Confrustion in Learning from Erroneous Examples: Does Type of Prompted Self-Explanation Make a Difference?

J. Elizabeth Richey[1[0000-0002-0045-6855]], Bruce M. McLaren[1], Miguel Andres-Bray[2], Michael Mogessie[1], Richard Scruggs[2], Ryan Baker[2[0000-0002-3051-3232]], and Jon Star[3[0000-0002-4830-3815]]

[1] Carnegie Mellon University, Pittsburgh PA 15213, USA
[2] University of Pennsylvania, Philadelphia PA 19104 USA
[3]Harvard University, Cambridge MA 02138
jrichey@andrew.cmu.edu

Abstract. Confrustion, a mix of confusion and frustration sometimes experienced while grappling with instructional materials, is not necessarily detrimental to learning. Prior research has shown that studying erroneous examples can increase students' experiences of confrustion, while at the same time helping them learn and overcome their misconceptions. In the study reported in this paper, we examined students' knowledge and misconceptions about decimal numbers before and after they interacted with an intelligent tutoring system presenting either erroneous examples targeting misconceptions (erroneous example condition) or practice problems targeting the same misconceptions (problem-solving condition). While students in both conditions significantly improved their performance from pretest to posttest, students in the problem-solving condition improved significantly more and experienced significantly less confrustion. When controlling for confrustion levels, there were no differences in performance. This study is interesting in that, unlike prior studies, the higher confrustion that resulted from studying erroneous examples was not associated with better learning outcomes; instead, it was associated with poorer learning. We propose several possible explanations for this different outcome and hypothesize that revisions to the explanation prompts to make them more expert-like may have also made them – and the erroneous examples that they targeted – less understandable and less effective. Whether prompted self-explanation options should be modeled after the shorter, less precise language students tend to use or the longer, more precise language of experts is an open question, and an important one both for understanding the mechanisms of self-explanation and for designing self-explanation options deployed in instructional materials.

Keywords: Erroneous Examples, Affect, Confusion, Frustration, Affect Detection, Learning Outcomes.

1 Introduction

Erroneous examples, or examples that illustrate typical student errors and misconceptions, can support performance, learning, and transfer [1-6]. Researchers have

hypothesized that erroneous examples derive their benefits through multiple processes, including helping students to recognize errors in their own work [7] and highlighting the underlying principles necessary for understanding correct solutions [2, 8]. However, no studies to our knowledge have examined the affective consequences of erroneous examples. In particular, it is unclear whether students experience confusion and frustration as they try to identify and understand errors that they might make themselves. To understand how and when erroneous examples are most effective, we examined the effects of erroneous examples on the affective state of confrustion – a combination of confusion and frustration – and its consequences for learning outcomes.

Many previous studies have shown a learning advantage when students are prompted to compare correct and incorrect examples [9] or explain and fix errors in incorrect examples [5, 10], in contrast to more typical worked-example study or problem-solving practice. In particular, studying erroneous examples may highlight for students the common errors that they are likely to make and discourage them from underestimating the difficulty of a problem. Erroneous examples have been shown to be particularly beneficial for supporting long-term learning and transfer [2, 5, 6].

Erroneous examples relate more broadly to research on desirable difficulties [11, 12] and productive failure [13, 14]. Experiencing difficulty on a task can increase engagement and mental effort and improve long-term learning outcomes [12, 13]. Critically, difficulty should directly relate to the concepts or procedures being taught. In other words, simply making a task difficult for the sake of difficulty likely will not improve learning; however, making the task difficult enough to require more effort or some initial failure can ultimately help the learner. Erroneous examples may operate through a similar mechanism by presenting students with solutions that they may have thought were correct, based on their own inaccurate knowledge and misconceptions.

Although emotions likely play a role in learning from difficulty or productive failure, the learner's affective experiences while struggling in such contexts has not, to our knowledge, been explored. The affective states of confusion and frustration are likely to be especially relevant in these contexts. Although confusion and frustration are theoretically distinct constructs, both have strong but mixed connections to learning [15-17]. Confusion is typically viewed as positive if the student believes it can be resolved [18-20], and it has been related to positive motivational experiences such as engagement and flow [21]. Frustration can arise when a student cannot resolve their confusion, and it can lead to disengagement and poor learning outcomes [22].

Despite the differences between confusion and frustration, research on affect detection has suggested there are predictive benefits to combining the two as a measure of confrustion. Affect detectors rely on students' interaction data from a learning system to determine the students' affective states. They can examine affect at a grain-size of about 20-second intervals and can predict immediate performance as well as long-term student outcomes [23, 24]. To create affect detectors, human coders first label data based on the absence or presence of an affective state [25-28]. After acceptable inter-rater reliability is established, machine learning algorithms identify behaviors in the learning system that correspond to the affect judgments made by human coders.

Confusion and frustration are often hard for an observer to distinguish based on the students' interactions with educational technology [22]. Perhaps for this reason, a study comparing the predictive value of confrustion assessed through affect detectors against separate measures of confusion and frustration found that confrustion was a more

accurate predictor of learning [22]. Consequently, in this study we combine these constructs and calculate a single measure of confrustion.

To explore the role of confrustion in learning from erroneous examples, we examined students' log interactions with an educational technology platform that presents a series of 32 erroneous examples and 16 practice problems targeting common decimal number misconceptions [29-31]. Previous research with this technology showed that students who corrected and explained erroneous examples performed better than students who solved and explained the same problems, but only on a delayed posttest [1, 5]. This suggests that studying and correcting erroneous examples might hinder – or at least not benefit – immediate performance but does lead to long-term learning.

If erroneous examples support learning by creating difficulty, students who are working through the materials might experience greater confrustion than students completing similar materials without erroneous examples. To examine this, we created affect detectors to assess students' experiences of confrustion [32]. A re-analysis of log files from the experiments reported in [1, 5] indicated that students did experience greater confrustion in the erroneous examples condition. However, confrustion was negatively related to performance. Confrustion thus does not appear to be beneficial on its own, but it may be a necessary consequence of the cognitively demanding learning processes supported by erroneous examples. In other words, students studying erroneous examples might learn more despite experiencing greater confrustion.

The current study aimed to replicate and build on the previous results in several ways. First, the data that was reanalyzed with affect detectors was collected more than six years ago. Use of educational technology has continued to increase in the time since those data were collected, potentially changing the ways students would view and interact with the materials. Erroneous examples have also gained prominence among teachers and instructional designers. As a result, students might be more accustomed to interacting with erroneous examples. For these reasons, we wanted to replicate both the learning and confrustion results with a new group of students.

Second, students in the erroneous example condition previously received several explanation prompts focused on each erroneous example and thus were prompted to do more self-explanation than students in the problem-solving condition [1, 5]. Self-explanation is a robust instructional technique [33, 34], and it is possible some of the benefits experienced by students in the erroneous example condition resulted from extra self-explanation. Additionally, the extra self-explanation may have contributed to students in the erroneous example condition spending nearly twice as much time as students in the problem-solving condition [1, 5]. To reduce the difference in time between conditions and decrease the chance that benefits were being derived from extra self-explanation, we removed one of the additional self-explanation prompts, leaving only one extra self-explanation prompt focused directly on making sense of the erroneous example.

Third, we saw opportunities to revise the self-explanation prompts to improve precision in the mathematical language used, and we worked with a math education expert (the seventh author of this paper) to make these changes. For example, the previous materials referred to the misconception that "longer decimals are larger," while it is more mathematically precise to express the idea as "decimal numbers with more digits to the right of the decimal point are greater in magnitude." While prior research has established that self-explanation can still be beneficial when students select or complete explanations using provided options within a computer-based learning

environment [35-36], instead of generating the explanations themselves, we know of no prior research that has explored the question of whether these provided explanations should be more similar to the less mathematically precise language students typically use or the more mathematically precise language of math experts. In the current study, we investigated whether the same learning benefits would be observed if students selected self-explanation prompts using more mathematically precise language.

Fourth, materials were updated to operate in HTML instead of Flash and to conform with modern look-and-feel instructional technology. For instance, prompted explanation boxes, a common interface feature in this educational technology system, were created with current HTML multiple-choice widgets. Using these revised materials, we sought to replicate previous results by testing the following hypotheses:

H1: Students in both conditions will improve in performance from pretest to posttest and from pretest to delayed test. We do not expect any of the changes made to the materials to disrupt the basic learning benefits of the intervention.

H2: Confrustion will be negatively related to performance, even when controlling for prior knowledge. We do not expect the changes to the materials to change the confrustion students experience, which related negatively to learning in prior studies.

H3: Students in the erroneous example condition will experience greater confrustion than students in the problem-solving condition. We made revisions aimed at simplifying the appearance of the erroneous example materials (modern look-and-feel) and to reduce extra text they had to read (elimination of extra self-explanation prompt text). However, we expect that the greater levels of confrustion come from the erroneous examples themselves and not from other features of the problem interface.

H4: Students in the erroneous example condition will perform better than students in the problem-solving condition on the delayed posttest. We do not expect any of the changes to disrupt the relative benefits of erroneous examples.

While we had no way of empirically testing the effect of revisions on the amount of time students required to complete the materials, we expected them to take less time on the materials overall and to show less of a time difference between conditions.

2 Methods

2.1 Participants

Participants were recruited from a suburban, public middle school (four teachers) and an urban, public elementary school (four teachers) in the metropolitan area of a northeastern U.S. city. Participants' parents provided written consent to collect and analyze students' data. Students completed materials as a part of their regular, in-class instructional activities. A total of 53 fifth-grade students and 134 sixth-grade students participated. Six fifth-grade students and seven sixth-grade students were dropped for failing to complete the materials in the allotted time. The final dataset included 174 students: 47 fifth-graders (30 male, 17 female; mean age 10.4) and 127 sixth graders (69 male, 58 female; mean age 11.2). Students were randomly assigned to conditions at the individual level, with 89 students assigned to the problem-solving condition and 85 students assigned to the erroneous example condition.

All students had previously learned about decimal numbers during their regular math

instruction (Common Core standard CCSS.Math.Content.5.NBT.A.3 for fifth grade; CCSS.Math.Content.6.NS.B.3 for sixth grade). To avoid introducing information that could affect students' performance, teachers were asked to refrain from providing decimal-number instruction or practice outside of the intervention during the study.

2.2 Materials

Materials were developed using the Cognitive Tutor Authoring Tool (CTAT) and delivered through Tutorshop, a learning management system for CTAT tutors that supports classroom deployment via web delivery [37]. We followed updated look-and-feel principles to revise the presentation of materials (see Fig. 1 and 2). Materials were aimed at addressing misconceptions while providing feedback and practice to students who had basic knowledge of decimal numbers. Both the erroneous example and problem-solving materials presented the same problems in the same order. Materials were organized into three-item sets of two erroneous example or problem-solving items with self-explanation, followed by one practice item without self-explanation. Practice items were the same across conditions. The materials included a total of 48 problems targeting different decimal number misconceptions. Tasks included number line placement, ordering by magnitude, addition, and completing a decimal number sequence.

Each erroneous example item presented a decimal number word problem with an incorrect solution from a hypothetical student (Fig. 2). Students worked through the problem in three steps and could not advance until they completed each step correctly. First, the student was prompted to explain the error in the example. Second, they corrected the error by solving the problem. Third, they explained the correct solution or relevant principles through two self-explanation prompts with multiple-choice solution options [35, 36]. The tutor provided feedback on all incorrect responses.

Each problem-solving item presented the same decimal number word problem but without an incorrect solution. Students worked through the problem in two steps and could not advance until they completed each step correctly. They solved the problem and then explained the solution or principle by answering two multiple-choice self-explanation questions. The tutor provided feedback on all incorrect responses.

The third problem in each set was a practice problem targeting the same misconception as the previous two erroneous example or problem-solving items. Practice problems consisted of only one step: solving the problem. Items were identical across conditions and were included to give students additional practice applying what they were learning in the materials. Previous research has shown that including practice problems immediately after example problems can improve learning outcomes [38].

Tests were administered on computers using the same educational technology platform as the intervention. There were 25 items on the pretest, posttest, and delayed posttest, with some items containing multiple parts and points. The tests were worth a total of 61 points, and test scores were computed as the number of points earned out of 61. We deployed three versions of the test with isomorphic problems, and test version order was counterbalanced across students. Items targeted the same misconceptions as the instructional materials and also included some transfer items that did not directly address a specific misconception. Items included tasks such as adding decimal numbers (e.g., 2.41 + 0.6 = ___) and identifying the largest or smallest decimal number from a list (e.g., 5.413, 5.75, 5.6). Transfer items targeted an understanding of decimal number

principles (e.g., “Is a longer decimal number larger than a shorter decimal number?”) and included new skills not covered in the intervention (e.g., “Select all of the following numbers that are equal to 0.43”).

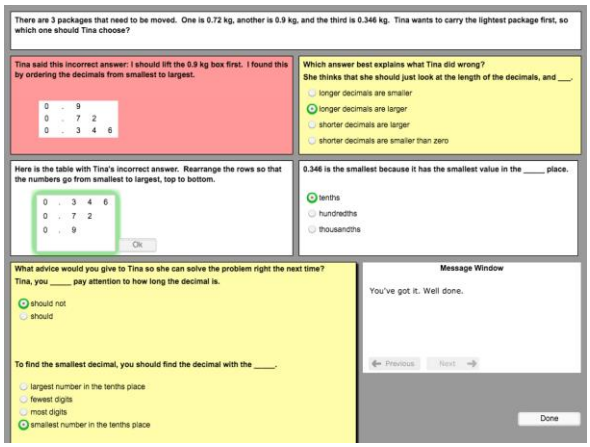

Fig. 1. A sample erroneous example problem from the original materials used in [1, 5]. This problem involves an ordering task and targets the “longer decimals are larger” misconception.

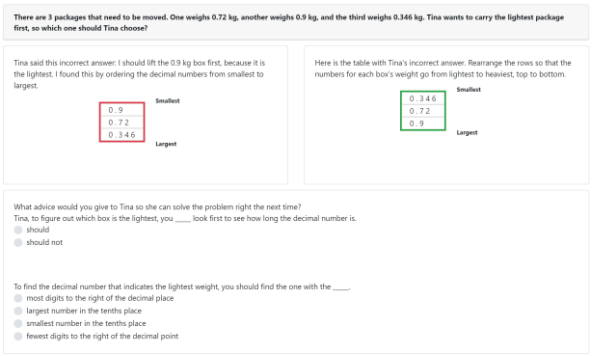

Fig. 2. A sample erroneous example problem from the current study, which has the same decimal content as Figure 1. Students still received a prompt to explain the error, but the prompt disappeared from the screen after it was answered.

2.3 Procedure

All materials were deployed during students’ regular math classes over the course of one week. Members of the research team were present each day to assist in administering the materials and ensure the protocol was followed. Students completed the pretest, instructional materials, and posttest at their own pace and their progress was saved each day. Students who completed the posttest before the end of the week were given math assignments by their teachers that did not target decimal number concepts. One week after the intervention, all students completed a delayed posttest.

2.4 Affect Detection

We developed affect detectors using labels from text replay coding, where segments of log files are pretty-printed and coded by humans. Those codes are input into machine learning algorithms to emulate the coders' judgments, based on prior studies that showed it was feasible to detect confrustion using this approach [27]. The detectors were built on log file data from 598 students across five middle schools collected in previous research with this educational technology platform [1, 5]; data related to the dropped self-explanation step in the previous version of the erroneous example condition were removed from the dataset before developing automated detectors, but remained included during text replay coding. Students' log files were broken down into individual clips for text replay coding, with each problem corresponding to a single clip. Two coders manually labeled 1,600 clips for confrustion based on holistic assessment of confrustion in the current task. For example, for multiple-choice problems, did the student spend a significant amount of time on a first, incorrect attempt and then make a subsequent incorrect attempt? For number line problems, did the student make two substantially distant, incorrect attempts (e.g., 0.3, then 1.1, then 1.8) or multiple incorrect attempts in both directions on the number line (e.g., 0.7, then 0.81, then 0.55)? As with most labels, these text replay labels are imperfect – we do not know if they genuinely capture the affective experience of confusion or frustration in all cases. As these labels are derived only from log files, unlike work that also considers facial expressions or posture [e.g. 25], these labels may in some cases capture only behavior *associated* with confrustion, rather than true confrustion. Agreement was computed after two coders separately labeled the same 129 clips, and results indicated high agreement ($\kappa = .82$, $p < .001$). The remaining 1,471 clips were coded independently by one of the two coders. The set of clips coded was stratified to equally represent both student cohorts, both conditions, and all four problem types.

We built the confrustion detector using the Extreme Gradient Boosting (XGBoost) classifier based on these labeled clips [39]. The classifier uses an ensemble technique that trains an initial, weak decision tree and calculates its prediction errors. It then iteratively trains subsequent decision trees to predict the error of the previous decision tree, with the final prediction representing the sum of the predictions of all the trees in the set. We determined that the detector could effectively infer students' confrustion ($\kappa = .82$, AUC = .92) based on 10-fold student-level cross-validation, which involved repeatedly building the model on some students' data and testing it on other students' data. Once effective detection was confirmed, we applied the detector to the new dataset (9,065 clips across 187 students). A total of 30 features were used to predict confrustion, and the importance of each feature was calculated as the proportion it contributed to the final model. The detector reported the probability that a student experienced confrustion on each problem; overall confrustion scores were computed as the average probability of confrustion across all problems.

3 Results

We report results in the order of our hypotheses. To examine whether students' performance improved as a result of the intervention (H1), we conducted a series of paired-samples t-tests separately by condition. Results indicated that students in the problem-

solving condition improved significantly from pretest to posttest, $t(88) = 6.83$, $p < .001$, $d = .44$, and from pretest to delayed test, $t(88) = 8.18$, $p < .001$, $d = .48$ (Table 1). Likewise, a paired-samples t-test indicated that students in the erroneous example condition improved significantly from pretest to posttest, $t(84) = 4.26$, $p < .001$, $d = .28$, and from pretest to delayed test, $t(84) = 5.29$, $p < .001$, $d = .37$.

Table 1. Proportional confrustion and test scores by condition.

Measure	Problem solving	Erroneous example
Confrustion	$M = .22$, $SD = .12$	$M = .33$, $SD = 12$
Pretest	$M = .53$, $SD = .22$	$M = .54$, $SD = .21$
Posttest	$M = .62$, $SD = .21$	$M = .59$, $SD = .21$
Delayed test	$M = .63$, $SD = .22$	$M = .62$, $SD = .23$

To test the relation between confrustion and performance (H2), we examined the correlation between the variables. Confrustion was negatively correlated with pretest performance, $r = -.64$, $p < .001$, posttest performance, $r = -.63$, $p < .001$, and delayed posttest performance, $r = -.62$, $p < .001$. To examine the relation between confrustion and performance when controlling for prior knowledge (H2), we tested a multiple regression including pretest and confrustion to predict posttest. The model was significant, $R^2 = .70$, $F(2, 171) = 195.66$, $p < .001$. Both confrustion, $\beta = -.186$, $p = .001$, and pretest, $\beta = .69$, $p < .001$, were significant predictors of posttest performance when holding the other factor constant. We applied the same multiple regression model to predict delayed posttest. The model was significant, $R^2 = .69$, $F(2, 171) = 188.22$, $p < .001$. Both confrustion, $\beta = -.15$, $p = .006$, and pretest, $\beta = .72$, $p < .001$, were significant predictors of posttest performance when holding the other factor constant. These results indicate that confrustion predicted test performance even when accounting for students' prior knowledge. In other words, the predictive value of confrustion was not merely a reflection of students' prior knowledge.

To examine the effect of condition on confrustion (H3), we conducted a one-way analysis of variance (ANOVA) that indicated students in the erroneous example condition experienced greater confrustion than students in the problem-solving condition, $F(1, 172) = 41.29$, $p < .001$, $d = 0.38$. To determine whether the relation of confrustion and test performance differed between conditions, we conducted a moderation analyses using PROCESS, an SPSS macro that uses 5000 bootstrap estimates to test mediation and moderation by creating confidence intervals for indirect effects [40]. We tested a PROCESS 1 model using condition as a moderator of the relation between confrustion and posttest performance and, separately, delayed test performance. For the immediate posttest, there was no significant interaction between confrustion and condition, $B = .29$, 95% CI [-.10, .67], and the inclusion of the interaction term did not explain significantly more variance in the model, $\Delta R^2 = .007$, $F(1, 170) = 2.11$, $p = .15$. For the delayed posttest, there was also no significant interaction between confrustion and condition, $B = .27$, 95% CI [-.14, .68], and the inclusion of the interaction term did not explain significantly more variance in the model, $\Delta R^2 = .006$, $F(1, 170) = 2.11$, $p = .19$. These results indicate that the relation between confrustion and performance did not differ between conditions.

To test the effect of condition on performance (H4), we conducted an ANOVA that revealed no differences between conditions on pretest, $F(1, 172) = 0.08$, $p = .77$, $d = 0.04$, posttest $F(1, 172) = 0.82$, $p = .37$, $d = 0.14$, or delayed posttest, $F(1, 172) = 0.17$, $p = .68$, $d = 0.06$ (Table 1). When controlling for pretest, an analysis of co-variance (ANCOVA) indicated that there was a significant effect of condition on posttest, $F(2, 171) = 4.10$, $p = .045$, $\eta_p^2 = .023$, with students in the problem-solving condition performing better. There was no effect of condition on delayed posttest when controlling for pretest, $F(2, 171) = 1.29$, $p = .26$, $\eta_p^2 = .008$. In other words, students in the problem-solving condition improved on the posttest significantly more than students in the erroneous example condition, but there were no differences in improvement on the delayed test. To understand the role of confrustion in this effect, we conducted ANCOVAs testing the effect of condition on test performance controlling for both confrustion and pretest. Results revealed no effect of condition on posttest, $F(3, 170) = 0.02$, $p = .90$, $\eta_p^2 < .001$, or on delayed posttest, $F(3, 170) = 0.35$, $p = .56$, $\eta_p^2 = .002$. This indicates that the variance in confrustion between conditions accounted for the condition effect on posttest improvement.

To understand other potential consequences of the revisions, we examined the amount of time students spent on the materials. An ANOVA indicated a significant difference in total time spent on the instructional materials, $F(1, 172) = 9.95$, $p = .002$, $d = 0.48$, with students in the erroneous example condition ($M = 62.90$, $SD = 23.24$) taking longer to complete the materials than students in the problem-solving condition ($M = 51.59$, $SD = 23.99$). This suggests that the extra self-explanation prompt that was eliminated from the erroneous examples was not responsible for the difference in times across conditions observed in previous studies.

4 Discussion

Unlike prior studies [1,5], students in the erroneous example condition performed *worse* than students in the problem-solving condition on the immediate posttest when controlling for the pretest, and there were no differences between conditions on the delayed posttest. While the results are reversed in terms of which condition performed better, there is a similar trend in the difference between posttest and delayed posttest. In prior studies, the benefits of erroneous examples emerged only on a delayed posttest, suggesting that students did not experience initial performance benefits but ultimately learned and were able to transfer knowledge better [1,5]. These previous results were consistent with other research on erroneous examples, which have tended to show the greatest benefit on delayed or transfer tests [2, 6, 10]. In the current study, students in the problem-solving condition showed an immediate performance advantage on the posttest but that advantage did not persist to the delayed posttest, suggesting that benefits from the problem-solving condition primarily affected performance and not the lasting, transferrable learning benefits that are typically most valued as an instructional goal. Thus, while results were inconsistent with prior work in the sense that students in the erroneous examples condition did not perform better on the delayed posttest, it was not a full reversal of effects as would have been seen if students in the problem-solving condition performed better on the delayed posttest.

We predicted that the benefits of erroneous examples would be robust enough to

persist despite several changes made to better align the conditions with one another and with the more precise mathematical language used by experts. While this prediction was not upheld, there are several possible explanations for the different outcome. First, students might be more accustomed to using instructional technology than they were when the original materials were tested six years ago. While we would not expect this to change the cognitive benefits of the instructional materials, it might reduce any confusion or frustration students would experience with the interface, such as understanding how to drag numbers to reorder them or select options from a drop-down menu. However, this idea is not supported by the time students spent on the materials. Students in the current study spent on average 50 to 60 minutes across conditions, while students in the previous studies spent on average 40 to 50 minutes across conditions [1,5].

Second, the elimination of the extra self-explanation prompt in the erroneous example condition might have reduced learning in that condition. We think this is unlikely, as students in the erroneous example condition still responded to three self-explanation prompts per question. However, only a direct comparison between versions with and without the additional prompt could provide conclusive evidence.

Third, and we think most likely, the shift to more mathematically precise language may have diminished the benefit of studying and explaining erroneous examples. Students on average spent 10 more minutes on the revised materials compared to the original ones. No other major changes were made to the content of the problem-solving materials, and the other major change to the erroneous examples condition involved *removing* materials. Students' prior knowledge in the current experiment was slightly lower than in prior studies (.53 current, .57 prior), which could cause an increase in the amount of time students needed. Nevertheless, the dramatic increase in time spent on materials supports the idea that students struggled more with reading the new explanation prompts and thus may not have benefitted from them as much. Erroneous example interventions typically instruct students to engage in an evidence-based learning activity to study and understand the erroneous examples, such as comparison [9] or explanation [10]. Without these instructionally robust activities to provide scaffolding, students may not pay as close attention to the erroneous examples or may fail to identify the underlying principles they represent. Put another way, if the mathematically precise language of the new explanation prompts was too difficult for students to understand, then the effect may have been similar to having no explanation prompts at all.

Future research should investigate these possible explanations empirically. We plan to attempt to replicate previous results by randomly assigning students to either the erroneous example or problem-solving conditions using either the original or revised materials, which will also permit a more direct comparison of times and performance across versions. Examination of students' own self-explanations has suggested that the act of engaging in self-explanation is beneficial even when explanations are flawed or mathematically imprecise [41]. Whether provided self-explanation options should be modeled after the imprecise language students tend to use or the more precise language of experts is an open question, and an important one both for understanding the mechanisms of self-explanation and for designing self-explanation options deployed in instructional materials. Since many instructional technologies use the self-explanation method of offering options from which students may choose [35-36], this is important question to resolve.

References

1. Adams, D., McLaren, B. M., Durkin, K., Mayer, R.E., Rittle-Johnson, B., Isotani, S., & Van Velsen, M. Using erroneous examples to improve mathematics learning with a web-based tutoring system. Computers in Human Behavior, 36C, 401-411 (2014).
2. Booth, J. L., Lange, K. E., Koedinger, K. R., & Newton, K. J. Using example problems to improve student learning in algebra: Differentiating between correct and incorrect examples. Learning and Instruction, 25, 24-34 (2013).
3. Borasi, R. Exploring mathematics through the analysis of errors. For the Learning of Mathematics, 7(3), 2-8 (1987).
4. Isotani, S., Adams, D., Mayer, R.E., Durkin, K., Rittle-Johnson, B., & McLaren, B.M. Can erroneous examples help middle-school students learn decimals? In: EC-TEL, pp. 181-195 (2011).
5. McLaren, B. M., Adams, D. M., & Mayer, R.E. Delayed learning effects with erroneous examples: A study of learning decimals with a web-based tutor. International Journal of Artificial Intelligence in Education, 25(4), 520-542 (2015).
6. Siegler, R. S. Microgenetic studies of self-explanation. Microdevelopment: Transition processes in development and learning, 31-58 (2002).
7. Siegler, R. S., & Chen, Z. Differentiation and integration: Guiding principles for analyzing cognitive change. Developmental Science, 11(4), 433-448 (2008).
8. Rushton, S. J. Teaching and learning mathematics through error analysis. Fields Mathematics Education Journal, 3(1), 4 (2018).
9. Durkin, K., & Rittle-Johnson, B. The effectiveness of using incorrect examples to support learning about decimal magnitude. Learning and Instruction, 22(3), 206-214 (2012).
10. Große, C. S., & Renkl, A. Finding and fixing errors in worked examples: Can this foster learning outcomes? Learning and Instruction, 17(6), 612-634 (2007).
11. Bjork, R. A. Memory and metamemory considerations in the training of human beings. In: Metcalfe, J., Shimamura, A. (eds.) Metacognition: Knowing about knowing, pp. 185 – 205. Cambridge, MA: MIT Press (1994).
12. Soderstrom, N. C., & Bjork, R. A. Learning versus performance: An integrative review. Perspectives on Psychological Science, 10(2), 176-199 (2015).
13. Kapur, M. Productive failure in learning math. Cognitive Science, 38(5), 1008-1022 (2014).
14. Kapur, M., & Bielaczyc, K. Designing for productive failure. Journal of the Learning Sciences, 21(1), 45-83 (2012).
15. D'Mello, S. A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. Journal of Educational Psychology, 105(4), 1082-1099 (2013).
16. Rodrigo, M. M. T., Baker, R. S., Jadud, M. C., Amarra, A. C. M., Dy, T., Espejo-Lahoz, M. B. V., ...Tabanao, E. S. Affective and behavioral predictors of novice programmer achievement. In: ACM SIGCSE, vol. 41, pp. 156–160, ACM Press, New York, NY (2009).
17. Schneider, B., Krajcik, J., Lavonen, J., Salmela-Aro, K., Broda, M., Spicer, J., … Viljaranta, J. Investigating optimal learning moments in U.S. and Finnish science classes. Journal of Research in Science Teaching, 53(3), 400–421 (2015).
18. D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. Confusion can be beneficial for learning. Learning and Instruction, 29, 153-170 (2014).
19. Lehman, B., D'Mello, S., Strain, A., Mills, C., Gross, M., Dobbins, A., ... & Graesser, A. Inducing and tracking confusion with contradictions during complex learning. International Journal of Artificial Intelligence in Education, 22(1-2), 85-105 (2013).

20. VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. Why do only some events cause learning during human tutoring? Cognition and Instruction, 21(3), 209-249 (2003).
21. D'Mello, S., & Graesser, A. Dynamics of affective states during complex learning. Learning and Instruction, 22(2), 145-157 (2012).
22. Liu, Z., Pataranutaporn, V., Ocumpaugh, J., Baker, R.S.J.D.: Sequences of frustration and confusion, and learning. In: EDM, pp. 114–120 (2013).
23. Kostyuk, V., Almeda, M.V., & Baker, R.S. Correlating affect and behavior in reasoning mind with state test achievement. In: LAK, 26 (2018).
24. Pardos, Z. A., Baker, R. S., San Pedro, M., Gowda, S. M., & Gowda, S. M. Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. Journal of Learning Analytics, 1(1), 107-128 (2014).
25. Ocumpaugh, J., Baker, R.S., & Rodrigo, M.M.T. Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences (2015).
26. Baker, R.S.J.d., Corbett, A.T., Wagner, A.Z. Human classification of low-fidelity replays of student actions. In: EDM-ITS, 29-36 (2006).
27. Lee, D.M.C., Rodrigo, M.M.T., Baker, R.S.J.d., Sugay, J.O., & Coronel, A. Exploring the relationship between novice programmer confusion and achievement. In: ACII (2011).
28. Sao Pedro, M. A., de Baker, R. S., Gobert, J. D., Montalvo, O., & Nakama, A. Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. User Modeling and User-Adapted Interaction, 23(1), 1-39 (2013).
29. Resnick, L. B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. Conceptual bases of arithmetic errors: The case of decimal fractions. Journal for research in mathematics education, 8-27 (1989).
30. Sackur-Grisvard, C. & Léonard, F. Intermediate cognitive organizations in the process of learning a mathematical concept: The order of positive decimal numbers. Cognition and Instruction, 2, 157-174 (1985).
31. Stacey, K. Travelling the road to expertise: A longitudinal study of learning. In: PME, vol. 1, pp. 19-36 (2005).
32. Authors (under review).
33. Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. Eliciting self-explanations improves understanding. Cognitive Science, 18, 439–477 (1994).
34. Rittle-Johnson, B., Loehr, A. M., & Durkin, K. Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM*, *49*(4), 599-611 (2017).
35. Johnson, C. I., & Mayer, R. E. Adding the self-explanation principle to multimedia learning in a computer-based game-like environment. Computers in Human Behavior, 26, 1246-1252 (2010).
36. Mayer, R. E., & Johnson, C. I. Adding instructional features that promote learning in a game-like environment. Journal of Educational Computing Research, 42, 241–265 (2010).
37. Aleven, V., McLaren, B.M., & Sewall, J. Scaling up programming by demonstration for intelligent tutoring systems development: An open-access website for middle school mathematics learning. IEEE Transactions on Learning Technologies, 2(2), 64-78 (2009).
38. Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. When problem solving is superior to studying worked examples. Journal of Educational Psychology, 93, 579-588 (2001).
39. Chen, T., & Guestrin, C. XGBoost: A scalable tree boosting system. In: ACM-SIGKKD, pp. 785-794 (2016).

40. Hayes, A. F. Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford Press, New York, NY (2013).
41. Hausmann, R. G., & Vanlehn, K. Explaining self-explaining: A contrast between content and generation. Frontiers in Artificial Intelligence and Applications, 158, 417-424 (2007).