Computers in Human Behavior 36 (2014) 401–411

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Using erroneous examples to improve mathematics learning with a web-based tutoring system

Deanne M. Adams [a], Bruce M. McLaren [b], Kelley Durkin [c], Richard E. Mayer [a,*], Bethany Rittle-Johnson [d], Seiji Isotani [e], Martin van Velsen [b]

[a] *Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106-9660, USA*
[b] *Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[c] *Department of Psychological and Brain Sciences, University of Louisville, Louisville, KY 40292*
[d] *Department of Psychology and Human Development, Vanderbilt University, Nashville TN 37203, USA*
[e] *Department of Computer Systems, University of São Paulo, ICMC-USP, Brazil*

ARTICLE INFO

Article history:

Keywords:
Erroneous examples
Problem solving
Mathematics learning
Computer-based tutors
Decimals

ABSTRACT

This study examines whether asking students to critique incorrect solutions to decimal problems based on common misconceptions can help them learn about decimals better than asking them to solve the same problems and receive feedback. In a web-based tutoring system, 208 middle school students either had to identify, explain, and correct errors made by a fictional student (erroneous examples group) or solve isomorphic versions of the problems with feedback (problem-solving group). Although the two groups did not differ significantly on an immediate posttest, students in the erroneous examples group performed significantly better on a delayed posttest administered one week later ($d = .62$). Students in the erroneous examples group also were more accurate at judging whether their posttest answers were correct ($d = .49$). Students in the problem-solving group reported higher satisfaction with the materials than those in the erroneous examples group, indicating that liking instructional materials does not equate to learning from them. Overall, practice in identifying, explaining, and correcting errors may help students process decimal problems at a deeper level, and thereby help them overcome misconceptions and build a lasting understanding of decimals.

1. Introduction

1.1. Objective

Students learn and understand mathematics at a deeper level when they are prompted to make judgments about what is wrong in other students' erroneous solutions. This proposal, which we call the *erroneous examples hypothesis*, is the focus of the present study. In particular, the objective of this experiment is to examine the effectiveness of mathematics instruction based on erroneous examples compared to traditional problem solving when both are incorporated into an online tutoring system.

To examine this hypothesis, we focus on the content area of decimals, because this is a critical sub-domain of mathematics, essential for almost all of more advanced mathematics. Students often have difficulty mastering decimals and misconceptions are common and persistent (Irwin, 2001; Resnick et al., 1989; Sackur-Grisvard & Léonard, 1985), with difficulties enduring even into adulthood (Putt, 1995; Stacey et al., 2001). We focus on online tutoring in order to contribute to evidence-based design principles for improving computer-based learning of mathematics. Participants in the problem-solving control group solved problems such as filling in the next two numbers in a sequence (i.e. 2.97, 2.98, 2.99, ___, ___). In contrast, participants in the erroneous examples group were presented with isomorphic questions in which a fictional student had solved the same problem incorrectly due to a misconception such as treating the two sides of the decimal as separate (i.e., 2.97, 2.98, 2.99, 2.100, 2.101) and were asked to identify and correct errors. This study investigated whether students who are exposed to erroneous examples (the erroneous examples group) would have a better learning outcome than students who get practice in solving the same problems (the problem-solving group).

* Corresponding author. Tel.: +1 805 893 2472.
E-mail addresses: deanne.adams@gmail.com (D.M. Adams), bmclaren@cs.cmu.edu (B.M. McLaren), kelley.durkin@gmail.com (K. Durkin), mayer@psych.ucsb.edu (R.E. Mayer), bethany.rittle-johnson@vanderbilt.edu (B. Rittle-Johnson), sisotani@icmc.usp.br (S. Isotani), vvelsen@cs.cmu.edu (M. van Velsen).

http://dx.doi.org/10.1016/j.chb.2014.03.053
0747-5632/© 2014 Elsevier Ltd. All rights reserved.

1.2. The potential of erroneous examples

According to cognitive theory of multimedia learning and cognitive load theory from which it is derived, the learner can engage in three kinds of cognitive processing during learning (Mayer, 2009, 2011; Moreno & Park, 2010; Sweller, 1999; Sweller, Ayres, & Kalyuga, 2011): *extraneous processing* (or extraneous load) is cognitive processing that does not support the instructional goal and is often caused by poor instructional design; *essential processing* (or intrinsic load) is cognitive processing required to mentally represent the material as presented and is caused by the complexity of the material; and *generative processing* (or germane load) is cognitive processing aimed at making sense of the presented material such as reorganizing it and integrating it with relevant prior knowledge and is caused by the learner's motivation to exert effort to understand the material. The challenge of instructional design is to promote generative processing, without burdening learners with too much essential and extraneous processing that overloads their limited working memory capacity.

Worked-out examples (also called *worked examples*), which consist of a problem statement, the steps taken to reach a solution, and the final solution, have been used effectively to help manage essential processing and decrease extraneous processing (Cooper & Sweller, 1987; McLaren, Lim, & Koedinger, 2008; McLaren & Isotani, 2011; Renkl, 2005, 2011; Renkl & Atkinson, 2010; Zhu & Simon, 1987). Worked examples achieve this by focusing the student's attention on the correct solution procedure to follow (Sweller, Ayres, & Kalyuga, 2011), which helps the student avoid searching their prior knowledge for solution methods and lessens extraneous processing. The freed up cognitive resources can then be used for generative processing, in particular, understanding and eventually automatizing the steps in a problem's solution procedure. In a classic experiment on worked examples, Cooper and Sweller (1987) found that learning to solve algebra equations by studying worked examples, paired with problems to solve, resulted in faster transfer test performance than simply solving all the same problems. A variety of subsequent studies and reviews have documented the superiority of instruction using worked examples (McLaren & Isotani, 2011; Renkl, 2005, 2011; Sweller et al., 2011).

However, one possible issue with worked examples is that although students may free up cognitive resources, this does not mean that the freed cognitive capacity will be used for generative processing (Renkl & Atkinson, 2010). One way to encourage generative processing is through self-explanation, in which learners explain the instructional materials to themselves. Chi, Bassok, Lewis, Reimann, and Glaser (1989) found that good problem solvers are more likely to generate self-explanation statements while thinking out loud when studying worked examples of physics problems. In addition, other research has shown the importance of explicitly prompting for self-explanation (Hausmann & Chi, 2002). While worked examples can be used to reduce processing demands, self-explanation prompts can be used to encourage deeper processing leading to better performance on transfer items (Atkinson, Renkl, & Merrill, 2003; Hausmann & Chi, 2002).

A less studied way to encourage deeper processing while using worked examples is to present students with *incorrect* (i.e., erroneous) examples. Erroneous examples have been studied by a few learning science researchers (e.g., Durkin & Rittle-Johnson, 2012; Große & Renkl, 2007; Siegler, 2002; Tsovaltzi et al., 2010) and involve most of the same steps as a worked example except one or more of the steps is incorrect. In these past studies, students typically must locate the error(s), explain the error(s), and then make appropriate corrections. Erroneous examples may encourage students to engage in generative processing, as they explain to themselves why a particular part of the problem is incorrect (Durkin & Rittle-Johnson, 2012). Erroneous examples may also help students focus on each step of a solution method separately to identify where the error occurred.

On the other hand, erroneous examples may also place additional processing demands on learners, overloading working memory with extraneous processing. For instance, a student working with an erroneous example may search for what is wrong in the example by trying to represent internally both the correct and incorrect solution steps and compare those steps (Große & Renkl, 2007). One possible way to relieve this processing demand is to highlight the error in the problem for the student (Große & Renkl, 2007). Tsovaltzi et al. (2010) found that providing erroneous examples of fraction problems, along with providing computer-generated help in finding and correcting the error, was more effective than presenting erroneous examples with no help. Given what is known from this past research, the erroneous examples materials used in the present experiment indicate where an error has occurred in the decimal problems.

Some of this past research suggests that erroneous examples can facilitate learning of mathematics. For example, Durkin and Rittle-Johnson (2012) found that students who compared worked and erroneous examples of decimal problems learned more than students who compared pairs of correct worked examples. Students who compared worked and erroneous examples were also twice as likely to discuss correct concepts as students who compared worked examples only. In addition, Siegler (2002) found that having students self explain both correct and incorrect examples of mathematical equality was more beneficial for learning a generalizable procedure than self-explaining correct examples only. In Kawasaki (2010), 5th grade students benefited when the teacher contrasted both a correct and an incorrect solution to a math problem, especially when the student also committed the same kind of error.

One issue with using erroneous examples for learning is that the prior knowledge level of the learner can interact with the effectiveness of the erroneous examples. Große and Renkl (2007) found that high prior knowledge (but not low prior knowledge) college students benefitted from lessons containing both correct and incorrect solutions to probability problems rather than seeing only correct solutions. In addition, the high prior knowledge students did not benefit from having errors highlighted, presumably because they were able to identify the errors on their own. Low prior knowledge individuals did significantly better when the errors were highlighted than when they were not. This research suggests that erroneous examples may not be effective for students who do not already have a basic grasp of the instructional material while high prior knowledge students may benefit by processing the information at a deeper level.

Similarly, if working memory is overloaded by the essential processing demands of the task, erroneous examples may not encourage deeper processing. This may have occurred, for instance, in an earlier study by our group (Isotani et al., 2011) with similar materials to those used in the present study. Middle school students learned decimals by either solving practice problems, or by having to piece together self-explanation sentences for either correct worked examples or a combination of correct and erroneous examples. There were no significant differences among the three groups on either an immediate posttest or delayed posttest. The possible benefits of erroneous examples may have been offset by a self-explanation interface that inadvertently induced cognitive load. The students were prompted to create self-explanations of why solutions are incorrect by being presented with the start of a sentence and then being asked to complete the sentence from two pull-down menus. Instead of focusing on the mathematic content, students may have devoted too much of their cognitive processing to selecting the correct sentence segments and reviewing the completed sentences. For the present study a simpler self-explanation interface was used by the erroneous examples group, aimed at minimizing extraneous processing while fostering generative processing.

Finally, it is important to note that much of the prior research on erroneous examples has focused on multi-step mathematics problems. Worked examples, which are excellent at showing multi-step problem solutions, were sometimes used as control conditions to examine the possible benefits of using erroneous examples. In contrast, the present study focuses on simpler decimal problems that (for the most part) require single-step solutions (e.g. comparing two decimals to decide which one is larger). Therefore, worked examples are not a good control condition for the present study; standard problem solving constitutes a better control to contrast with the effects of erroneous examples. Furthermore, previous studies, including the classic work of Cooper and Sweller (1987), have used problem solving as the control condition in a variety of learning situations and an analysis of mathematics textbooks indicates that presenting students with problems to solve is still the *de facto* standard in mathematics instruction (Hiebert et al., 2005; Mayer, Sims, & Tajika, 1995). Therefore, in order to produce a clear comparison between erroneous examples and the instructional approach most commonly used for decimal instruction, we focused on problem solving as the control group in the present study.

1.3. Learning decimals

Persistent misconceptions in students' decimal knowledge must be overcome so students can move on to more advanced mathematics. Yet, teaching students decimals is complicated by the fact that teachers are not always aware of common misconceptions and may misattribute incorrect answers to the wrong underlying misunderstandings (Stacey et al., 2001). Pre-service teachers in their study were aware of the misconception that longer decimals are larger, i.e., what Isotani et al. (2011) have called Megz; however, few were aware of the shorter-is-larger misconception (i.e., called Segz) and often make those errors themselves.

Based on an extensive literature review, Isotani et al. (2011) created a taxonomy of misconceptions that represents 17 distinct misconceptions. The present study focuses on four of these misconceptions, the ones that prior research has shown are most common and contributory to other misconceptions: Megz ("longer decimals are larger", e.g., 0.23 > 0.7), Segz ("shorter decimals are larger", e.g., 0.3 > 0.57), Negz ("decimals less than 1.0 are negative" e.g., .07 goes to the left of 0 on a number line), and Pegz ("the numbers on either side of a decimal are separate and independent numbers", e.g., 11.9 + 2.3 = 13.12). Our approach in this study is to address these common misconceptions directly; all of the problems presented to students in our study target one or more of these four misconceptions.

1.4. Present study

For the current study we streamlined the materials from Isotani et al. (2011) to help learners more easily focus on explaining and fixing errors in erroneous examples. The materials were altered so that students only had to complete sentences with a single choice for their self-explanation statements, as opposed to having to complete a sentence with two pull-down choices. Taking a cue from work by Johnson and Mayer (2010), in which students performed better on an embedded transfer test when they selected an explanation rather than having to generate one on their own, we propose that providing the explanation statements for misconceptions, rather than having learners generate their own, will relieve processing demands.

In addition, in order to focus the present study on a comparison of erroneous examples to the most common control, as well as to maximize statistical power, we simplified the design to two groups: erroneous examples and problem solving. Thus, this study examines whether erroneous examples can encourage deeper understanding than problem solving.

Our assumption is that three basic conditions are necessary to lead to learning from erroneous examples. First, to avoid embarrassment and demotivation, the errors should be examples of *another* student's errors, not their own. Second, the erroneous examples should be interactive and engaging. Using computer-based materials, students are prompted for explanations, asked to find and correct errors, and given feedback. Finally, enough guidance and structure should be provided to minimize extraneous processing and manage essential processing during learning with erroneous examples.

Taking those three points into consideration, the two groups in this study were presented with isomorphic problems, but with different ways of interacting with those problems. Students in the erroneous examples group were presented with an incorrect solution, prompted to explain and correct the error and reflect on the correct answer, and received feedback on their responses. Students in the problem-solving group were asked to solve the same problems, reflect on the correct answers and explain their solution, and received feedback on their work. The additional steps in the erroneous examples condition of explaining and correcting the error were intended to improve learning outcomes by encouraging learners to engage in generative processing of decimal principles. A comparison of the steps presented by the two conditions can be seen in Fig. 1.

1.5. Theory and predictions

From the perspective of generative learning theory (Mayer, 2009, 2011), the theoretical rationale for using erroneous examples is that they offer enough challenge to foster generative processing in learners—that is, cognitive processing aimed at making sense of the material by mentally reorganizing it and connecting it with relevant prior knowledge—while providing enough structure and guidance to not overload the learner with extraneous processing. In contrast, problem solving can cause learners to engage in so much extraneous processing that they fail to abstract the underlying solution principle used for solving the problem (Sweller et al., 2011). Based on research and theory on generative learning techniques in the science of learning, deep learning may be most sensitive to delayed rather than immediate tests (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013).

Hypothesis 1. Thus, we expect the erroneous examples group to engage in deeper cognitive processing during learning than the problem-solving group, yielding superior posttest performance, particularly on delayed tests as has been found with other generative techniques such as the testing effect (Dunlosky et al., 2013; Johnson & Mayer, 2009).

Hypothesis 2. By virtue of their training in evaluating student solutions, we also expect the erroneous examples group to outperform the problem-solving group on correctly assessing their level of confidence for their solutions of posttest problems.

2. Method

2.1. Participants and design

The participants were 208 middle-school students (101 boys and 107 girls) from the same Pittsburgh area school. One hundred

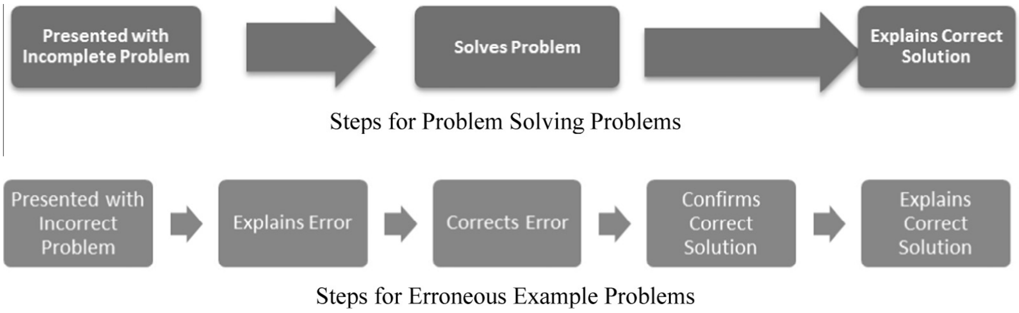

Presented with Incomplete Problem
Solves Problem
Explains Correct Solution
Steps for Problem Solving Problems
Presented with Incorrect Problem
Explains Error
Corrects Error
Confirms Correct Solution
Explains Correct Solution
Steps for Erroneous Example Problems

Fig. 1. Comparison between the sequences of steps in the two experiment conditions.

and five students were in the 6th grade and 103 were in the 7th grade. The students' ages ranged from 11 to 13 years old (M = 11.99, SD = .72). In a between subjects design, 100 students served in the erroneous examples group and 108 students served in the problem-solving group.

2.1.1. Materials and apparatus

The computer-based materials consisted of three isomorphic versions of an online 46-item decimal assessment test (with the three tests referred to as A, B, and C and used as pretest, immediate posttest, and delayed posttest, with counterbalanced orderings), a demographic questionnaire, an evaluation questionnaire, and two versions of an online lesson on decimals (i.e., the erroneous examples lesson and the problem-solving lesson) both implemented using intelligent tutor authoring software (Aleven, McLaren, Sewall, & Koedinger, 2009).

Corresponding items on the three 46-item decimal assessment tests were equivalent difficulty but contained different cover stories and values. Question types included placing decimals on a number line (i.e. "Place .7 on a number line between 0 and 1"), putting a given list of decimal numbers in order ("Please list .345, .34, and .4 in order from largest to smallest"), providing the next two decimal numbers in a sequence ("3.97, 3.98, 3.99, ____, ____), and answering true/false statements ("All decimals that have 0 in the ones place are negative"). All three tests had a total of 50 points possible; some of the problems included multiple parts that were scored separately. The 46 questions were designed to assess whether students held any of the four misconceptions before and after the intervention. To increase the accuracy in evaluating students' misconceptions, multiple question types addressed the same misconception to determine whether answers made by students were due to possible guessing or simple computation mistakes. For a subset of 15 of these items, spanning an equal number of the four target misconceptions, students were asked to rate how sure they were that their answer was correct using a 5-point Likert scale ranging from "Not at all sure" (1) to "Very sure" (5).

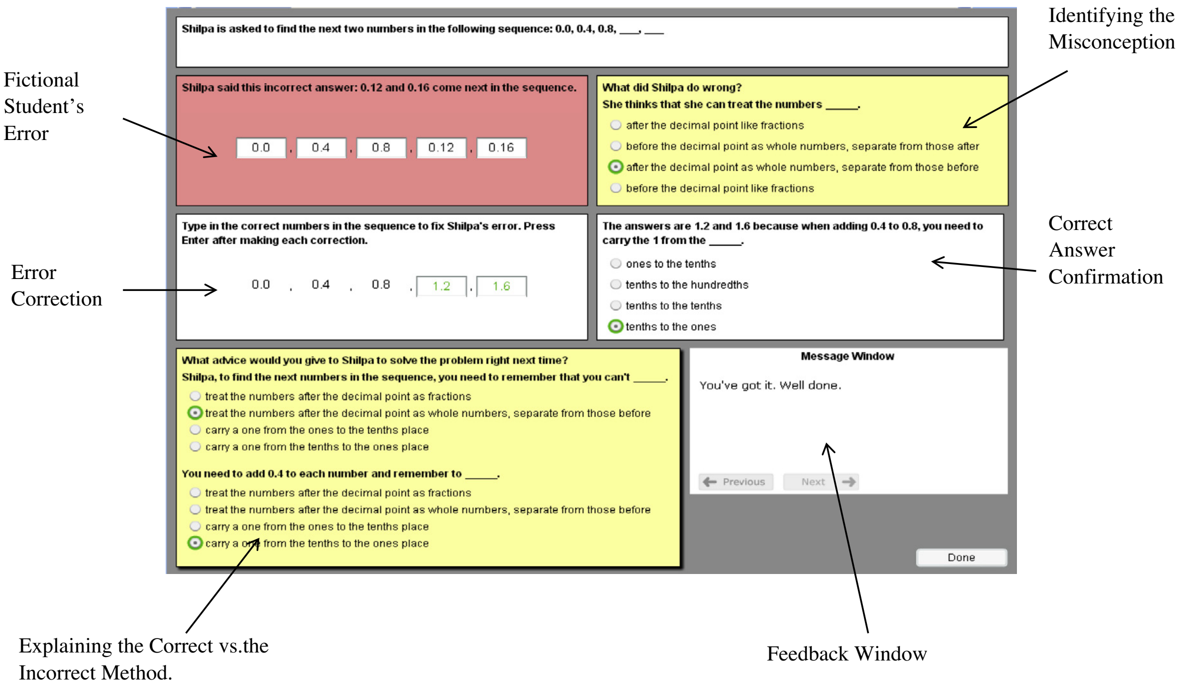
The demographic questionnaire, presented to all students before working on the online lesson, solicited basic information about age, gender, and grade level.

The evaluation questionnaire, presented to all students just after finishing the intervention, asked students how they felt about the lesson using a 5-point Likert scale ranging from "Strongly agree" (1) to "Strongly disagree" (5). The questionnaire included 10 items, which were combined into 4 categories: how well students liked the lesson, ease of interacting with the tutor, positive feeling about math, and perceived difficulty of the lesson. For how well students liked the lesson we averaged the ratings for two items, "I would like to do more lessons like this," and "I liked doing this lesson." For ease of interacting with the tutor we averaged four items: "I liked the way the material was presented on the screen," "I liked the way the computer responded to my input," "I think the interface of the system was confusing," and "It was easy to enter my answer into the system." For whether the tutor gave the student a positive feeling about math two items were combined: "This lesson made me feel more like I am goof at math," and "This lesson made me feel that math is fun." Lastly, for perceived difficulty of the lesson, two items were combined: "The material in this lesson was difficult for me," and "I worked hard on understanding the materials in this lesson."

Each of the two decimal interventions (i.e., the erroneous examples lesson and the problem solving lesson) was comprised of a total of 36 problems, organized into 12 blocks of three related problems, with each group targeted at one of the four misconception types. Table 1 shows the intervention format including the

Table 1
Problem order for the problem solving and erroneous examples conditions. Each row represents problems that are isomorphic between conditions except every third question which are the identical embedded test questions.

Supported Problem Solving (PS)	Erroneous Examples (ErrEx)
1. Megz supported PS 1	1. Megz ErrEx 1
2. Megz supported PS 2	2. Megz ErrEx 2
3. Megz PS 1	3. Megz PS 1
4. Segz supported PS 1	4. Segz ErrEx 1
5. Segz supported PS 2	5. Segz ErrEx 2
6. Segz PS 1	6. Segz PS 1
7. Pegz supported PS 1	7. Pegz ErrEx 1
8. Pegz supported PS 2	8. Pegz ErrEx 2
9. Pegz PS 1	9. Pegz PS 1
10. Negz supported PS 1	10. Negz ErrEx 1
11. Negz supported PS 2	11. Negz ErrEx 2
12. Negz PS 1	12. Negz PS 1
13. Megz supported PS 3	13. Megz ErrEx 3
14. Megz supported PS 4	14. Megz ErrEx 4
15. Megz PS 2	15. Megz PS 2
16. Segz supported PS 3	16. Segz ErrEx 3
17. Segz supported PS 4	17. Segz ErrEx 4
18. Segz PS 2	18. Segz PS 2
19. Pegz supported PS 3	19. Pegz ErrEx 3
20. Pegz supported PS 4	20. Pegz ErrEx 4
21. Pegz PS 2	21. Pegz PS 2
22. Negz supported PS 3	22. Negz ErrEx 3
23. Negz supported PS 4	23. Negz ErrEx 4
24. Negz PS 2	24. Negz PS 2
25. Megz supported PS 5	25. Megz ErrEx 5
26. Megz supported PS 6	26. Megz ErrEx 6
27. Megz PS 3	27. Megz PS 3
28. Segz supported PS 5	28. Segz ErrEx 5
29. Segz supported PS 6	29. Segz ErrEx 6
30. Segz PS 3	30. Segz PS 3
31. Pegz supported PS 5	31. Pegz ErrEx 5
32. Pegz supported PS 6	32. Pegz ErrEx 6
33. Pegz PS 3	33. Pegz PS 3
34. Negz supported PS 5	34. Negz ErrEx 5
35. Negz supported PS 6	35. Negz ErrEx 6
36. Negz PS 3	36. Negz PS 3



Shilpa is asked to find the next two numbers in the following sequence: 0.0, 0.4, 0.8, ___, ___

Shilpa said this incorrect answer: 0.12 and 0.16 come next in the sequence.

0.0 , 0.4 , 0.8 , 0.12 , 0.16

What did Shilpa do wrong?
She thinks that she can treat the numbers _____.

○ after the decimal point like fractions
○ before the decimal point as whole numbers, separate from those after
● after the decimal point as whole numbers, separate from those before
○ before the decimal point like fractions

Type in the correct numbers in the sequence to fix Shilpa's error. Press Enter after making each correction.

0.0 , 0.4 , 0.8 , 1.2 , 1.6

The answers are 1.2 and 1.6 because when adding 0.4 to 0.8, you need to carry the 1 from the _____.

○ ones to the tenths
○ tenths to the hundredths
○ tenths to the tenths
● tenths to the ones

What advice would you give to Shilpa to solve the problem right next time?
Shilpa, to find the next numbers in the sequence, you need to remember that you can't _____.

○ treat the numbers after the decimal point as fractions
● treat the numbers after the decimal point as whole numbers, separate from those before
○ carry a one from the ones to the tenths place
○ carry a one from the tenths to the ones place

You need to add 0.4 to each number and remember to _____.

○ treat the numbers after the decimal point as fractions
○ treat the numbers after the decimal point as whole numbers, separate from those before
○ carry a one from the ones to the tenths place
● carry a one from the tenths to the ones place

Message Window

You've got it. Well done.

Previous
Next
Done
Fictional Student's Error
Identifying the Misconception
Error Correction
Correct Answer Confirmation
Explaining the Correct vs.the Incorrect Method.
Feedback Window

Fig. 2. Sample Pegz sequence completion problem from the erroneous condition.

order in which the targeted misconception question types were presented. The intervention questions were more complex than the relatively straightforward assessment questions. Thus, while the raw number of questions vary between the assessments and the intervention, the depth and time required to complete the intervention questions was considerably greater than the assessment questions.

For the erroneous examples intervention, each of the 12 blocks consisted of a series of three problems. All three problems in a block involved the same misconception. The first two problems in each block asked students to evaluate an erroneous example, whereas the third problem required students to solve a problem that commonly produced the same misconception, with feedback on correctness provided. The erroneous examples problems contained up to 5 components (not including the problem statement) for the students to interact with.

Fig. 2 shows an erroneous example. It starts with a problem in which a fictional student has incorrectly filled in the next two numbers in a sequence by committing the Pegz misconception and treating the two sides of the decimal point as separate. In the top left box students read the error made by the fictional student in the word problem. After pressing a "Next" button students were asked to identify what the fictional student had done wrong from a list of 3–4 options, one of which was the misconception exhibited by that student. In the left middle panel students were asked to correct the mistake. This involved either placing a decimal correctly on a number line, correctly adding two decimals, correcting an incorrect sequence of decimals, or correctly ordering a list of decimals. In the right middle panel participants explained why the new answer was correct or confirmed the correct solution. Finally, in the bottom left panel the students were asked to give advice on how to solve the problem correctly. For every panel that required students to make a selection, feedback was provided (with the answer turning green[1] for correct answers, or red for incorrect answers). Students also received text feedback from a message window in the bottom right corner of the intervention screen. Messages included encouragement for students to try incorrect steps again (e.g., "Can you try that again? That answer is not correct") or "success" feedback to continue onto the next step or problem after correctly solving a step (e.g., "Nice. On to the next step . . .").

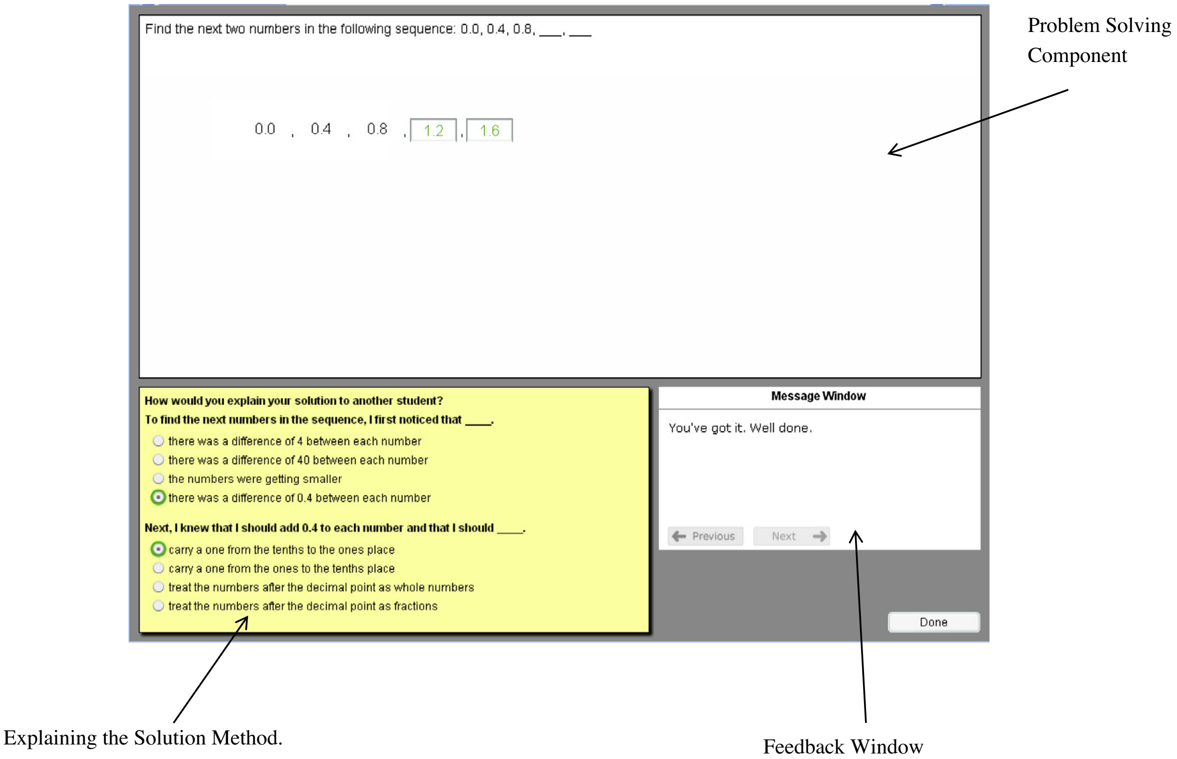
In the problem-solving intervention, students were given the same problems as those in the erroneous examples lesson except they were asked to solve the problems on their own. Fig. 3 includes the problem-solving version of the sequence completion problem from Fig. 2. These problems were also arranged in groups of three, isomorphic in content to the erroneous examples lesson sequence. For the first two problems of each group, students solved the problem, received feedback on correctness, and then were prompted to explain their solution. The explanation statements, which were multiple-choice questions, included one correct explanation, along with misconception distracters. Students in this group also received feedback from a message window in the bottom right panel as well as green/red feedback on their solution and multiple-choice explanation questions. On the third problem of each group the students were given simple correct/incorrect feedback but were not required to explain their solution.

The study took place in the school's computer lab, with lab time replacing the students' regular math class. The grades students received on the tests they took were averaged and used for a real class grade on decimals (and the students knew the test grades from the study would be used for a real class grade). Students worked on Apple computers with Internet connectivity.

2.2. Procedure

The students were randomly assigned to either the problem-solving group or the erroneous examples group. Students were also randomly assigned to receive one of the six possible pretest/posttest/delayed-posttest orderings (ABC, ACB, BAC, BCA, CAB, CBA). The students received a total of five 43-min sessions to complete the entire sequence of materials (i.e., pretest, demographic questionnaire, intervention, evaluation questionnaire, immediate

[1] For interpretation of color in Fig. 2, the reader is referred to the web version of this article.



Find the next two numbers in the following sequence: 0.0, 0.4, 0.8, ___, ___

0.0 , 0.4 , 0.8 , 1.2 , 1.6

How would you explain your solution to another student?

To find the next numbers in the sequence, I first noticed that ____.

○ there was a difference of 4 between each number
○ there was a difference of 40 between each number
○ the numbers were getting smaller
● there was a difference of 0.4 between each number

Next, I knew that I should add 0.4 to each number and that I should ____.

● carry a one from the tenths to the ones place
○ carry a one from the ones to the tenths place
○ treat the numbers after the decimal point as whole numbers
○ treat the numbers after the decimal point as fractions

Message Window

You've got it. Well done.

Fig. 3. Sample Pegz sequence completion problem from the problem solving condition.

posttest, delayed-posttest). Students completed the pretest during the first session. During the second and third session students completed their respective versions of the intervention as well as the evaluation questionnaire. During the 4th session students completed the immediate posttest. Finally, one week later, the students returned to the computer lab at the school for the fifth and final session, in which they took the delayed posttest. In every case, if the student finished early in that session, they were given non-decimal math homework to work on.

3. Results

3.1. Are the groups equivalent on basic demographic characteristics and prior knowledge measures?

Based on *t*-tests and chi-square results, the two groups did not differ significantly (with $p < .05$) on average age, proportion of boys and girls, or proportion of 6th and 7th graders.

Concerning pretest score, due to an error in data logging for four of the test problems, the data for those problems was removed from the pretest, immediate posttest, and delayed posttest scores making the total possible score out of 46. The first column of Table 2 shows the mean (and standard deviation) of the problem-solving group (PS) and the erroneous examples group (ErrEx) on the pretest. Despite random assignment, a *t*-test showed that the ErrEx group performed significantly better on the pretest than the PS group, $t(206) = 3.045$, $p = .003$, so subsequent analyses were based on using pretest score as a covariate as well as other techniques to mitigate this issue, as recommended by statisticians to statistically correct for selection bias (Hayes, 1994; Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). We also applied other recommended techniques that offer converging ways of statistically equating the groups on pretest score—removing outliers based on pretest score and partitioning groups into high and low pretest-score subgroups.

Table 2
Test performance for the two conditions.

Test	Groups			
	Problem solving		Erroneous examples	
	M	*SD*	*M*	*SD*
Pretest	24.68	(9.42)	28.69	(9.58)
Posttest	29.07	(9.48)	32.58	(8.95)
Delayed posttest	31.06	(9.20)	36.23	(7.47)*

Note. Asterisk indicates ErrEx group is significantly greater than PS group at $p < .05$.

3.2. Do the groups differ on learning outcomes?

The second row of Table 2 shows the mean (and standard deviation) on the immediate posttest for the PS and ErrEx groups. An ANCOVA with pretest score as a covariate revealed that the ErrEx and PS group did not perform significantly differently from one another on the immediate posttest, $F(1,205) = .768$, $MSE = 34.97$, $p = .382$, $d = .38$. The third row of Table 2 shows the mean (and standard deviation) on the delayed posttest for the PS and ErrEx groups. An ANCOVA with pretest score as a covariate showed that students in the ErrEx students scored significantly higher than students in the PS group on the delayed posttest, $F(1,205) = 9.896$, $MSE = 349.08$, $p = .002$, $d = .62$. Thus, the major empirical finding in this study is that the erroneous example students had higher learning gains than the problem-solving students on the delayed posttest but not the immediate posttest, suggesting that the effects of the erroneous example training are more significant over time.

To examine whether grade level affected performance on the immediate posttest and delayed posttest an ANCOVA with pretest as a covariate found no significant differences between 6th and 7th grades on either the immediate posttest, $F(1,203) = .111$, $MSE = 5.095$, $p = .739$, or the delayed posttest, $F(1,203) = .010$, $MSE = .350$, $p = .921$. There were also no significant interactions between condition and grade level for either the immediate posttest, $F(1,203) = .038$, $MSE = 1.761$, $p = .845$ or delayed posttest, $F(1,203) = .267$, $MSE = 9.511$, $p = .606$. Thus, we conclude that the erroneous examples training was equally effective for the 6th and 7th graders.

In addition to using ANCOVA to statistically equate the groups for pretest score, we conducted an ANOVA after removing students who scored one standard deviation above or below the overall mean on the pretest ($M = 26.61$, $SD = 9.696$). This left 69 students in the ErrEx group and 71 in PS group. Using this smaller better matched sample, a *t*-test showed that the pretest score of the ErrEx group ($M = 26.06$, $SD = 6.80$) did not differ significantly from the pretest score of the PS group ($M = 24.99$, $SD = 6.23$), $t(138) = .973$, $p = .332$, $d = .16$. The immediate posttest score of the ErrEx group ($M = 31.09$, $SD = 8.01$) did not differ significantly from that of the PS group ($M = 29.54$, $SD = 8.239$), $t(138) = 1.13$, $p = .261$, $d = .19$. However, on the delayed posttest, students in the ErrEx group scored significant higher ($M = 35.14$, $SD = 6.74$) than did the students in the PS group ($M = 31.58$, $SD = 7.71$), $t(138) = 2.911$, $p = .004$, $d = .49$. Thus, the same pattern of results favoring the ErrEx group on a delayed posttest was replicated with a smaller, better-matched sample that did not have the pretest differences. This analysis provides converging evidence based on recommended techniques for mitigating differences in pretest score that erroneous examples was a more effective instructional method than problem solving for learning about decimals.

Although not the primary focus of the study, we also note that overall students improved their test performance from pretest and immediate posttest, $t(207) = -8.06$, $p < .001$, $d = 0.44$, and from immediate posttest to delayed posttest, $t(207) = -8.23$, $p < .001$, $d = 0.31$. The increase from immediate posttest to delayed posttest may be an example of a testing effect in which the act of taking a test can help improve learning.

3.3. Are the group differences in learning outcome greater for students with low or high prior knowledge?

To determine whether the intervention had a different effect for students with high versus low prior knowledge we conducted an additional analysis. First, we classified students based on a median split on pretest score, with 107 students classified as low prior knowledge (i.e., pretest score from 8 to 25 points) and 101 students classified as high prior knowledge (i.e., pretest score from 26 to 45 points). Although this approach loses some information, it allows for a third way of mitigating the effects of unequal pretest scores in a way that is not biased in favor of the predicted group differences. Table 3 shows the mean (and standard deviation) for each group on the pretest, immediate posttest, and delayed posttest separately for low prior knowledge students and high prior knowledge learners. For high prior knowledge students, there was no difference for pretest scores between the two conditions, $t(99) = 1.65$, $p = .103$; however, for low prior knowledge students the ErrEx group had a non-significant trend of scoring higher than the PS group, $t(105) = 1.96$, $p = .052$.

For low prior knowledge students, to examine whether immediate posttest scores or the delayed posttest scores differed between the two groups, ANCOVAs with pretest as a covariate were conducted. Low prior knowledge participants in the ErrEx and PS groups did not perform significantly different on the immediate posttest, $F(1,105) = 2.33$, $MSE = 169.43$, $p = .13$, $d = .30$; however, the ErrEx condition performed significantly better on the delayed posttest, $F(1,105) = 8.95$, $MSE = 588.61$, $p = .003$, $d = .59$. The same pattern was found for students classified as having high prior knowledge with no significant difference on the immediate posttest, $F(1,99) = 1.08$, $MSE = 40.91$, $p = .301$, $d = .21$, but significantly higher performance on the delayed posttest favoring the ErrEx group, $F(1,99) = 7.58$, $MSE = 165.39$, $p = .007$, $d = .55$. Thus, the ErrEx group outscored the PS group on the delayed posttest for both low and high prior knowledge learners, yielding results consistent with other techniques for mitigating unequal pre-test scores. Similar results were found when participants were broken down into three equal groups (high, low, and average) as well as into 4 groups.

To determine whether prior knowledge level interacted with lesson condition, an ANCOVA with pretest score as a covariate was conducted. There were no significant interactions between prior knowledge level (high versus low) and lesson condition (PS versus ErrEx) for either immediate posttest scores, $F(1,203) = .55$, $MSE = .01$, $p = .46$, or delayed posttest scores, $F(1,203) = 1.94$, $MSE = .03$, $p = .17$. These results suggest that the erroneous examples lesson was just as effective for participants with low prior knowledge as it was for participants with high prior knowledge.

These analyses suggest that using erroneous examples to learn a relatively simple area of math, i.e., decimals, may be successful, regardless of prior knowledge. This result differs from that of Große and Renkl (2007) in which higher prior knowledge students benefited more from erroneous examples, but in a more complex mathematics domain (i.e., statistics) and with older learners (college students). This might have occurred because the erroneous examples students, both low and high prior knowledge, were encouraged to engage in more generative processing than the problem solving students, through the prompted explanation and correction of errors.

3.4. Does gaming behavior affect learning through problem solving or erroneous examples

One possible issue with using an interactive intelligent tutor, especially one that prompts for many multiple-choice responses (as the explanation statements did in this study), is that students may progress through the intervention without seriously engaging with the materials, clicking quickly through the answers while receiving visual feedback, and, consequently, not learning. This behavior has been referred to as *gaming the system* (Baker, Corbett, & Koedinger, 2004). Students who game the system take

Table 3
Test scores of high and low prior knowledge individuals in both lesson conditions.

Test	Groups							
	Low prior knowledge				High prior knowledge			
	PS ($N = 63$)		ErrEx ($N = 44$)		PS ($N = 45$)		ErrEx ($N = 56$)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Pretest	17.71	(3.89)	19.25	(4.01)	34.40	(5.36)	36.11	(5.03)
Posttest	24.24	(8.82)	26.8	(8.06)	35.84	(5.33)	37.13	(6.75)
Delayed-posttest	26.30	(8.53)	31.07	(7.47)*	37.71	(5.06)	40.29	(4.34)*

Note: Asterisk indicates ErrEx groups is significantly greater than PS group at $p < .05$.

advantage of elements of the system to progress quickly through the intervention, while only paying superficial attention to the lesson. To determine whether a student was gaming the system the amount of time the student took to complete the lesson along with the number of incorrect selections the student made was examined. A *gaming score* was calculated per student – dividing the total lesson time for that student by the number of incorrect choices made by the student. Students who gamed the system should have smaller gaming scores whereas students who seriously engaged and struggled with the problems, without rushing and rapidly responding, should have larger gaming scores. Due to a varying number of total possible responses between the two intervention conditions, median splits were conducted within each condition separately to determine gamers versus non-gamers, yielding 50 gamers and 50 non-gamers for the erroneous examples condition and 54 gamers and 54 non-gamers for the problem solving condition. Table 4 includes the means and standard deviations for the pretest, posttest, and delayed posttest for gamers and non-gamers for each condition.

When examining learning outcome differences between gamers and non-gamers, non-gamers scored significantly higher on the pretest than gamers, $t(206) = -8.277$, $p < .001$, $d = 1.15$. To look at how performance differed on the immediate and delayed posttest with regards to condition and gaming, an ANCOVA with pretest as a covariate showed a main effect of gaming in which non-gamers scored significantly higher on both the immediate posttest, $F(1,203) = 13.75$, $MSE = 588.94$, $p < .001$, $d = 1.15$, and the delayed posttest, $F(1,203) = 29.73$, $MSE = 902.05$, $p < .001$, $d = 1.33$. Condition still showed a significant main effect favoring the ExErr group only on the delayed posttest favoring the ErrEx group, $F(1,203) = 16.11$, $MSE = 488.64$, $p < .001$, but not on the immediate posttest, $F(1,203) = 1.80$, $MSE = 77.19$, $p = .18$. There was a non-significant interaction trend between condition and gaming behavior for the delayed posttest, $F(1,203) = 3.76$, $MSE = 114.12$, $p = .054$, but no interaction for the immediate posttest, $F(1,203) = .672$, $MSE = 28.78$, $p = .413$. These results are consistent with the claim that students who game the system are less likely to learn the material in the lesson, and the erroneous example training is more effective than problem solving training on the delayed posttest for both gamers and non-gamers.

3.5. Do students differ in misconceptions they have?

To determine the types of errors students made during the intervention and whether they were consistent with our target misconceptions, the students' behavior was modeled using a Bayes Net of decimal misconceptions (Goguadze, Sosnovsky, Isotani, & McLaren, 2011). These models are updated when students take the A, B, C pretests described above. The Bayes Net represents the misconceptions that a student might have – Megz, Segz, Pegz, and Negz – and is updated based on test questions that precisely probe for each of the misconceptions. Our A, B, C isomorphic tests contain 9 Megz problems, 10 Segz problems, 10 Pegz problems, and 9 Negz problems, after the 4 buggy problems were eliminated (there are also 8 problems that are targeted at a more general misconception called Regz, which contributes to all of the other misconceptions). Students can either get these problems correct, in which case the probability of the targeted misconception drops, they can get them incorrect in an unexpected way, in which case the misconception in the Bayes Net is only partially increased, or they can get them incorrect in a way that provides direct evidence for the misconception, in which case the misconception in the Bayes Net is increased more. Some of the misconception problems have possible answers that can indicate more than one misconception. The details of the Bayes Net are discussed in Goguadze et al., 2011. Our approach was inspired by the similar implementation of Stacey, Sonenberg, Nicholson, Boneh, and Steinle (2003).

Given how the Bayes Net of each of the 208 students in the present study were updated, we calculated mean probabilities over all misconceptions as Segz = 0.37; Megz = 0.31; Pegz = 0.15; Negz = 0.15. Furthermore, we created *misconception profiles* for all of the students, based on the order of probability of each of the misconceptions for that student. For instance, a student with a Megz probability of 0.92, Segz probability of 0.75, Pegz probability of 0.32 and Negz probability of 0.20 would have a misconception profile of Megz > Segz > Pegz > Negz. Table 5 summarizes the misconception profiles of all the students by most prominent misconception, i.e., the misconception that has the highest probability. As can be seen, the students were reasonably well distributed across the most prominent misconception categories, but there are stark differences in the mean values. Note that students who displayed the Megz misconception ("longer decimals are larger") and Segz misconception ("shorter decimals are larger") as their most likely misconception, show a very high probability for actually having those misconceptions (see bold items in rows 1 and 2), while the students who displayed the Pegz ("each side of the decimal is separate and independent") and Negz ("decimals less than 1.0 are negative") misconceptions as most likely, show a much lower probability for actually having those misconceptions (see bold items in rows 3 and 4). Furthermore, the pretest scores of the Megz and Segz students are dramatically lower than the Pegz and Negz students, and the *other* possible misconceptions of the Megz and Segz students have a much high probability than those of the Pegz and Negz students.

3.6. Do the groups differ on their awareness of misconceptions?

In addition to measuring the prevalence of students' misconceptions, the strength of students' misconceptions was also calculated. First, the mean confidence level for each student was calculated using the data from the confidence scales. Overall, after including pretest confidence as a covariate in the ANCOVA model, there were no differences between groups for level of confidence at immediate posttest or delayed posttest, $F(1,205) = 0.16$, $MSE = .224$, $p = .693$ and $F(1,205) = 1.49$, $MSE = .249$, $p = .224$, respectively. Students' responses were then categorized by confidence level and accuracy, which led to four response categories: high confidence error, low confidence error, low confidence

Table 4
Test scores of students classified by gaming behavior according to their condition.

Test	Groups							
	Gamers				Non-gamers			
	PS (*N* = 54)		ErrEx (*N* = 50)		PS (*N* = 54)		ErrEx (*N* = 50)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Pretest	20.52	(6.78)	23.14	(7.83)	28.83	(9.9)	34.24	(7.83)
Posttest	24.41	(8.86)	27.88	(8.52)	33.74	(7.64)	37.28	(6.64)
Delayed-posttest	25.8	(8.29)	31.76	(7.65)	36.31	(6.76)	40.7	(3.67)

Table 5
Summary of the misconception profiles of all 208 students.

Most prominent misconception	Number of students	Pretest average score (out of 46)	Megz	Segz	Pegz	Negz	General misconception profile
Megz	42	18.1	**0.97**	0.55	0.28	0.19	Megz > Segz > Pegz > Negz
Segz	60	19.3	0.34	**0.89**	0.10	0.21	Segz > Megz > Negz > Pegz
Pegz	58	33.9	0.03	0.02	**0.22**	0.03	Pegz > Megz > Negz > Segz
Negz	48	34.3	0.00	0.00	0.00	**0.19**	Negz > Pegz > Megz > Segz

correct, and high confidence correct. Students' responses were categorized as being low confidence if they were a 1 or 2 on the 5-point scale and high confidence if they were a 4 or 5 on the 5-point scale. High confidence errors were most likely strongly held misconceptions, while low confidence errors were most likely due to guessing.

At pretest, there were no significant differences between groups for any of the response types (p's > .06) except for high confidence correct responses, $F(1,206) = 11.33$, $MSE = .061$, $p = .001$, $d = .44$. Students in the ErrEx condition made high confidence correct responses more often ($M = 44.33\%$, $SD = 26.16$) than students in the PS condition ($M = 32.78\%$, $SD = 23.35$). ANCOVA models were then run with each of the four response types as an outcome measure for immediate posttest and delayed posttest, with the pretest rate of the respective response type as a covariate to control for any pretest differences. These analyses indicated that on the immediate posttest, students in the ErrEx group were less likely to make high confidence errors (i.e., strong misconception errors), $F(1,205) = 4.42$, $MSE = .016$, $p = .037$, $d = -.36$. Students in the ErrEx group only made such errors 13.93% of the time ($SD = 11.76$), while students in the PS condition made them 18.89% of the time ($SD = 15.48$). On the delayed posttest, there was a nonsignificant trend in which students in the ErrEx group were less likely to make high confidence errors ($M = 14.60\%$, $SD = 15.06$) than students in the PS condition ($M = 20.00\%$, $SD = 16.98$), $F(1,205) = 3.52$, $MSE = .020$, $p = .062$, $d = -.34$. Also on the delayed posttest, students in the ErrEx group were more likely to make high confidence correct responses ($M = 62.20\%$, $SD = 26.21$) than students in the PS condition ($M = 47.72\%$, $SD = 26.62$), $F(1,205) = 5.20$, $MSE = .042$, $p = .024$, $d = .55$. There were no other significant differences between conditions for the different response types (p's > .272).

These response categorizations based on confidence level and accuracy were also used to measure students' ability to accurately assess their own knowledge (i.e., knowledge calibration). In this case, students' answers were considered well calibrated if they had low confidence in errors and high confidence in correct responses, but they were considered poorly calibrated if they had high confidence in errors and low confidence in correct responses. The mean proportions of well-calibrated answers were calculated for each student for the pretest, immediate posttest, and delayed posttest. There was a significant difference between groups for calibration on the pretest, $F(1,205) = 8.531$, $MSE = .039$, $p = .004$, $d = .41$. Students in the ErrEx condition gave well-calibrated responses more often ($M = 52.44\%$, $SD = 20.45$) than students in the PS condition ($M = 44.41\%$, $SD = 19.22$). To control for these differences, pretest calibration was used as a covariate in all subsequent models on calibration. Using an ANCOVA model that included pretest calibration as a covariate, there was no significant difference between conditions for calibration on the immediate posttest, $F(1,205) = 0.78$, $MSE = .036$, $p = .379$, $d = .32$. Students in the ErrEx condition gave well-calibrated responses about as often ($M = 59.07\%$, $SD = 23.82$) as students in the PS condition ($M = 51.98\%$, $SD = 20.52$). However, there was a significant difference between groups for calibration on the delayed posttest, even after controlling for pretest calibration, $F(1,205) = 5.48$, $MSE = .041$, $p = .020$, $d = .49$. Students in the ErrEx condition gave well-calibrated responses more often ($M = 65.39\%$, $SD = 23.62$) than students in the PS condition ($M = 53.81\%$, $SD = 23.37$). Overall, the results are consistent with the prediction that the ErrEx group developed better skill at making evaluative judgments for solving decimal problems than the PS group.

3.7. Do the groups differ on their satisfaction with the online lesson?

To examine whether students' evaluations of the two lessons differed four scales were created to assess different aspects of the lesson: lesson enjoyment, perceived material difficulty, feelings of math efficacy, and ease of interface use. The PS condition students reported that they liked the lesson significantly more than the ErrEx students $t(206) = -2.37$, $p = .02$, $d = -0.33$, as well as finding it easier to interact with the tutor interface, $t(206) = -2.69$, $p = .008$, $d = -0.37$. PS students were also more likely to report that the lesson gave them more positive feelings about math, $t(206) = -2.05$, $p = .04$, $d = -028$. There was no significant difference between the two groups on perceived difficulty of the lesson materials between the two groups, $t(206) = -.71$, $p = .48$. Overall, the results show that the PS students liked the lesson more than the ErrEx students, even though previous analysis showed that the ErrEx group performed better on the delayed pretest than the PS group. Clearly, liking instructional materials does not translate into learning.

4. Discussion

4.1. Empirical contributions

Consistent with Hypothesis 1, the primary finding is that the ErrEx group significantly outperformed the PS group on the delayed posttest, both when pretest score was used as a covariate ($d = .62$) and when outliers were eliminated ($d = .49$). However, the superiority of the ErrEx group over the PS group ($d = .38$) did not reach statistical significance on the immediate posttest. This pattern of results is a major new contribution to the research literature on erroneous examples and is consistent with some research on other generative learning methods, such as the testing effect (Dunlosky et al., 2013; Johnson & Mayer, 2009), in which the effects of generative learning methods tend to be strongest on delayed posttests. It is also important to note that the ErrEx group significantly outperformed the PS group on the delayed posttest both for low prior knowledge learners ($d = .59$) and high prior knowledge learners ($d = .55$).

Consistent with Hypothesis 2, another important finding that represents a new contribution to the research literature is that the ErrEx group demonstrated better ability to judge the correctness of their answers. In short, erroneous examples helped students make fewer errors related to strongly held misconceptions and have more accurate knowledge calibration ($d = .49$). This confirms our prediction and complements past findings that exposing students to erroneous examples can reduce the frequency of misconceptions (e.g., Durkin & Rittle-Johnson, 2012). By directly addressing students' misconceptions, students may be forced to recognize that their previous conceptions were incorrect and that they need to accommodate a new conception of the material

(Eryilmaz, 2002). Erroneous examples may have improved students' knowledge calibration by helping students label incorrect concepts as wrong and correct concepts as right by emphasizing critical features of erroneous and correct examples (Van den Broek & Kendeou, 2008). This improved knowledge calibration can help people learn by making them aware of weaknesses in their knowledge and by motivating them to attend to information that could strengthen this knowledge (Cunningham, Perry, Stanovich, & Stanovich, 2004).

Results also showed that students did best in the erroneous examples group if they engaged in deeper processing of the material, by taking their time and making fewer answer attempts, and did not ‘‘game the system.'' In contrast, students performed worse if they were in the problem solving condition and engaged in gaming behavior by quickly clicking through answer choices to find the correct solution. Overall, students who gamed the system were the least likely to actually learn the material.

Finally, the PS group reported liking the lesson more than the ErrEx group although liking did not translate into better learning.

4.2. Theoretical contributions

The results on the delayed posttest are consistent with the idea that erroneous examples encourage learners to process the material more deeply during learning—more specifically, erroneous examples prime generative processing without creating extraneous processing. We infer these processes from posttest performance, so future work is needed to directly assess cognitive processing during learning. Furthermore, the results on the calibration measures are consistent with the idea that erroneous examples help students build skills that allow them to better judge the adequacy of their own solutions. Overall, these results suggest that students taught with erroneous examples may have had a deeper learning experience that persists over time. In particular, when the erroneous example treatment is designed to not overload the learner's cognitive system, it appears to encourage generative processing (i.e., deeper cognitive processing aimed at organizing the material and relating it to relevant prior knowledge) and the development of metacognitive skills, such as how to evaluate the adequacy of a solution procedure.

4.3. Practical contributions

The present study provides evidence for instructional design based on what can be called the *erroneous examples principle*: People learn more deeply when they are asked to critique the incorrect solution procedures of others. For example, when teaching how to solve mathematics problems, students can benefit from being asked to critique an incorrect worked example by articulating what is wrong in the student's conceptualization of the problem and comparing the wrong approach with a correct one. The erroneous example principle complements the well-established worked example principle (Sweller, Ayres, & Kalyuga, 2011; Renkl, 2011). The present study suggests that training with erroneous examples is effective when extraneous load is minimized and when the learner ultimately is encouraged to explain the correct steps. The appropriate balance between erroneous examples and problem solving – how much and precisely when to present each type of material – is a matter for additional research.

4.4. Methodological contributions

Our study also found evidence that middle school students hold misconceptions such as thinking longer decimals are larger or Megz (as is so with whole numbers) or that shorter decimals are larger or Segz (as is so with shorter denominators in fractions), confirming prior mathematics education findings (Irwin, 2001; Resnick et al., 1989; Sackur-Grisvard & Léonard, 1985). By taking these results into account, later versions of this intervention will be adapted by focusing on problems that might correct these misconceptions. An adaptive tutoring system could also use the misconception profile for each student in order to create an intervention curriculum that fits the individual needs of each learner. If the system identifies that a student holds a particular misconception then more problems can be added from the available problem sets that address that misconception. Future studies will be conducted to explore whether an adaptive intervention might further increase learning outcomes with erroneous examples.

4.5. Limitations and future directions

One limitation of the present study is that we did not include a correct worked examples condition. The reasons for this are straightforward. First, in the present study we wanted to compare the most common ecological control condition – that of students solving problems – to the much less typical learning experience of working with erroneous examples. Second, analysis of the instructional materials from the Isotani et al. (2011) study showed us that erroneous examples and problem solving were more comparable from a cognitive load perspective. As designed, they both require active problem solving – in the case of erroneous examples, finding the error and correcting it; and in the case of problem solving, generating the solution from the given problem – something worked examples do not require. Third, as we mentioned in the introduction, this work focused on simpler decimal problems that require single-step solutions and, therefore, worked examples may not be a useful control condition. Studying just the worked examples without any active problem solving may become redundant, thus decreasing the amount of mental effort students put into the lesson. Nevertheless, to more completely compare instructional approaches, in a future study it would be useful to include a worked examples condition.

The use of a multiple-choice format for responding during learning has the disadvantage of enabling students to game the system by trying responses until one works, but has the advantage of minimizing cognitive load and guiding learner's choices so that learners can focus on content of the lesson. Future research is needed to examine response formats that reduce cognitive load during learning without encouraging students to game the system.

In addition, although this study included a one-week delayed test, it would be useful in future work to examine whether the effects persist over a longer term. In summary, this study provides evidence that presenting students with erroneous examples along with prompts to analyze, explain, and correct the errors can facilitate the learning of decimals from a computer-based tutor.

Acknowledgments

This project was supported by Award No. R305A090460 (PI: B. McLaren) from the Institute of Education Sciences (IES) of the U.S. Department of Education. We also thank the Pittsburgh Science of Learning Center, NSF Grant 0354420, for technical support of our work.

References

Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2009). A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education, 19*(2), 105–154.

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Combining fading with prompting fosters learning. *Journal of Educational Psychology, 95*, 774–783.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004) Detecting student misuse of intelligent tutoring systems. *In Proceedings of the 7th international conference on intelligent tutoring systems* (pp. 531–540).

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, R., & Glaser, R. (1989). Self explanations: How students study and used examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

Cooper, G., & Sweller, J. (1987). The effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79*, 347–362.

Cunningham, A. E., Perry, K. E., Stanovich, K. E., & Stanovich, P. J. (2004). Disciplinary knowledge of K-3 teachers and their knowledge calibration in the domain of early literacy. *Annals of Dyslexia, 54*, 139–167.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 5–58.

Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction, 22*, 206–214.

Eryilmaz, A. (2002). Effects of conceptual assignments and conceptual change discussions on students' misconceptions and achievement regarding force and motion. *Journal of Research in Science Teaching, 39*, 1001–1015.

Goguadze, G., Sosnovsky, S., Isotani, S., & McLaren, B. M. (2011). Evaluating a Bayesian student model of decimal misconceptions. *In Proceedings of the 4th international conference on educational data mining (EDM 2011), Netherlands* (pp. 301–306).

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction, 17*(6), 612–634.

Hausmann, R. G. M., & Chi, M. T. H. (2002). Can a computer interface support self-explanation? *International Journal of Cognitive Technology, 7*, 4–14.

Hayes, W. L. (1994). *Statistics*. New York: Holt, Rinehart & Winston.

Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., et al. (2005). Mathematics teaching in the United States today (and tomorrow): Results from the TIMSS 1999 video study. *Education Evaluation and Policy Analysis, 27*(2), 111–132.

Irwin, K. C. (2001). Using everyday knowledge of decimals to enhance understanding. *Journal for Research in Mathematics Education, 32*(4), 399–420.

Isotani, S., Adams, D., Mayer, R. E., Durkin, K., Rittle-Johnson, B., & McLaren, B. M. (2011). Can erroneous examples help middle-school students learn decimals? *In Proceedings of the sixth European conference on technology enhanced learning: Towards ubiquitous learning (EC-TEL-2011)*.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*, 621–629.

Johnson, C. I., & Mayer, R. E. (2010). Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior, 26*, 1246–1252.

Kawasaki, M. (2010). Learning to solve mathematics problems: The impact of incorrect solutions in fifth grade peers' presentations. *Japanese Journal of Developmental Psychology, 21*(1), 12–22.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed). New York: Cambridge University Press.

Mayer, R. E. (2011). *Applying the science of learning*. New York: Pearson.

Mayer, R. E., Sims, V., & Tajika, H. (1995). A comparison of how textbooks teach mathematical problem solving in Japan and the United States. *American Educational Research Journal, 32*, 443–460.

McLaren, B. M., & Isotani, S. (2011). When is it best to learn with all worked examples? In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education (AIED-2011). Lecture Notes in Computer Science* (pp. 222–229). Berlin: Springer.

McLaren, B. M., Lim, S., & Koedinger, K. R. (2008). When and how often should worked examples be given to students? New results and a summary of the current state of research. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

Moreno, R., & Park, B. (2010). Cognitive load theory: Historical development and relation to other theories. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 9–28). Cambridge: Cambridge University Press.

Putt, I. J. (1995). Preservice teachers ordering of decimal numbers: When more is smaller and less is larger! *Focus on Learning Problems in Mathematics, 17*(3), 1–15.

Renkl, A., & Atkinson, R. K. (2010). Learning from worked-out examples and problem solving. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 91–108). Cambridge: Cambridge University Press.

Renkl, A. (2005). The worked-out examples principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 229–245). New York: Cambridge University Press.

Renkl, A. (2011). Instruction based on examples. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction*. New York: Routledge.

Resnick, L. B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. (1989). Conceptual bases of arithmetic errors: The case of decimal fractions. *Journal for Research in Mathematics Education, 20*(1), 8–27.

Sackur-Grisvard, C., & Léonard, F. (1985). Intermediate cognitive organizations in the process of learning a mathematical concept: The order of positive decimal numbers. *Cognition and Instruction, 2*, 157–174.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs*. Washington, DC: American Educational Research Association.

Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31–58). Cambridge, NY: Cambridge University.

Stacey, K., Helme, S., Steinle, V., Baturo, A., Irwin, K., & Bana, J. (2001). Preservice teachers' knowledge of difficulties in decimal numeration. *Journal of Mathematics Teacher Education, 4*(3), 205–225.

Stacey, K., Sonenberg, E., Nicholson, A., Boneh, T., & Steinle, V. (2003). A teacher model exploiting cognitive conflict driven by a Bayesian network. In P. Brusilovsky, A. T. Corbett, & F. D. Rosis (Eds.), *User modeling 2003: Proceedings of the ninth international conference* (pp. 352–362). PA, USA: Johnstown.

Sweller, J. (1999). *Instructional design in technical areas*. Camberwell, Australia: ACER Press.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer.

Tsovaltzi, D., Melis, E., McLaren, B. M., Meyer, A.-K., Dietrich, M., & Goguadze, G. (2010). Learning from erroneous examples: When and how do students benefit from them? *Proceedings of the European conference on technology enhanced learning, LNCS* (vol. 6383). Heidelberg: Springer.

Van den Broek, P., & Kendeou, P. (2008). Cognitive processes in comprehension of science texts: The role of co-activation in confronting misconceptions. *Applied Cognitive Psychology, 22*, 335–351. http://dx.doi.org/10.1002/acp.1418.

Zhu, X., & Simon, H. A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction, 4*(3), 137–166.