



PEER-REVIEWED RESEARCH

Make Meaningful Feedback Possible at Scale with AI

Overview of the paper: A Large Scale Randomized Control Trial
Showing LLM Generated Feedback Helps Low-Knowledge Middle
School Math Students with Short-Term Learning

DATE

September 2026

AUTHORS

Eamon Worden

ABOUT ASSISTMENTS

ASSISTments is a [research-proven digital math learning tool](#) shown to increase student learning by 73%. As students complete their work, they receive immediate feedback and teachers get real time, actionable insight into where students need help. ASSISTments is actively incorporating Artificial Intelligence (AI) to amplify teacher insight, deepen student learning, and make math practice smarter, more responsive, and more equitable. Our Learning Briefs series shares these findings openly, detailing the questions we set out to answer, the methods and data used, and the key insights that guide our next steps.

AI-generated feedback is as good as teacher-written feedback and affordable enough to support every mistake.

What the Study Found

This study tested the potential benefits of providing students with an AI-generated, targeted feedback message based on the wrong answer they submitted.

Comparing targeted feedback to no targeted feedback

Students who received the targeted AI message were more likely to correct the problem on their next attempt without asking for a hint. Across 87,270 problem attempts, students who received AI feedback targeted to their exact error had higher odds of answering the problem correctly on their next attempt than students who only saw a message that their attempt was incorrect. This pattern held regardless of how well the student had performed earlier in the assignment.

On the next problem, the benefit went entirely to the students with more unfinished learning.

Students with more intensive learning needs improved significantly: their odds of success on the next problem were higher than those of similar students who only received correctness feedback on the previous problem.

We determined who had the most need by dividing students into three groups based on how likely they were to benefit from more support, using their scores on the last 5 problems as the measure. After checking whether the immediate revision was correct, we checked whether each student got the next problem right on the first try, without requesting a hint or explanation.

Students whose prior performance indicated they were mostly on track did not improve significantly; however, their future performance did not decrease either.

LLM-generated feedback helped students with more need for support but did not hurt students with minimal need for support.

Comparing teacher-written feedback to AI-written feedback

AI-written feedback performed as well as teacher-written feedback. In a separate study running at the same time, students were randomly assigned to receive either a teacher-written message or an AI-written one, when both were available. This comparison was based on 12,950 attempts. We did not observe performance differences between those who received teacher feedback and those who received LLM feedback. This suggests that our AI-generated feedback is likely similar in quality to teacher-written feedback. We grounded the prompt in examples of teacher-written feedback, so the AI's messages reflect the style and approach teachers already use.

What was different was the cost. We conservatively estimate it takes about 2 minutes for a teacher to write a targeted feedback message, so writing 3,862 pieces of feedback would take about 16 days of work for one person. The AI did it in 28 minutes and costs only 5 cents per message with proprietary LLMs. If we pay someone \$45/hour to write feedback messages, it would cost over \$5,500 compared to \$193 for the AI. These messages are stored and can be reused whenever a student submits the same wrong answer.

TEACHER-WRITTEN • 3,862 MESSAGES

\$5,500+

~16 days of work for one person

AI-WRITTEN • 3,862 MESSAGES

\$193

28 minutes

What the Feedback Actually Looked Like

The AI followed these rules: Keep feedback concise. Name the specific mistake. Point toward the next step. Never give the answer.

EXAMPLE 1

Problem: A bakery used 30% more sugar this month than last month. If the bakery used 560 kilograms of sugar last month, how much did it use this month?

Correct answer: 728 kilograms

Student's answer: 168 kilograms

Feedback: 168 kg is 30% of the total 560 kg of sugar. This month, the bakery used 30% more than the 560 kg of sugar.

EXAMPLE 2

Problem: A car travels 60 miles in 1 hour. How long will it take to travel 15 miles? Give your answer in minutes.

Correct answer: 15 minutes

Student's answer: 900 minutes

Feedback: 60 miles in 1 hour is the same as 60 miles in 60 minutes. Since 15 miles is smaller than 60 miles, it will take less than 60 minutes to travel 15 miles.

How the Study Was Run

Selecting the wrong answers that needed targeted feedback

Because thousands of students had solved these problems in ASSISTments for over 5 years, we had extensive data on which wrong answers were common. We selected 653 seventh-grade problems from the Illustrative Mathematics curriculum and found 3,862 wrong answers across them, each answered at least 100 times over the last 5 years.

Crafting the AI to write good, targeted feedback

We collaborated with two former teachers, each with more than seven years of classroom experience, and worked with them throughout five rounds of prompt revisions. Some major rules were:

- never reveal the answer,
- acknowledge what the student got right,
- give one targeted hint,
- use the problem's own vocabulary,
- stop at three sentences.

Once the prompt was ready, we used DeepSeek to write all 3,862 messages in 28 minutes. Then we used GPT-4.1-mini to review every message for math errors, teaching quality, and a seventh-grade reading level. Anything it flagged was discarded. The two former teachers then reviewed 100

randomly selected messages by hand and found one math error, which they removed; because the error rate was ~1%, we felt confident adding AI-generated feedback at scale.

Running the study

From August 2025 through January 31, 2026, students did their normal classwork and homework. Each time a student opened a problem, the system randomly assigned them to see an AI-feedback message or not. Two comparisons ran at once. In total, 21,478 students participated; some students participated in both studies.

Study 1: AI message vs. nothing

20,706 students in 1,577 classrooms, 87,270 problem attempts. The control group saw only a mark indicating the answer was wrong.

Study 2: AI message vs. a teacher's message

6,055 students in 594 classrooms, 12,950 problem attempts. The control group saw a teacher-written message for that same wrong answer.

Success was measured in two ways:

- Did the student get that problem right on the attempt right after receiving the targeted feedback?
- Did the student get the next problem right on the first try, without asking for a hint or explanation?

We account for student knowledge by including how many of their previous five problems they got right as a covariate, since these problems are often the same or very similar skills. That score separated students into three groups:

0–1 of 5

May benefit from more support

2–3 of 5

Moderate support needs

4–5 of 5

Minimal support needs

Oversight

The analysis plan was publicly registered before the data were examined, and WPI's Institutional Review Board reviewed the study. Every message students saw carried a visible notice that it may have been written by AI, along with a button to report it.

Worden, E., Dang, L., Lim, W. C. I., Miller, S., Zhang, J., Haim, A., Sales, A., Gurung, A., & Heffernan, N. (2026). A Large Scale Randomized Control Trial Showing LLM Generated Feedback Helps Low-Knowledge Middle School Math Students with Short-Term Learning. In *Proceedings of the Thirteenth ACM Conference on Learning @ Scale* (pp. 361–372).

Read the peer-reviewed paper, presented at the Conference on Learning at Scale, Seoul, 2026:

<https://doi.org/10.1145/3774398.3811594>

Data, analysis code, and the full prompt: <https://osf.io/q7jmy/overview>