

When More Dialogue Does Not Help: Evaluating Instructional Efficiency in LLM- Supported Reflective Learning

Bruce M. MCLAREN^{a*}, Aditi HAIMAN^a, Daniel POLLOCK^a, J. Elizabeth RICHEY^b, Esma N. KAHVECI^c, Alessandro PAGANO^d, Marshall AN^a, Mahboobeh MEHRVARZ^a & Hayden STEC^a

^a*Carnegie Mellon University, USA*

^b*University of Pittsburgh, USA*

^c*University of Massachusetts Dartmouth, USA*

^d*University of Bari Aldo Moro, Italy*

*bmclaren@cs.cmu.edu

Abstract: Reflection and self-explanation help learners connect procedural and conceptual knowledge, yet digital learning environments often provide limited opportunities for structured reflection. Recent advances in Large Language Models (LLMs) enable personalized Socratic dialogue, but it remains unclear whether richer, longer interaction improves learning enough to justify its additional instructional cost. We investigated this question in a digital mathematics learning game for 5th- and 6th-grade students through two classroom studies. Study 1 (N = 191) compared multiple-choice self-explanation, open-ended self-explanation, and multi-turn LLM-supported Socratic dialogue. All three conditions produced significant learning gains, but increasingly dialogic reflection did not improve learning, while the open-ended and Socratic dialogue conditions required substantially more instructional time. Study 2 (N = 33) examined whether LLM-generated personalization improved learning when dialogue was reduced to a single personalized Socratic question. Again, learning gains and instructional time were comparable across conditions. For upper-elementary and middle-school learners, extended dialogue and personalization did not yield measurable benefits beyond concise reflection prompts. In sum, the findings suggest that richer AI interaction is not inherently better instruction, and that new AI-enabled pedagogies should be evaluated against strong existing instructional baselines while considering instructional efficiency.

Keywords: Large Language Models, Socratic Dialogue, Reflection, Self-Explanation, Digital Learning Games, Instructional Efficiency

1. Introduction

Reflection is widely recognized as an important mechanism for promoting conceptual understanding in mathematics. Asking learners to explain why a solution works, justify a strategy, or identify misconceptions can strengthen connections between procedural and conceptual knowledge and improve transfer (Chi et al., 1989; Rittle-Johnson & Loehr, 2017; Wylie & Chi, 2014). Consequently, reflective prompts and self-explanation have become common instructional strategies in intelligent tutoring systems and other AI-supported learning environments.

Recent advances in Large Language Models (LLMs) have enabled new opportunities to extend reflective learning through personalized Socratic dialogue. Rather than presenting fixed prompts, LLMs can generate context-sensitive questions that respond to learners' prior answers, misconceptions, and reasoning processes (Kasneci et al., 2023). Recent reviews suggest that LLMs show considerable promise for tutoring, feedback, and conversational learning support while also emphasizing the need for rigorous empirical evaluation against

established instructional methods rather than technology demonstrations alone (Kasneci et al., 2023; Lo, 2023). Despite this growing interest, relatively little empirical work has examined whether extended LLM-supported dialogue produces greater learning than simpler reflective supports, particularly for younger learners in authentic classroom settings.

Whether richer reflective dialogue improves learning is theoretically unclear. Cognitive Load Theory suggests that additional reading, writing, and reasoning demands may impose extraneous cognitive load, particularly for novice learners (Sweller, 1988; Sweller et al., 2011). In contrast, the ICAP framework predicts that learning improves as learners progress from passive to active, constructive, and interactive forms of engagement, suggesting that multi-turn Socratic dialogue may improve learning when it promotes deeper constructive or interactive cognitive engagement than traditional self-explanation (Chi & Wylie, 2014). Thus, richer, longer dialogue offers both potential costs and potential benefits, making its educational value an important empirical question.

As GenAI becomes easier to integrate into educational software, educators increasingly need evidence that multi-turn dialogue provides educational value beyond well-designed non-LLM alternatives. More broadly, this reflects the **Assistance Dilemma**: determining how much instructional support should be provided without reducing productive cognitive engagement (Koedinger & Alevan, 2007). Accordingly, we evaluate LLM-supported reflection in terms of both learning outcomes and instructional efficiency.

To investigate this question, we embedded reflective scaffolding within *Decimal Point*, a digital mathematics learning game for 5th- and 6th-grade students. Study 1 compared three increasingly interactive forms of reflection: multiple-choice self-explanation (**MCQ-SE**), open-ended self-explanation (**OE-SE**), and multi-turn LLM-supported Socratic Dialogue for Reflection (**SDR**). Because Study 1 manipulated both dialogue length and personalization simultaneously, Study 2 explored whether any potential benefits of personalization became apparent when the dialogue was reduced to a single personalized Socratic question.

We investigate three questions:

- **RQ1:** Do increasingly dialogic forms of reflection improve learning?
- **RQ2:** What instructional costs accompany these increasingly interactive reflection formats?
- **RQ3:** Does LLM-generated personalization improve learning beyond non-personalized self-explanation when dialogue length is controlled?

Rather than asking whether LLMs can generate richer dialogue, we ask a more important question: **When does richer, longer dialogue provide sufficient instructional value to justify its additional cost?**

2. Method

Participants and Design. The research consisted of two classroom studies with 5th- and 6th-grade students from schools in the northeastern United States. Both studies used the same digital learning game, assessments, and procedures during regular mathematics instruction.

Study 1 included 282 participants, of whom 191 completed all instructional materials and assessments. Students were randomly assigned to:

- **MCQ-SE:** multiple-choice self-explanation;
- **OE-SE:** open-ended self-explanation; and
- **SDR:** multi-turn LLM-supported Socratic Dialogue for Reflection.

Study 2 was an exploratory follow-up (48 participants; 33 completed) comparing a single personalized Socratic question (SQ) with OE-SE.

Learning Environment. Both studies used *Decimal Point*, a digital mathematics learning game designed to improve students' understanding of decimal numbers and common decimal misconceptions (McLaren, 2024; McLaren et al., 2017). Students completed a sequence of mini-games involving number-line estimation, decimal comparison, ordering, sequencing, and arithmetic operations. Reflection activities were embedded after selected game episodes. The instructional content and gameplay remained identical across experimental conditions. Only the form of post-problem reflection differed.

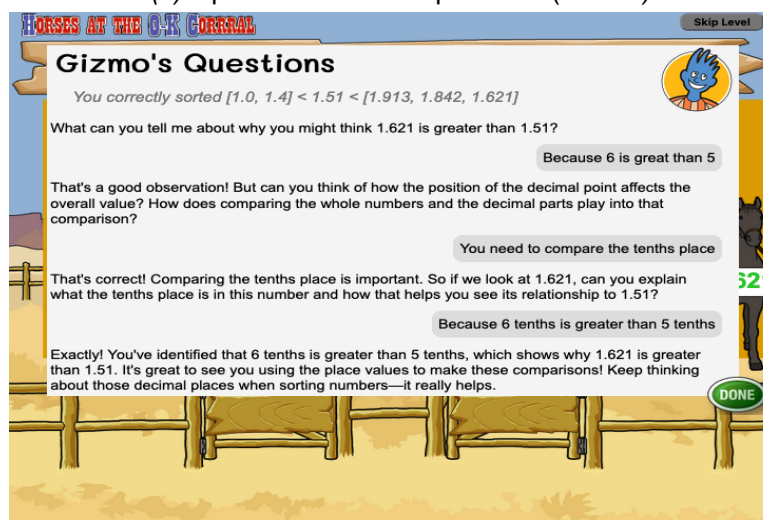
Figure 1 illustrates the three Study 1 reflection conditions, representing an ordered progression from selecting an explanation to generating an explanation to engaging in adaptive multi-turn dialogue. Reflection followed the same mathematical problems at identical points during gameplay; only its form differed.



1(a) Multiple-Choice Question Self-Explanation (**MCQ-SE**)



1(b) Open-Ended Self-Explanation (**OE-SE**)



1(c) LLM-supported Socratic Dialogue for Reflection (**SDR**)

Figure 1. Post-problem-solving reflection for the same decimal item in *Decimal Point* across the three conditions.

Experimental Conditions. The three Study 1 conditions represented increasing levels of learner generation and interaction. In the **MCQ-SE** condition, students selected among several candidate explanations for their solution. In the **OE-SE** condition, students wrote a brief explanation describing why their solution was correct. In the **SDR** condition, students engaged in a three-turn Socratic dialogue generated by an LLM. After each learner response, the system generated a follow-up question intended to probe reasoning.

The SDR implementation used GPT-4o-mini (OpenAI, 2024) because of its low latency, allowing conversational interaction without substantial delays during classroom use. The prompt supplied to the model included the current problem, the student's previous responses, identified misconceptions, and instructions to generate age-appropriate Socratic questions following Stevens and Collins (1977). Conversation history was maintained across dialogue turns.

Study 2 was designed in response to Study 1 to determine whether the null findings reflected extended dialogue rather than LLM-generated personalization. Because of district technology requirements, Gemini 2.5 Flash-Lite (Gemini Team, 2025; Google Cloud, 2026) was used instead of GPT-4o-mini in Study 2. Aside from eliminating additional dialogue turns, prompting followed the same instructional strategy as Study 1.

Measures. Learning was measured using isomorphic pretests, immediate posttests, and delayed posttests assessing both procedural fluency and conceptual understanding of decimal numbers. The assessment instruments have been used extensively in previous evaluations of *Decimal Point*.

Instructional efficiency was assessed using total time spent completing the intervention. Because one of the primary goals of this work was to evaluate the educational value of extended LLM-supported interaction, learning outcomes were interpreted jointly with instructional time rather than in isolation, consistent with instructional efficiency perspectives in educational research.

3. Results

Study 1: Learning Outcomes. All three reflection conditions produced significant learning gains (Table 1). A mixed-measures ANOVA revealed a significant main effect of time, indicating improvement from pretest to posttest and delayed posttest (all $p < .001$). However, neither the main effect of condition nor the interaction between condition and time was significant. **Students learned equally well across the three reflection conditions.**

Table 1. Means and Standard Deviations on Tests (Scored out of 100) by Condition

Condition	N	Pretest <i>M</i> (<i>SD</i>)	Posttest <i>M</i> (<i>SD</i>)	Delayed Posttest <i>M</i> (<i>SD</i>)
SDR	67	45.1 (19.8)	51.8 (19.7)	54.3 (20.5)
OE-SE	58	48.1 (20.2)	53.8 (18.2)	56.2 (21.0)
MCQ-SE	66	47.8 (19.2)	54.7 (18.9)	56.3 (17.4)

These findings answer **RQ1**: increasingly dialogic forms of reflection did not produce greater learning gains.

Study 1: Instructional Efficiency. Although learning outcomes were equivalent across conditions, instructional time differed significantly. Students in the OE-SE and SDR conditions required approximately **25% more instructional time** than students in the MCQ-SE condition to complete the intervention ($p = .011$ and $p = .012$, respectively), with no significant difference between OE-SE and SDR. Thus, richer, longer reflective interactions increased instructional time without improving learning.

Post-intervention questionnaires measuring engagement, situational interest, enjoyment, self-efficacy, challenge, meaningfulness, mathematics self-concept, and test anxiety showed no significant differences between conditions.

These findings answer **RQ2**: more interactive reflection increased instructional cost but did not improve learning.

Study 2: Personalization Versus Dialogue Length. Because Study 1 combined multi-turn dialogue with personalization, it was unclear whether the absence of learning benefits reflected the cognitive demands of extended dialogue or the limited value of personalization itself. Study 2 therefore isolated personalization while minimizing dialogue length. Students were assigned either a single personalized Socratic question (SQ) or a traditional open-ended self-explanation.

Consistent with Study 1, both conditions produced significant learning gains over time, but no significant differences were observed between conditions. Total instructional time also did not differ significantly. Although the personalized Socratic question was generated from each student's prior problem-solving behavior and inferred misconceptions, it produced learning outcomes comparable to those obtained using a standard open-ended self-explanation.

Because Study 2 included only 33 analyzed students, these findings should be interpreted as exploratory. Nevertheless, they provide preliminary evidence that LLM-generated personalization alone may not improve learning beyond concise non-personalized reflection prompts.

Together, these findings answer **RQ3**: when interaction length was controlled, LLM-generated personalization did not produce measurable learning benefits.

4. Discussion

We investigated whether richer and longer LLM-supported reflective dialogue provides educational benefits beyond simpler forms of reflection in a classroom-oriented digital mathematics learning game. Across two classroom studies, students consistently demonstrated significant learning gains regardless of whether reflection was delivered through multiple-choice self-explanation, open-ended self-explanation, multi-turn Socratic dialogue, or a single personalized Socratic question. However, the more interactive approaches generally required additional instructional time without producing greater learning. Together, these findings identify an important boundary condition for the use of LLM-supported reflective dialogue with upper-elementary and middle school learners. Importantly, the second study suggests that the null findings from Study 1 cannot be attributed solely to the cognitive demands of extended dialogue; reducing the interaction to a single personalized question also failed to improve learning.

One plausible explanation is that increasing dialogic interaction does not necessarily produce the deeper cognitive engagement needed to improve learning. Cognitive Load Theory suggests that instructional designs requiring additional reading, writing, and reasoning may impose extraneous cognitive demands on younger learners, offsetting potential benefits (Sweller, 1988; Sweller et al., 2011). Similarly, the ICAP framework predicts that dialogue should outperform simpler reflection only when it successfully promotes constructive or interactive engagement (Chi & Wylie, 2014). In the present studies, it appears that extended dialogue did not consistently translate into deeper learning processes.

An informal review of SDR transcripts suggested several possible explanations for these findings, including inconsistent probing of misconceptions and limited adaptation to minimal learner responses. A more systematic qualitative analysis is an important direction for future work.

More broadly, these findings suggest that evaluations of LLM-supported educational technologies should focus on **instructional value**, not simply technological capability (Kasneci et al., 2023). LLMs enable personalized conversation, but longer, richer interaction does not inherently improve learning. Accordingly, instructional innovations should be evaluated against strong pedagogical baselines, not merely by how sophisticated the AI is.

Study 2 should be interpreted as exploratory because of its limited sample size. Future work should examine older learners, spoken interaction, adaptive dialogue strategies, and improved probing of misconceptions.

5. Conclusion

LLMs have enabled new opportunities to incorporate conversational reflection into educational software. The present work demonstrates that these capabilities should not be evaluated solely by how sophisticated the interaction is, but by whether the interaction produces meaningful educational benefits beyond simpler instructional alternatives.

Across two classroom studies, increasingly dialogic and personalized forms of reflection did not improve learning relative to concise self-explanation prompts, while often requiring additional instructional time. For upper-elementary mathematics learning, these findings suggest that brief, carefully designed reflection may provide a better balance between effectiveness and efficiency than extended LLM-supported dialogue.

As generative AI becomes more deeply integrated into educational technologies, instructional innovations should be evaluated against strong pedagogical baselines rather than by how sophisticated the AI is. Identifying when LLM-supported dialogue genuinely adds educational value and when simpler reflection is sufficient remains an important challenge for future research.

References

- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2), 145–182.
- Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243.
- Gemini Team. (2025). *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*. arXiv. <https://doi.org/10.48550/arXiv.2507.06261>
- Google Cloud. (2026). Gemini 2.5 Flash-Lite. *Google Cloud Documentation*.
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Koedinger, K.R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educ. Psychol. Rev.* 19(3), 239–264. <https://doi.org/10.1007/s10648-007-9049-0>
- Lo, C. K. (2023). What Is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410.
- Mayer, R. E. (2019). Computer games in education. *Annual Review of Psychology*, 70, 531–549.
- McLaren, B.M. (2024). *Decimal Point: A decade of learning science findings with a digital learning game*. In: Ilic, P., Casebourne, I., Wegerif, R. (eds) *Artificial Intelligence in Education: The Intersection of Technology and Pedagogy*. Intelligent Systems Reference Library, vol 261. Springer, Cham. https://doi.org/10.1007/978-3-031-71232-6_9.
- McLaren, B.M., Adams, D.M., Mayer, R.E., & Forlizzi, J. (2017). A computer-based game that promotes mathematics learning more than a conventional approach. *International Journal of Game-Based Learning (IJGBL)*, 7(1), 36-56. <https://doi.org/10.4018/IJGBL.2017010103>
- OpenAI. (2024). GPT-4o Mini: Advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>
- Rittle-Johnson, B., & Loehr, A. M. (2017). Eliciting explanations: Constraints on when self explanation aids learning. *Psychonomic Bulletin Review*, 24, 1501–1510.
- Stevens, A. L., and Collins, A. (1977). The goal structure of a Socratic tutor. In *Proceedings of the 1977 annual conference*, 256–263. ACM.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285.
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. Springer.
- Wylie, R., & Chi, M. T. H. (2014). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 413–432). Cambridge University Press.