ORIGINAL RESEARCH

The effect of worked examples on student learning and error anticipation in algebra

Christina Areizaga Barbieri[1] · Julie L. Booth[2] · Kreshnik N. Begolli[3] ·
Nicholas McCann[2]

Abstract

The present study examines the effectiveness of incorporating worked examples with prompts for self-explanation into a middle school math textbook. Algebra 1 students ($N=75$) completed an equation-solving unit with textbooks either containing the original practice problems or in which a portion of those problems were converted into a combination of correct, incorrect, and incomplete examples. Students completed pre- and posttest measures of algebraic feature knowledge, equation-solving skills, and error anticipation. Example-based textbook assignments increased students' equation-solving skills and their ability to anticipate errors one might make when solving problems. Differences in students' anticipation of various types of errors are also examined. Error anticipation, a particular form of *negative knowledge*, is a potentially important skill that relates to algebraic feature knowledge and equation-solving skills.

Keywords Worked examples · Algebra learning · Mathematics education · Error anticipation · Negative knowledge · Algebraic thinking

Extensive research in cognitive science has demonstrated learning benefits of confronting errors (e.g., Metcalfe, 2017) in various ways (e.g., error detection, generation, correction, explanation) and from various sources (e.g., one's own errors, others' errors, errors made by a fictitious person). Negative information may be given more weight and elicit deeper levels of cognitive processing from students (Peeters & Czapinski, 1990). Yet education research demonstrates that errors are commonly avoided in mathematics classrooms (Stevenson & Stigler, 1994; Tulis, 2013). When errors *are* addressed in mathematics classrooms, the most common method used is one in which the student who made the error is not asked to correct it themselves. Rather, the error is fielded to another student in the classroom, typically one that the teacher knows will have the correct answer, and *that*

✉ Christina Areizaga Barbieri
  barbieri@udel.edu

[1] School of Education, University of Delaware, 113 Willard Hall Education Building, Newark, DE 19716, USA

[2] College of Education and Human Development, Temple University, Philadelphia, PA, USA

[3] School of Education, University of California, Irvine, Irvine, CA, USA

student is asked to correct the error. This approach is most typical in mathematics class-rooms (Tulis, 2013). Oser and Spychiger (2005) refer to this approach as the *Bermuda triangle of error correction* because presumably the opportunity for the student who made the initial error to learn from it is lost. Tulis (2013) also noted that discouraging or embarrassing remarks regarding errors made by students are more often heard in mathematics classrooms than other subject areas.

Though many scholars in mathematics education note that discussing errors in the classroom can provide prime opportunities for deep learning and refinement of principled knowledge (Borasi, 1994), systematic research demonstrates that this approach does not seem to be common practice in mathematics classrooms. In the United States, teachers tend to shy away from talking about errors (Lannin et al., 2006) at least in part due to the fear that their students will adopt the errors in their own problem solving (Santagata, 2004). Compared to other subject areas, student errors are the *most* common in mathematics classrooms but the least often discussed (Tulis, 2013). This finding coupled with the consideration that high school mathematics proficiency is generally poor (Castle, 2014) and struggles in algebra in particular are commonplace (Cangelosi et al., 2013) highlight the need to consider the potentially great learning opportunities that are lost when errors are ignored or glossed over. Although these findings highlight a general classroom context of error avoidance, particularly in mathematics, the literature on learning from errors reveals that they can be a promising learning tool.

Overoye and Storm (2015) review literature demonstrating that learning tasks are particularly beneficial when they induce uncertainty in students and provoke them to attempt to resolve this uncertainty. They explain that tasks that provoke uncertainty fall in two general categories: uncertainty through inquiry and uncertainty through contradiction. The current study focuses on a method that relies on both inquiry and contradiction. Inquiry-based uncertainty requires learners to produce or generate information and involves tasks such as testing, problem-solving, and interrogative questioning (i.e., asking students questions while they study, such as in self-explanation). Testing (e.g., Karpicke & Roediger, 2008), interrogative questioning (e.g., Dunlosky et al., 2013), and generation (e.g., Bertsch et al., 2007) have each been shown to improve learning. In fact, error generation (i.e., making a mistake on an assessment) leads to improvements in memory when accompanied with corrective feedback (Pashler et al., 2007).

Contradiction-based uncertainty involves tasks that present information which is contradictory to the learners' current knowledge and reveal a mismatch between what the student believes is accurate and what is truly accurate. Making mistakes and confusion induce contradiction-based uncertainty. Testing, generation, and interrogative questioning can consequentially reveal contradictions but this experience can also be purposefully provoked by creating confusion (e.g., D'Mello & Graesser, 2014) and presenting students with discrepant events (e.g., Gorsky & Finegold, 1994). Though none of the aforementioned studies addressed the specific skills of algebraic feature knowledge or negative knowledge, the findings do address a range of skills that are conceptually-based (e.g., explanations) and procedurally-based (e.g., multiplication). This suggests that the role of uncertainty may be applicable to learning in general for a range of outcomes. We focus here on a method that is particularly relevant to mathematics that leads to uncertainty both through inquiry and through contradiction: studying and explaining mathematical errors.

Two established theories provide unique yet complementary explanations for how studying errors can be useful for learning. Ohlsson's (1996) theory suggests that explaining why an error is wrong can help learners identify the particular features of the problem that make the solution incorrect. This focus on what makes a solution incorrect

can lead to refinement of problem-solving skills and remediation of misconceptions. In addition, Siegler's (1996) *overlapping waves theory* maintains that individuals know and use a variety of (correct and incorrect) strategies for solving problems, and those strategies compete for use each time a problem is encountered. Studying errors can be an effective way of helping learners accept that those particular strategies are wrong and prompt them to construct and strengthen other, correct strategies. Consistent with these theories, recent research has found that studying and explaining errors is indeed beneficial to learning. For example, having students think about and correct their own errors can lead to greater engagement and improved problem-solving skill (Cherepinsky, 2011; Henderson & Harper, 2009).

Studying the errors made by others may be even more effective than studying one's own errors (Yerushalmi & Polingher, 2006), in part because this strategy exposes students to multiple perspectives other than their own (Siegler & Chen, 2008). Further, limitations in prior knowledge can make error detection challenging. For example, Groβe and Renkl (2007) demonstrated that students with low prior knowledge of the learning material benefitted less from instructional materials that required them to find errors made within a problem than those with high prior knowledge. There is some evidence to suggest that teachers often do not believe their lower performing students would benefit from discussing a common misconception, even when compared to a correct way of solving the problem (Begolli et al., 2018). They expect that this is something most suitable for their higher performing students. However, recent work reveals particular benefits of studying errors for those with low prior knowledge in materials that highlight errors (Barbieri & Booth, 2016, 2020). Considering these findings, one may expect that a struggling student is less likely to benefit from instructional tasks requiring them to detect their own error than one the requires them to study and explain highlighted errors that are strategically designed to target an important mathematical misunderstanding as represented in other students' work. Highlighting an error for the student may help to alleviate the students' dependence on prior knowledge, as suggested by Groβe and Renkl (2007).

It may also be useful to consider whether learners should study real students' errors or those made by a fictitious student. Yerushalmi and Polingher (2006) conducted a classroom study with high school physics students in an attempt to assess the effectiveness of these two methods for guiding students in evaluating errors. One group was presented with fictional students' statements and was asked to identify mistakes in fictional students' work, explain why they were incorrect, and correct the mistakes. Another group identified, explained, and corrected errors made by students on previous exams. Most students correctly identified errors in fictional students' work whereas very few did so for actual student work. Teacher-constructed or researcher-constructed errors presented as those made by a fictitious student may thus be more effective at leveraging errors for learning. They can be strategically formulated to address common student misconceptions rather than errors that may not represent an important misunderstanding but rather an error made due to carelessness. As such, the current study focuses on learning from errors in a particular context: Examples of fictitious student's work.

The use of incorrect worked examples, either alone or in combination with correct examples, is one method for incorporating learning from errors into mathematics classrooms. Incorrect worked examples are worked out problem solutions that display a common error made by a fictitious student. When using incorrect examples, students are typically asked to study and explain the error, and to sometimes also correct it, before moving on to complete a similar practice problem. These examples have been found to improve mathematics learning as a form of practice with instruction when used alone (Adams et al.,

2014; Barbieri & Booth, 2020) or in combination with correct examples (Begolli & Richland, 2016; Booth et al. 2013; Durkin & Rittle-Johnson, 2012).

There is a large body of work demonstrating learning benefits of correct worked examples alone compared to problem-solving practice on conceptually and procedurally-based skills such as algebraic equation solving, knowledge of mathematical principles and features (Carroll, 1994; Pol et al., 2009; Ward & Sweller, 1990). Further, the use of worked examples in general is recommended by the *Institute of Education Sciences* (IES) Practice Guide (Pashler et al., 2007). More recent work focuses on the benefits of *a combination* of correct and incorrect examples as an effective replacement of open-ended problem-solving practice. Booth et al. (2015) found that a series of sequentially-presented correct and incorrect examples improved students' conceptual knowledge more than traditional problem-solving practice, especially for those with low prior conceptual knowledge. This study (Booth et al., 2015) did not include conditions with correct and incorrect examples alone so it is not clear whether these specific benefits of the combination of the two example types are what was driving the effects or whether one type may be prompting the learning benefits. However, other work points to the inclusion of incorrect examples being particularly effective in worked examples practice materials. Durkin and Rittle-Johnson (2012) found that, compared to correct examples alone, comparison of correct *and* incorrect examples has been found to reduce mathematical misconceptions, increase classroom discussion of correct concepts, and improve students' procedural skill and conceptual understanding (Durkin & Rittle-Johnson, 2012). Further, when used alone, incorrect examples have been found to improve algebraic equation-solving compared to a problem-solving control (Barbieri & Booth, 2020), reduce encoding errors of algebraic equations compared to correct examples alone (Booth et al., 2013) and increase *negative knowledge* of fractions compared to correct examples alone (Heemsoth & Heinze, 2014).

Negative knowledge, or the knowledge of incorrect strategies and concepts (i.e., what *doesn't* work) has been found to be an important skill for problem-solving (Gartmeier et al., 2008). Oser and Spychiger (2005) define negative knowledge as, "what something is not (in contrast to what it is), and how something does *not* work (in contrast to how it works), which strategies do *not* lead to the solution of complex problems (in contrast to those, that do so), and why certain connections do *not* add up (in contrast to why they add up)". Most work on negative knowledge has focused on the domain of workplace performance and expertise in professional fields (Ericsson et al., 2006; Gruber & Palonen, 2007). However, the theory and implied learning mechanisms are highly relevant to mathematics, as suggested by some recent work (e.g., Heemsoth & Heinze, 2014). Negative knowledge improves problem-solving success by increasing one's certainty in the correct procedure while allowing one to simultaneously avoid incorrect procedures. Ohlsson's theory of learning from errors emphasized the importance of identifying problem features that make a solution incorrect. Though Ohlsson did not specifically term this negative knowledge, this is indeed a form of such knowledge. As previously noted, Siegler (1996, 2002) emphasized the importance of studying errors as a method for encouraging learners to accept the strategies shown as incorrect. This, too, is a form of negative knowledge though not termed as such. Research on methods of increasing negative knowledge is limited, with only one published study to date demonstrating improvements in negative knowledge through studying incorrect worked examples alone (Heemsoth & Heinze, 2014). In this study, which is the only published worked example study to date that directly measured negative knowledge, incorrect worked examples alone were more effective at improving negative knowledge of fraction computations and modeling than correct examples alone. A combination condition (correct and incorrect examples) was not employed. No published study to date

examines effects of a combination of correct and incorrect examples on negative knowledge. Considering the contrastive nature of negative knowledge (i.e., what *doesn't* work compared to what *does*), as defined by Oser and Spychiger (2005), and the greater implicit opportunity for comparison when presented with both correct and incorrect examples as opposed to incorrect examples alone, it is possible that a combination of correct and incorrect examples will also significantly improve negative knowledge in comparison to a problem-solving control by allowing students to anticipate errors likely to be made in specific problem types.

We acknowledge that there is a body of literature focusing on error management in organizational contexts with a specific focus on action errors (e.g., Frese & Keith, 2015). Action errors can be defined as accidental changes from plans or goals, and are typically studied in work settings (Frese & Zapf, 1994; van Dyck et al., 2005). For example, action errors in the medical field (e.g., administering the wrong medication to a patient) are one specific type that is studied in relation to error management. The current study is not situated within an organizational setting and as such does not focus on action errors. In addition to the obvious differences in field, error management is distinct from negative knowledge in that negative knowledge focuses on incorrect strategies and concpets whereas error management is an umbrella term for the way that employees perceive, cope with, and learn from their own errors (Keith & Frese, 2008). However, both areas of work share the same view that errors, whether action errors in the workplace or problem-solving errors in an equation, may be used to improve learning and performance.

The current study

The current study has three aims. First, we aim to replicate prior work that demonstrates procedural and conceptual improvements of correct and incorrect worked examples compared to problem-solving controls (Barbieri & Booth, 2016, 2020; Lee & Chen, 2015; Retnowati et al., 2010). We hypothesize that our correct and incorrect examples condition will outperform the problem-solving control on both our conceptual (i.e., algebraic feature knowledge) and procedural (i.e., equation-solving) measures. That is, we expect to replicate previous findings showing that a combination of correct and incorrect worked examples supplemented with written self-explanation prompts are effective learning tools for students learning algebra (Booth et al., 2015). Worked examples are thought to be a more effective form of practice than problem-solving alone because, rather than having to use attentional and working memory processes to figure out the correct solution, learners can devote their cognitive resources to understanding the reasoning behind the steps taken in the example to then be relied upon in future problems. In the present study, correct, incorrect, and partially completed (i.e., faded) examples were simultaneously incorporated into a commonly used practice workbook to be used as the experimental manipulation, compared to the original practice workbook. The errors we focus on within incorrect worked examples are those that students commonly make when learning the material. That is, we target errors that may be indicative of a misunderstanding or misconceptions of a conceptual feature or strategy needed to complete the problem type, rather than simple arithmetic errors. For example, concluding that 5 is the sum of $3+3$ is the type of careless error that we avoid targeting as errors like this are unrelated to algebra learning (Booth et al., 2014).

We hypothesize that the worked examples-supplemented practice workbook will be more effective than the original problem-solving practice workbook. We investigate differences in workbook effectiveness for more conceptually-focused knowledge vs. more computationally-focused skills such as solving multi-step equations. Our conceptually-focused outcome measure is of a specific understanding within the broader umbrella term of conceptual understanding: algebraic feature knowledge, which refers to understanding the meaning of specific features within algebraic equations such as the equals sign (e.g., McNeil, 2014).

The second purpose of the current study is to investigate whether students' ability to *anticipate* the types of errors students might make when solving equations is related to their algebraic feature knowledge and equation-solving skills in algebra. We conceptualize the skill to anticipate potential errors for specific problem types as one form of negative knowledge. A prior study on the effect of incorrect examples on negative knowledge employed a multiple-choice test which asked students to indicate which step taken was incorrect in a sample problem (Heemsoth & Heinze, 2014). In the current study, we use open-ended questions to assess students' ability to anticipate errors most likely to be made by students solving an algebraic question. This differs from the measure used by Heemsoth and Heinze (2014) by requiring students to anticipate the error themselves as well as use some judgement on what the most common error would be. To gauge students' understanding of the material, reported errors are then categorized as problem-specific or general. Students' ability to report problem-specific errors was considered to be an indication of their deeper understanding of the material. Thus, rather than simply marking off which step is incorrect in a problem already solved as done in prior work, we provide an added dimension to our measure of negative knowledge by focusing on whether students anticipate *problem-specific* errors. As our error anticipation measure is a new method for measuring negative knowledge and negative knowledge is a relatively new construct to consider in relation to core mathematical competencies, we explore whether this measure is related to more conceptually- (i.e., feature knowledge) and computationally-based (i.e., equation solving) skills, which are each critical in mathematics (NMAP, 2008). Due to the sparse literature on negative knowledge we take an exploratory approach to this second aim and thus, while we do expect negative knowledge to be related to algebra knowledge generally, we do not have specific hypotheses of whether our error anticipation measure would be more strongly related to feature knowledge or equation-solving.

The third purpose of this study was to examine whether a combination of correct and incorrect examples, presented sequentially, increases students' ability to anticipate errors in comparison to a problem-solving control. We hypothesize that our correct and incorrect examples condition will significantly improve students' ability to anticipate the types of errors others might make when solving equations compared to the problem-solving control. We expect that the experience of explaining a combination of correct and incorrect examples which highlight challenging problem features could increase attention to errors to a greater extent than the experience of solving practice problems within the control condition. This heightened attention to errors may then translate to improvements in error anticipation over the course of the study. Though learners complete individual problem-solving practice with the goal of correctness (in solution and procedure), this experience is unlikely, without immediate feedback, to prompt students to consider or anticipate common errors to the same extent that direct exposure to errors with explanation prompts. Heemsoth and Heinze (2014) found that incorrect worked examples alone were more effective at improving negative knowledge of fraction computations and modeling than correct examples alone. Considering that negative knowledge is contrastive in nature (i.e., what

doesn't work as opposed to what *does* work), we expect that our combination of correct and incorrect examples will lead to greater improvements in error anticipation, a specific form of negative knowledge, than the problem-solving control.

Methods

Participants

Seventy-five 8th grade Algebra I students ($n = 37$ Example condition; $n = 38$ Control condition) from an inner-ring suburban middle school in the Midwestern United States participated in the study (55% female, 45% male; 59% Black, 21% White, 15% American Indian/Alaskan, 4% Asian, and 1% classified as other ethnicities). Age was not provided by the district, but US 8th graders are typically between 13 and 14 years old. All Algebra I classes utilized the *Connected Mathematics Project 2* Curriculum (CMP2; Lappan et al., 2006). CMP2 includes rich, problem-based investigations during classroom lessons, and provides a variety of practice problems for students to solve afterward as classwork and/or homework. This study took place during the *Say It With Symbols* unit, which focuses on understanding symbols in algebraic equations. Power analyses using G*Power (Faul et al., 2007) revealed that our sample size of 75 provided power of 0.95 to detect a medium effect on our outcome measures.

Materials

Workbooks

The current study used one textbook within the *Connected Mathematics 2* curriculum titled 'Say It With Symbols', which focuses mainly on developing understanding of equivalent expressions and equations. Students are expected to demonstrate skills such as "model situations with symbolic statements", "determine if different symbolic expressions are mathematically equivalent", "recognize how and when to use symbols to display relationships, generations, and proofs." The *Connected Mathematics* 2 curriculum (Lappan et al., 2006) is problem-centered and promotes inquiry-based instruction, published by Pearson and aligned with Common Core State Standards-Mathematics (CCSS-M). CMP emphasizes connections between mathematical ideas and real-world applications. Each textbook includes several multi-step *Investigations* or problems to solve. For example, in Investigation 1 in the textbook used in the current study (i.e., Say It With Symbols), students focus on equivalent expressions.

CMP materials prompt instructors to encourage classroom discussion of students' different strategies. CMP lessons are structured as follows: first, the instructor ensures students understand the mathematical nature of the *Investigation* or problem. Students are then asked to invent and share different strategies for finding the solution to the problem, with teacher guidance. After sharing invented strategies, teachers close the lesson with a summary discussion of the strategies and ideas presented and the mathematical connections made. Each *Investigation* ends with a series of ACE (Applications-Connections-Extensions) problems that teachers assign either for homework or in-class practice. ACE problems are termed as such because they are *transfer* problems in that

they are similar, but *not* identical, to those presented during the lesson and require students to connect and extend their new knowledge to different problems and contexts.

The experimental modification of the workbook used in the current study was made solely to ACE homework problems, not to the *Investigations* themselves which are held constant across conditions. Thus, the control and experimental workbooks were identical with the exception of approximately 50% (129) of ACE problems being replaced with worked examples. Sample problems can be seen in Fig. 1 in their original form (Control condition) and in worked example form (Example condition).

Correct worked examples	**26.** $5 - 6(x + 10) - 4$ Dorothy solved this problem in a correct way. Look at her work, and then answer the question below. $5 - 6(x + 10) - 4$ $5 - 6x + (-60) - 4$ $5 - 4 - 60 - 6x$ $-59 - 6x$ Why didn't Dorothy multiply -4 by -6?
Incorrect worked example	**c.** Suppose the student council sells 100 T-shirts. What is the profit? Eric found the income and expenses correctly, but he made a mistake when he tried to calculate the profit. Look at his work, and then answer the question below. $I = 12(100) + 150$ $I = 1200 + 150$ $I = 1350$ $E = 250 + 4.25(100)$ $E = 250 + 425$ $E = 675$ $P = 675 - 1350$ $P = -675$ ✗ What does Eric think is the formula for profit? How could he have figured out from his answer that he didn't compute it correctly?
Faded worked example	**b.** $(x - 5)(x + 5)$ Raven started to solve this problem, but she got stuck. Fill in the correct operations in the circles and the correct terms in the squares. Then finish solving the problem. $(x - 5)(x + 5)$ $x^2 \; ● ■ ● ■ - 25$
Note: The Control condition solved the same problems but without the examples.

Fig. 1 Experimental workbook examples

Three types of worked examples simultaneously replaced a portion of the practice ACE problems in the unrevised textbook: (1) Correct examples, in which a fictitious student's correct problem-solving was displayed, (2) Incorrect examples, in which a fictitious student's incorrect problem-solving was displayed and marked as such, and (3) Faded examples, in which a fictitious student's correct but incomplete problem-solving was displayed and students were prompted to complete the problem. Many of these examples was paired with written self-explanation prompts to aid their usage and focus students' attention on key features of the example, connecting procedures shown to underlying concepts.

Algebraic feature knowledge

To examine students' algebraic feature knowledge, we used 21 items that focused on the meanings of different terms in an equation, identification of equivalent expressions, and categorization of functions as linear, quadratic, or exponential. The percentage of these items answered correctly was computed for each student at pretest and at posttest. Internal consistency was high at posttest ($\alpha = 0.879$). Sample items are displayed in Fig. 2.

Solving multi-step equations

To examine equation-solving skill, we used nine items which asked students to solve multi-step equations, simplify expressions using the distributive property, and evaluate formulas
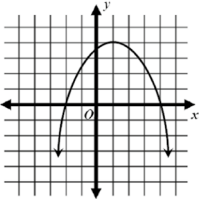
		What type of function is it?		
a.	$y = x^2 + 6$	Linear	Quadratic	Exponential
b.	$\begin{array}{c c c c c} x & 2 & 4 & 6 & 8 \\ y & 1 & 4 & 7 & 10 \end{array}$	Linear	Quadratic	Exponential
c.	$y = 2^x + 6$	Linear	Quadratic	Exponential
e.	$\begin{array}{c c c c c} x & 3 & 4 & 5 & 6 \\ y & 11 & 18 & 27 & 38 \end{array}$	Linear	Quadratic	Exponential
f.	$y = 2x + 6$	Linear	Quadratic	Exponential
g.		Linear	Quadratic	Exponential

Fig. 2 Sample algebraic feature knowledge items

Table 1 Solving multi-step equation items

1. A pump is used to empty a swimming pool. The equation $w = -200(t - 6)$ represents the gallons of water w that remain in the pool t hours after pumping starts
(a) How many gallons of water are pumped out of the pool each hour?
(b) How long will it take to empty the pool?
(c) How many gallons of water are in the pool at the start?

Table 2 Sample problem-specific errors anticipated for $5x - 2 = 8$

Problem-specific errors	Sample student responses
Variables	"Thinking that the x is a multiplication sign and that it is 5*-2=8"
Like terms	"Ignore the x and do $5 - 2$."
Negatives	"Subtracting 2 on both sides instead of adding 2."
Equals sign	"They might only do a step on one side and forget to do it with the other side (unbalanced equation)"
Operations	"Forget to combine, add, or subtract the numbers"

at given values. The percentage of these items answered correctly was computed for each student at pretest and at posttest. Internal consistency was high at posttest ($\alpha = 0.816$). Sample items are displayed in Table 1.

Error anticipation

To evaluate ability to anticipate errors that others might make when solving multi-step equations, we utilized one item which asked students what mistakes they thought a seventh-grader might make in solving the equation $5x - 2 = 8$. In total, students were asked to identify two potential errors. Student responses were coded first in terms of whether or not the provided responses were problem-specific or general, and then by the type of error referenced if they were problem-specific. Problem-specific errors included mistakes involving variables (e.g., handle the coefficient separately from the variable), like terms (e.g., subtract 2 from 5x), negative signs (e.g., subtract 2 from both sides instead of adding 2), equals sign (e.g., perform an operation to one side and not the other), and operations (e.g., adding two numbers instead of multiplying). General errors typically included errors about general problem-solving strategies or that were sometimes nonsensical (e.g., "putting random numbers"; "not bringing the negative down"; "the number might equal something else like the 7th grader can add it wrong". For each type of error, students were scored in terms of whether at least one of their responses fit in that category. Table 2 provides an example of the coding scheme. All responses were coded by an independent coder.

Procedure

We conducted a quasi-experimental study (i.e., students in pre-existing groups/classrooms assigned to condition) in real-world Algebra I classes using CMP2 Curriculum (CMP2; Lappan et al., 2006). Prior to beginning their unit, all students took the paper-and-pencil pretest. The test included three subscales or types of items: algebraic feature knowledge,

solving multi-step equations, and error anticipation. Middle school students were assigned to the experimental Example and Control conditions according to their rostered section of Algebra I. The four sections were taught by two teachers, each having one treatment and one control class. Assignment at the classroom level (as opposed to within-class) was necessary as students needed to use the workbooks either in class and for homework over an extended period of time to prevent diffusion of treatment and avoid contamination effects. As the participating teachers were asked to use the workbooks as they normally would and teachers vary in their instructional style and pace, duration of the study varied by teacher (but not by condition). However, all classrooms completed the study over the course of approximately two months from pre-to post-test in the early spring of the school year.

Instruction for the Example and Control classrooms was kept constant within teacher (e.g., Teacher A provided the same lesson to both her Example class and her Control class). All instructors used the *Connected Mathematics 2 Curriculum* in their classrooms. The only difference between conditions occurred when students were to work on their practice problems in their workbooks, students in the Example classes were given an adapted version of the *Say It With Symbols* book. In the adapted book, approximately 50% of the practice problems were replaced with a correct, incorrect, or partial (i.e., faded) example of a solution to that problem. Participating teachers were asked to use the textbooks as they normally would in their classroom and to assign problems as they would normally (for in-class or homework practice after instruction). Teachers could assign as many or as few practice problems as they desired as long as the same items were assigned to both their Example and Control groups. We provide a description of how this curriculum is designed to be used in the Materials section. When the unit was complete, students took the paper-and-pencil posttest, which was identical to the pretest.

Results

Descriptive statistics

First, to establish baseline balance between conditions at pretest, we compared conditions on key study variables and found no significant differences. Descriptive statistics by condition are presented in Table 3 and demonstrate equivalence. Participants were nested in four classrooms. Intra-class correlations (ICC) were calculated on posttest data and were

Table 3 Tests of equivalency between conditions at pretest

	Total sample		Worked examples		Control		Difference test
	M	*SD*	*M*	*SD*	*M*	*SD*	
Algebraic feature knowledge	51.00	14.03	53.24	12.48	48.82	15.23	$F[1,73]=1.89, p=0.173$, *ns*
Solving multi-step equations	31.41	22.96	30.97	22.31	31.84	23.88	$F[1,73]=0.026, p=0.871$, *ns*
Anticipating problem-specific errors	0.95	0.82	0.78	0.79	1.11	0.83	$F[1,73]=2.96, p=0.090$, *ns*
Female	54.67		48.65		60.52		$\chi^2[1]=1.07, p=0.302$, *ns*
Underrepresented racial or ethnic minority (URM)	74.32		72.97		75.68		$\chi^2[1]=0.071, p=0.790$, *ns*

generally low. The ICC for posttest equation-solving scores by classroom was 0.0612 and 0.0223 for posttest feature knowledge scores. Thus, due to the low dependency of posttest scores on cluster and the few number of clusters, all further analyses were conducted at the student level.

Plan of analysis

The current study had three main aims. The first aim was to demonstrate that our worked examples condition showed greater procedural (i.e., equation-solving) and conceptual (i.e., algebraic feature knowledge) improvements than the problem-solving control. We address this through a series of split plot Time (pretest, posttest)×Condition (Worked examples workbook, Control workbook) analyses of variance on the two aforementioned outcome measures (i.e., Feature knowledge and Equation-solving).

The second aim was to investigate whether students' ability to anticipate the types of errors students might make when solving equations is related to their algebra skills. As this is a new area of research we did not pose specific hypotheses on whether and if error anticipation would be more strongly related to equation-solving or feature knowledge. We take an exploratory approach in addressing this aim through a series of correlations. We present correlations using this measure in Table 4.

The third aim of the current study was to examine whether the worked examples condition, which includes a combination of correct and incorrect examples, improves students' ability to anticipate errors significantly more than the problem-solving control. As the current study is a novel attempt at using error anticipation to measure negative knowledge, we address this aim using a series of mixed ANOVAs (Time by ANOVA) that compare improvements from pre- to posttest by condition on three different error anticipation outcome measures: (1) the number of *problem-specific* errors anticipated overall, (2) the number of *types* of *problem-specific* errors anticipated overall (i.e., how variable the anticipated *problem-specific* errors were), and (3) the number of *problem-specific* errors anticipated within each error category. That is, we first assess whether the Example workbooks led students to anticipate more problem-specific errors overall. However, this is a measure of quantity of errors anticipated, not type. Thus, students could anticipate two unique errors

Table 4 Mean number of errors anticipated by type

Error category	Example				Control			
	Pretest		Posttest		Pretest		Posttest	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
General errors	0.46	0.73	0.08	0.36	0.29	0.57	0.16	0.49
Problem-specific errors	0.78	0.79	1.57	0.77	1.11	0.83	1.45	0.83
Problem-specific errors by type								
Variable	0.38	0.59	0.43	0.60	0.47	0.60	0.32	0.53
Terms	0.14	0.42	0.43	0.55	0.13	0.34	0.45	0.65
Negatives	0.16	0.37	0.24	0.43	0.21	0.41	0.32	0.53
Equals sign	0.05	0.23	0.08	0.28	0.11	0.39	0.05	0.23
Operations	0.03	0.16	0.19	0.40	0.13	0.34	0.18	0.39
Other Problem-specific	0.03	0.16	0.19	0.46	0.05	0.23	0.13	0.34

that are of the same type of error (e.g., two different errors that both involve like terms). Therefore, we also assess whether the Example workbooks led students to anticipate more types of errors than the Control workbooks by evaluating differences in the quantity of categories that the errors anticipated were nested in. Last, we assess whether there were differences in quantity of errors anticipated within each of the six error categories.

Correction for multiple tests

Because the current study was proposed to include planned multiple comparisons within the mixed ANOVAs noted above, corrections are not necessary (Armstrong, 2014; Perneger, 1998). Additionally, exploratory analyses that do not test specific hypotheses but rather provide suggestions for future work do not require corrections. Thus, we make no corrections to interpreting significance levels for our exploratory goals addressed in Aim 2. Aim 1 includes one parsimonious model to assess improvements in feature knowledge and equation-solving skills simultaneously so no corrections are needed. However, we have opted to take a more conservative approach in interpreting our findings for Aim 3 which include several analyses on three different yet related error anticipation outcome measures. We employed Benjamini and Hochberg's (1995) correction procedure for multiple tests to interpret the three time×condition interactions within the mixed ANCOVAs that assess differential improvements by condition on (1) the number of problem-specific errors made overall, (2) the number of type of problem-specific errors made, and (3) the error types made. This approach decreases the False Discovery Rate (FDR), or the expected proportion of the rejected null hypotheses which are incorrectly rejected. Unlike the classic Bonferroni correction (Bonferroni, 1936), which adjusts the alpha level once to use for all comparisons, the BH correction adjusts the alpha level down to an increasingly conservative cutoff, using an ordered set of obtained m p values, only after each statistically significant result and not after nonsignificant results. After finding the largest p value that satisfies $p_k \leq \frac{k}{m}\alpha$, all tests with smaller p values are declared significant. BH corrections were applied to the three time×condition interactions within the mixed ANOVAs with adjusted alpha levels of 0.05, 0.033, and 0.017. In addition, we adopt the most conservative approach—the classic Bonferroni correction—to interpret follow-up tests that serve to explicate the significant interactions found within Aim 3 with a consistent p-value of 0.017.

Research aim 1: Differential improvements by condition on equation-solving and feature knowledge

To examine the effectiveness of the Examples workbooks for improving students' algebraic feature knowledge and equation-solving skill, we conducted a 2 (Condition: Example vs. Control)×2 (Time: Pretest vs. Posttest)×2 (Measure: feature knowledge vs. equation-solving) mixed ANOVAs. The analysis yielded a main effect of time (F [1, 73]=23.50, $p < 0.001$, $\eta_p^2 = 0.24$), with students performing better at posttest ($M = 52\%$) than at pretest ($M = 41\%$). There was also a main effect of measure (F [1, 73]=52.97, $p < 0.001$, $\eta_p^2 = 0.42$), with students performing better on feature knowledge items ($M = 52\%$) than on equation-solving items ($M = 41\%$). There was a significant time by measure interaction, F (1, 73)=15.43, $p < 0.001$, $\eta_p^2 = 0.17$, with students improving more from pretest to posttest on equation-solving items (31% to 48%) than algebraic feature knowledge (51% to 55%). There was a significant interaction between time and condition, F (1, 73)=4.66, $p = 0.034$, $\eta_p^2 = 0.06$, revealing that students in the treatment group improved more from pretest to

posttest across measures (42% to 57%) than students in the control group (40% to 46%). The three-way interaction between time, measure, and condition did not reach statistical significance, $F(1, 73) = 3.35$, $p = 0.071$, $\eta_p^2 = 0.04$, suggesting that the differential improvements by condition did not vary by measure.

Research aim 2: Exploring the relationship between error anticipation and algebra performance

Although prior work has established the importance of equation-solving and feature knowledge for algebraic learning, the current study is one of the first to propose error *anticipation* as a useful skill. Therefore, it is important to examine whether students' error anticipation abilities were related to their feature knowledge and equation-solving skills. Correlations between error anticipation scores both at pretest and at posttest with the corresponding feature knowledge and equation-solving scores are presented in Table 5. Anticipating general errors at either pre- or posttest is not correlated with equation-solving or feature knowledge. However, both the overall quantity and number of types of problem-specific errors anticipated at pre- and posttest does significantly and positively correlate with equation-solving and feature knowledge scores. This indicates an important differentiation between anticipating generaland problem-specific errors and suggesting error anticipation as a potentially important skill for algebraic competency.

Research aim 3: Impact of worked examples on error anticipation skills

Three 2 (Condition) × 2 (Time) mixed ANOVAs were conducted on (1) the number of problem-specific errors anticipated overall, (2) the number of *types* of problem-specific errors anticipated overall, and (3) the number of problem-specific errors anticipated within each error category. First, to examine whether the Example workbooks led students to anticipate more problem-specific errors overall, we conducted a 2 (Condition: treatment vs. control) × 2 (Time: pretest vs. posttest) ANOVA on number of problem-specific errors anticipated. As explained in our Plan of Analysis, we adopt BH adjusted alpha levels of $p < 0.05$, 0.033, and 0.017 to interpret the results of the three Time × Condition interactions,

Table 5 Bivariate correlations between error anticipation, algebraic feature knowledge, and solving multi-step equations

	Pretest measures		Posttest measures	
	Algebraic feature knowledge	Solving multi-step equations	Algebraic feature knowledge	Solving multi-step equations
Number of Problem-specific Errors Anticipated	0.230*	0.325**	0.445**	0.421**
Number of General Errors Anticipated	0.032	− 0.222	− 0.112	− 0.116
Types of Problem-specific Errors Anticipated	0.240*	270*	0.403**	0.403**

** $p < 0.01$; * $p < 0.05$

starting with the first statistically significant result being compared to $p < 0.05$ and the next being compared to a more conservative cut-off of $p < 0.033$, and so on.

There was a significant main effect of time, $F (1, 73) = 25.21$, $p < 0.001$, $\eta_p^2 = 0.26$, with students anticipating more problem-specific errors at posttest ($M = 1.51$) than pretest ($M = 0.95$). The interaction between time and condition was not significant, $F (1, 73) = 3.88$, $p = 0.053$, $\eta_p^2 = 0.05$), with students in the Example condition showing similar improvements in their ability to anticipate problem-specific errors between pre- and posttest.

Then, to examine whether the Example condition led students to anticipate more *types* of problem-specific errors, we conducted a 2 (Condition: treatment vs. control) $\times$ 2 (Time: pretest vs. posttest) ANOVA on number of *types* of problem-specific errors anticipated. There was a significant main effect of time, $F (1, 73) = 23.92$, $p < 0.001$, $\eta_p^2 = 0.26$), with students anticipating more types of problem-specific errors at posttest ($M = 1.39$) than pretest ($M = 0.87$). There was a significant time by condition interaction, $F (1, 73) = 4.77$, $p = 0.032$, $\eta_p^2 = 0.06$), with students in the Example condition showing significantly more increases in the types of problem-specific errors anticipated between pre- ($M = 0.70$) and posttest ($M = 1.46$) than control (Pretest $M = 1.03$; Posttest $M = 1.32$) with a medium effect size ($g = 0.59$). This is displayed in Fig. 3.

To gain a better understanding of the types of errors being anticipated at pretest and posttest and how they differed over time as well as between condition, we conducted a follow-up 2 (Condition: treatment vs. control) $\times$ 2 (Time: pretest vs. posttest) $\times$ 6 (Error type: variable, like terms, negative signs, equals signs, operations, and other problem-specific errors) mixed MANOVA. We should note for the reader that the components of the analysis which collapsed across error type (i.e., main effect of time and the time by condition interaction) were mathematically identical to those in the first mixed ANOVA presented previously. Therefore, they are not repeated here. This particular analysis was conducted to assess the main effects and interactions involving *error type*; specifically, the main effect of error type, the interaction between error type and time, and the three-way interaction between error type, time, and condition. Thus, this is what is presented below.

There was a main effect of error type, $F (5, 69) = 10.57$, $p < 0.001$, $\eta_p^2 = 0.43$, demonstrating that there were significant differences in the number of error types anticipated

Fig. 3 Number of error types anticipated from pre- to posttest by condition
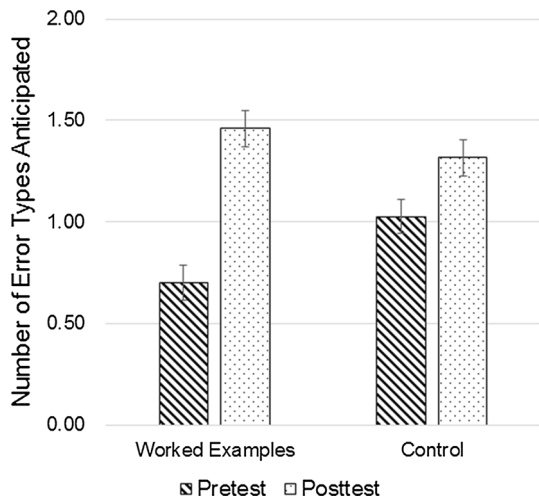
overall. Post-hoc tests with Bonferroni correction revealed that *variable*, *like terms*, and *negative sign* errors were more likely to be anticipated than *equals sign* and *other* problem-specific errors; variable errors were also more commonly anticipated than operations errors (see Fig. 4).

The analysis yielded a significant interaction between time and error category, F (5, 69) = 2.44, p = 0.043, η_p^2 = 0.15), that demonstrated that the number of errors anticipated differed from pre- to posttest by category type. Follow-up paired-sample t-tests with Bonferroni correction (critical value of $p < 0.017$ for three comparisons) revealed that the frequency of anticipating one type of errors increased from pretest to posttest for *Like terms* (12% to 39%; $t(74) = 4.37$, $p < 0.001$), but not for *operations* (8% to 19%; $t(74) = 2.04$, $p = 0.045$), or *other* problem-specific errors (4% to 15%; $t(74) = 2.38$, $p = 0.020$). The three-way interaction between time, error, and condition, was not significant, F (5, 69) = 0.258, $p = 0.934$, η_p^2 = 0.02), indicating that although the number of errors anticipated significantly differed over time by error type, and the number of errors anticipated overall differed slightly over time by condition, the number of errors anticipated by category did not significantly differ over time by condition. Therefore, for simplicity, Fig. 4 displays percentages of error types anticipated collapsed across conditions out of all errors anticipated at pre- and posttest. Table 4 displays descriptive information on the mean number of errors anticipated by error type at pre- and posttest by condition.

Discussion

Our first aim was to replicate prior work that demonstrates that studying a combination of correct and incorrect worked examples and answering written self-explanation prompts improves both conceptually (i.e., algebraic feature knowledge) and computationally-focused (solving multi-step equations) algebra competencies more than the problem-solving control. Algebra students in both conditions showed significant improvements in algebraic feature knowledge and skills in solving multi-step equations from pre- to posttest. These findings are somewhat consistent with prior work demonstrating improvements in algebra understanding on a composite measure of conceptual and procedural skill after working with example-based assignments (Booth et al., 2015). Therefore, it seems that supplementing a workbook that focuses primarily on underlying features (i.e., structure) of
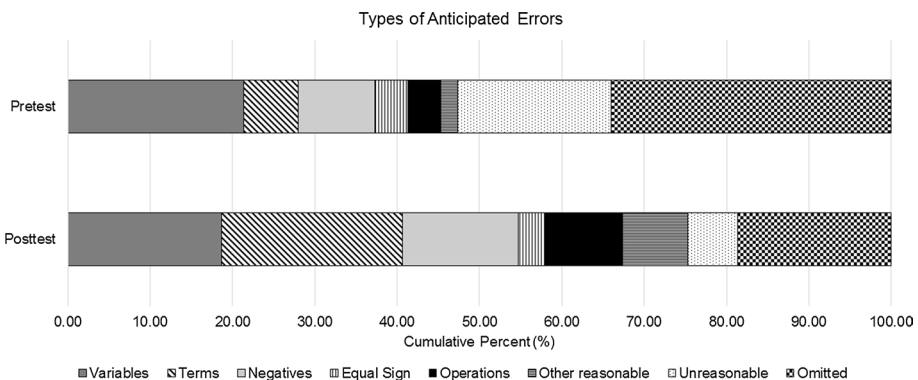

Fig. 4 Types of errors anticipated from pre to posttest

problems with correct and incorrect worked examples improved students' equation-solving skills as well as their learning of algebraic features more than the original (Control) workbook. Future work should explore the potentially interactive effects between curriculum focus and example-type used in practice.

The second aim of the current study was to assess whether a particular type of negative knowledge, one's ability to anticipate problem-specific errors that a student may make when solving a multi-step equation, was a potentially important skill relating to other algebra competencies (i.e., algebraic feature knowledge and equation-solving skills). We present preliminary findings that suggest that error anticipation may be an important skill for algebra competency. Students' ability to anticipate problem-specific errors (but not general errors) at both pre- or posttest is significantly and positively correlated with equation-solving and feature knowledge. More specifically, both the overall quantity of problem-specific errors anticipated as well as the variation of errors anticipated (i.e., different types of errors anticipated) correlate with equation-solving and feature knowledge scores. *Negative knowledge*, or the knowledge of incorrect strategies and concepts has been found to be an important skill for problem-solving (Gartmeier et al., 2008) but much of this work is in the domain of workplace performance (Ericsson et al., 2006; Gruber & Palonen, 2007). Though it is logical that knowing what *not* to do and what *doesn't* work would impact problem-solving in mathematics, and in algebra in particular, our study is the first to establish its significant relationship with algebra knowledge.

In addressing Aim 3, we also take this exploration a step further by considering whether working with a combination of correct and incorrect examples can hone this skill over the course of just one algebra unit. Algebra students in both conditions showed significant improvements in their ability to anticipate problem-specific errors that a hypothetical student may make when solving a multi-step equation from pre- to posttest. Students across conditions also showed improvements from pre- to posttest on how varied the errors that they anticipated were. That is, they were more likely to anticipate unique types of errors rather than errors within the same category. This was especially true for students who used the Example workbooks in that they showed significantly more variation in error types anticipated than those students exposed to the control workbooks. Using a distinct measure of negative knowledge relating to fractions, Heemsoth and Heinze (2014) found that incorrect examples in particular improved negative knowledge. Our example condition included both incorrect and correct examples. It is possible that only including incorrect examples may have bolstered the effects of examples on negative knowledge of algebra. Alternatively, it is possible that a single unit was not enough to bring about significant change and that an increased dosage of example-based learning may have brought about more substantial effects on negative knowledge. Future work should explore the impact of incorrect examples and correct examples in isolation on negative knowledge within algebra as well as consider the necessary dosage of examples needed to foster change. Longitudinal work can also establish whether *improvements* in negative knowledge lead to long-term improvements in algebra understanding.

The current study has several methodological strengths. First, the study demonstrates high ecological validity being conducted in real-world classrooms. We were able to detect effects of the experimental manipulation even amongst the noise that exists in everyday classrooms. It is possible that a laboratory experiment conducted to reach the same aims may have had even larger effects. Another strength was the equivalence between conditions on key study measures at pretest. This assures us that differences found between conditions in learning are not due to preexisting differences between conditions but rather due to our experimental manipulation. Additionally, our between class assignment ensures that there

was no carryover effect between students within the same classroom, as is often the risk when using within class assignment. Relatedly, to our surprise, the effects of cluster were negligible allowing us to focus on student-level analyses and confirming that class differences cannot explain our results. These features give us greater confidence that our findings reflect true effects of our example-based assignments.

Despite these strengths, the study has several limitations. Our sample size limited the complexity of analyses that could answer more nuanced questions, such as those relating to either mediating or moderating effects. We also did not have detailed information on how the assignments were used within the classroom. Methods for using assignments may impact their effectiveness in fostering learning and change. Future work can address these issues by using a larger sample and conducting classroom observations to help quantify some of the methods that may impact assignment effectiveness. Lastly, because our Example condition used a combination of correct and incorrect examples, it is unclear whether the effects of the Example condition are due to the combination of the two or whether one type of example drove the differences. Future work may address this by examining correct and incorrect examples in isolation and in combination to determine whether the effects of using both types of examples on learning are equivalent, additive, or multiplicative, especially with consideration to our new outcome of error anticipation. Using a computerized tutor Booth et al. (2013) compared the effects of correct and incorrect examples in isolation to the effects of their combination (as well as a problem-solving control) and found no differences between the effects of the three worked examples conditions on procedural or conceptual algebra knowledge. However, error anticipation was not measured in said study. Thus, varying conditions and outcomes make definitive conclusions difficult.

Scholarly significance

Results from the present study replicate and extend prior studies on the effectiveness of worked examples in mathematics learning by demonstrating that worked examples can be effectively incorporated into workbooks, leading to improved equation-solving and algebraic feature knowledge over problem-solving practice alone. In addition, practice containing worked examples, many of which prompt students to reflect on incorrect procedures, increased the likelihood that students could anticipate errors that others might make. If teachers want to make learning from errors a more prominent part of their classrooms but do not have access to—or time to create—relevant error-centered lessons, it may be desirable to have students think about potential errors on their own and reflect on why those anticipated errors are problematic. Introducing assignments with incorrect examples earlier in the process may train students to anticipate such errors on their own.

The present study also revealed differences in the types of errors students tend to anticipate. Errors dealing with variables, like terms, and negative signs were the most frequently anticipated across time points, and anticipation of like terms, operations, and other problem-specific errors were most likely to increase after students gained more knowledge about the content area. Interestingly, one of the types of highly-anticipated errors in the present study—those involving negative signs—have previously been shown to be highly prevalent in equation-solving activities, but that the other two highly-anticipated error types in the present study—variables and terms errors—were not among the most prevalent errors in Algebra I (Booth et al., 2014). This indicates that students were not just anticipating errors they themselves were likely to make; further study is needed to determine how students come up with the errors they anticipate. Getting students to anticipate a wider

variety of errors may require further intervention targeted at helping them first notice less anticipated errors.

Data availability The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Declarations

Conflicts of interest The authors declare that they have no conflict of interests.

References

Adams, D. M., McLaren, B. M., Durkin, K., Mayer, R. E., Rittle-Johnson, B., Isotani, S., & Van Velsen, M. (2014). Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior, 36*, 401–411.

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*(5), 502–508.

Barbieri, C., & Booth, J. L. (2016). Support for struggling students in algebra: Contributions of incorrect worked examples. *Learning and Individual Differences, 48*, 36–44.

Barbieri, C. A., & Booth, J. L. (2020). Mistakes on display: Explaining displayed errors refines algebraic equation solving. *Applied Cognitive Psychology, 34*, 862–878.

Begolli, K. N., & Richland, L. E. (2016). Teaching mathematics by comparison: Analog visibility as a double-edged Sword. *Journal of Educational Psychology, 108*(2), 194–213.

Begolli, K. N., Richland, L. E., Lyons, E., Jaeggi, S., & Matlen, B. (2018). Executive functions in learning mathematics by comparing representations: Incorporating everyday classrooms into the science of learning. *Thinking and Reasoning.* https://doi.org/10.1080/13546783.2018.1429306

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (methodological), 57*(1), 289–300.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*(2), 201–210.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche e Commerciali Di Firenze, 8*, 3–62.

Booth, J. L., Barbieri, C., Eyer, F., & Paré-Blagoev, E. J. (2014). Persistent and pernicious errors in algebraic problem solving. *The Journal of Problem Solving, 7*(1), 3.

Booth, J. L., Cooper, L. A., Donovan, M. S., Huyghe, A., Koedinger, K. R., & Paré-Blagoev, E. J. (2015). Design-based research within the constraints of practice: AlgebraByExample. *Journal of Education for Students Placed at Risk (JESPAR), 20*(1–2), 79–100.

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

Borasi, R. (1994). Capitalizing on errors as "springboards for inquiry": A teaching experiment. *Journal for Research in Mathematics Education, 25*, 166–208.

Cangelosi, R., Madrid, S., Cooper, S., Olson, J., & Hartter, B. (2013). The negative sign and exponential expressions: Unveiling students' persistent errors and misconceptions. *The Journal of Mathematical Behavior, 32*(1), 69–82.

Carroll, W. M. (1994). Using worked examples as an instructional support in the algebra classroom. *Journal of Educational Psychology, 86*(3), 360–367.

Castle, C. (2014, Sep 10). University of Iowa center looks to start high school STEM academy. *Press - Citizen* Retrieved from http://search.proquest.com/docview/1561031112?accountid=14270.

Cherepinsky, V. (2011). Self-reflective grading: Getting students to learn from their mistakes. *Primus, 21*(3), 294–301.

D'Mello, S., & Graesser, A. (2014). Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta Psychologica, 151*, 106–116.

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*(1), 4–58.

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

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *Handbook on Expertise and Expert Performance*. Cambridge University Press.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Frese, M., & Keith, N. (2015). Action errors, error management, and learning in organizations. *Annual Review of Psychology, 66*, 661–687.

Frese, M., & Zapf, D. (1994). Action as the core of work psychology: A German approach. *Handbook of Industrial and Organizational Psychology, 4*(2), 271–340.

Gartmeier, M., Bauer, J., Gruber, H., & Heid, H. (2008). Negative knowledge: Understanding professional learning and expertise. *Vocations and Learning, 1*(2), 87–103.

Gorsky, P., & Finegold, M. (1994). The role of anomaly and of cognitive dissonance in restructuring students' concepts of force. *Instructional Science, 22*(2), 75–90.

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.

Gruber, H. & Palonen, T. (2007). Learning in the workplace: new developments in the relation between working and learning. In Gruber, Hans and Palonen, Tuire, (Eds.) *Learning in the Workplace: New developments. Research in Educational Sciences*, 32. Finnish Educational Research Association (FERA) (pp. 11–18). ISBN 978-952-5401-32-5.

Heemsoth, T., & Heinze, A. (2014). The impact of incorrect examples on learning fractions: A field experiment with 6th grade students. *Instructional Science, 42*(4), 639–657.

Henderson, C., & Harper, K. A. (2009). Quiz corrections: Improving learning by encouraging students to reflect on their mistakes. *Physics Teacher, 47*(9), 581–586.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968.

Keith, N., & Frese, M. (2008). Effectiveness of error management training: A meta-analysis. *Journal of Applied Psychology, 93*(1), 59.

Lannin, J., Townsend, B., & Barker, D. (2006). The reflective cycle of student error analysis. *For the Learning of Mathematics, 26*(3), 33–38.

Lappan, G., Fey, J., Fitzgerald, W., Friel, S., & Philips, E. (2006). *Connected Mathematics 2*. Prentice Hall (School Division).

Lee, C., & Chen, M. (2015). Effects of worked examples using manipulatives on fifth graders' learning performance and attitude toward mathematics. *Journal of Educational Technology & Society, 18*(1), 264–275.

McNeil, N. M. (2014). A change–resistance account of children's difficulties understanding mathematical equivalence. *Child Development Perspectives, 8*(1), 42–47.

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology, 68*, 465–489.

Mueller, M. M., Palkovic, C. M., & Maynard, C. S. (2007). Errorless learning: Review and practical application for teaching children with pervasive developmental disorders. *Psychology in the Schools, 44*(7), 691–700.

National Mathematics Advisory Panel. (2008). *Foundations for Success: The Final Report of the National Mathematics Advisory Panel*. US Department of Education.

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review, 103*(2), 241–262.

Oser, F., & Spychiger, M. (2005). *Lernen ist schmerzhaft. Zur Theorie des Negativen Wissens und zur Praxis der Fehlerkultur [Learning is Painful. On the Theory of Negative Knowledge and the Practice of Error Culture]*. Beltz.

Overoye, A. L., & Storm, B. C. (2015). Harnessing the power of uncertainty to enhance learning. *Translational Issues in Psychological Science, 1*(2), 140–148.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing Instruction and Study to Improve Student Learning. IES Practice Guide. NCER 2007–2004*. National Center for Education Research.

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology, 1*(1), 33–60.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ, 316*(7139), 1236–1238.

Pol, H. J., Harskamp, E. G., Suhre, C. J., & Goedhart, M. J. (2009). How indirect supportive digital help during and after solving physics problems can improve problem-solving abilities. *Computers & Education, 53*(1), 34–50.

Retnowati, E., Ayres, P., & Sweller, J. (2010). Worked example effects in individual and group work settings. *Educational Psychology, 30*(3), 349–367.

Santagata, R. (2004). "Are you joking or are you sleeping?" Cultural beliefs and practices in Italian and U.S. teachers' mistake-handling strategies. *Linguistics and Education, 15*, 141–164.

Siegler, R. S. (1996). *Emerging Minds: The Process of Change in Children's Thinking*. Oxford University Press.

Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition Processes in Development and Learning* (pp. 31–58). Cambridge University Press.

Siegler, R. S., & Chen, Z. (2008). Differentiation and integration: Guiding principles for analyzing cognitive change. *Developmental Science, 11*(4), 433–453.

Stevenson, H., & Stigler, J. W. (1994). *Learning Gap: Why Our Schools Are Failing and What We Can Learn from Japanese and Chinese Educ*. Simon and Schuster.

Tulis, M. (2013). Error management behavior in classrooms: Teachers' responses to student mistakes. *Teaching and Teacher Education, 33*, 56–68.

Van Dyck, C., Frese, M., Baer, M., & Sonnentag, S. (2005). Organizational error management culture and its impact on performance: A two-study replication. *Journal of Applied Psychology, 90*(6), 1228.

Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7*(1), 1–39.

Yerushalmi, E., & Polingher, C. (2006). Guiding students to learn from mistakes. *Physics Education, 41*(6), 532–538.