Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1656

Exploring the Potential of Multimodal AI to Support Collaborative Mathematics Learning in Elementary Classrooms

Mingyu Feng
WestEd

Melissa Lee
WestEd

Vivian Hsu
WestEd

Matt Miller
OKO Labs

TJ Marston
OKO Labs

This study examines the potential of multimodal artificial intelligence (AI) to facilitate collaborative mathematics learning in elementary classrooms. We conducted a cluster-randomized study involving 20 teachers and 1,040 students to investigate the use of OKO, an AI-supported learning environment designed to promote peer interaction during mathematical problem solving. Unlike AI systems that provide direct instructional support to individual students, OKO uses speech and visual information to facilitate group discussions, encourage mathematical explanations, and support participation. The study draws on classroom observations, teacher interviews, system-generated collaboration indicators, student surveys, and mathematics achievement data. Preliminary findings suggest that AI-supported facilitation can promote mathematical discussion and peer collaboration, with suggestive but inconclusive evidence of benefits for mathematics achievement. The full manuscript, including detailed analyses and findings, is currently under peer review.

VERSION: November 2026

Suggested citation: Feng, Mingyu, Melissa Lee, Vivian Hsu, Matt Miller, and TJ Marston. (2026). Exploring the Potential of Multimodal AI to Support Collaborative Mathematics Learning in Elementary Classrooms. (EdWorkingPaper: 26-1656). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/es8v-j755

Exploring the Potential of Multimodal AI to Support Collaborative Mathematics Learning in Elementary Classrooms

Mingyu Feng, Melissa Lee, Vivian Hsu
mfeng@wested.org, mlee3@wested.org, vhsu@wested.org
WestEd

Matt Miller, TJ Marston
matt.miller@okolabs.ai, tj.marston@okolabs.ai
OKO Labs

Abstract: This study examines the potential of multimodal artificial intelligence (AI) to facilitate collaborative mathematics learning in elementary classrooms. We conducted a cluster-randomized study involving 20 teachers and 1,040 students to investigate the use of OKO, an AI-supported learning environment designed to promote peer interaction during mathematical problem solving. Unlike AI systems that provide direct instructional support to individual students, OKO uses speech and visual information to facilitate group discussions, encourage mathematical explanations, and support participation. The study draws on classroom observations, teacher interviews, system-generated collaboration indicators, student surveys, and mathematics achievement data. Preliminary findings suggest that AI-supported facilitation can promote mathematical discussion and peer collaboration, with suggestive but inconclusive evidence of benefits for mathematics achievement. **The full manuscript, including detailed analyses and findings, is currently under peer review.**

Introduction

Collaborative learning can help students develop understanding by explaining their reasoning, considering peers' ideas, resolving disagreements, and solving problems together (Barron, 2000; Roschelle et al., 2010). However, simply placing students in groups does not ensure productive collaboration. Students need guidance to stay focused on shared learning goals, explain their thinking, respond to one another, and participate meaningfully (Dillenbourg & Tchounikine, 2007; Fischer et al., 2013). In mathematics, these interactions are particularly important because explaining and comparing different approaches can help students refine their reasoning (Roschelle et al., 2010).

Teachers play an important role in facilitating productive group interactions, but it can be difficult for one teacher to attend to several small groups at the same time. Computer-supported collaborative learning (CSCL) environments can help by structuring group activities and prompting particular forms of interaction. For example, collaboration scripts can organize students' roles and activities to promote knowledge sharing and reasoning (Dillenbourg & Tchounikine, 2007; Fischer et al., 2013). In elementary mathematics, Roschelle et al. (2010) found that technology designed to elicit individual responses, reveal differences in students' answers, and prompt group discussion increased questioning, explanation, and discussion of disagreements and improved mathematics learning.

Recent advances in multimodal artificial intelligence may allow technology to take a more active role in facilitating collaboration. Multimodal learning analytics can draw on data such as speech, video, gaze, and other behavioral signals to examine participation and interaction during collaborative

learning (Zhou et al., 2024; Ouyang & Bai, 2026). An AI-supported system can potentially use information about an ongoing interaction to decide when and how to intervene—for example, by drawing in a quieter student, prompting a student to explain their reasoning, or providing support when a group is stuck. However, there is still limited evidence about how this type of AI-supported facilitation works when used repeatedly in classrooms. This study asks: **Can a multimodal AI facilitator support productive collaborative mathematics learning in elementary classrooms?**

We examine this question in a classroom study of **OKO**, a multimodal AI-supported environment for small-group mathematics learning. OKO engages students in a consensus-building mathematics game in which students first solve a problem independently and then discuss the problem when their answers differ. The system uses students' speech and visual information, including gestures and indicators of engagement, to inform facilitation moves such as inviting quieter students to contribute, prompting students to explain their thinking, or offering a hint when a group is stuck. OKO's design builds on the Consensus Game and prior research on technology-supported group explanation and feedback (Roschelle et al., 2010).

We studied OKO from January to April 2026 in elementary mathematics classrooms. Twenty teachers and 1,040 students participated in a cluster-randomized study, with teachers assigned to treatment or control conditions. Teachers varied in how they implemented OKO, including whether they used it with their full class or selected students. Using classroom observations, teacher interviews, system-generated collaboration indicators, student surveys, and standardized mathematics assessments, we address three research questions:

1. What patterns of collaborative mathematical activity are observed when students use a multimodal AI-supported collaborative learning environment?
2. How do indicators of students' collaborative mathematical participation change over repeated use of the environment?
3. What evidence of promise is there that participation in AI-supported collaborative learning is associated with mathematics learning?

Theoretical Framework

Productive collaboration in mathematics

The benefits of collaborative learning depend on the quality of students' interactions. Productive groups attend to, acknowledge, and build on one another's ideas, whereas less successful groups may fail to establish shared attention or engage with ideas introduced by peers (Barron, 2000). Effective collaboration also involves questioning, explaining and justifying ideas, and responding to others' reasoning (Soller, 2001). These processes make students' thinking available to their peers and create opportunities to develop shared understanding.

Explanation is particularly important in mathematics. Giving explanations and using help received from peers both related to student learning (Webb et al., 2006). Differences in students' answers can also prompt students to reconsider their thinking. Roschelle et al. (2010), for example, found that students in a technology-supported collaborative condition engaged in more questioning, explanation, and discussion of disagreements than students who worked individually and also demonstrated stronger mathematics learning. Productive mathematical collaboration therefore involves more than reaching a shared answer; how students explain, respond to, and build on one another's reasoning is central to the learning process.

Scaffolding collaborative interaction

These forms of interaction do not always emerge spontaneously. Participation can be uneven, and groups may reach agreement without discussing the reasoning behind their answers. CSCL research has therefore examined ways to structure collaborative activity. Collaboration scripts can organize roles, activities, and sequences of interaction to elicit productive forms of participation and knowledge sharing (Dillenbourg & Tchounikine, 2007; Fischer et al., 2013).

The challenge is to provide enough structure without unnecessarily constraining the interaction. Students already have expectations and routines for working with peers, which shape how they respond to external guidance (Fischer et al., 2013). The support a group needs may also change during an activity. Students who are already explaining and comparing ideas may need little intervention, while groups with uneven participation or limited discussion may benefit from additional guidance. Dillenbourg and Tchounikine (2007) describe this tension in terms of the flexibility of CSCL scripts: support needs to guide interaction while leaving room for groups to adapt their collaborative activity.

AI and multimodal support for collaboration

Intelligent systems can use information about ongoing group interaction to provide more responsive support. Early work by Soller (2001), for example, examined how information from students' conversations could be used to identify problems in collaborative interaction and provide targeted support. Multimodal learning analytics has expanded the range of information available for understanding group processes. Speech, video, gaze, gesture, and other behavioral data can provide complementary information about participation and interaction during collaborative learning (Ouyang & Bai, 2026; Zhou et al., 2024).

Generative AI introduced additional roles for technology in collaborative learning. Recent studies have embedded large language models in group activities to answer questions, provide feedback, summarize discussions, or contribute to collaborative problem solving (An et al., 2024; Feng, 2025; Yin et al., 2025). A recent review found that generative AI can support knowledge development, communication, coordination, reflection, and feedback during group work, while also identifying risks of overreliance on AI and reduced peer interaction (Wei & Perkins, 2026). In much of this emerging work, however, the AI functions as a resource or participant with which students interact directly. Feng (2025), for example, found that five of six groups developed an AI-centered interaction pattern in which students interacted more with the chatbot than with their peers.

This raises a different CSCL question: whether AI can support collaboration without becoming the central partner in the interaction. Rather than contributing ideas or answers as another group member, an AI system can use information about students' interactions to decide when facilitation may be useful, while leaving the substantive collaborative work to the students. Multimodal information is especially relevant to this role because it captures what students stay and how they participate. This approach remains relatively underexplored, particularly in elementary education, as much of the recent research on generative AI-supported collaboration has focused on higher education (Wei & Perkins, 2026). The present study examines this less studied role for AI in elementary mathematics classrooms.

OKO: A multimodal AI facilitator for small group collaborative learning

OKO is a web-based learning environment for small-group collaborative learning in elementary through high school classrooms. Teachers use a dashboard to create groups, select mathematics content, and review student activity. Post-session reports provide teachers with information about performance, participation, time in the activity, and mathematics standards. Students participate from individual

devices in groups of two to five. The system combines speech recognition and generation, large language models, and computer vision. Information from students' speech and visual behavior, including gestures, body position, and indicators of attention, informs facilitation moves such as inviting a quieter student to contribute, asking a student to explain their reasoning, or providing support when a group is stuck.

The activity examined in this study was OKO's consensus-building game, which uses standards-aligned mathematics items spanning K–12, with the deepest coverage in grades 3–10. . The game builds on the Consensus Game developed by Miguel Nussbaum and subsequently studied in the TechPALS project (Roschelle et al., 2010) with the addition of OKO's multimodal facilitation layer where the speech and vision input drives when and how the system intervenes. Each problem begins with an individual response. Students answer a multiple-choice mathematics question without discussing their answers with their group, and the system does not provide immediate feedback. Once all students have responded, OKO determines whether the group has reached consensus. If their answers differ, students see the options selected by group members and are prompted to discuss the problem. This sequence gives students an opportunity to develop an initial solution independently before comparing their reasoning with their peers. If the group continues to disagree, the system provides additional support, including removing an incorrect answer option. The round ends when the group reaches the correct answer or completes the available attempts, after which OKO provides brief feedback and moves to the next problem.

Figure 1 illustrates how OKO is set up in the classroom; students participate from individual devices while working as a small group. Figure 2 shows the student interface. Each student first responds independently to a mathematics problem (Figure 2a). When students disagree or when the system identifies a need for support, the AI facilitator can prompt students to explain their reasoning or engage with one another (Figure 2b). At the end of each question, students receive feedback from OKO, along with a collaboration score for how well they worked together as a group (Figure 3).

Figure 1

Classroom Small-Group Configuration of OKO



Figure 2

Student View During Problem Solving (a) and AI Facilitation (b)

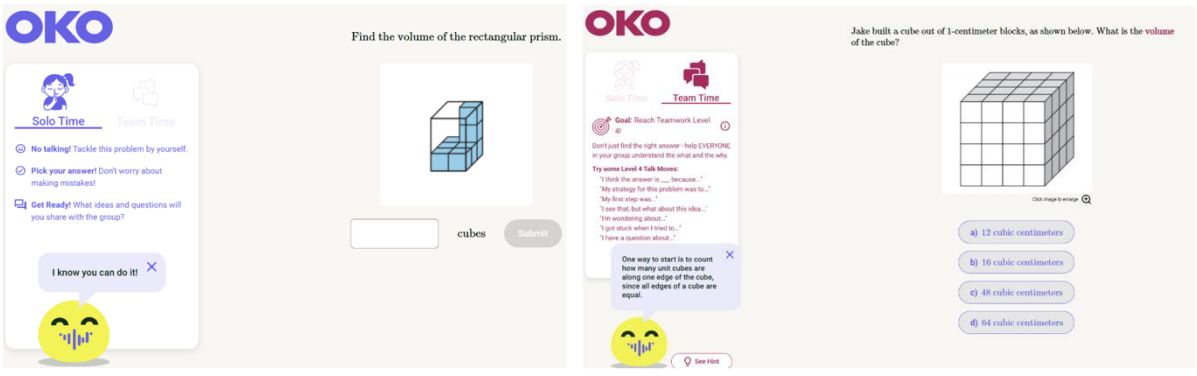


Figure 3

Student View of OKO's Feedback and Collaboration Score



Methods

Sample and Study Design

The study used a cluster-randomized design in which teachers were the unit of assignment. Twenty mathematics teachers from 11 schools in one Tennessee public school district in the United States participated. Two teachers taught fourth grade and 18 taught fifth grade. The participating teachers served 1,040 students. The sample was approximately evenly divided by gender and included students from a range of racial and ethnic backgrounds. Nine percent were identified as English learners, 22% as economically disadvantaged, and 17% received special education services.

Teachers were randomly assigned within school and grade blocks to either the treatment or control condition, resulting in 10 teachers in each condition. Students followed the assignment of their teachers. Treatment teachers were asked to use OKO for 15–30 minutes per session, two to three times per week, over approximately 10 weeks. Control teachers continued using existing mathematics software in their classrooms.

Study Procedures

The study took place from January through April 2026. Before implementation, all participating teachers attended a 45-minute study orientation. Treatment teachers also received an introduction to OKO and its core features. Although the planned implementation period was approximately 10 weeks, teachers used OKO for an average of 11.2 weeks (range = 8–13 weeks). Students typically worked in groups of two to five. Implementation varied across treatment classrooms. Some teachers used OKO with their full class, whereas others selected a subset of students whom they considered appropriate for the activity. Across the 500 students assigned to treatment classrooms, 255 students used OKO during the study. We refer to these students as *OKO users* when reporting analyses based on actual use.

Data Sources

We used multiple data sources to examine collaborative activity and mathematics learning.

Classroom observations. Researchers observed all 10 treatment classrooms and four control classrooms in March 2026. Observations lasted approximately 15–40 minutes. In treatment classrooms, researchers documented how OKO was implemented, students' engagement and persistence, interactions with peers, mathematical discussion, teacher facilitation, and indications of student learning.

Teacher interviews. At the end of the study, six treatment teachers and three control teachers participated in structured interviews lasting approximately 10–20 minutes. Treatment teachers were

selected to represent relatively low, moderate, and high levels of OKO use. Interviews addressed implementation, usability, instructional fit, student engagement, collaboration, and perceived learning.

System-generated collaboration scores. When students discussed a problem, OKO assigned each student a collaboration score from 1 to 4 based on transcripts of their speech. A large language model tagged each student's contributions for specific discourse moves, and rules converted those tags into a score. Scores reflected two dimensions. Reasoning ranged from stating answers, to describing procedures, to explaining why a solution works. Reciprocity captured how students engaged with peers, such as revoicing a peer's idea, disagreeing respectfully, or asking for help. The highest score required conceptual explanation combined with reciprocity. Problems resolved without discussion were not scored. Each student's problem-level scores were averaged within each session. For the longitudinal analysis, each student's session scores were then averaged within each week in which they used OKO. Because the scores were generated by OKO and were not validated against human coding, we treat them as indicators of collaborative participation rather than as a validated measure of collaboration.

Student surveys. Students completed pre- and post-surveys measuring mathematics motivation and attitudes toward collaboration. Collaboration was assessed using the Teamwork Scale for Youth (Lower-Hoppe et al., 2017), an eight-item measure of students' perceptions of and confidence in working collaboratively with peers. Items were rated on a five-point scale, and the scale showed acceptable internal consistency in this sample (Cronbach's α = .77). Mathematics motivation was assessed using the Beliefs, Engagement, Attitude, and Math Motivation Scale (BEAMMS; Orosco, 2016)**,** a 10-item measure designed for elementary students. Students selected between paired statements reflecting more positive or negative attitudes toward mathematics; higher scores indicated more positive mathematics attitudes and motivation. The scale showed good internal consistency in this sample (Cronbach's α = .81).

Mathematics achievement. Student mathematics achievement was measured using the NWEA MAP Growth mathematics assessment (Meyer et al., 2025), which the district administered as part of its regular benchmark testing. Winter 2026 scores served as the pretest and spring 2026 scores as the posttest. The primary outcome was the MAP Growth RIT score, a vertically scaled score that allows achievement to be compared across testing occasions. We also examined students' growth relative to their NWEA projected growth. Excess growth was calculated as observed growth minus projected growth, with positive values indicating that a student grew more than projected and negative values indicating growth below projection (NWEA, 2025).

Analytic Approach

Given the pilot nature of the study and the limited number of randomized teachers (n = 20), we used a two-sided significance threshold of α = .10 to identify preliminary evidence of promise. We report exact p-values, effect sizes, and 90% confidence intervals to support interpretation of the findings.

For RQ1, we analyzed classroom observation notes and treatment-teacher interviews to identify recurring patterns in students' collaborative mathematical activity. The analysis focused on mathematical explanation and reasoning, peer help, participation, persistence, and breakdowns in collaboration. We compared patterns across observations and interviews to identify themes supported by both data sources.

For RQ2, we examined changes in system-generated collaboration scores across students' active weeks of OKO use. We first examined trends in weekly scores and then estimated a linear mixed-effects growth model, with repeated weekly observations nested within students. The model included student background characteristics and allowed students to vary in their initial scores and rates of

change over time. We also compared post-study Teamwork Scale for Youth scores between treatment and control students using a two-level model that controlled for baseline Teamwork Scale scores. This analysis examined whether changes in situated indicators of collaboration were accompanied by differences in students' broader perceptions of their collaborative skills.

For RQ3, we first estimated the intent-to-treat effect of assignment to OKO using two-level hierarchical linear models (HLMs) with students nested within teachers. Spring NWEA MAP mathematics scores were predicted from treatment assignment, controlling for winter MAP scores and student background characteristics. We used the same modeling approach to examine excess growth, controlling for winter MAP scores.

Because not all students assigned to the treatment condition used OKO, we also conducted exploratory analyses comparing students who used OKO with students in control classrooms. Teachers selected which treatment students used OKO, so these comparisons do not preserve the randomized design. To assess the sensitivity of the findings to observed selection differences, we used coarsened exact matching (CEM; Iacus et al., 2012) to match OKO users with unexposed comparison students on baseline MAP scores, gender, race/ethnicity, English learner status, economic disadvantage, and special education status. We then estimated CEM-weighted two-level models for spring achievement and excess growth.

Results

RQ1: What patterns of collaborative mathematical activity are observed when students use a multimodal AI-supported collaborative learning environment?

Classroom observations and teacher interviews showed that students frequently engaged in mathematical discussion and reasoning during OKO sessions. All six treatment teachers interviewed reported hearing students engage in math talk and reasoning with one another. Teachers described students asking peers for help and explaining their thinking rather than relying only on the teacher. One teacher described students saying to one another, "I don't know how to do that. Could you show me how to do this?" and then helping each other work through the problem. Observations showed similar patterns: students discussed OKO problems, explained their reasoning, corrected errors, and worked through challenging problems together. These behaviors provided evidence that students were engaged not only socially but also in cognitive work through explanation, error correction, strategy use, and collaborative problem solving.

Peer support was especially evident when students encountered difficulty. In one observed classroom, students repeatedly tried to help group members who did not understand a problem, explaining their thinking to move the group forward. In another classroom, students tried multiple strategies before arriving at a solution. In these cases, students worked not only on their own solutions but also to help the group make progress.

Students were also generally engaged during OKO sessions. Four of the six treatment teachers interviewed said their students were consistently excited to use OKO and would ask whether they were going to use it that day. Classroom observations were broadly consistent with these reports. In two classrooms, students wanted to continue using OKO after the scheduled session had ended. One teacher described the combination of engagement and mathematical interaction:

> *"My kids never complained. Not one time. Do you know how huge that is? Do you know how huge it is to have all your students talking about math, doing math, working in groups. I loved how the kids loved it."*

The same teacher noted that students were communicating and working throughout the activity and described being able to recognize quieter students for participating.

Participation also differed across students and groups. Students who were shy, less confident in mathematics, or less comfortable collaborating sometimes contributed less, while higher-performing students often took active roles in solving problems and supporting peers. Teachers also described instances in which students who rarely shared their thinking during regular whole-class instruction were more willing to participate during OKO activities.

Some observations also showed challenges in sustaining collaboration. In a few groups, students initially discussed their reasoning but eventually shifted to telling peers to "just put" a particular answer rather than continue working through the problem together. Group composition also mattered. In one classroom, some students were dissatisfied with their assigned groups and refused to participate, which reduced both engagement and persistence during the activity. Productive collaboration was therefore not sustained equally across all groups or throughout every session.

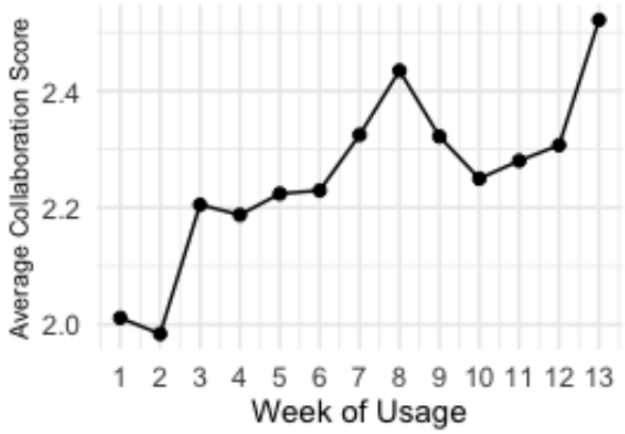
In contrast, teachers from control classrooms reported using Waggle, IXL, and Khan Academy. Observations of control classrooms showed that students were generally engaged, motivated, and persistent during mathematics instruction, but their interactions were primarily teacher-directed. Students frequently responded to teachers' questions and participated in whole-class discussions but rarely discussed mathematical content with one another.

RQ2: How do indicators of students' collaborative mathematical participation change over repeated use of the environment?

System-generated collaboration scores showed an overall increase across students' active weeks of OKO use. As shown in Figure 3, average weekly scores generally trended upward over the implementation period, and a monotonic trend test indicated a significant positive trend, $p = .002$.

Figure 3

Average Student Collaboration Scores Across Active Weeks of OKO Use (n = 255)

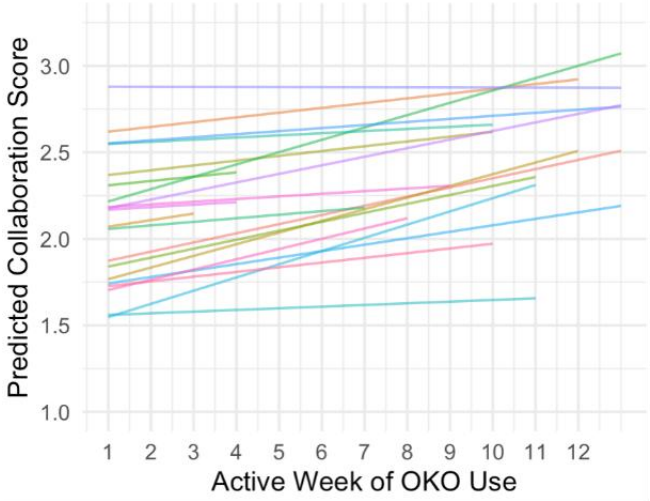


The mixed-effects growth model showed a similar pattern. The model-adjusted mean collaboration score in students' first active week was 2.08 on the 1–4 scale. Collaboration scores increased by an estimated 0.03 points per active week of use (standard error [SE] = 0.005, $p < .001$). Students varied in both their initial collaboration scores and their rates of change over time (random-intercept SD = 0.41; random-slope SD = 0.039). Figure 4 illustrates this variation using model-predicted

trajectories for a random sample of 20 students. Some students showed relatively little change, while others showed more substantial increases across their active weeks of use.

Figure 4

Model-Predicted Collaboration Score Trajectories for a Random Sample of Students (n = 20)



The Teamwork Scale for Youth showed a different pattern. After controlling for baseline scores, there was no statistically significant difference between treatment and control students on post-study teamwork scores. The adjusted mean was 4.05 in both groups (p = .976, Hedges' g = .004). Among OKO users, the adjusted mean was 4.06 compared with 4.04 for control students, and this difference was also not statistically significant (p = .862, g = .019).

RQ3: What evidence of promise is there that participation in AI-supported collaborative learning is associated with mathematics learning?

The randomized study had high outcome-data retention, with spring NWEA MAP Growth mathematics scores available for 998 of the 1,040 students (96%). Overall attrition was 4.04%, with differential attrition of 0.31 percentage points, meeting the What Works Clearinghouse criteria for low attrition (WWC, 2022). Among students with both winter and spring scores, the baseline difference between treatment and control groups was small (p = .34, Hedges' g = .191). Implementation varied across treatment classrooms.

According to the OKO platform data, among OKO users, students averaged nine weeks of use, sessions ran about 17.5 minutes roughly twice weekly, and seven of ten teachers met the study's minimum weekly threshold in at least 70% of their implementation weeks.

The primary intent-to-treat analysis showed positive, though not statistically significant, differences between students assigned to treatment and control classrooms. After adjusting for baseline achievement and student background characteristics, the adjusted mean spring NWEA MAP score was 220.27 for treatment students and 218.86 for control students, a difference of 1.40 RIT points (SE = 1.11, p = .218, Hedges' g = .074). Treatment students also showed greater excess growth than control students, with adjusted means of 2.24 and 0.79 RIT points above projected growth, respectively. The estimated difference was 1.46 RIT points (SE = 1.01, p = .166, g = .183).

Because only teacher-selected students assigned to the treatment condition actually used OKO, we also examined outcomes among students who used the system. In this exploratory comparison, OKO users had an adjusted spring MAP score of 221.37 compared with 218.92 for control students, a

difference of 2.46 RIT points ($p = .047$, $g = .132$). The excess-growth results showed the same pattern. OKO users averaged 3.21 RIT points above projected growth, compared with 0.75 among control students, a difference of 2.46 RIT points ($p = .033$, $g = .312$). These results indicate that students who used OKO showed stronger mathematics outcomes than control students, although the comparison is exploratory because teachers selected which students used OKO after random assignment.

The matched analysis tested whether this pattern remained after reducing observed baseline differences between OKO users and comparison students. CEM produced a sample of 225 OKO users and 363 comparison students with strong baseline balance; the standardized mean difference in winter MAP scores was 0.02. In the CEM-weighted analysis, OKO users had an adjusted spring MAP score of 225.70 compared with 223.98 for matched comparison students, an estimated difference of 1.72 RIT points (SE = 1.09, $p = .129$, $g = .114$). For excess growth, OKO users averaged 3.20 RIT points above projection compared with 1.41 for matched comparison students, a difference of 1.79 RIT points (SE = 1.09, $p = .118$, $g = .244$). The matched analysis used a more restricted, weighted sample than the unmatched comparisons; therefore, the adjusted means should not be directly compared across analyses. The estimated differences remained positive but were smaller than those observed in the unmatched user-level comparisons.

Across the analyses, the mathematics outcome estimates were consistently positive, although their magnitude and statistical significance varied. Overall, the pattern provides early suggestive evidence of promise for mathematics learning among students who used OKO, while also indicating that the certainty of the estimates depends on the analytic approach.

Discussion and Implications

AI as a facilitator of student-student collaboration

The findings suggest that AI can support peer interaction without becoming the main partner in the activity. This role differs from much of the recent work on generative AI-supported collaboration, where students interact directly with an AI agent that provides ideas, feedback, or information (An et al., 2024; Feng, 2025). In OKO, the AI supports interaction among students while leaving the mathematical work with the group.

Teachers also described peer interaction as giving students opportunities to assess their own understanding. Listening to classmates' explanations sometimes led students to reconsider how well they understood a concept, and teachers described students teaching and learning from one another. This may be one advantage of using AI as a facilitator rather than a partner: students still have to engage with one another's reasoning.

Group interaction was not uniform. Some students participated more than others, and some groups shifted from explanation to answer giving or disengaged because of group composition. These differences point to a key design challenge for AI-supported collaboration: deciding when a group needs support, what kind of prompt is likely to help, and when the system should stay in the background. Effective facilitation may need to respond to both the mathematical work and the social dynamics of the group.

Situated collaboration and broader perceptions of teamwork

System-generated collaboration scores increased over repeated OKO use, but the Teamwork Scale showed no difference between treatment and control students. Several factors may explain the difference. One possible explanation is that the measures capture different aspects of collaboration. The OKO score reflects reasoning and reciprocity during specific mathematics sessions, whereas the

Teamwork Scale measures broader perceptions of teamwork. Students may change how they participate in a particular activity without changing how they view themselves as collaborators more generally.

The measures also differ in proximity to the activity. OKO scores were generated during each session, while the Teamwork Scale was administered before and after the study. The session-level measure may therefore be more sensitive to short-term changes. The increase in OKO scores may partly reflect students becoming more familiar with the activity and with the forms of participation encouraged by the system. Over time, students may have learned when to explain their reasoning, respond to peers, or offer help.

These explanations are hypotheses. Future studies should compare system-generated measures with independent coding of student interaction and use measures closer to the collaborative activity, such as mathematics-based situational judgment tasks or brief post-session surveys. These measures could help distinguish changes in situated collaborative behavior from changes in students' broader perceptions of teamwork.

Evidence of promise for mathematics learning

Mathematics outcome estimates were positive across the intent-to-treat, user-level, and matched analyses, although their magnitude and statistical significance varied. The exploratory user-level analysis produced the largest differences, while the matched analysis retained positive but smaller estimates.

The implementation pattern is important for interpreting these findings. Only about half of the students assigned to treatment classrooms used OKO because teachers often selected students whom they considered well suited to the intervention. Teacher selection creates challenges for estimating the effect of actual OKO use, but it also reflects how the program was implemented in practice. The user-level findings therefore provide information about outcomes under a targeted implementation model rather than universal classroom use.

This pattern raises a question for future research: whether OKO is more effective when offered broadly or when teachers use their professional judgment to target students or groups. Examining both approaches could help identify which students benefit most and under what conditions.

Conclusion

This study provides encouraging evidence that multimodal AI can support collaborative mathematics learning while keeping student-student interaction at the center of the activity. Classroom observations and teacher interviews documented frequent mathematical explanation, peer support, and collaborative problem-solving during OKO sessions, while system-generated collaboration scores increased with repeated use. Together, these patterns suggest that AI-supported facilitation can create opportunities for productive mathematical interaction among peers. Students' self-reported teamwork, however, did not show corresponding improvements, suggesting that changes in collaborative participation during OKO activities may not extend to broader perceptions of teamwork. Mathematics achievement estimates were consistently positive across analyses, although the randomized intent-to-treat estimates were not statistically significant. Exploratory comparisons showed significantly higher achievement among OKO users than control students, but these differences were smaller and no longer statistically significant after matching on observed baseline characteristics. Overall, these findings highlight the potential of AI to facilitate collaborative learning and suggest possible benefits for mathematics achievement, while underscoring the need for further research to determine whether and under what conditions this approach improves student learning.

Several limitations should be considered when interpreting these findings. The study included only 20 randomized teachers and was not fully powered to detect modest effects. Implementation also varied across classrooms, and only about half of students assigned to treatment classrooms used OKO. A larger, fully powered study with more consistent implementation would provide more precise estimates of mathematics effects.

Teacher selection also limits causal interpretation of the user-level findings. OKO users differed from other treatment and control students at baseline. Matching improved comparability on observed characteristics but cannot account for unobserved factors related to both selection and outcomes. Future studies should examine both universal and targeted implementation and whether outcomes differ by student characteristics, group composition, and level of exposure.

The study also provides only a partial view of collaborative processes. Classroom observations captured examples of mathematical reasoning, peer support, persistence, and breakdowns in collaboration, but they were not designed for detailed interaction or discourse analysis. The system-generated collaboration score provided repeated information about reasoning and reciprocity but was not independently validated against human coding. Future studies could combine multimodal system data with systematic coding of student discourse and interaction to examine how specific facilitation moves relate to subsequent group behavior.

The study was conducted in one district and primarily in fifth-grade classrooms. Research across grades, mathematics content, schools, and student populations is needed to assess the generalizability of findings. Future work should also examine when AI facilitation is useful, for whom, and under what group and classroom conditions.

Acknowledgments

The research reported here was supported by the U.S. Department of Education, through the SBIR program, Grant# [blinded], [A Foundation] and [B] The opinions expressed are those of the authors and do not represent views of the funders.

References

An, S., Li, Y., Yang, Y., Ma, Y., Cheng, G., & Xu, G. (2024). Developing an LLM-empowered agent to enhance student collaborative learning through group discussion. *Proceedings of the 32nd International Conference on Computers in Education*. https://doi.org/10.58459/icce.2024.4953

Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *The Journal of the Learning Sciences, 9*(4), 403–436.

Dillenbourg, P., & Tchounikine, P. (2007). Flexibility in macro-scripts for computer-supported collaborative learning. *Journal of Computer Assisted Learning, 23*(1), 1–13. https://doi.org/10.1111/j.1365-2729.2007.00191.x

Feng, S. (2025). Group interaction patterns in generative AI-supported collaborative problem solving: Network analysis of the interactions among students and a GAI chatbot. *British Journal of Educational Technology, 56*(5), 2125–2145. https://doi.org/10.1111/bjet.13611

Fischer, F., Kollar, I., Stegmann, K., & Wecker, C. (2013). Toward a script theory of guidance in computer-supported collaborative learning. *Educational Psychologist, 48*(1), 56–66. https://doi.org/10.1080/00461520.2012.748005 Taylor & Francis Online

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*(1), 1–24. https://doi.org/10.1093/pan/mpr013

Kraft, M.A. (2023). The effect size benchmark that matters most: Education interventions often fail. *Educational Researcher*, 52(3), 183-187.

Lower-Hoppe, L. M., Newman, T. J., & Anderson-Butcher, D. (2017). Validity and reliability of the Teamwork Scale for Youth. *Research on Social Work Practice, 27*(6), 716–725. https://doi.org/10.1177/1049731515589614

Meyer, P., Johnson, J., Li, X., Hu, A., & Hall, C. (2025). *MAP Growth technical report: 2024–2025.* NWEA.

NWEA. (2025). *2025 MAP Growth norms technical manual*. NWEA Research Report. NWEA.

Orosco, M. J. (2016). Measuring elementary student's mathematics motivation: A validity study. *International Journal of Science and Mathematics Education, 14*(5), 945–958. https://doi.org/10.1007/s10763-015-9632-7

Ouyang, F., & Bai, X. (2026). A systematic review of multimodal learning analytics in computer-supported collaborative learning. *Computers & Education, 247*, 105574. https://doi.org/10.1016/j.compedu.2026.105574 ScienceDirect

Roschelle, J., Rafanan, K., Bhanot, R., Estrella, G., Penuel, W. R., Nussbaum, M., & Claro, S. (2010). Scaffolding group explanation and feedback with handheld technology: Impact on students' mathematics learning. *Educational Technology Research and Development, 58*(4), 399–419. https://doi.org/10.1007/s11423-009-9142-9

Soller, A. (2001). Supporting social interaction in an intelligent collaborative learning system. International Journal of Artificial Intelligence in Education, 12, 40–62.

Webb, N. M., Nemer, K. M., & Ing, M. (2006). Small-group reflections: Parallels between teacher discourse and student behavior in peer-directed groups. *The Journal of the Learning Sciences, 15*(1), 63–119. https://doi.org/10.1207/s15327809jls1501_8

Wei, Y., & Perkins, M. (2026). Generative AI and student collaboration: A scoping review of group work processes, outcomes, and risks. *International Journal for Educational Integrity, 22*, Article 8. https://doi.org/10.1007/s40979-026-00217-x

What Works Clearinghouse. (2022). What Works Clearinghouse procedures and standards handbook (Version 5.0). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Handbooks

Yin, X., Ye, J., Yu, S., Li, H., Liu, Q., & Zhao, G. (2025). The association between groups' interactions with the Visual-GenAI learning analytics feedback and student engagement in CSCL. *Computers & Education, 239*, 105434. https://doi.org/10.1016/j.compedu.2025.105434

Zhou, Q., Suraworachet, W., & Cukurova, M. (2024). Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions. *Education and Information Technologies, 29*, 1071–1098. https://doi.org/10.1007/s10639-023-12315-1