

Efficacy of Student–AI Co-Authored Math Word Problems in an Intelligent Tutoring System

Kole Norberg, April Murphy, Stephen Fancsali, and Steve Ritter

Carnegie Learning
436 Seventh Avenue, Suite 200
Pittsburgh, PA 15219
knorberg1@carnegielearning.com

Abstract

We evaluated student ($N = 1,361$) interest and performance on student-AI co-authored math word problems. Performance matched or exceeded standard problems. Students rated peer-authored problems more often, especially when authorship was disclosed. Liking predicted first-attempt accuracy when problems required greater textual engagement, supporting interest-based context personalization.

1 Background

Mathematics story problems require students to coordinate mathematical reasoning with the interpretation of narrative text (Koedinger and Nathan, 2004). As a result, contextual familiarity with the narrative can influence reading comprehension, cognitive load, and motivation and thereby have an effect on student performance (Walkington et al., 2015). A student unfamiliar with the vocabulary or cultural assumptions in a problem may face demands unrelated to the intended mathematical construct, so performance differences across problems may partly reflect access to context rather than mathematical knowledge (Walkington et al., 2018).

Context personalization aims to reduce these construct-irrelevant demands, and aligning problems with student interests has been found to improve accuracy, speed, and engagement (Bernacki and Walkington, 2018; Clinton and Walkington, 2019; Walkington, 2013). These benefits appear inconsistently, however, sometimes producing small or null effects (Fancsali and Ritter, 2014) and sometimes distracting students when extraneous details fail to integrate with the mathematics (Harp and Mayer, 1998; Magner et al., 2014). In addition to requiring the appropriate grain size of personalization, personalizing problem context for each student can present a challenge due to students' wide variety of interests (Walkington, 2026; Walkington and Hayata, 2017).

Generative AI offers a means of producing a broader, more responsive range of contexts. The AI Math Personalization Tool (AMPT) supported students in co-authoring story problems around their interests while preserving the mathematical structure required by the curriculum (Norberg et al., 2026). After writing problems with AMPT and learning their problems would be added to the Intelligent Tutoring System they used in their math class (MATHia), students' sense of belonging in mathematics improved (Norberg et al., 2025b). Although the authoring approach succeeded as a method to both engage students and generate larger and more varied samples of problem contexts, it was not clear whether these student–AI co-authored problems would pose additional difficulties for students who did not author them. Language models are prone to writing longer texts with less familiar vocabulary (Norberg et al., 2023) both of which can decrease performance in mathematics (Abedi and Lord, 2001; Norberg et al., 2025a,c; Walkington et al., 2018). Further, some contexts preferred by students while co-authoring the problems might not be familiar or interesting to a broader cohort. Therefore, in this study we tested if the student-AI authored problems created any additional challenges to students over standard problems.

A further question concerned authorship itself. Co-authoring problems with AMPT raised students' sense of belonging in mathematics (Norberg et al., 2025b), and students who later solve those problems might also respond to knowing that a classmate wrote them. Thus, we tested how notifications of authorship affected students' ratings and performance on the problems.

2 Aims

We examined a large-scale deployment of student co-authored problems to ask (a) whether students who did not write a problem perform as well on it

as on standard problems, (b) whether telling students that a peer authored the problem changes their performance and how they rate the problem context, (c) whether students like the same problem context, and (d) whether liking a problem context predicts performance on it.

3 Sample

Middle school students (initial $n = 1,452$) working through MATHia (an Intelligent Tutoring System) as standard educational practice were randomly assigned using UpGrade (Ritter et al., 2020) to three conditions: Co-Authored, Co-Authored + Notification, or Control. Six percent of students (91) assigned to a co-authored condition never encountered a co-authored problem and were dropped. The dataset comprised 2,049 problems (63% co-authored) across seven lessons on the topics of probability, algebra, and ratios. Lessons within a topic share a format, so we report results by topic. Random assignment occurred within topic which resulted in some students ($n = 141$) participating in more than one condition across separate topics. Across conditions, there were 525 control, 497 co-authored, and 480 co-authored-plus-notification student-condition enrollments.

4 Method

In the co-authored conditions, students saw a mixture of co-authored and standard problems; in the notification condition, co-authored problems carried the message “This problem was authored by a MATHia student”; control students received no co-authored problems. In both co-authored conditions, students could rate problems with a thumbs up or down (Figure 1). Because we hypothesized that rating alone may affect students’ experience, control was kept as a true business as usual condition and did not provide an option for rating.

In MATHia, a problem contains multiple steps, each associated with a skill the student is working to master. Students cannot move to the next problem until every step is answered correctly, so a student may make several errors on one step. MATHia traces each student’s mastery of a skill with Bayesian knowledge tracing (Corbett and Anderson, 1995; Ritter et al., 2007). Once all skills in a lesson are mastered, the student is said to have mastered the lesson. In this study, if a student reached 25 problems without mastering all of the skills the lesson concluded but with a flag

noting that the lesson was not mastered. Problems within a subtopic differ in difficulty according to the skills they assess, so each co-authored scenario was spread across problems assessing different skills following the same procedure for creating standard MATHia problems. For example, two problems could use the same scenario but assess different relations or skills.

All lessons utilized stories as part of the context; however, in the algebra topic the standard MATHia problems follow a highly predictable structure in which the problem can be answered by searching the text for the numbers and a few words which explain the relationship. In the ratio problems, on the other hand, the structure of the text is more varied. See Table 1 for descriptive statistics concerning the texts in each topic. We considered this variation in our models to understand if effects may vary based on how similar the texts within a topic are to each other.

Ratings asked “Do you like this problem’s storyline” and provided a thumbs up and thumbs down option. Ratings were optional and could be submitted before or after solving. 84.2% were submitted before solving. Although we asked students to rate the storyline, students might also be influenced by their performance on the problem. For this reason, we restricted analysis involving ratings to those provided before the student attempted to solve any steps in the problem and further controlled for problem difficulty. Problem difficulty was measured as each problem’s error rate across all students who have encountered it in MATHia, defined independently of this study’s conditions.

Lesson mastery was modeled among all student-lessons. Time to mastery, error rates, and first-attempt correctness were modeled only among the students who graduated the lesson, to avoid conflating these results with the graduation model. Students could complete more than one lesson. Therefore, lesson-level models used crossed random intercepts for students (nested within school) and lesson. Problem-level models used crossed random intercepts for students (nested within school) and problems (nested within lesson). Log odds are reported in tables but back transformed to odds ratios within the text.

Topic type	Words	Sentence length	Reading ease	Hapax ratio	Vocabulary variance
Picture algebra	66.9 / 81.6	10.3 / 10.3	68.7 / 72.8	0.13 / 0.20	0.62 / 0.80
Ratio and proportion	67.0 / 78.5	13.5 / 14.7	76.4 / 73.9	0.16 / 0.25	0.86 / 0.86
Probability	47.4 / 61.2	12.7 / 13.8	83.3 / 75.1	0.37 / 0.44	0.87 / 0.89

Table 1: Text Measures by Topic Type

Note. Each cell gives standard / peer-authored values. Words is words per problem, sentence length is tokens per sentence, and reading ease is Flesch reading ease. Hapax ratio is the proportion of a problem's words that appear only once. Vocabulary variance is the mean pairwise distance between problems' content-word sets within a cell, where 1 indicates that two problems share no content words.

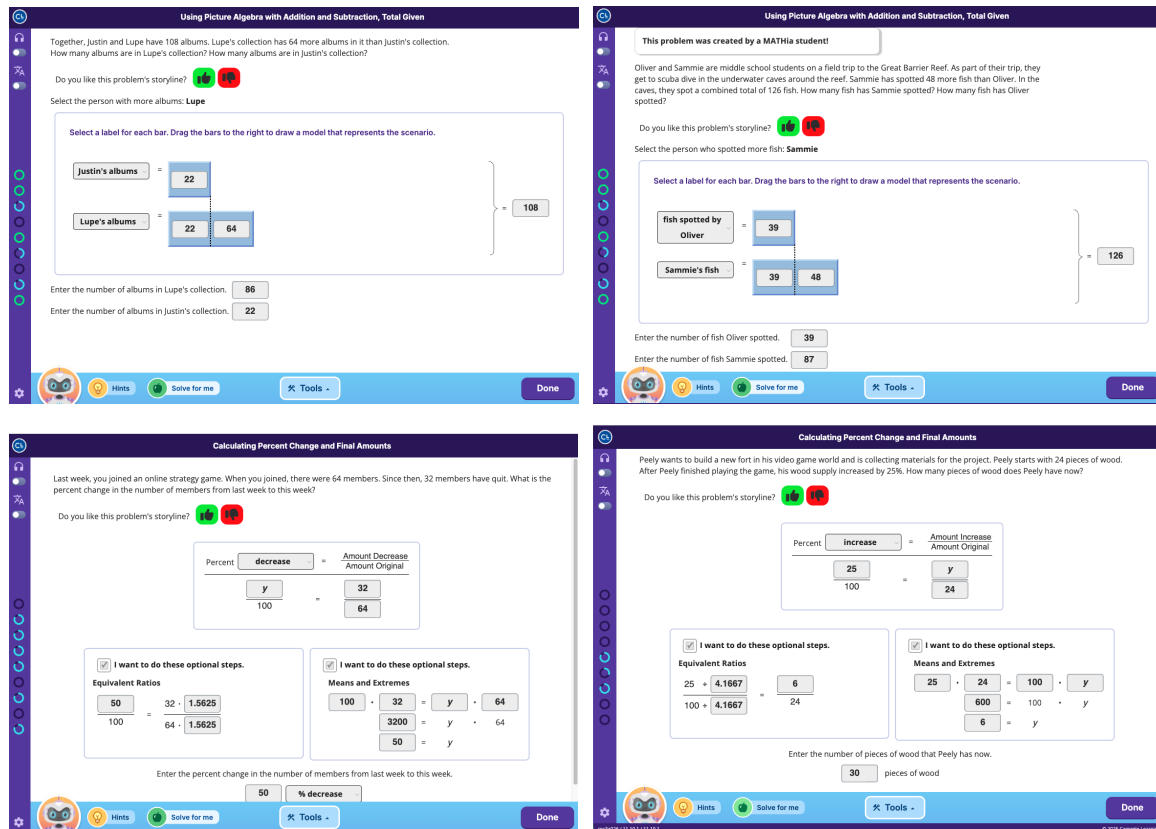


Figure 1: Standard (left) and peer-authored (right) picture-algebra problems with notification (top) and ratio-and-proportion problems without notification (bottom).

5 Results

5.1 Lesson-level outcomes

Lesson mastery rates in the co-authored conditions did not reliably differ from control nor did errors per step among students who mastered the lesson (all p -values $> .45$). However, students in the Co-Authoring + Notification condition took 2.68 min longer than students in the Control condition to master a topic, 95% CI [0.38, 4.98], $t(1851.3) = 2.28$, $p = .023$. Students in the Co-Authoring condition did not significantly differ in time to mastery from students in the Control condition, $p = .293$ (Table 2).

5.2 First attempt success at a step

The prior outcomes reflect broad lesson-level effects, but students in the Co-Authoring and Co-

Authoring + Notification conditions saw a mixture of peer-authored and standard problems. In this section we consider how often students answered a step correctly on their first attempt for standard as compared to peer-authored problems. The model contains a term for problem type and a term for condition. The intercept is a standard problem in the Control condition, the condition terms give standard problems in each co-authoring condition relative to that intercept, and the problem-type term gives peer-authored problems relative to the standard problems in the same condition.

The problem-type term compares the two kinds of problems a student saw in the same lesson. Students had 1.12 times greater odds of answering a step correctly on their first attempt for peer-authored problems than for standard problems, 95%

Predictor	<i>B</i>	<i>SE</i>	<i>z</i> / <i>t</i>	<i>p</i>
<i>Mastery (logistic)</i>				
Intercept	3.74	0.59	6.38	< .001
Co-Authoring	0.22	0.30	0.74	.459
Co-Authoring + Notification	−0.04	0.29	−0.15	.877
<i>Lesson-pooled error rate (linear; errors per step)</i>				
Intercept	0.19	0.02	12.34	< .001
Co-Authoring	0.004	0.01	0.53	.598
Co-Authoring + Notification	0.005	0.01	0.74	.459
<i>Time to mastery (linear; minutes)</i>				
Intercept	29.01	4.78	6.07	< .001
Co-Authoring	1.22	1.16	1.05	.293
Co-Authoring + Notification	2.68	1.18	2.28	.023

Table 2: Mixed-Model Estimates for Lesson-Level Mastery, Time to Mastery, and Pooled Error Rate

Note. Reference = Control. Mastery uses all student-lessons. Time and pooled-error models include mastered lessons only. Wald *z* and Satterthwaite *t* tests are reported.

CI [1.07, 1.18], $z = 4.45$, $p < .001$.

The condition terms compare performance on standard problems across conditions. Students in the Co-Authoring condition had slightly lower odds of answering correctly than students in the Control condition, $OR = 0.94$, 95% CI [0.88, 1.00], $z = -2.00$, $p = .046$. Performance did not significantly differ between the Co-Authoring + Notification and Control conditions, $p = .213$. See Table 3.

5.3 Ratings

Due to a technical error, ratings for the probability topic were not recorded and cannot be analyzed. Because the control condition contained no rating opportunity, the models in this section use only the two co-authoring conditions. To control for differences in rating and liking based on difficulty, each model includes problem difficulty, the problem's error rate across all students who have encountered it in MATHia.

5.3.1 Rating frequency

Students rated 45.3% of eligible problems, and they rated peer-authored problems more often than standard problems in both conditions, with a larger difference when the notification was present. With the notification present, students rated 50.4% of peer-authored problems and 42.8% of standard problems, $OR = 1.81$, 95% CI [1.51, 2.16], $z = 6.47$, $p < .001$. Without it, students rated 45.7% of peer-authored problems and 41.6% of standard problems, $OR = 1.26$, 95% CI [1.07, 1.49], $z = 2.72$, $p = .007$. The difference was 1.43 times larger with the notification, 95% CI [1.13,

1.83], $z = 2.94$, $p = .003$ (Table 4).

Students also rated harder problems more often. Each increase of 0.10 in a problem's error rate raised the odds of a rating by a factor of 1.11, 95% CI [1.04, 1.18], $z = 3.15$, $p = .002$.

5.3.2 Rating across the lesson

Students rated problems more often early in the lesson and tapered off as the lesson went on, but a notification of peer-authorship reduced the slope. Among peer-authored problems, the odds that a student rated a problem fell by a factor of 0.89 with each successive problem in the Co-Authoring condition, 95% CI [0.87, 0.92], $z = -6.98$, $p < .001$, a decline from 55.7% on the first problem to 33.9% by the tenth. Under the notification the odds fell by a factor of 0.96, 95% CI [0.93, 0.98], $z = -3.08$, $p = .002$, a decline from 52.5% to 43.9%. The two slopes differed, $OR = 1.07$, 95% CI [1.03, 1.12], $z = 3.08$, $p = .002$, while the slopes for standard problems did not significantly differ across the two conditions, $p = .655$ (Figure 2; Table 4).

5.3.3 Liking

Students gave a thumbs up to 59.3% of the problems they rated. They did not significantly favor peer-authored problems over standard problems, $p = .238$. Students in the two conditions also did not significantly differ in how positively they rated problems, $p = .850$. Liking did not significantly vary with problem difficulty, $p = .387$. Ratings did grow more positive as a lesson went on, $OR = 1.03$ per successive problem, 95% CI [1.00, 1.05], $z = 2.35$, $p = .019$.

Predictor	<i>B</i>	<i>SE</i>	<i>z</i>	<i>p</i>
Intercept	1.77	0.12	15.13	< .001
Problem-Type (Peer-Authored)	0.12	0.03	4.45	< .001
Co-Authoring Condition	−0.07	0.03	−2.00	.046
Co-Authoring + Notification Condition	−0.04	0.03	−1.24	.213

Table 3: First-Attempt Correctness

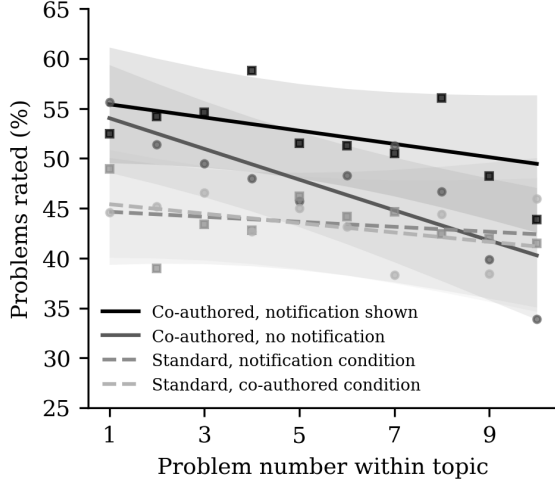


Figure 2: Percentage of problems rated by problem number within a lesson, for peer-authored and standard problems in each co-authoring condition, with observed points and logistic fits.

A null logistic model attributed 53.5% of the variance in liking to students (intercept variance = 3.79) and almost none to scenarios (0.009; 0.1%; Table 4). Whether a problem was liked depended overwhelmingly on who was rating, not on which scenario appeared.

5.4 Liking predicts success

A thumbs up submitted before solving predicted higher first-attempt success even when controlling for problem difficulty, but only in topic areas where the text of the problem varied more across problems (Figure 3; Table 5). In the ratio and proportion topics, students answered a step correctly on the first attempt at 1.20 times the odds when they had rated the problem positively, 95% CI [1.03, 1.40], $z = 2.36$, $p = .018$, a difference of 4.7 points in observed accuracy, 84.4% against 79.7%. In picture algebra, where the text is less varied across problems, liking did not predict first-attempt accuracy, $p = .055$, and observed accuracy was 89.3% and 89.4%. The interaction of liking with topic type was reliable, $OR = 1.33$, 95% CI [1.11, 1.60], $z = 3.05$, $p = .002$, in a model that also contained problem difficulty, problem type, problem position,

Predictor	<i>B</i>	<i>SE</i>	<i>z</i>	<i>p</i>
<i>Rating a problem</i>				
Intercept	−1.08	0.48	−2.26	.024
Problem Type	0.23	0.09	2.72	.007
Notification	−0.31	0.17	−1.77	.077
Problem Number	−0.07	0.01	−8.81	< .001
Problem Difficulty	1.04	0.33	3.15	.002
Problem Type × Notification	0.36	0.12	2.94	.003
<i>Rating a peer-authored problem across the lesson</i>				
Intercept	−0.58	0.48	−1.19	.233
Problem Number	−0.11	0.02	−6.98	< .001
Notification	−0.22	0.23	−0.97	.332
Problem Difficulty	0.69	0.38	1.83	.067
Problem Number × Notification	0.07	0.02	3.08	.002
<i>A rating being positive</i>				
Intercept	0.87	0.33	2.61	.009
Problem Number	0.03	0.01	2.35	.019
Problem Type	−0.10	0.08	−1.18	.238
Notification	0.03	0.18	0.19	.850
Problem Difficulty	−0.39	0.46	−0.87	.387
<i>Variance components for liking</i>				
	Variance		ICC	
Student	3.79		.54	
Scenario	0.01		.001	
Residual (logistic)	3.29		.46	

Table 4: Rating and Liking Across a Lesson

Note. Reference = Co-Authoring without notification. The first model uses all problem encounters ($n = 10,021$), the second peer-authored encounters ($n = 5,325$), and the third the ratings students submitted ($n = 4,540$). The variance components come from a null model over the same ratings with random intercepts for student and scenario where the residual is fixed at $\pi^2/3 = 3.29$.

and condition. Structural differences between the topic types may explain this interaction. Ratio and proportion problems are longer and share few content words with one another, while picture algebra problems restate one relation in a fixed frame and share extensive vocabulary, with a text variety of 0.62 against 0.86 for ratio and proportion (Table 1). The peer-authored problems introduce new variance in vocabulary to the picture algebra problems which may have reduced any effect of liking on performance.

6 Discussion

Context personalization has been studied as something done for a student, with the system selecting or generating a setting to match that student's interests (Bernacki and Walkington, 2018; Clinton and

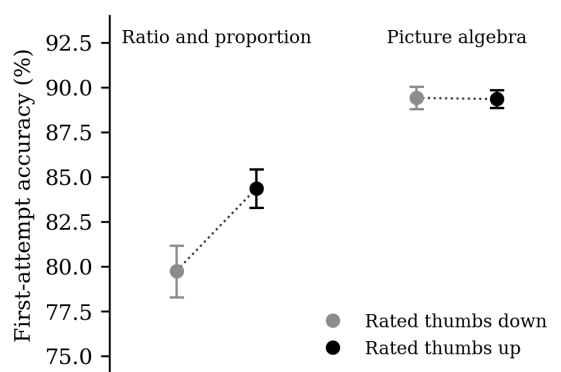


Figure 3: First-attempt accuracy on steps of problems rated before solving, by rating and topic type, with 95% Wilson intervals.

Predictor	<i>B</i>	<i>SE</i>	<i>z</i>	<i>p</i>
Intercept	2.33	0.21	11.28	< .001
Thumbs Up	−0.10	0.05	−1.92	.055
Ratio and Proportion	−0.80	0.26	−3.07	.002
Peer Auth Problem	0.01	0.04	0.30	.762
Problem Number	0.09	0.01	15.83	< .001
Problem Difficulty	−2.13	0.20	−10.65	< .001
Notify Condition	0.11	0.06	1.91	.056
Thumbs Up × Ratio	0.29	0.09	3.05	.002
Thumbs Up, ratio topics	0.18	0.08	2.36	.018

Table 5: First-Attempt Correctness by Liking and Topic Type

Note. Fit at the step level on 31,787 first attempts where the rating occurred before solving. Picture algebra is the reference topic type, so the Thumbs Up row is the liking slope in picture algebra; the final row is the liking slope in the ratio and proportion topics, a linear combination of that row and the interaction.

Walkington, 2019; Walkington, 2013; Yadav et al., 2023). In this study, we examined how students performed on and responded to problems co-authored by other students with generative AI. We previously showed that the co-authoring was beneficial to students' sense of belonging (Norberg et al., 2025b). Here we showed that using the student-AI generated problems did not have detectable harmful effects on performance. Students who did not write the problems actually had higher accuracy on first attempts at a step than on standard problems in the same section and mastery rates across conditions were similar.

Telling students that a peer wrote a problem affected how often they rated the problem, suggesting a greater interest in engaging with problems written by peers. Nevertheless, they were not more inclined to like the problem their peer wrote, and there was a time cost to including the attribution

in the problem. Students in the Co-Authored + Notification condition took longer to reach mastery perhaps because of the time it took to rate the problems, read the notifications, or both.

Adding the liking button revealed that student performance was higher when they liked a story even controlling for the problem difficulty but that this effect only held under specific contexts. When the problems in a lesson each had very similar texts, liking the problems did not have a notable effect on performance. As students become accustomed to reading the same text frame, their reading becomes more efficient (Myslín and Levy, 2016; Pickering and Ferreira, 2008). We speculate that this reduces any effect of liking the context. Supporting this, there was a positive effect of liking in the ratios and proportions lessons where the texts were more varied. Liking a problem context may reflect familiarity, which can also improve reading comprehension (Walkington et al., 2015; Walkington, 2013). This association may matter more when problems within a lesson vary more in structure and wording.

6.1 Limitations

Although this study examined a large sample drawn from MATHia's broadly representative student population, we could not determine whether the results differed across demographic groups because demographic data were unavailable. Future studies should examine whether student-AI co-authored problems affect different groups of students differently.

A second limitation concerns the outcomes available during standard educational practice. Previous research found that co-authoring improved the authors' sense of belonging in mathematics (Norberg et al., 2025b). The present study tested whether students might receive a similar affective benefit from knowing that a peer helped write the problems they solved. Ratings and accuracy provide only indirect evidence of such a benefit. Pre- and post-study attitude measures would be needed to test this possibility directly.

Finally, textual variation was not manipulated independently of topic type. The interaction therefore shows that the association between liking and performance differed between ratio and proportion and picture algebra, but it does not explain why. The findings identify textual variability and reading demand as possible explanations that should be tested directly in problems varying independently on these features.

7 Conclusions

Student–AI co-authored math problems generalized well in a large-scale deployment. Performance on the co-authored problems matched and even slightly outperformed standard problems. When lessons required more attention to individual texts, students also showed stronger performance on problems that they liked even when controlling for the difficulty of the problem. However, preferences for specific contexts were strongly idiosyncratic; students showed little agreement regarding which problems they liked best, reinforcing the case for individualized rather than broadly targeted personalization. Together, these findings support student–AI co-authoring as a scalable way to diversify problem contexts while preserving mathematical measurement.

Acknowledgments

This work was supported, in whole or in part by the Gates Foundation [INV-063849] and [INV-072481]. The conclusions and opinions expressed in this work are those of the author(s) alone and shall not be attributed to the Foundation. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 License has already been assigned to the Author Accepted Manuscript version that might arise from this submission.

References

- Jamal Abedi and Carol Lord. 2001. [The language factor in mathematics tests](#). *Applied measurement in education*, 14(3):219–234.
- Matthew L. Bernacki and Candace Walkington. 2018. [The role of situational interest in personalized learning](#). *Journal of Educational Psychology*, 110(6):864–881.
- Virginia Clinton and Candace Walkington. 2019. [Interest-enhancing approaches to mathematics curriculum design: Illustrations and personalization](#). *The Journal of Educational Research*, 112(4):495–511.
- Albert T. Corbett and John R. Anderson. 1995. [Knowledge tracing: Modeling the acquisition of procedural knowledge](#). *User Modeling and User-Adapted Interaction*, 4(4):253–278.
- Stephen E. Fancsali and Steven Ritter. 2014. [Context personalization, preferences, and performance in an intelligent tutoring system for middle school mathematics](#). In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, pages 73–77.
- Shannon F. Harp and Richard E. Mayer. 1998. [How seductive details do their damage: A theory of cognitive interest in science learning](#). *Journal of Educational Psychology*, 90(3):414–434.
- Kenneth R. Koedinger and Mitchell J. Nathan. 2004. [The real story behind story problems: Effects of representations on quantitative reasoning](#). *The Journal of the Learning Sciences*, 13(2):129–164.
- Ulrike I. E. Magner, Rolf Schwonke, Vincent Alevén, Octav Popescu, and Alexander Renkl. 2014. [Triggering situational interest by decorative illustrations both fosters and hinders learning in computer-based learning environments](#). *Learning and Instruction*, 29:141–152.
- Mark Myslín and Roger Levy. 2016. [Comprehension priming as rational expectation for repetition: Evidence from syntactic processing](#). *Cognition*, 147:29–56.
- Kole Norberg, Husni Almoubayyed, and Stephen Fancsali. 2025a. [Linguistic features predicting math word problem readability among less-skilled readers](#). In *Proceedings of the 18th International Conference on Educational Data Mining*, pages 457–464, Palermo, Italy. International Educational Data Mining Society.
- Kole Norberg, Husni Almoubayyed, Stephen E Fancsali, Logan De Ley, Kyle Weldon, April Murphy, and Steve Ritter. 2023. [Rewriting math word problems with large language models](#). *Grantee Submission*.
- Kole Norberg, April Murphy, Logan De Ley, Ethan Shafran Moltz, Husni Almoubayyed, and Steve Ritter. 2025b. [Using generative ai to foster student sense of belonging in mathematics](#). In *Artificial Intelligence in Education*, volume 15882 of *Lecture Notes in Computer Science*, pages 188–195. Springer.
- Kole Norberg, April Murphy, and Steve Ritter. 2026. [Ampt: A tool for personalizing math learning with generative ai](#). In *GenAI in Novel Educational Applications: Practices of Integrating GenAI in the K-12 Classroom, Volume 1*, pages 73–97. Springer.
- Kole A Norberg, Husni Almoubayyed, Logan De Ley, April Murphy, Kyle Weldon, and Steve Ritter. 2025c. [Rewriting content with gpt-4 to support emerging readers in adaptive mathematics software](#). *International Journal of Artificial Intelligence in Education*, 35(2):587–626.
- Martin J. Pickering and Victor S. Ferreira. 2008. [Structural priming: A critical review](#). *Psychological Bulletin*, 134(3):427–459.
- Steve Ritter, April Murphy, and Stephen Fancsali. 2020. [Managing group random assignment in upgrade](#). In *Proceedings of the First Workshop on Educational A/B Testing at Scale (at Learning@ Scale 2020)*.
- Steven Ritter, John R. Anderson, Kenneth R. Koedinger, and Albert Corbett. 2007. [Cognitive tutor: Applied research in mathematics education](#). *Psychonomic Bulletin & Review*, 14(2):249–255.

- Candace Walkington. 2026. [The implications of generative artificial intelligence for mathematics education](#). *School Science and Mathematics*, 126(4):356–365.
- Candace Walkington, Virginia Clinton, Steven N. Ritter, and Mitchell J. Nathan. 2015. [How readability and topic incidence relate to performance on mathematics story problems in computer-based curricula](#). *Journal of Educational Psychology*, 107(4):1051–1074.
- Candace Walkington, Virginia Clinton, and Pooja Shivraj. 2018. [How readability factors are differentially associated with performance for students of different backgrounds when solving mathematics word problems](#). *American Educational Research Journal*, 55(2):362–414.
- Candace Walkington and Carole A. Hayata. 2017. [Designing learning personalized to students’ interests: Balancing rich experiences with mathematical goals](#). *ZDM*, 49(4):519–530.
- Candace A. Walkington. 2013. [Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes](#). *Journal of Educational Psychology*, 105(4):932–945.
- Gautam Yadav, Ying-Jui Tseng, and Xiaolin Ni. 2023. [Contextualizing problems to student interests at scale in intelligent tutoring system using large language models](#). *arXiv preprint arXiv:2306.00190*.