Self-Explanation of Worked Examples Integrated in an Intelligent Tutoring System Enhances Problem Solving and Efficiency in Algebra

Nicholas A. Vest (navest@wisc.edu)
University of Wisconsin-Madison, 1202 West Johnson Street, Madison, WI 53706, USA

Elena M. Silla (esilla@udel.edu)
University of Delaware, 16 West Main Street, Newark, DE 19716, USA

Anna N. Bartel (anbartel@wisc.edu)
University of Wisconsin-Madison, 1202 West Johnson Street, Madison, WI 53706, USA

Tomohiro Nagashima (tnagashi@cs.cmu.edu)
Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA

Vincent Aleven (aleven@cs.cmu.edu)
Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA

Martha W. Alibali (mwalibali@wisc.edu)
University of Wisconsin-Madison, 1202 West Johnson Street, Madison, WI 53706, USA

Abstract

One pedagogical technique that promotes conceptual understanding in mathematics learners is self-explanation integrated with worked examples (e.g., Rittle-Johnson et al., 2017). In this work, we implemented self-explanations with worked examples (correct and erroneous) in a software-based Intelligent Tutoring System (ITS) for learning algebra. We developed an approach to eliciting self-explanations in which the ITS guided students to select explanations that were conceptually rich in nature. Students who used the ITS with self-explanations scored higher on a posttest that included items tapping both conceptual and procedural knowledge than did students who used a version of the ITS that included only traditional problem-solving practice. This study replicates previous findings that self-explanation and worked examples in an ITS can foster algebra learning (Booth et al., 2013). Further, this study extends prior work to show that guiding students towards conceptual explanations is beneficial.

Keywords: learning; self-explanation; worked examples; Intelligent Tutoring System; middle-school algebra

Introduction

How can instruction foster learners' acquisition of deep understanding in mathematics? And how can technology-based learning environments, such as Intelligent Tutoring Systems, support this learning? Deep understanding of mathematics involves several distinct types of knowledge, including knowledge of fundamental concepts, knowledge of how to solve problems, and understanding of the connections between them (Crooks & Alibali, 2014; Hiebert & LeFevre, 1986).

Intelligent Tutoring Systems (ITSs) are computer-based programs that administer lessons and learning activities to students. ITSs support learning across various domains (for a meta-analysis, see Ma et al., 2015). Many studies have provided evidence that practice in an ITS can support procedural understanding of mathematics (Ma et al., 2015). However, current ITSs are less successful at promoting gains in conceptual understanding (e.g., Long & Aleven, 2017; Pane et al., 2014; but see Aleven & Koedinger, 2002). In this research, we extended and tested an ITS for equation solving in algebra, with the broad goal of creating an ITS that would foster gains in conceptual understanding.

One pedagogical technique that has been shown to support gains in conceptual knowledge in a range of domains is *self-explanation*. Self-explanation involves generating explanations of to-be-learned material for oneself, in an effort to more deeply process that material (Chi et al., 1994). Many studies have documented the value of self-explanation as means to help students learn and retain new material (for a review, see Rittle-Johnson et al., 2017). In a foundational study, Chi (1994) prompted some students to provide self-explanations as they read a brief text about the circulatory system. Students who produced self-explanations retained more information and generated more accurate inferences based on the material than students who did not produce self-explanations. Other studies have documented the value of self-explanation in mathematics (Barbieri & Booth, 2020; Barbieri et al., 2019; Hilbert et al., 2008; Rittle-Johnson, 2017), including in ITSs (Aleven & Koedinger, 2002).

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

In the context of mathematical problem solving, some research has suggested that self-explanation can *potentiate* other sorts of learning activities, enhancing their benefits for conceptual knowledge. For example, instruction that involves both strategy comparison and prompts to self-explain yields greater benefits for learning than instruction that involves comparison on its own (Sidney & Alibali, 2015). Similarly, instruction that involves self-explanations of worked examples or problems steps yields greater benefits for learning than similar instruction without self-explanation prompts (Aleven & Koedinger, 2002; Barbieri et al., 2019).

In general, self-explanation is thought to be effective because it engages constructive processes, such as identifying inconsistencies, filling in knowledge gaps, integrating different knowledge elements, and monitoring understanding (e.g., Roy & Chi, 2005). However, the quality of self-explanations also matters. High-quality self-explanations–ones that demonstrate inference generation or knowledge integration–are associated with greater benefits for learning than lower-quality self-explanations, such as simple restatements or paraphrases (Wylie & Chi, 2014).

Given the established benefits of high-quality self-explanation for building conceptual understanding, we sought to integrate activities that would elicit high-quality self-explanations into an Intelligent Tutoring System for early algebra. Building on previous research with similar aims (e.g., Booth et al., 2013), we extended an ITS so it incorporates worked examples produced by hypothetical students, and prompts learners to explain the bases of (correct or erroneous) problem-solving steps taken by these hypothetical students. Rather than have students "build" self-explanations from pieces (as in Booth et al., 2013)–a process that some students find challenging and laborious–we drew on previous studies that showed that *selecting* possible explanations from a menu is a practical, time-effective, and straightforward way to elicit explanations from students (Rittle-Johnson et al., 2017), especially within ITSs.

Although it is not known whether the cognitive processes involved in *selecting* explanations are the same as those involved in *generating* explanations, past research has documented benefits of selecting explanations for student learning (e.g., Rau et al., 2015; Rittle-Johnson et al., 2017). In designing the self-explanation activities for the ITS, we based the set of explanation choices that we offered on self-explanations that were generated by middle-school students in a one-on-one tutorial interaction in a pilot study (Bartel et al., 2020). As might be expected, student-generated self-explanations varied widely in their *quality*, and many student-generated self-explanations did not incorporate relevant concepts. In our ITS, we included choice options that aligned with students' typical explanations–including non-conceptual explanations–but when learners selected non-conceptual explanations, the ITS prompted them to select a second explanation that invoked key concepts.

In brief, in this work we test the effectiveness of an ITS that incorporates an approach to self-explanation of worked examples that involves (1) students selecting possible explanations, and (2) students receiving encouragement to consider conceptually rich explanations, if they initially select explanations that are not conceptually rich. We compare this tutor to a baseline tutor that does not include self-explanation activities, and we evaluate participants' gains in both procedural skill and conceptual understanding. We hypothesized that students who studied worked examples and who provided self-explanations in addition to solving problem-solving items would perform better than students who received only problem-solving items on measures of procedural and conceptual knowledge, and that these students would also show enhanced performance on problem-solving items in the tutor (i.e., less time spent per step, fewer incorrect steps, fewer hint requests).

Method

Participants

Participants were 175 middle-school students recruited via an online database and via word of mouth. Six participants were excluded due to technical issues (e.g., computer malfunctions, n = 5, and incomplete session, n = 1). Two additional participants were excluded for having tutor interactions (e.g., length of time per steps) that were three standard deviations above the mean. Thus, the final analytic sample consisted of 167 students (M age = 12.81 years, SD age = 0.76 years; 57 6th grade, 73 7th grade, 36 8th grade, one declined to respond). Of the 167 participants in the final sample, 128 were White, 23 were biracial, six were Black/African American, six were Asian, one was Native Hawaiian/Pacific Islander, and three declined to report race and ethnicity. Ninety-nine of the students identified as male, 64 as female, three as non-binary, and one declined to report their gender. Ninety-three students reported they were in advanced math, 73 reported they were not in advanced math, and one declined to report. Participants were compensated 15 USD in the form of a gift card, cash, or check after completing the study.

Design and Procedure

Data were collected as part of a study assessing the effectiveness of a range of interventions on students' conceptual and procedural knowledge of algebra. Participants completed the study in a virtual setting, and the sessions were conducted by trained experimenters.

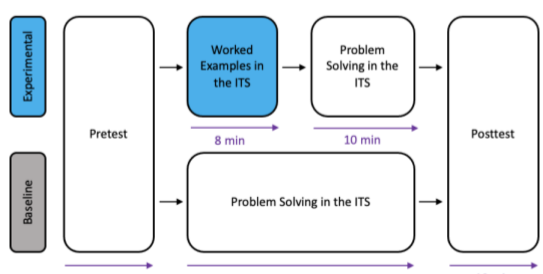

Figure 1: Study procedure

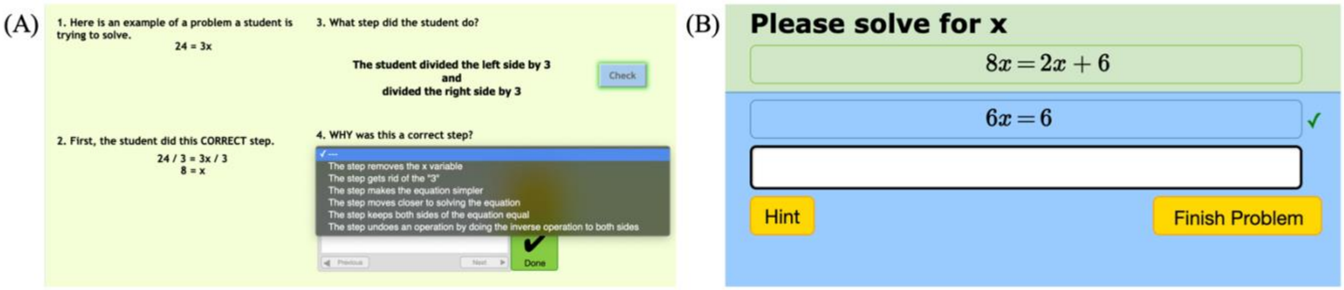

Figure 2: Participants in the experimental condition received both (A) worked examples and (B) problem solving. Participants in the baseline condition only received (B). Both were presented in the Intelligent Tutoring System.

Each session lasted for about one hour. Participants were randomly assigned to one of five conditions. In four of these conditions, participants completed both worked examples with self-explanation prompts and practice problems within the ITS; these conditions varied in whether participants also saw visual representations (yes/no) or engaged in warm-up activities (yes/no). Preliminary analyses showed that the visual representation and warm-up manipulations had little impact on student performance. Thus, for purposes of this paper, we collapsed these conditions into a unified experimental condition, in which all students used a version of the ITS that included self-explanations of worked examples (n = 134). We compared students in the unified experimental condition against students in a baseline condition who used an ITS that included problem-solving activities but that did not include self-explanations or worked examples (n = 33).

This study was preregistered on the Open Science Framework. The preregistration includes many analyses that fall outside the scope of the current report. Here, we focus specifically on comparing students who used a version of the ITS that included self-explanations of worked examples and students who used a version of the ITS that did not include these activities (Hypothesis 1 in the preregistration; see Figure 1 for a schematic of the study procedure).

Measures of Learning: Pretest and Posttest Participants completed an online pretest and isomorphic posttest that assessed algebra knowledge. Specifically, these tests assessed students' procedural knowledge (3 items) and conceptual knowledge (8 items) of basic algebra. Items assessing procedural knowledge measured students' abilities to solve linear equations, whereas items assessing conceptual knowledge measured students' understanding of underlying concepts in algebra, such as understanding inverse operations and doing the same thing to both sides of the equation when solving problems The posttest contained two additional transfer items. Items were adapted from prior literature (Fyfe et al., 2018; Nagashima et al., 2020; Rittle-Johnson et al., 2011). Some items had multiple parts and were thus scored accordingly. Participants were given 11 minutes to work on the pretest, and 13 minutes to work on the posttest.

Measures of Performance in the Intelligent Tutoring System Participants then solved problems in an Intelligent Tutoring System (ITS). The ITS consisted of two sections: worked examples (unified experimental condition only) and problem solving (all conditions; see Figure 1). Before each section, students watched a short instructional video. In both the worked examples and problem-solving activities, students received immediate feedback on their responses. They also could request scaffolded hints from the tutor at any time.

Participants in the unified experimental condition were presented with correct and incorrect worked examples (with a maximum of 8 problems). In each worked example, students were asked to use a drop-down menu to provide explanations about what operation a hypothetical student performed at a specific step of the equation, as well as to identify the conceptual basis of the step (see Figure 2A). Students could select from two conceptually-focused explanations (e.g., "the step keeps both sides of the equation equal" in Figure 2), two procedurally-focused explanations (e.g., "the step makes the equation simpler" in Figure 2), and two incorrect explanations (e.g., "the step removes the x variable" in Figure 2). Unique to this tutor was that students *had* to choose a conceptual response in order to advance. If students chose a response that was procedural or incorrect, they were asked to choose another response (even if the procedural explanation was, in fact, correct) via a prompt (e.g., "That's true but does not tell *why* the student did this step.")

Participants in both the unified experimental condition and the baseline condition were then presented with linear equations to solve (e.g., $3x + 8 = 11$; max. 11 problems) in increasing levels of difficulty. Participants typed their response for each problem-solving step of the equation into the ITS and received immediate feedback (Figure 2B). To keep time consistent across conditions, participants in the experimental condition had 10 minutes to complete the worked example activities and 10 minutes to complete these problem-solving items, while participants in the baseline condition had 20 minutes to complete the problem-solving item.

Table 1: Procedural (max: 3) and conceptual (max: 15) pretest and posttest scores and transfer (max: 2) posttest scores, with standard deviations in parentheses.

	Procedural		Conceptual		Transfer	
Condition	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
Baseline	2.12 (0.86)	2.33 (0.74)	6.91 (3.17)	9.06 (3.65)		0.64 (0.74)
Experimental	2.17 (0.97)	2.51 (0.74)	7.55 (3.89)	9.74 (3.63)		0.75 (0.84)

Demographic Questions Parents were sent a demographic questionnaire prior to the study session. Questions included age, grade, gender, math level in school, and self-reported socioeconomic status.

Results

Effects on Learning

In this section, we report the effect of the intervention on three measures of learning: procedural knowledge, conceptual knowledge, and transfer. Table 1 presents average pretest and posttest scores on each measure.

Procedural knowledge We first examined the effect of the intervention on procedural learning. Recall that we hypothesized that students who received worked examples and self-explanations in addition to problem-solving items (i.e., the unified experimental condition) would perform better than students who received only problem-solving items (i.e., the baseline condition). To analyze the data, we constructed a linear regression with procedural posttest score as the dependent variable and procedural pretest score, condition (coded: baseline = -.5; experimental = .5), grade level (coded: 6th grade = -1; 7th grade = 0; 8th grade = 1), and number of problem-solving items attempted in the ITS as independent variables. We included grade level and number of problem-solving items completed in the ITS as covariates to account for algebra experience and for the number of problem-solving items to which students were exposed. We chose to control for number of problem-solving items attempted to zero in on whether increases in performance were a result of students' self-explanations of the worked examples or because they were able to solve more problems and potentially learn more from the problem-solving condition.

Students in the experimental condition scored higher on the procedural posttest than students in the baseline condition, $\beta = 0.28$, $F(1, 161) = 5.32$, $p = 0.022$, indicating that students who generated self-explanations benefited more than those who simply solved a comparable number of problems. However, it should be noted that the effect of condition was non-significant if the covariate (number of problem-solving items attempted) was not included in the model, $\beta = 0.16$, p = 0.148. Students with higher pretest scores scored higher on the procedural posttest, $F(1, 161) = 48.8$, $p < 0.001$, as did students who attempted more problems in the ITS, $F(1, 161) = 7.26$, $p = 0.008$.

Conceptual Knowledge We next examined the effect of the intervention on conceptual knowledge. We constructed a linear regression with conceptual posttest score as the dependent variable and conceptual pretest score, condition, grade level, and number of problem-solving items attempted in the ITS as independent variables. Again, grade level and number of problem-solving items completed were included as covariates to account for algebra experience and exposure to problem-solving items in the ITS.

As hypothesized, students in the experimental condition scored higher on the conceptual knowledge posttest than students in the baseline condition, $\beta = 1.23$, $F(1, 161) = 7.18$, $p = 0.008$, indicating that students who generated self-explanations gained more conceptual knowledge than those who simply solved a comparable number of problems. Once again, the effect of condition was non-significant if the covariate (number of problem-solving items attempted) was not included in the model, $\beta = 0.35$, p = 0.459. Students with higher conceptual knowledge at the pretest scored higher on the conceptual knowledge posttest, $F(1, 161) = 90.62$, $p < 0.001$, as did students who attempted more problems in the ITS, $F(1, 161) = 29.23$, $p < 0.001$.

Procedural Transfer Because transfer items were not included in the pretest, we could not test for pre to posttest improvement. However, we tested the effect of condition on transfer. We constructed a linear regression with transfer score as the dependent variable and procedural pretest score, condition, grade level, and number of problem-solving items attempted in the ITS as independent variables. We also included the procedural pretest score in the model because it most closely resembled the transfer items. There was not a significant effect of condition; however, the pattern of findings aligned with those reported above ($\beta = 0.27$, $F(1, 161) = 3.41$, $p = 0.067$). There were significant main effects of the procedural pretest, $F(1, 161) = 12.41$, $p < 0.001$, and number of problem-solving items attempted in the ITS, $F(1, 161) = 12.64$, $p < 0.001$.

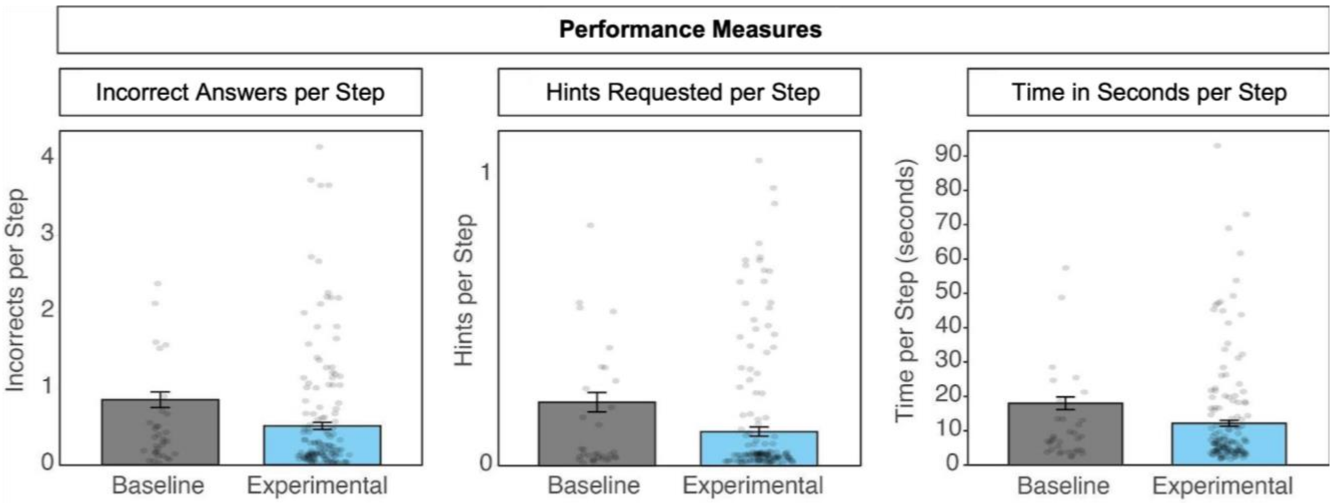

Figure 3: Each performance measure organized by condition. Error bars reflect standard error.

Effects on Performance in the ITS

To investigate students' performance in the ITS, we analyzed log data collected by the ITS during problem-solving items. Specifically, we explored the total number of problems attempted, the average number of incorrect attempts at each problem-solving step, the average number of hints requested at each step, and the amount of time spent on each step. These are standard measures investigated in the ITS literature (Long & Aleven, 2013). To examine whether learners in the baseline or experimental condition exhibited more efficient learning, we conducted four separate linear regressions with each of the performance measures in the ITS. In each model, condition, pretest score (procedural and conceptual separately), and grade level were included as independent variables. Additionally, we included the number of problems attempted in the ITS as an independent variable in three of the models (the ones in which it was not the dependent variable, because the number of problems solved was strongly/moderately correlated with each of the other dependent variables).

Number of Problems Attempted Students in the baseline condition solved more problems ($M = 8.97$, $SD = 2.67$) than students in the experimental condition ($M = 6.83$, $SD = 3.29$), $\beta = -2.42$, $F(1, 161) = 29.30$, $p < 0.001$, presumably because students in the baseline condition received more time to complete the problems than students in the experimental condition. Moreover, procedural pretest scores, $F(1, 161) = 39.50$, $p < 0.001$, and conceptual pretest scores, $F(1, 161) = 35.86$, $p < 0.001$, were both positively associated with number of problems attempted in the ITS.

Incorrect Attempts per Step Overall, students made about one incorrect attempt per two steps (M per step = 0.58, SD per step = 0.8). Controlling for pretest (procedural and conceptual separately), grade, and problems attempted in the ITS, students in the unified experimental condition exhibited fewer incorrect attempts per step than students in the baseline condition, $\beta = -0.34$, $F(1, 160) = 8.75$, $p = 0.004$ (see Figure 3). Students who attempted more problems also made fewer incorrect attempts per step, $\beta = -0.19$, $F(1, 160) = 103.31$, $p < 0.001$.

Number of Hints per Step Controlling for pretest (procedural and conceptual), grade, and number of problems attempted in the ITS, students in the unified experimental condition requested fewer hints per step than those in the baseline condition, $\beta = -0.10$, $F(1, 160) = 6.97$, $p = 0.009$ (see Figure 3), and number of problems attempted in the ITS was inversely related to the number of hints used, $\beta = -0.10$, $F(1, 160) = 52$, $p < 0.001$.

Average time spent per step On average, students spent 13.35 seconds on each step ($SD = 15.49$). Controlling for pretest (procedural and conceptual), grade, and number of problems attempted in the ITS, students in the baseline condition spent more time on each step, $\beta = -5.83$, $F(1, 160) = 7.84$, $p = 0.006$; see Figure 3. Number of problems attempted in the ITS was inversely related with the average time spent per step, $\beta = -5.83$, $F(1, 160) = 108.02$, $p < 0.001$.

Discussion

In the current study, we investigated whether a new self-explanation task integrated with worked examples, in which students were guided towards conceptual explanations, influenced performance and learning in middle-school students learning algebra with an Intelligent Tutoring System. Our findings indicate that, indeed, this form of intervention helped students gain conceptual and procedural knowledge of algebra over and above a problem-solving control. Moreover, students who studied worked examples and provided explanations solved problems faster, asked for fewer hints, and made fewer mistakes within the ITS than those in the baseline condition.

This study confirms earlier work that showed that worked examples with self-explanation can enhance learning within an ITS (e.g., Salden et al., 2010). Prior research suggests that self-explanation helps learners integrate to-be-learned information with prior knowledge, resulting in deeper understanding of the content (Bisra et al., 2018; Rittle-Johnson & Loehr, 2017).

This study also extends past work on self-explanations and worked examples (e.g., Booth et al., 2013) through the design of an ITS that guides students towards conceptually-focused explanations, and by demonstrating that this new format for selecting self-explanations is effective. Like previous efforts (Burr et al., 2020; Rittle-Johnson & Loehr, 2017), this intervention has menu-based explanations with correctness feedback, but unlike some previous efforts, students are asked for two-step explanations that ask for the operation *and* the conceptual justification. A special feature of the second explanation step is that, included among the menu options (for the conceptual justification) are correct procedural explanations. These explanations do not "count" as correct, but they do give the system an opportunity to give feedback stating that these explanations do not get at *why* the step is justified, so they may help the student learn how conceptual and procedural explanations differ. In this way, this version of the ITS may also help students recognize that–in general–they should think about, not only *what* to do, but *why* it is correct.

Moreover, this intervention led to improvements on posttest scores as well performance measures in this ITS. These findings suggest that self-explanations and worked examples affect both problem-solving accuracy and problem-solving efficiency. In future work, researchers should explore the relations between learning measures (e.g., pre- to posttest gains) and ITS performance measures.

Our findings do not specify the nature of the cognitive processes elicited by the self-explanation task or how these processes may have yielded the observed benefits of self-explanation. It is worth noting that our task involved selecting potential explanations from a menu, rather than generating explanations "from scratch", and our system also did not accept solely procedural explanations, but rather encouraged students to consider why steps were correct. It is possible that the mechanism of action for this type of self-explanation may differ from that for self-explanations that are spontaneously generated. To elucidate these mechanisms, future work that involves collecting talk-aloud protocols as students perform the self-explanation task would be valuable.

We acknowledge several limitations of this study. First, the baseline and experimental conditions had dramatically unequal numbers of students, due to the design of the larger experiment. We recognize this may violate assumptions about equal variance between samples, but we believe that our findings hold value as they correspond with the findings of previous research. Moreover, this experiment was conducted remotely during the COVID-19 pandemic. Given the unique context of the study, it may not be warranted to generalize conclusions to more typical settings. Lastly, the sample of students in this study was fairly homogeneous and made up primarily of White students, and it included many students who were above grade level in mathematics. Future studies are needed to investigate the impact of this intervention with students from a wider variety of backgrounds.

Conclusion

In brief, this study replicates past findings that self-explanations with worked examples can promote both procedural and conceptual understanding, and it introduces a new approach to eliciting such explanations within an ITS. Like a human tutor, our new version of the ITS encourages students to provide more conceptually rich explanations, if they initially provide less rich ones. In so doing, this new ITS supports students in focusing on the conceptual basis of their problem-solving steps, supporting both performance and learning.

Acknowledgements

This research was supported by NSF Award #1760922 and by the Institute of Education Sciences, U.S. Department of Education, through award #R305B15003 to the University of Wisconsin-Madison. The opinions expressed do not represent views of NSF or IES. We thank Tyler Tommasi, Brandan Risen, Holden Manhart, Mikayla Petersdorff, Zach Buehler, and Alan Zhou for their help designing the intervention, collecting, coding, and analyzing data.

References

Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive science*, *26*, 147-179.

Barbieri, C. A., & Booth, J. L. (2020). Mistakes on display: Incorrect examples refine equation solving and algebraic feature knowledge. *Applied Cognitive Psychology*, *34*, 862-878.

Barbieri, C. A., Miller-Cotto, D., & Booth, J. L. (2019). Lessening the load of misconceptions: Design-based principles for algebra learning. *Journal of the Learning Sciences*, *28*, 381-417.

Bartel, A.N., Silla, E.M., Vest, N.A., Nagashima, T., Tang, Y., Aleven, V., & Alibali, M.W. (2020, June). Do tape diagrams promote a focus on conceptual principles? Evidence from equation solving with an Intelligent Tutoring System. In T. T. Wong (Chair), *Principle knowledge in mathematics: its development, cognitive predictors, and potential interventions*. Symposium presented at the Annual Meeting of the Mathematical and Cognition Learning Society Conference.

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, *30*, 703-725.

Booth, J. L., Lange, K. E., Koedinger, K. R., & Newton, K. J. (2013). Using example problems to improve student learning in algebra: Differentiating between correct and incorrect examples. *Learning and Instruction*, *25*, 24-34.

Burr, S. M. D. L., Douglas, H., Vorobeva, M., & Muldner, K. (2020). Refuting Misconceptions: Computer Tutors for Fraction Arithmetic. *Journal of Numerical Cognition*, *6*, 355-377.

Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439-477.

Crooks, N. M., & Alibali, M. W. (2014). Defining and measuring conceptual knowledge in mathematics. *Developmental Review*, *34*, 344-377.

Fyfe, E. R., Matthews, P. G., Amsel, E., McEldoon, K. L., & McNeil, N. M. (2018). Assessing formal knowledge of math equivalence among algebra and pre-algebra students. *Journal of Educational Psychology*, *110*, 87.

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 1-27). Hillsdale, NJ: Erlbaum.

Hilbert, T. S., Renkl, A., Schworm, S., Kessler, S., & Reiss, K. (2008). Learning to teach with worked-out examples: A computer-based learning environment for teachers. *Journal of computer assisted learning*, *24*, 316-332.

Long, Y., & Aleven, V. (2013, July). Supporting students' self-regulated learning with an open learner model in a linear equation tutor. In *International conference on artificial intelligence in education* (pp. 219-228). Springer, Berlin, Heidelberg.

Long, Y., & Aleven, V. (2017). Enhancing learning outcomes through self-regulated learning support with an open learner model. *User Modeling and User-Adapted Interaction, 27*, 55-88.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*, 901.

Nagashima, T.,, Bartel, A. N., Silla, E. M., Vest, N. A., Alibali, M. W., & Aleven, V. (2020). Enhancing conceptual knowledge in early algebra through scaffolding diagrammatic self-explanation. In M. Gresalfi & I. S. Horn (Eds.), *14th International Conference of the Learning Sciences (pp. 35-43)*. Nashville, TN: International Society of the Learning Sciences.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis*, *36*, 127-144.

Rau, M. A., Aleven, V., & Rummel, N. (2015). Successful learning with multiple graphical representations and self-explanation prompts. *Journal of Educational Psychology*, *107*, 30.

Rittle-Johnson, B., & Loher, A. M. (2017). Eliciting explanations: Constraints on when self-explanation aids learning. *Psychonomic Bulletin & Review, 24*(5), 1501-1510.

Rittle-Johnson, B., Loehr, A. M., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM*, *49*, 599-611.

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology*, *103*(1), 85.

Roy, M., & Chi, M. T. (2005). The self-explanation principle in multimedia learning. *The Cambridge handbook of multimedia learning*, 271-286.

Salden, R. J. C. M., Koedinger, K. R., Renkl, A., & McLaren, B. M. (2010). Accounting for beneficial effects of worked examples in tutored problem solving. *Educational Psychology Review*, *22*, 379-392.

Sidney, P. G., & Alibali, M. W. (2015). Making connections in math: activating a prior knowledge analogue matters for learning. *Journal of Cognition and Development*, *16*, 160-185.

Wylie, R., & Chi, M. T. (2014). 17 The Self-Explanation Principle in Multimedia Learning. *The Cambridge handbook of multimedia learning*, 413.