifact assignment.

B.2 Personalized Peer Construction

Peer-profile construction. The system used a topological enumeration algorithm over the task-specific cognitive graph to generate prerequisite-consistent states for Leading and Struggling Peers, and sampled the required number of each according to the group composition. Struggling Peers were also allowed to master selected foundational competencies that the learner had not yet demonstrated, creating opportunities for reciprocal tutoring.

Artifact assignment. Artifacts were assigned in three passes. First, the system covered the learner’s mastery boundary, intermediate learning paths, and target competencies with artifacts compatible with each peer profile. Second, it added role-consistent incomplete or flawed artifacts to support correction and learning-by-teaching. Third, it supplemented the portfolios so that each peer had at least two artifacts, relevant terminal competencies were represented, and the group contained ten artifacts overall at least. When several assignments were possible, the priorities were candidate scarcity, cognitive fit, distributing repeated competency coverage across peers, balancing portfolio sizes, and novelty in mathematical behavior and modality.

Figure 11 provides two additional construction examples: a Mixed Group for the task of representing 17 (Figure 11a) and a Leading Group for an addition task (Figure 11b).

B.3 Adaptive Conversational Agency

Activity intensity. To estimate interruption cost from interface behavior, the activity monitor grouped client events into four intensity levels. No-intensity events included opening or closing a chat panel and completing a replay; low-intensity events included navigating among peers, artifacts, and panels or pausing a replay; high-intensity events included listening to a message, playing or seeking within a solution replay, and changing a chat attachment; and extreme-intensity events included composing, recording, or sending a message and actively creating or submitting an artifact. These levels controlled how long the system continued to treat the learner as engaged: more intensive activity prolonged the waiting period, and extreme activity could cancel a pending proactive message.

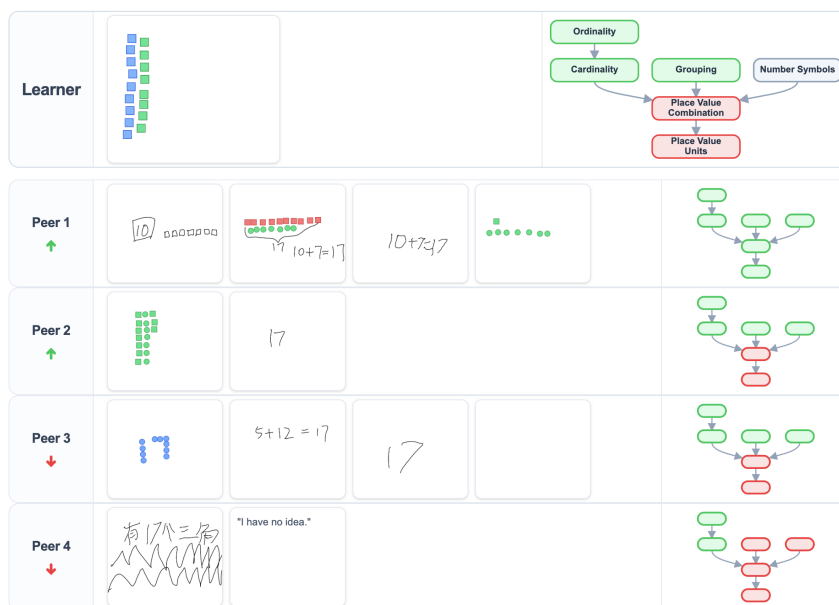
Proactive follow-ups. During an ongoing group topic, sustained silence could trigger a *pickup follow-up*, in which a peer re-entered the discussion with a more accessible invitation grounded in the current topic. Before the learner first responded, the system allowed up to two pickup follow-ups; after a response, it allowed one further pickup. If the learner continued to remain silent after the available pickups—or during a private topic, where pickups were not used—an *end follow-up* produced a brief, concluding peer message without posing another question and closed the topic. Follow-ups were withheld while the relevant chat channel contained an unread message, preserving an opportunity for the learner to inspect the existing messages first.

Message-generation agents. The formal system used four message-generation agents, summarized in Table 7. The three chat agents used interaction-specific context and output constraints, while the Topic Agent followed the learner-centered orchestration described in the main text. All four used Gemini 3.6 Flash with a temperature of 0.

C User Study

C.1 Child Participants

Before the study, parents completed a brief background questionnaire reporting the child’s birth month, gender, grade, prior use of math-learning platforms, usage frequency of electronic devices and conversational AI, regular extracurricular mathematics-learning experiences, and current mathematics progress. Among the 22 formal-study participants, 15 used



(a) A Mixed Group with two Leading and two Struggling Peers for a task about representing 17.

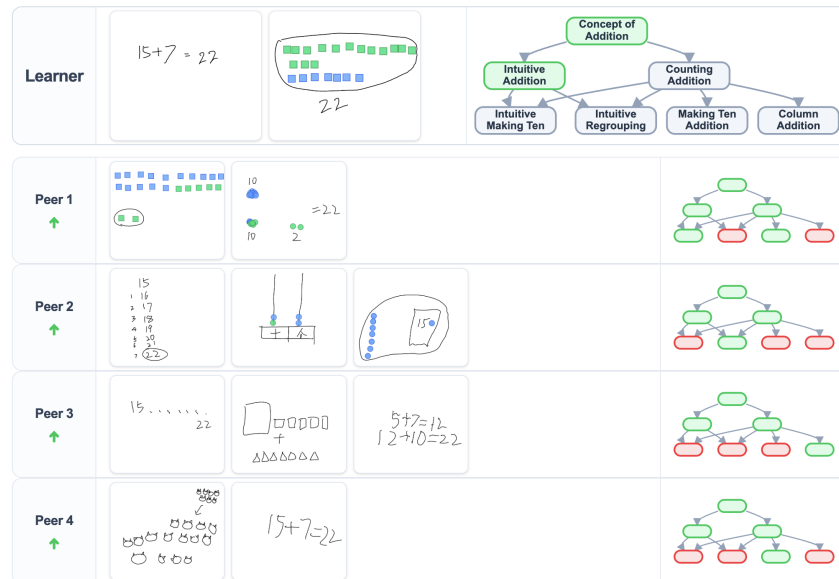
(b) A Leading Group with four Leading Peers for the addition task $15 + 7$.

Fig. 11. Additional examples of personalized peer construction across tasks and group compositions.

Table 7. Message-generation agents used in the formal system.

Agent	Trigger and Inputs	Outputs
Group Reactive Agent	A learner message in group chat; the current topic, peer profiles and artifacts, and group-chat history.	One or more peer replies with optional artifact attachments, together with whether to continue or close the topic.
Private Reactive Agent	A learner message in private chat; the selected peer’s profile and artifacts, the current topic, and relevant private- and group-chat context.	One or more replies from that peer with optional artifact attachments, together with whether to continue or close the topic.
Follow-up Agent	Sustained silence during an ongoing topic; the interaction state, current topic, relevant channel history, and peer profiles and artifacts.	A peer message with optional artifact attachments, realized as either a pickup or end follow-up according to the interaction state.
Topic Agent	An idle period with no ongoing topic; the task and cognitive graph, user and peer profiles and artifacts, prior topic history, and complete chat history.	The next topic type, channel, speaker or facilitation role, message, and referenced artifacts.

electronic devices daily, 5 weekly, and 2 rarely; 11 used conversational AI frequently, 7 occasionally, and 4 had never used it. 11 had used mathematics-learning platforms, 17 had participated in some form of regular extracurricular or home mathematics learning, and 7 were learning ahead of their completed grade.

Table 8 reports the formal-study participants. Grade denotes the school year completed before the study in summer break; K denotes the final kindergarten year. Task tiers correspond to Table 9.

Table 8. Demographics and Study Assignments of Child Participants.

ID	Grade	Gender	Age	Cond.	Task Tier	Device Use	Conversational AI Use	Prior Math Platform Use	Ahead of Grade
U1	1	Male	7y5m	B	3	Daily	Occasional	No	No
U2	1	Male	7y7m	A	3	Rarely	Occasional	No	Yes
U3	1	Male	7y8m	A	2	Daily	Frequent	Yes	No
U4	K	Female	6y6m	A	1	Daily	Frequent	Yes	No
U5	1	Female	7y10m	A	3	Rarely	Occasional	No	No
U6	1	Male	7y4m	B	3	Daily	Frequent	Yes	Yes
U7	K	Female	6y7m	B	1	Daily	Frequent	Yes	No
U8	K	Female	6y9m	A	2	Daily	Occasional	No	Yes
U9	K	Male	6y2m	A	1	Daily	Occasional	Yes	No
U10	1	Male	8y2m	B	3	Daily	Frequent	No	Yes
U11	K	Male	6y2m	B	1	Daily	Frequent	Yes	No
U12	K	Female	6y9m	B	1	Weekly	Frequent	Yes	Yes
U13	K	Male	6y0m	A	1	Daily	Never	Yes	No
U14	K	Male	6y0m	B	1	Weekly	Never	Yes	No
U15	K	Female	6y2m	A	1	Daily	Frequent	No	Yes
U16	1	Male	7y6m	B	2	Daily	Never	No	No
U17	1	Female	7y4m	A	3	Weekly	Occasional	No	No
U18	1	Female	7y9m	B	3	Daily	Frequent	No	No
U19	1	Female	7y7m	A	3	Daily	Frequent	No	No
U20	K	Male	6y6m	B	2	Daily	Occasional	Yes	Yes
U21	1	Female	7y4m	B	3	Weekly	Occasional	No	No
U22	K	Female	6y11m	A	1	Weekly	Frequent	Yes	No

C.2 Mathematics Content and Tasks

The user study focused exclusively on *Subtraction with Regrouping*. Each participant completed one of the three content levels shown in Table 9. The initial task was used for the pre-test and learning, while the corresponding quiz task was used for the post-test. Across the three pre-test and learning tasks, the artifact repository comprised 130 retained and preprocessed artifacts: 92 from the OpenCD dataset, 6 collected during the formative study, and 32 authored by

the researchers to supplement strategies underrepresented among the child-contributed artifacts. Across U1–U22, 62 unique artifacts were assigned to at least one peer portfolio: 18 of 45 in Tier 1, 19 of 48 in Tier 2, and 25 of 37 in Tier 3.

Table 9. Mathematics Content and Tasks Used in the User Study.

Content	Pre-test and Learning Task	Post-test Task
Subtraction within 20	“How can you calculate $13 - 7$?”	“How can you calculate $15 - 8$?”
Subtraction within 30	“How can you calculate $24 - 15$?”	“How can you calculate $26 - 8$?”
Subtraction within 100	“How can you calculate $45 - 28$?”	“How can you calculate $62 - 37$?”

C.3 Post-study Questionnaire and Interview

The child-facing questionnaire used labeled five-point pictorial faces adapted from the Smileyometer [33]; labels were customized to each construct while retaining an ordered negative-to-positive response. The researcher read or clarified items when needed without suggesting an answer.

Peer liking. Children rated each peer from strongly dislike to strongly like and then ranked the four peers from most to least liked.

Perceived mathematics level. Children rated each peer’s mathematics level from very low to very high, ranked the four peers, and placed themselves within that ranking.

Social pressure. Children rated four statements: “After seeing their solutions, I felt less capable than them”; “After our discussion, I felt less capable than them”; “During the discussion, I worried that my method was poor”; and “During the discussion, I was reluctant to share my ideas”. Each item used five labeled responses from endorsement to the opposite position.

Experience. Children compared learning this way with regular classes on two five-point scales: much more relaxed to much more tiring, and much more interesting to much more boring.

The structured interview followed a common set of questions aligned with the questionnaire; the researcher used brief clarification prompts to elicit reasons or concrete examples. The questions were: (1) Did you learn any new methods, and where did they come from? (2) Which of your own and the peers’ methods did you like most? (3) Did you help peers, and do you like teaching peers? (4) Did you ask peers for help, and do you like doing so? (5) How did this experience compare with the ways you usually learn mathematics? and (6) Which parts did you like or dislike?

C.4 Analysis of Child-study Data

Message analysis. For learner messages, two researchers first discussed an initial set of coding dimensions and independently coded all messages from four children. They then reconciled their interpretations and finalized a codebook covering content domain (mathematical learning, peer-interaction coordination, or off-task conversation) and conversational role (initiating, responding, or continuing the learner’s own prior message). Each response was additionally linked to the specific prior message it answered. One researcher applied the codebook to the remaining learner messages, after which the second researcher reviewed the complete coding.

For the larger set of non-learner messages, the researchers defined a categorical codebook covering learning relevance, question versus non-question, open- versus closed-ended questions, and the target of group-chat questions (another peer, the whole group, or the learner). Given the volume of these messages, their comparatively standardized and explicit system-generated language, and the direct binary or three-way classification criteria, researchers used GPT-5.6-sol

to apply these codes before they reviewed the resulting classifications. Channel, topic, and follow-up sequence were obtained directly from system metadata. A group-chat topic was eligible for the follow-up response analysis when the learner had not yet contributed to that topic, or when the learner’s first message was coded as a response linked to a message within that topic.

Artifact analysis. One researcher initially coded the pre- and post-test artifacts. Both researchers then jointly reviewed every artifact from the beginning and coded its representational forms, calculation strategies, and correctness based on the mathematical behavior features synthesized in OpenCD [113]. Comparing each child’s two artifacts yielded three non-exclusive forms of positive change: adopting a new calculation strategy, introducing a new visual representation, and improving correctness. For additional artifacts created during learning, we aligned camera video, screen recordings, and system logs to reconstruct the surrounding interaction. We coded typical events preceding creation, actions immediately following it, and content relationships to the child’s own or peers’ earlier artifacts and chat suggestions.

Interview analysis. Because the interviews followed a shared set of structured questions, we conducted question-guided qualitative content analysis: responses were organized by interview question and recurring explanations and examples were grouped within each question. Quotations reported in the paper were translated from Chinese into English.

Quantitative analysis. Log-derived behavioral measures and questionnaire responses were summarized descriptively. Between-condition comparisons used two-sided, tie-aware exact Mann–Whitney U tests, with Cliff’s δ as the effect size. Comparisons of response rates across message characteristics used two-sided, user-stratified exact permutation tests to account for messages nested within children. Full test specifications and results are reported in Appendix D.1.

C.5 Post-hoc Expert Evaluation

C.5.1 Evaluation Dataset and Persona-centered Baseline. At each intervention point, the learner-centered candidate was the message delivered during the user study. The persona-centered candidate was generated using the interaction state immediately before that intervention, including the task, available artifacts, and conversation history.

We implemented the persona-centered baseline as two sequential LLM calls. The first selected the peer most likely to initiate a topic based on the peers’ cognitive profiles, available artifacts, recent participation, and the current conversational context. The second prompted the selected peer to generate a topic from its own perspective, jointly determining the message and attached artifacts from its profile and relevant chat history. Both calls used Gemini 3.6 Flash with a temperature of 0.

C.5.2 Review Interface and Rating Instrument. The custom interface (Figure 12) replayed each session chronologically. At each intervention point, experts were provided with the conversation history up to that point. They were then presented with two candidate topic messages in a randomized order, with the condition identities hidden. The interface displayed each message’s sender, channel, and attached artifacts. Experts independently rated both messages before the actual subsequent trajectory was revealed for the next turn.

Each topic was rated for its overall pedagogical appropriateness within the displayed context. The five-point scale ranged from 1 (*very inappropriate*) to 5 (*very appropriate*). For ratings of 3 or below, experts were required to select at least one predefined reason or provide a written explanation. The predefined tags included: too difficult for the learner, too easy for the learner, limited learner participation, inconsistent with the peer’s cognitive profile, potentially misleading, and repetitive.

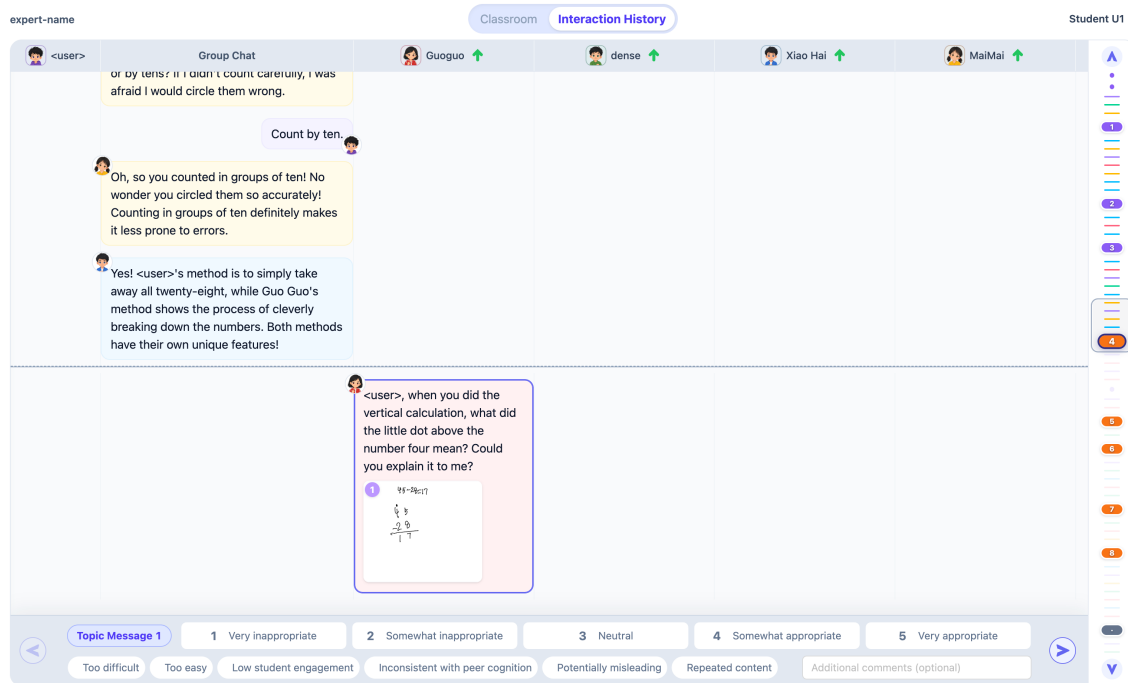


Fig. 12. Expert-review interface. The interaction history preceding a topic appears above the divider, while one blinded candidate topic appears below it. Experts rated each candidate on a five-point appropriateness scale and could select reason tags or enter a comment.

C.5.3 Review Procedure. The review followed a three-phase protocol:

(1) *Onboarding and Calibration.* We first introduced the review task, rating scale, reason tags, and replay interface. The experts then independently reviewed 3 participant sessions. Subsequently, they met with the researchers to compare their judgments and discuss their interpretations of the scale and tags. The researchers clarified rating criteria and procedural questions, while the experts resolved discrepancies in the calibration cases. This established shared rating criteria for the remaining sessions.

(2) *Independent Review.* The remaining 19 sessions contained 133 intervention points, yielding 266 candidate messages. To ensure blind review, both experts rated the candidate messages before viewing the subsequent trajectory at each intervention point. The mean absolute difference between the two experts' pre-consensus ratings was 0.387. Because 85.5% of the individual ratings were 4 or 5, the concentrated marginal distributions may partly explain the moderate weighted κ despite high raw agreement.

(3) *Consensus Meeting.* After both independent reviews were completed, the experts used a comparison view showing their scores, selected tags, and comments side by side. They jointly revisited every candidate message for which their scores differed by more than one point and adjusted their ratings until the difference was no greater than one point. During this meeting, they also briefly explained to the researchers each message they regarded as inappropriate. The two resulting ratings were then averaged to obtain the final score for each message. For the reasons for inappropriateness, all distinct reasons identified by either expert were retained. The researchers interpreted and organized these reasons

with reference to the experts' selected tags, written comments, and meeting explanations. The final analysis utilized these scores and reasons across all 22 participant sessions.

D Results

D.1 Statistical Details for Results Reported in the Main Text

Peer pressure and perceived mathematical standing. Across the four peer-pressure items, the condition means ranged from 3.2 to 3.5 in Condition A ($SD = 0.6\text{--}1.3$) and from 3.5 to 4.2 in Condition B ($SD = 0.8\text{--}1.3$), with no significant differences ($U = 41.5\text{--}52.0$, all $p \geq .149$). Children's mean ratings of their peers' mathematical levels were comparable (A: 4.4 ± 0.6 , 3.3–5.0; B: 4.1 ± 0.5 , 3.3–5.0), with no significant between-condition difference ($U_A = 80.5$, $p = .191$). Children also did not rank themselves lower in Condition B (A: 2.6 ± 1.3 , 1–5, $n = 9$; B: 1.8 ± 1.1 , 1–4, $n = 11$; $U = 66.5$, $p = .207$; lower ranks indicate higher perceived standing). These questionnaire comparisons used two-sided, tie-aware exact Mann–Whitney U tests with exhaustive permutation of the observed condition labels. Each comparison included 11 children per condition, except the self-ranking comparison, for which two responses in Condition A were unavailable ($n_A = 9$, $n_B = 11$).

Significant between-condition results. The following statistics correspond to the significant between-condition results reported in the first three Results subsections. Each comparison used a two-sided, tie-aware exact Mann–Whitney U test, with the null distribution obtained by exhaustively permuting the condition labels across the 22 children ($n_A = n_B = 11$; $\binom{22}{11} = 705,432$ assignments). We report exact p values. Cliff's δ is oriented as Condition A minus Condition B.

- *Additional solution creation.* Three children in Condition A created 7 additional solutions, whereas eight children in Condition B created 22. The per-child counts were 0.6 ± 1.5 (0–5) in Condition A and 2.0 ± 1.7 (0–4) in Condition B. The exact test yielded $U_A = 31.0$, $p = .0370$, and Cliff's $\delta = -.488$.
- *Private topics.* Sessions in Condition A contained 21 private topics in total, compared with 9 in Condition B. The per-child counts were 1.9 ± 0.7 (1–3) in Condition A and 0.8 ± 0.8 (0–2) in Condition B. The exact test yielded $U_A = 101.5$, $p = .0060$, and Cliff's $\delta = .678$.
- *Private-chat messages.* Sessions in Condition A contained 242 private-chat messages in total, compared with 79 in Condition B. The per-child counts were 22.0 ± 21.6 (2–71) in Condition A and 7.2 ± 8.9 (0–30) in Condition B. The exact test yielded $U_A = 93.5$, $p = .0289$, and Cliff's $\delta = .545$.
- *Willingness to learn with the same peers again.* Children in Condition A selected 3.8 ± 0.6 peers (2–4), compared with 2.5 ± 1.6 peers (0–4) in Condition B. The exact test yielded $U_A = 94.0$, $p = .0152$, and Cliff's $\delta = .554$.

Expert evaluation of topic generation.

- *Appropriateness ratings.* Across 157 paired topic rounds, the two experts' post-consensus ratings were averaged for each message. Learner-centered topics received 4.694 ± 0.585 points (median = 5.0), compared with 4.360 ± 0.784 (median = 4.5) for the persona-centered alternatives. A linear mixed-effects model of the paired rating difference, fitted by restricted maximum likelihood with a random intercept for child, estimated a learner-centered advantage of $\beta = 0.338$ ($SE = 0.091$), $t(21) = 3.705$, $p = .0013$, 95% CI [0.148, 0.527].

Message characteristics and learner responses. The following comparisons used two-sided, user-stratified exact permutation tests. The procedure fixed each child's category-specific message counts and total number of responses, permuted category labels within child, and tested the response-rate difference. Both comparisons included 22 children,

of whom 20 contributed nondegenerate permutation information. Wilson 95% CIs are message-level descriptive intervals and do not account for within-child clustering.

- *Questions versus non-question messages.* Questions received responses in 187 of 442 eligible cases (42.3%, 95% CI [37.8%, 47.0%]), compared with 52 of 275 non-question messages (18.9%, 95% CI [14.7%, 24.0%]). The response-rate difference was 23.4 percentage points ($p = 2.58 \times 10^{-13}$).
- *Closed- versus open-ended questions in group chat.* Closed-ended questions received responses in 74 of 183 eligible cases (40.4%, 95% CI [33.6%, 47.7%]), compared with 48 of 174 open-ended questions (27.6%, 95% CI [21.5%, 34.7%]). The response-rate difference was 12.9 percentage points ($p = .0105$).