Rewriting Math Word Problems to Improve Learning Outcomes for Emerging Readers: A Randomized Field Trial in Carnegie Learning's MATHia*

Husni Almoubayyed[1], Rae Bastoni[2], Susan R. Berman[1], Sarah Galasso[1], Megan Jensen[1], Leila Lester[2], April Murphy[1], Mark Swartz[1], Kyle Weldon[1], Stephen E. Fancsali[1], Jess Gropen[2] and Steve Ritter[1]

[1] Carnegie Learning, Inc., Pittsburgh PA 15219, USA
[2] CAST, Lynnfield MA 01940, USA

Abstract. We present a recent randomized field trial delivered in Carnegie Learning's MATHia's intelligent tutoring system to a sample of 12,374 learners intended to test whether rewriting content in a selection of so-called "word problems" improves student mathematics performance within this content, especially among students who are emerging as English language readers. In addition to describing facets of word problems targeted for rewriting and the design of the experiment, we present an artificial intelligence-driven approach to evaluating the effectiveness of the rewrite intervention for a sub-population of learners of interest. We hypothesize that the intervention may be especially effective to emerging readers using MATHia. Data about students' reading ability is generally neither collected nor available to MATHia's developers. Instead, we rely on a recently developed neural network predictive model that infers whether a student is an emerging reader. We present the results of the intervention on a variety of performance metrics in MATHia and compare performance of the intervention group to the entire user base of MATHia, as well as by comparing likely emerging readers to those who are not inferred to be emerging readers. We conclude with areas for future work using these kinds of more comprehensive models of learners.

Keywords: machine learning, A/B testing, intelligent tutoring systems, reading ability, middle school mathematics

1 Introduction

A growing body of research has found connections between math learning outcomes and reading comprehension (see, e.g., [5], [6], [3], [10]). Recent work seeks

* Appearing in The 24th International Conference on Artificial Intelligence in Education (AIED 2023): Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky

to develop more comprehensive models of learners as they use adaptive learning software, including efforts to build models that incorporate reading ability while students use adaptive software for mathematics instruction (e.g., [7], [1]). Such work recognizes that students draw on skills outside of the target domain as they receive instruction and practice skills (e.g., drawing reading skills while doing math, especially in word problems). With more comprehensive learner models, adaptive learning software like intelligent tutoring systems (ITSs) could go beyond existing approaches that adapt based on skills within the target domain (i.e., math) to adapt to factors like students' reading ability. More generally, these observations raise questions about the nature of supports that might be used in an ITS for mathematics to adapt to student reading ability.

2 MATHia and UpGrade

MATHia (formerly Cognitive Tutor [8]) is an ITS for middle- and high-school mathematics. Currently used by over half a million students across the United States, MATHia content is presented to students organized by topic in "workspaces," which either take the form of "Concept Builders" or "Mastery Workspaces." Concept Builders present instructional content and interactive, exploratory tools along with a fixed sequence of multi-step problems that introduce students to new materials (e.g., vocabulary terms) and develop students' conceptual understanding of target material. In Mastery Workspaces, students work through complex, multi-step problems as they work toward mastery of each of a set of knowledge components (KCs; [4]) or skills associated with the workspace. Mastery is determined using Bayesian Knowledge Tracing [2]), and problems are selected that emphasize KCs that a student has yet to master. We refer to students who successfully achieve mastery of all the KCs in a workspace as "graduated." Conversely, we refer to students as "promoted" when they are moved on to the next workspace in a curriculum sequence if they encounter a pre-set maximum number of problems but still fail to achieve mastery of all KCs.

UpGrade is a free and open source platform for conducting randomized field trials (sometimes referred to as "A/B tests") in educational software applications [9]. UpGrade enables large-scale randomized field trials in real classroom settings by integrating with EdTech software applications like MATHia and allowing researchers to manage experimental design and logistics through a simple, web-based user interface. UpGrade then communicates with the EdTech application to randomly assign appropriate experimental conditions to learners using the application. By integrating UpGrade with MATHia, we are able to deliver instructional interventions across Carnegie Learning's sizeable customer base, within multiple math topic areas (or workspaces).

3 Predicting Reading Ability

The student's first interaction with MATHia is a Concept Builder known as the Pre-Launch Protocol. This introductory activity prepares the student for work-

ing with MATHia, and is not particularly related to mathematics. Almoubayyed et al, 2023 [1] developed a neural-network based model to predict the end-of-year English Language Arts (ELA) scores of students based on student performance in the Pre-Launch Protocol, by training it on a sample of end-of-year ELA scores. Due to the difficulty in obtaining ELA scores for a large-scale study (typically supplied by individual school districts through data sharing agreements), we use this predictive model to predict which students are emerging readers. We use the predictions of the model to assess the impact of rewriting word problems on students identified as emerging readers.

We use a version of the predictive model that is trained on predicting the probability that a student would pass the end-of-year ELA exam (the model has an AUC of around 0.8, please see [1] for model details). In this study, we define emerging readers as those whose probability of passing their ELA test is in the bottom quartile. We do not retrain the model as a binary classifier of students in or outside of the bottom quartile of reading ability. The reason for this choice is that the training set in [1] comes from a single school district, and it can be unclear what thresholds must be used to represent the bottom quartile of the student population that uses MATHia nationally – especially for students in different states with different exams. We define the bottom quartile for each workspace independently and for the sample in each condition independently.

4 Rewriting Word Problems

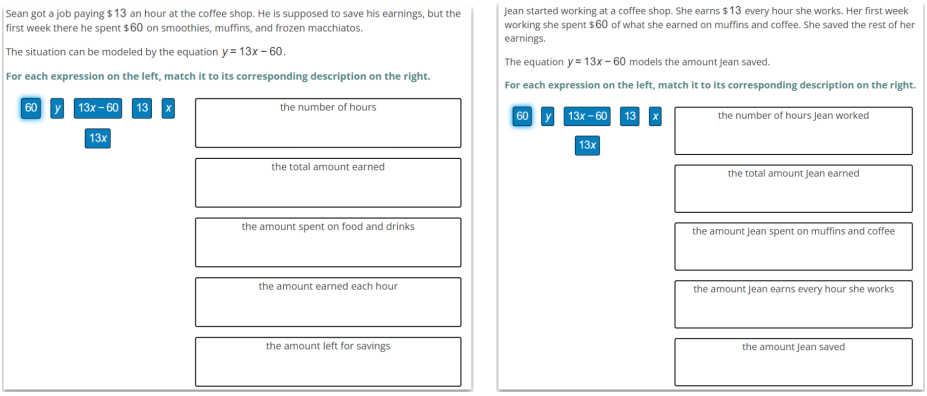

Fig. 1: An example of an unmodified word problem (left) in MATHia and the rewritten version (right).

Two of MATHia's Mastery Workspaces, including 200 problems based on 30 scenarios were chosen to be rewritten: "Analyzing Models of Two Step Linear Relationships" in Middle School Course 2 (Carnegie Learning's Grade 7 math

sequence), which we refer to as 'Integers'; and "Analyzing Models of Linear Relationships" in Middle School Course 3 (Carnegie Learning's Grade 8 math sequence), which we refer to as 'Rationals.' These two workspaces were chosen due to the high correlation of student performance in these workspaces with students' ELA end-of-year state test scores, compared to the correlation with students' end-of-year state test math scores. Correlational relationships were based on a historical dataset for which ELA test scores were made available to researchers. Test score data was provided the district to Carnegie Learning according to data sharing agreements between Carnegie Learning and the district that allows for the use of these data for research purposes.

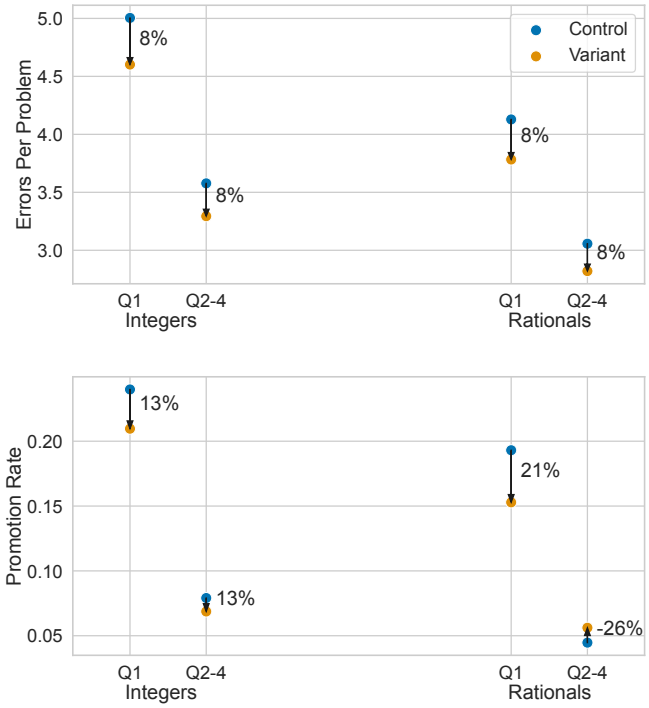

Fig. 2: The results of the metrics of the average Errors Per Problem and Promotion Rate. Q1 indicates emerging readers (predicted to be in the bottom quartile of reading scores), with Q2-4 indicating the rest of the students. Q1 students show improvement in all cases, with the percent change annotated. While the percentage change is 8% in all cases for Errors Per Problem, the absolute changes are larger for Q1 students.

The content was rewritten, relying on a set of principles developed by an internal instructional design team. This guideline had two specific goals: (a) us-

ing only recognizable content (e.g., everyday objects or phenomena that appear in word problems) that is easily understandable and relevant to students across the country, and (b) using clear and precise language, that supports students in visualizing and connecting meaning across sentences. Explicitly, the rewrites aimed to preserve the underlying mathematical content difficulty. Rewriting was carried out by staff from Carnegie Learning and CAST, with every rewrite reviewed for quality assurance by a different person than the writer. An example of a problem before and after the rewrite process is provided in Fig 1.

5 Evaluation Across Reading Abilities

The study was randomized with equal probability of a student receiving the control or the variant condition as they encounter one of the target workspaces. It was deployed to the entire user base for a time period of 7 weeks, with a substantial number of students (14,767) enrolling in the experiment. Out of the students that enrolled in the experiment, we only use data for students that completed their assigned workspace and the Pre-Launch Protocol – a total of 12,374 students. In the Integers workspace, there were 4,113 in the variant and 3,894 in the control condition. In the Rationals workspace, there were 2,230 students in the variant and 2,137 in the control sample.

We use a variety of metrics to compare the results between the control and variant sample for each of the two workspaces. For all of these metrics, a smaller value indicates a better outcome. These metrics are:

- *Promotion Rate*: The number of students who were promoted in a sample (failed to master any skill in a workspace) over the number of all students in that sample.
- *Time Spent*: The total time spent by each student in a sample is computed, then the median is taken over all the students in that sample. We use the median here due to outliers with very large time values (e.g., when a student has MATHia open but is not paying attention). We report this metric independently for all students, and also for students who were 'graduated', or mastered all skills in the relevant workspace.
- *Total Errors*: The total number of errors that a student makes in the relevant workspace. The average is taken over all students in a sample.
- *Total Problems Completed*: The average of the total problems completed over all students in a sample. We report this metric independently for all students and for graduated students. For each workspace in this study, students that are promoted complete 25 problems.

The results of the evaluation metrics are presented in Figures 2 and 3. In all Quartile 1 (Q1, predicted to be emerging readers) and almost all Q2-Q4 (predicted not to be emerging readers) cases, the variant has better metric outcomes than the control. The improvement for Q1 is always better (sometimes dramatically) or at least equal to the improvement for Q2-Q4. In particular, the results show that for the population of students that were predicted by the

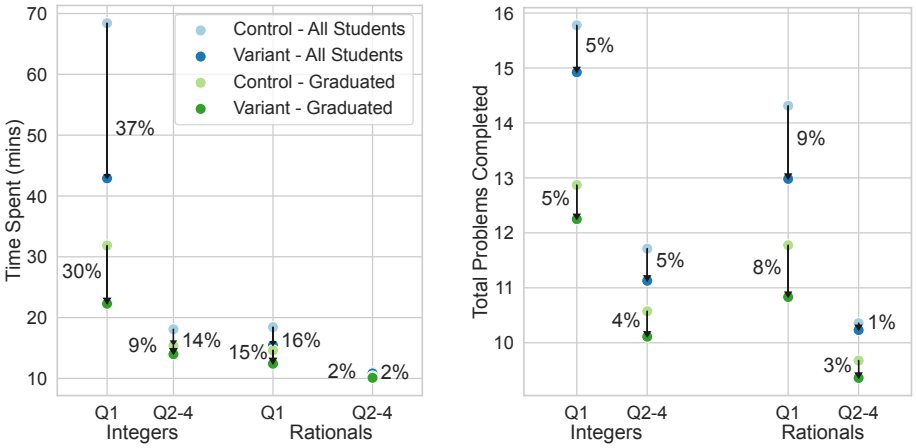

Fig. 3: The results of the Time Spent and Total Problems Completed metrics for all students (with percent change annotated to the right of each arrow) and graduated students (with percent change annotated to the left) independently. In all cases, there is an improvement for students receiving the rewritten problems, and this improvement is larger for Q1 students.

machine learning model from [1] to be emerging readers, rewriting word problems following the guideline described in Section 4 results in making 8% fewer errors. Additionally 13-21% more emerging readers that received the rewritten problems were able to master all the skills in the targeted workspaces. Out of the Q1 students who master all skills in the targeted workspaces, they do so in 15-30% less time, completing 5-8% fewer problems, which means they can spend more time on other material. We hope that these outcomes will result in higher end-of-year exam scores, but we leave that to future studies.

6 Conclusions

Using comprehensive learner models has the potential to help adaptively target and deliver supports to students who need them. In this study, we applied a machine learning model to classify emerging readers in MATHia, a math ITS, based on an introductory activity. We found that rewriting word problems with simple guidelines for added clarity and relevance led to large improvements in several performance metrics for students predicted to be emerging readers. More emerging readers were able to master all skills in the workspaces, do so in fewer problems, spend much less time in the workspaces, and make fewer errors. These improvements were not as high, and occasionally non-existent or of potentially negative impact for other students, highlighting the importance of exploring the adaptive delivery of these kinds of supports (i.e., to likely emerging readers).

In future studies, we will explore whether these and similar improvements and adaptive reading supports also positively affect end-of-year exam outcomes for these students. However, even at the level of specific workspaces, we found that these improvements could save emerging readers a significant amount of valuable time.

Acknowledgements

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through grant R324A210289 to Center for Applied Special Technology (CAST). The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

1. Almoubayyed, H., Fancsali, S.E., Ritter, S.: Instruction-embedded assessment for reading ability in adaptive mathematics software. In: Proceedings of the 13th International Conference on Learning Analytics and Knowledge. LAK '23, Association for Computing Machinery, New York, NY, USA (2023)
2. Anderson, J.R., Corbett, A.T.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction **4**, 253–278 (1995)
3. Fuentes, P.: Reading comprehension in mathematics. The Clearing House **72**(2), 81–88 (1998), http://www.jstor.org/stable/30189563
4. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive Science **36**(5), 757–798. https://doi.org/10.1111/j.1551-6709.2012.01245.x
5. Koedinger, K.R., Nathan, M.J.: The real story behind story problems: Effects of representations on quantitative reasoning. Journal of the Learning Sciences **13**(2), 129–164 (2004). https://doi.org/10.1207/s15327809jls1302_1
6. Krawitz, J., Chang, Y.P., Yang, K.L., Schukajlow, S.: The role of reading comprehension in mathematical modelling: improving the construction of a real-world model and interest in germany and taiwan. Educational Studies in Mathematics **109**, 337–359 (2022)
7. Richey, J.E., Lobczowski, N.G., Carvalho, P.F., Koedinger, K.: Comprehensive views of math learners: A case for modeling and supporting non-math factors in adaptive math software. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) Artificial Intelligence in Education. pp. 460–471. Springer International Publishing, Cham (2020)
8. Ritter, S., Anderson, J.R., Koedinger, K., Corbett, A.T.: Cognitive tutor: Applied research in mathematics education. Psychonomic Bulletin & Review **14**, 249–255 (2007)
9. Ritter, S., Murphy, A., Fancsali, S.E., Fitkariwala, V., Lomas, J.D.: Upgrade: An open source tool to support a/b testing in educational software. In: Proceedings of the First Workshop on Educational A/B Testing at Scale. EdTech Books (2020)

10. Vilenius-Tuohimaa, P.M., Aunola, K., Nurmi, J.: The association between mathematical word problems and reading comprehension. Educational Psychology **28**(4), 409–426 (2008). https://doi.org/10.1080/01443410701708228