Computers & Education 139 (2019) 173–190

Contents lists available at ScienceDirect

Computers & Education

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

More confusion and frustration, better learning: The impact of erroneous examples

J. Elizabeth Richey[a,*], Juan Miguel L. Andres-Bray[b], Michael Mogessie[a], Richard Scruggs[b], Juliana M.A.L. Andres[b], Jon R. Star[c], Ryan S. Baker[b], Bruce M. McLaren[a]

[a] Carnegie Mellon University, United States
[b] University of Pennsylvania, United States
[c] Harvard University, United States

ARTICLE INFO

Keywords:
Erroneous examples
Affect
Confusion
Frustration
Affect detection
Learning outcomes

ABSTRACT

Prior research suggests students can sometimes learn more effectively by explaining and correcting example problems that have been solved incorrectly, compared to problem-solving practice or studying correct solutions. It remains unclear, however, what role students' affect might play in the process of learning from *erroneous examples*. Specifically, it may be that students experience greater confusion and frustration while studying erroneous examples, but that their confusion and frustration lead to greater learning. We analyzed student log data from previously published research comparing erroneous example instruction of decimal number mathematics to problem-solving instruction in a computer-based intelligent tutoring system. We created and applied affect detectors for a combination of confusion and frustration ("confrustion") and compared the role of confrustion across conditions. As predicted, students in the erroneous example condition experienced greater confrustion while working through the instructional materials. However, contrary to predictions, confrustion was negatively correlated with posttest and delayed posttest performance across conditions, though less so for the erroneous example condition. Given that students in the erroneous example condition performed better on the delayed posttest than students in the problem-solving condition, it appears they learned more *despite* also experiencing greater confrustion rather than *because* of it. Results suggest that learning from erroneous examples may be an inherently more confusing and frustrating process than traditional problem solving. More generally, this research demonstrates that logging student actions at a step-by-step problem-solving level and analyzing those logs to infer affect can be a powerful way to investigate learning.

Incorporating examples into instruction is a common pedagogical technique that has been studied extensively in cognitive and educational psychology. Research has often focused on instructional principles for implementing examples to make them more effective (see Atkinson, Derry, Renkl, & Wortham, 2000; Wittwer & Renkl, 2010, for reviews). This line of research has identified learning benefits from several uses of examples, including worked examples (i.e., an example with the solution steps provided; Renkl, 1997; Van Gog, Kester, & Paas, 2011; Ward & Sweller, 1990), examples with instructional explanations (i.e., worked examples with

* Corresponding author. Human-Computer Interaction Institute, 5000 Forbes Ave, Carnegie Mellon University, Pittsburgh, PA 15213, United States.
E-mail address: jelizabethrichey@cmu.edu (J.E. Richey).

https://doi.org/10.1016/j.compedu.2019.05.012
Received 1 November 2018; Received in revised form 20 May 2019; Accepted 21 May 2019
Available online 25 May 2019
0360-1315/

conceptual explanations provided along with each step; Renkl, 2002), and erroneous examples (i.e., worked examples that incorporate at least one incorrect solution step; Booth, Lange, Koedinger, & Newton, 2013; McLaren, van Gog, Ganoe, Karabinos, & Yaron, 2016; Siegler & Chen, 2008; Tsovaltzi, Melis, & McLaren, 2012). However, examples vary in their effectiveness and efficiency depending on learner characteristics (e.g., students' prior knowledge), as well as which learning outcomes are considered (e.g., procedural knowledge, near vs. far transfer).

Erroneous examples may be particularly effective for addressing misconceptions, or students' inaccurate conceptual beliefs (Durkin & Rittle-Johnson, 2012; Siegler, 2002, pp. 31–58). Misconceptions tend to be difficult to change and, when they involve more than single incorrect beliefs, are often resistant to direct refutation (Brown, 1992; Chi, 2008). Their deep, conceptual nature also means they tend to disrupt students' learning across a wide range of new topics within a domain and, if unaddressed, can significantly diminish a student's progress in more advanced concepts (Booth et al., 2013; Hiebert & Wearne, 1985; Steinle & Stacey, 2004). Identifying instructional techniques for addressing and correcting misconceptions is a theoretically and pedagogically important endeavor that has been the subject of much research and debate (Smith, DiSessa, & Roschelle, 1994; Vosniadou, 2012, pp. 119–130; for many perspectives, see Sinatra & Pintrich, 2003; Vosniadou, 2009).

Studying erroneous examples might appear to risk reinforcing students' misconceptions or introducing an inaccurate understanding; however, exploring students' errors can play an important pedagogical role in mathematical discussions (Borasi, 1987; Rushton, 2018). There is evidence that showing students the hypothetical errors of others can foster reflection, helping students to recognize and correct errors in their own work (Booth et al., 2013; Durkin & Rittle-Johnson, 2012; Groβe & Renkl, 2007; Siegler & Chen, 2008). Other research has shown that comparing students' own incorrect mental models to accurate models and prompting them to self-explain the differences can lead to greater learning gains than explaining only a correct model (Gadgil, Nokes-Malach, & Chi, 2012). These results appear to contradict intuitions that showing students incorrect examples might strengthen existing misconceptions or introduce new errors, particularly when the instructional materials clearly identify the error in the example.

Although there is evidence that deeply engaging with incorrect knowledge can help students revise their misconceptions, understanding the mechanisms underlying these learning processes still requires greater investigation. While cognitive factors have been considered in many of the studies mentioned above, less research has examined the degree to which other factors may be instrumental in learning from errors. For example, Melis (2004) proposed that studying erroneous examples could encourage students to engage in metacognition as they sought to understand why an example was incorrect, while also improving motivation by encouraging a learning-oriented approach to errors. Given the existing literature suggesting that affect more generally plays a role in learning (e.g., Baker, D'Mello, Rodrigo, & Graesser, 2010; Efklides, 2011; Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011), it may be the case that affect is a factor in understanding whether students learn from erroneous examples.

In particular, a student attempting to explain an incorrect example or to reason through their misconception in the face of evidence of its incorrectness may experience confusion, or even frustration. While these affective experiences are often thought to be unpleasant (e.g., Baker, D'Mello, Rodrigo, & Graesser, 2010), there is increasing evidence that confusion often precedes successful learning (D'Mello, Lehman, Pekrun, & Graesser, 2014; Lehman et al., 2013). As such, understanding the potential role confusion and frustration play when a student is learning from incorrect examples may help us to better understand the underlying learning processes that are occurring.

In this paper, we investigate whether erroneous examples that reflect common misconceptions lead students to experience more confusion and frustration than simply solving practice problems, and whether confusion and frustration more generally accompany, or perhaps even support, the process of learning. We investigate these questions by re-analyzing previously published datasets from a study investigating the effects of erroneous examples on learning (Adams et al., 2014; McLaren, Adams, & Mayer, 2015). In these datasets, erroneous examples were deployed through a computer-based tutoring system, which provides detailed process data on student interactions with the materials. We view this research as both basic, aimed at understanding the cognitive and affective processes around learning from erroneous examples, and use-inspired, aimed at creating opportunities to personalize student learning based on their affective or cognitive states. For example, once the mechanisms of learning from erroneous examples are better understood, the materials within the tutor could be customized to guide students toward productive affective or cognitive processes or intervene when unproductive affective or cognitive states arise. In the following sections, we review prior research on erroneous examples; the relations between confusion, frustration and learning; hypothesized mechanisms for erroneous examples' effectiveness; and the log-based detection of affect employed in this study.

1. Erroneous examples

While worked examples have proven to be an effective instructional approach in many situations, one common shortcoming is their passive nature (Atkinson et al., 2000; Kalyuga, Chandler, Tuovinen, & Sweller, 2001). The most straightforward instructional use of a worked example involves asking students to study a step-by-step solution. However, this approach may not be optimal, since students may forget the steps they studied before they have an opportunity to apply the steps themselves (Trafton & Reiser, 1993). More generally, passive instructional activities tend to promote shallow learning (Chi, 2009). A more typical and more effective instructional approach involves pairing worked examples with practice problems and removing or "fading" the support provided by the worked examples as students progress through the materials (Atkinson, Renkl, & Merrill, 2003). Worked examples may be even more effective when students are prompted to explain the examples; the quality of students' explanations have been found to predict learning (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Renkl, 1997; Renkl & Atkinson, 2002). Prompted explanation of worked examples has also been shown to be effective in the context of intelligent tutoring systems (McLaren, Lim, & Koedinger, 2008), the focus of the current paper. Other research has shown that prompting students to compare worked examples leads to learning benefits

(Rittle-Johnson & Star, 2009). Results from these studies suggest that simply seeing and solving examples is not sufficient to promote learning; students must engage more deeply with the examples in a way that prompts them to identify key principles and understand why steps are correct.

Even when students engage with worked examples through explanation or comparison, there is still the risk that students who are exposed only to correct solutions may miss opportunities to test their understanding and the limits of the examples, or to identify areas where they might still be confused. Erroneous examples have been deployed as a means of increasing student engagement with worked examples and highlighting common errors that students tend to make, to prevent students from underestimating the difficulty of a problem or procedure. Students are typically prompted to identify and explain the errors, and then to correct them (e.g., Groβe & Renkl, 2007). Correcting, comparing, or explaining errors has been found to be particularly effective for conceptual learning; for example, several studies have shown that studying erroneous examples improved performance on deeper measures of learning, such as conceptual understanding and far transfer, but not on more shallow knowledge measures like near transfer (Booth et al., 2013; Siegler, 2002, pp. 31–58). However, these benefits may not be equal for all students. Students with higher knowledge seem to benefit more from erroneous examples than other students when little to no scaffolding is provided (Groβe & Renkl, 2007), indicating that a basic understanding of a domain is important before students are exposed to erroneous examples. Other research found no effect of prior knowledge when students engaged in more scaffolded comparison of correct and incorrect examples, suggesting that erroneous examples can be beneficial to all learners when the instructional task includes higher levels of support (Durkin & Rittle-Johnson, 2012).

A related area of research has examined productive failure as a means of preparing students for learning a new topic (Kapur, 2016). This instructional design involves a problem-solving phase during which students generate (usually unsuccessfully) a solution procedure for a novel type of problem, followed by an instructional phase during which students are taught the correct or canonical solution strategies (Kapur & Bielaczyc, 2012). Research has shown that productive failure leads to greater conceptual learning and mental effort than vicarious failure, in which the learner examines another student's incorrect solution (Kapur, 2016). Both productive failure and vicarious failure have led to greater learning outcomes than introducing a new concept through direct instruction first (Kapur, 2014). While the productive failure research might seem to suggest that problem solving would be more effective than studying erroneous examples, the productive failure approach has been examined in situations where students are learning a new concept for the first time, and where these students reliably produce incorrect responses. In the current study, as in most prior research on erroneous examples, students have already been introduced to the target concepts and are practicing problems for which misconceptions frequently cause incorrect responses. For this reason, and based on the prior erroneous examples literature, we expect students to learn more from studying and correcting typical errors than from solving problems on their own.

2. Confusion, frustration, and learning

The last several decades have seen an explosion of scientific interest in academic emotions and affect during learning (Calvo & D'Mello, 2010; Wu, Huang, & Hwang, 2015). A range of studies have found evidence that differences in learner affect are associated with differences in student learning outcomes in the short-term (Pekrun, Goetz, Titz, & Perry, 2002; Rowe, Shores, Mott, & Lester, 2011) and in outcomes as distant as choosing to attend college years later (San Pedro, Baker, Bowers, & Heffernan, 2013). There has been particular interest in confusion and frustration, as confusion and frustration have shown varying correlations to learning across studies. Some studies find strong positive correlations between confusion or frustration and learning (D'Mello et al., 2014; Lehman et al., 2013), whereas other studies find strong negative correlations to learning (Rodrigo et al., 2009; Schneider et al., 2015). One possible explanation for the inconsistency in results is that not all instances of confusion and frustration are identical. For example, Liu, Pataranutaporn, Ocumpaugh, and Baker (2013) found that brief confusion and frustration were associated with positive outcomes, whereas extended confusion and frustration were associated with negative outcomes. This led Liu and colleagues to hypothesize that confusion and frustration signal that the learner is engaged in the type of productive struggle that is necessary for learning, but that if confusion and frustration are left unresolved, learning does not occur. However, there has been relatively limited work to operationalize what productive struggle looks like behaviorally during the process of learning. One exception to this is work by Kai, Almeda, Baker, Heffernan, and Heffernan (2018) in which they differentiated productive struggle from unproductive struggle. However, this work looked at learning and behavior over longer time periods and without relating the struggle to affective states.

A related perspective is seen in D'Mello and Graesser (2012), who hypothesized that a positive state of confusion transitions into increasingly negative frustration, and then boredom, if it is not resolved. More recently, Shute et al. (2015), working with data from an educational game, proposed the existence of two paths that lead to learning: one through engaged concentration and the other through confusion. They also found that frustration was negatively correlated with boredom, suggesting that students must be engaged in order to be frustrated. Indeed, confusion—appropriately used—can be a positive instructional intervention. Lehman et al. (2013) found that inducing confusion through contradictory information led to better learning outcomes. This result was replicated by D'Mello et al. (2014), who also found that students only appeared to learn from contradictory information if they experienced confusion.

In the D'Mello et al. study, self-reports of confusion were not predictive; only behavioral indicators of confusion predicted learning. This is important because the behavioral indicators used in this research, as well as in other research on behavioral indicators of confusion (e.g., Lee, Rodrigo, Baker, Sugay, & Coronel, 2011), could be associated with frustration as well as confusion. When Liu et al. (2013) investigated the relation between the duration of confusion and learning, they also investigated the relation between the duration of frustration and learning and found that these two sets of patterns (confusion duration → learning, frustration duration → learning) looked very similar. Longer confusion or frustration was associated with poorer outcomes; brief confusion or

frustration was associated with better outcomes. In fact, treating confusion and frustration as the same construct within these analyses led to stronger associations with learning outcomes than considering them separately. This led them to hypothesize that confusion and frustration might represent two points on the same continuum, which they referred to as *confrustion* (Liu et al., 2013).

3. Confrustion and erroneous examples

There has been limited research thus far on confusion and frustration in the specific context of erroneous examples. However, there is evidence that students working with erroneous examples typically take longer and experience considerable uncertainty when first encountering erroneous examples (Adams et al., 2014; McLaren et al., 2015; Siegler, 2002, pp. 31–58). Other research has found that the use of erroneous examples increases cognitive load and learning time (Heitzmann, Fischer, & Fischer, 2018). With practice, however, these students become more efficient and accurate than students engaged in problem solving or explaining correct examples (Siegler, 2002, pp. 31–58), eventually completing post-test materials more quickly than students who were not exposed to erroneous examples while learning (Tsovaltzi et al., 2012). Thus, while erroneous examples may initially lead to confrustion and errors, it seems that when students are able to resolve their confrustion, they may acquire deeper, more flexible knowledge. This is consistent with the view that examining errors provides an opportunity for students to dig deeper into a concept and differentiate conditions for appropriate strategy use, often leading to more conceptually rich solution strategies (Borasi, 1994; Siegler, 2002, pp. 31–58). Erroneous examples could be considered a specific mechanism for creating cognitive conflict, which has been shown to support conceptual change and the revision of misconceptions when conflict is delivered in a way that is motivating and meaningful to students (Limón, 2001).

However, it may not be the case that all students are able to overcome the initial confusion associated with explaining incorrect solutions. Groβe and Renkl (2007) found that studying a mix of correct and erroneous examples led to better far transfer than studying only correct examples, but only for students with high levels of relevant prior knowledge. For students with less prior knowledge, the mix of correct and erroneous examples reduced performance compared to seeing only correct examples. This suggests that the amount of difficulty experienced by students solving erroneous examples, which may lead to confusion and/or frustration, may be important. Students who experience enough confusion to engage more deeply with the example may acquire more robust conceptual knowledge, but only if they possess sufficient prior knowledge and motivation to resolve their confusion. As such, learning from erroneous examples may be mediated by the degree of difficulty and confusion students experience.

4. Automated detection of confusion and frustration

As students interact with educational technology, they experience an array of affective states that impact their performance and learning (D'Mello, 2013). Work over the last decade has established that it is possible to create *affect detectors* that can determine a student's affective state (albeit imperfectly) at any point during interaction with a learning system, solely from the student's interaction with the system. It is also possible to detect affect from physical and physiological sensors (e.g., Muldner, Burleson, & VanLehn, 2010), but it is more difficult to scale the use of these sensors to larger groups of students or deploy them in classroom settings. Researchers have designed models that can detect confusion and frustration solely from interaction data for a variety of learning systems (D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008; Baker et al., 2012; Baker & Inventado, 2014; DeFalco et al., 2018; Kostyuk, Almeda, & Baker, 2018; Liu et al., 2013; Pardos, Baker, San Pedro, Gowda, & Gowda, 2014; Paquette, Baker, et al., 2014). These detectors have also been successful at predicting longer-term student outcomes (Kostyuk et al., 2018; Pardos et al., 2014).

Affect detectors have been used to study affect in fine-grained detail, at a grain-size of around 20-second intervals (D'Mello & Graesser, 2010; Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2013; Pardos et al., 2014). It is also possible to determine students' transitions between affective states through these detectors (Botelho, Baker, Ocumpaugh, & Heffernan, 2018). Multiple studies have shown that the patterns of these transitions can predict differences in student learning, more so than the states on their own (e.g., Lee et al., 2011; Liu et al., 2013).

The process of creating a detector for an affective state almost always starts with first obtaining “ground truth”—human-labeled data that show the presence or absence of the affective state in question (Baker & Inventado, 2014)—for a sufficiently large sample of data. These labels, which are verified for acceptable inter-rater reliability (Ocumpaugh, Baker, & Rodrigo, 2015), are then used to develop detectors, using machine-learning algorithms to identify the in-system behaviors that correspond to the human judgments of affect. Most commonly, ground truth labels are created through human observation protocols (e.g., BROMP; Ocumpaugh et al., 2015), where coders are personally present and code for affect. Video data has also been coded to study learner affect (Sinha, Bai, & Cassell, 2017). Coders using these methods are able to base their coding on both the students' body language and facial expressions, and their interaction behaviors. In some retrospective analysis cases, however, where coders were not physically present and no video was obtained, it is also possible to obtain ground truth using an alternate approach, text replay coding (Baker, Corbett, & Wagner, 2006). In this method, coders base their affect coding on log data gathered on the students' interaction with the intervention. Text replay coding involves breaking down the existing data set into text replays, or clips, each either spanning a specific amount of time, a specific number of transactions, or delineated by start or end events. Human coders then look at all the interactions within a clip and decide whether the student displayed a specific behavior or affective state. Text replays have been found useful for labeling gaming the system (Baker et al., 2006), scientific inquiry skills (Sao Pedro, de Baker, Gobert, Montalvo, & Nakama, 2013), and confrustion (Lee et al., 2011), but have not yet been used to study other affective states. In the specific case of confrustion, it is not feasible to differentiate confusion from frustration within text replays, but trained coders have achieved good inter-reliability at determining whether either of these affective states is present, from the visible behavior of struggling with the material over multiple responses

Table 1
Misconceptions examined in current study.

Name	Misconception	Example
Megz	Decimal numbers with more digits to the right of the decimal point are larger in magnitude than those with fewer digits	.625 is larger than .82
Segz	Decimal numbers with fewer digits to the right of the decimal point are larger in magnitude than those with more digits	.62 is larger than .825
Pegz	The two sides of the decimal point are viewed as separate numbers	1.9 + 0.2 = 1.11
Negz	Decimal numbers between 0 and 1 are smaller in magnitude than 0	.06 is placed on the left side of the number line, at −0.6

(i.e., Lee et al., 2011). In other words, it is possible to accurately code for confrustion with text replays, but not for confusion or frustration separately.

5. Decimal misconceptions

Given the role erroneous examples can play in helping students recognize and correct errors in their thinking, correcting and explaining erroneous examples may be particularly effective when students have existing misconceptions about the content of the examples. Within the mathematical domain of decimal fractions (or decimals), students have been found to have several well-documented misconceptions, largely based on inappropriately transferring existing knowledge about integers and fractions to decimal fractions (Desmet, Gregoire, & Mussolin, 2010; Stacey, Helme, & Steinle, 2001; Stacey & Steinle, 1998). The materials used in the erroneous examples tutor focus on four common misconceptions that have been identified in students' knowledge of decimal fractions and that contribute to many errors in tasks with decimals (Resnick et al., 1989; Sackur-Grisvard & Léonard, 1985; Stacey, 2005). We label these misconceptions with names created by Isotani et al. (2011): Megz (mega numbers misconception), Segz (shorter numbers misconception), Pegz (misconception on each side of the "peg"), and Negz (negative numbers misconception; Table 1). These misconceptions have been shown to persist throughout grade school and into adulthood, and they have been observed even in pre-service mathematics teachers (Putt, 1995; Stacey et al., 2001).

The materials in the present study targeted and measured these four misconceptions because they have been observed to occur at different rates depending on students' ages and the sequence in which students have learned mathematical concepts, such as whether they are learning decimal concepts before or after fractions (Resnick et al., 1989; Steinle, 2004). As a result, some misconceptions may create greater levels of confusion and frustration, and some may be more resistant to correction through practice. In the case of late elementary and early middle school students, all four of the misconceptions targeted in these materials are typical, but the Megz misconception is the most common (Isotani et al., 2011; Sackur-Grisvard & Léonard, 1985).

6. Present analysis

In this paper, we analyze log data from previously published research comparing erroneous example instruction of decimal number mathematics to more conventional problem-solving instruction (Adams et al., 2014; McLaren et al., 2015). We create and apply affect detectors for confrustion and compare the role of confrustion across conditions, as well as its relations with learner characteristics and different learning outcomes. We test the following research questions and hypotheses:

Do the erroneous example and problem-solving groups differ in their levels of confrustion while working through the instructional materials? We hypothesize that students in the erroneous example condition will experience greater confrustion. The productive struggle generated as students try to understand errors is often identified as a key mechanism in explaining how erroneous examples support learning (Siegler, 2002, pp. 31–58). This productive struggle should create greater confusion and frustration (i.e., *confrustion*) as students study erroneous examples compared to solving problems.

Does confrustion predict learning outcomes? We hypothesize that greater levels of confrustion will be associated with positive learning outcomes. Although prior research on the relations between confusion, frustration, and learning have produced mixed results (D'Mello et al., 2014; Lehman et al., 2013; Rodrigo et al., 2009; Schneider et al., 2015), the confusion and frustration generated by erroneous examples are expected to result from the same features of erroneous examples that facilitate learning (i.e., an answer that students might initially consider correct presented as incorrect, and the ensuing struggle to make sense of the incorrect information). For this reason, combined with previous results showing that confrustion that is resolved is associated with better learning (Liu et al., 2013), we expect confrustion to be associated with better learning, possibly more so in the erroneous example condition (which promotes confrustion but also provides the support needed to resolve it) than in the problem-solving condition.

Do confrustion levels differ based on the misconception targeted? Within the erroneous example condition, we predict that students will experience high levels of confrustion across all types of problems. When working on erroneous examples targeting misconceptions that they hold, students are likely to experience confrustion when trying to correct the problem. On the other hand, when working on erroneous examples targeting misconceptions that they do not hold, students may experience confrustion when trying to explain how a student could produce such an error. In contrast, within the problem-solving condition we predict that students will experience high levels of confrustion only on items targeting misconceptions that they hold. Therefore, in the problem-solving condition, we hypothesize that confrustion will be greatest on the problems targeting the Megz misconception, which has

been found to be the most common and pervasive misconception in this content area (Isotani et al., 2011).

Do confrustion levels decline as students work through the complete set of materials? We hypothesize that the difference in confrustion across conditions will be concentrated in the first half of items in the instructional materials. This is consistent with prior research showing that students are initially slower and less certain when working with erroneous examples, but that they eventually become more efficient and effective (Siegler, 2002, pp. 31–58; Tsovaltzi et al., 2012). During the second half of instruction, students in the erroneous example condition may have resolved their initial confusion and frustration, leading to similar or lower levels of confrustion compared to the problem-solving condition. If confrustion remains greater in the erroneous example condition, it will indicate that students' confrustion is coming not from the novelty of working with erroneous examples, but instead from the relative difficulty of understanding and correcting hypothetical errors compared to problem solving.

7. Methods

The current study analyzes data previously collected through a series of studies investigating the impact of erroneous examples on students' learning of decimal fraction concepts (Adams et al., 2014; McLaren et al., 2015). We analyzed interaction log data collected during the prior studies to examine the role confrustion played in students' learning across conditions. Below, we describe the methodological details of the prior studies, as previously reported in the original publications.

7.1. Participants and design

Data were collected across three semesters over a two-year period at five urban and suburban schools in the metropolitan area of a northeast U.S. city. One to two sixth-grade math teachers at each school participated in the study, and students from all sections of those teachers' courses completed the materials as part of their regular instructional activities. A total of 787 students participated. Students were dropped from the study if they failed to complete all materials within the allotted time (n = 119), or if they were assigned to a piloted adaptive condition that is not included in the analyses reported here (n = 68). In the rare event that students participated twice as a result of repeating a grade and thus completing the experiment in both the first and second school years, data from their second completion of the materials were dropped (n = 2). As a result, the final dataset included 598 students (305 females, 293 males) with a mean age of 11.75 years old. The experiment had a between-subjects design, with students randomly assigned at the individual level to either the erroneous examples (ErrEx) or problem-solving (PS) condition.

7.2. Materials

All materials were developed using CTAT, the Cognitive Tutor Authoring Tools (Aleven et al., 2016), and delivered through Tutorshop, a learning management system that supports web delivery for classroom deployment of CTAT tutors (Aleven, McLaren, & Sewell, 2009). Materials were developed in consultation with a mathematics education expert to target four common misconceptions about decimal numbers (Table 1). Although decimal number operations are typically introduced in fifth grade in the United States (Common Core standard CCSS.Math.Content.5.NBT.A.3), they are also a learning objective in sixth grade (Common Core standard CCSS.Math.Content.6.NS.B.3). The materials are not aimed at introducing decimal numbers for the first time, but rather to address the misconceptions that many students hold after first learning about decimal numbers (Stacey et al., 2001), and to give students more extensive practice toward developing decimal number fluency. All students across conditions saw problems in the same order, and problems were organized into sequences of three problems (two intervention problems and one practice problem) targeting the same misconception. A more detailed description of materials is available in the original papers reporting results from these studies (Adams et al., 2014; McLaren et al., 2015).

Erroneous examples intervention materials. A series of 32 problems[1] were written to address four common, well-documented misconceptions about decimal numbers (Table 1; Isotani et al., 2011; Stacey & Steinle, 1998). The problems included sorting decimal numbers in order of magnitude, placing decimal numbers on a number line, completing a sequence of decimal numbers, and adding two decimal numbers. For each erroneous example, students were presented with a decimal number word problem and an incorrect solution provided by a hypothetical student (Fig. 1). They were informed that the solution was incorrect, and were prompted to correct the solution. They also responded to a series of three to four multiple-choice questions in which they explained the hypothetical student's error, the correct solution, and the relevant underlying principles. Although self-explanation prompts frequently require students to construct their own explanations (e.g., Chi, de Leeuw, Chiu, & LaVancher, 1994; McNamara, 2004), the questions were designed as multiple-choice selections to promote self-explanation without creating significant working memory demands. Prior research in computer-based instructional environments has shown this style of self-explanation to be effective (Johnson & Mayer, 2010; Mayer & Johnson, 2010). Students received feedback at each step to indicate whether their responses were correct or incorrect; for any incorrect steps, the student was prompted to correct the errors and could not proceed without correction.

PS intervention materials. For each ErrEx problem, a version of the same problem was created to target the same content and misconception, but without an erroneous example for students to study (Fig. 2). Instead, students read the problem text, solved the problem, and responded to one or two multiple-choice questions asking them to explain the correct solution and underlying principle.

[1] Students from one school (n = 208) saw only 24 intervention problems, as the final eight problems were added to the materials the following year to give students additional learning opportunities as part of the study.

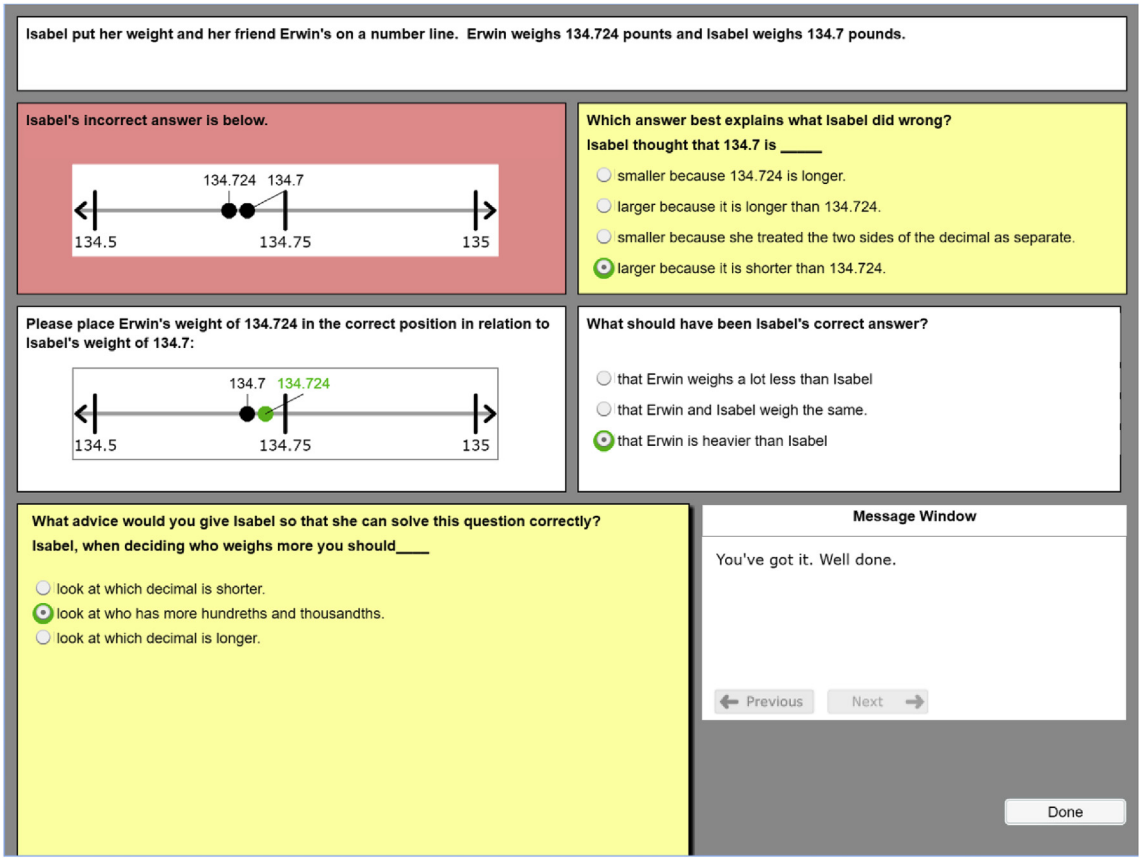

Fig. 1. Example of an erroneous-example item focused on a Segz misconception (shorter decimals are larger).

As in the ErrEx condition, students received feedback at each step to indicate whether their responses were correct or incorrect, and they were prompted to correct any errors before proceeding to the next step.

Practice problems. Sixteen practice problems[2] were included in both the ErrEx and PS materials; every third problem in the materials was a practice problem, and practice problems were identical across conditions. Practice problems targeted the same misconception as the preceding two problems and provided feedback on accuracy, but no explanation prompts were included in the practice problems. Practice problems were included based on worked-examples literature that has shown that students benefit from tackling practice problems after studying worked examples (Kalyuga et al., 2001; McLaren et al., 2008; Renkl & Atkinson, 2003).

Tests. Three isomorphic versions of a 46-item test were created to assess students' decimal number knowledge and misconceptions about decimals. Tests were used as pre-, post-, and delayed post-test measures, with the version order counterbalanced across students. We used pretest scores as a measure of prior knowledge; all references to prior knowledge in our results refer to these pretest scores. Items were written to target different decimal number misconceptions (9 Megz, 10 Segz, 10 Pegz, and 9 Negz); an additional 8 questions targeted decimal number knowledge that was not directly related to one of the four misconceptions. Examples of each item are shown in Table 2. All items were multiple choice or fill-in-the-blank, and each item had only one correct answer. After 15 of the items, dispersed throughout the 46 items, students were asked to rate their confidence in their responses on a scale of 1 (not at all sure) to 5 (very sure). Test performance was reported as a percentage.

7.3. Procedure

Data collection occurred over a period of 6 day at each school. Students completed the pretest, intervention materials, and posttest during the first five of the six days, as part of normal instructional activities in their math classes. Classes typically lasted between 45 and 60 min each day, depending on school schedules. Members of the research team were present throughout the tests and intervention to guide the activities and provide technical support, but they did not provide instruction or content support to students. Students were instructed to work at their own pace. Students were given scratch paper and the option to write out their work, but they were not allowed to use calculators or to collaborate with others for any part of the experiment. If they finished all of the

[2] There were 12 practice problems for the students from one school (n = 208) who received only 24 learning problems.

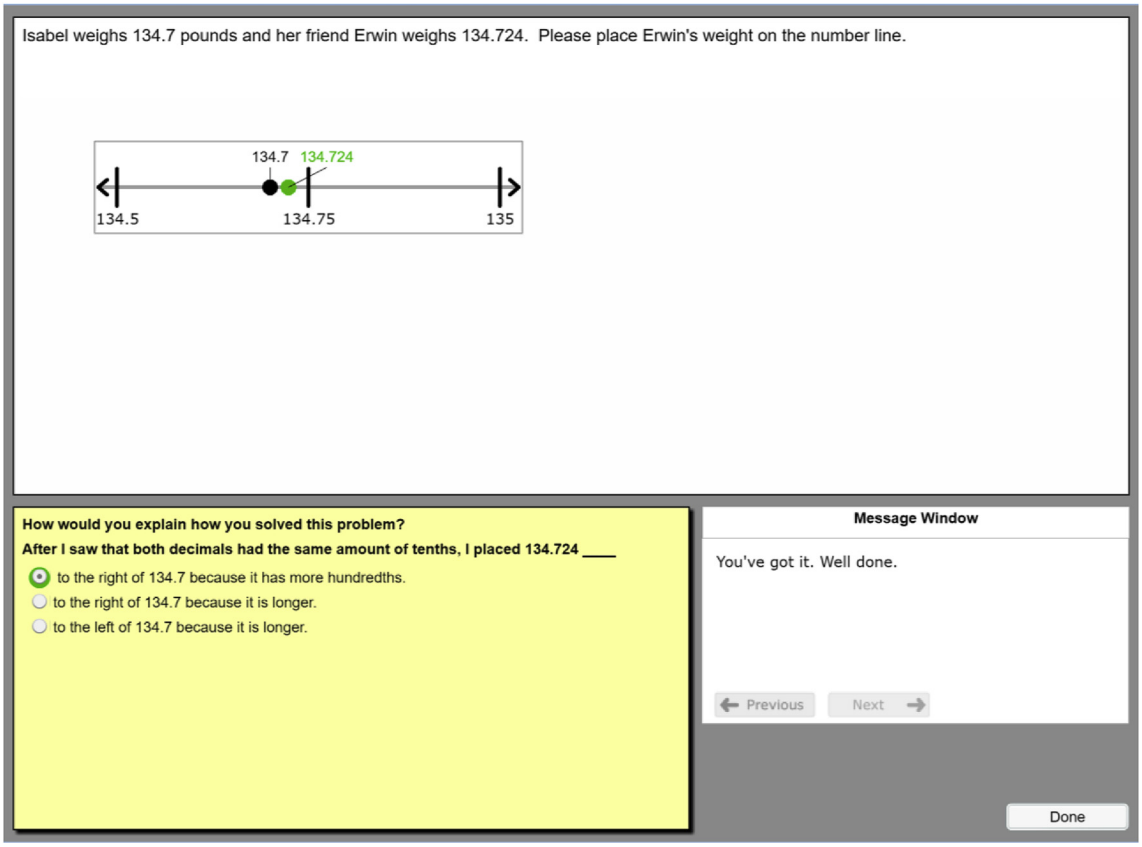

Fig. 2. Example of a PS item focused on a Segz misconception (shorter decimals are larger).

Table 2
Example test items targeting each decimal misconception.

Item type	Example problem
Megz	Which number is largest: 0.12, 0.101, 0.2
Segz	Place the following numbers in order from largest to smallest: 0.899, 0.89, 0.8, 0.8997
Pegz	22.70 + 0.4 = ?
Negz	Are the following numbers listed in order from smallest to largest? 0.1, 0.4, 0, 1.0

materials before the end of the initial five-day period, they were assigned unrelated coursework that did not involve decimal numbers. One week after the initial five days, students were asked to complete the delayed post-test. Teachers agreed not to cover any decimal number concepts in their own instruction or in any assignments given to students during the data collection period.

Affect detection. Our first attempt to apply affect detection involved using existing detectors that had been built using interaction data from a different tutor, MathTutor (Aleven, McLaren, & Sewall, 2009), which was also implemented using CTAT. Both MathTutor and the decimal tutor that was used to gather the data in the current study were implemented on the same platform, so we decided to test whether these detectors could be applied to the current study's data. Once the detectors were applied, however, we found very low, unrealistic proportions of all states: 0% incidence of off-task behavior, 3.86% incidence of boredom, 0.03% incidence of confusion, and 0.06% incidence of frustration. Upon further investigation, we found that these detectors were heavily reliant on hints, a feature commonly used in CTAT tutors but not present in the decimal tutor. Because of this limitation, we decided that the detectors were not directly generalizable to the current study's dataset, and thus built new detectors using text replay coding, which was discussed earlier. We chose text replay coding over quantitative field observations or video coding because log data was already available, and past evidence has shown that confrustion detection from text replay coding on log data is feasible (e.g., Lee et al., 2011). Confrustion was coded in the text replays rather than coding confusion and frustration separately, due to the difficulty humans have in distinguishing these affective states from each other in log data, and the theoretical linkages between these affective states, as discussed earlier (Liu et al., 2013).

In our current study, two coders manually labeled a sample of 1600 problem-level clips for confrustion. The two coders, the second and fifth authors of this paper, each had multiple publications and considerable research experience in affect and affective

Line	Duration (sec)	Step Name	Outcome	Input
1	0	.	.	34:23.8
2	5	answer.nextButton ButtonPressed	Correct	-1
3	5	_root showWhyQ	Correct	-1
4	17	whyQChoices UpdateRadioButton	Correct	whyQ.p2Answer1: R
5	17	_root showMiddle1	Correct	-1
6	6	middlepanel1.p3numline ButtonPressed	InCorrect	134.798
7	6	middlepanel1.p3numline ButtonPressed	InCorrect	134.863
8	1	middlepanel1.p3numline ButtonPressed	InCorrect	134.947
9	1	middlepanel1.p3numline ButtonPressed	InCorrect	134.875
10	3	middlepanel1.p3numline ButtonPressed	InCorrect	134.62
11	3	middlepanel1.p3numline ButtonPressed	InCorrect	134.574
12	1	middlepanel1.p3numline ButtonPressed	InCorrect	134.542
13	3	middlepanel1.p3numline ButtonPressed	InCorrect	134.937
14	0.5	middlepanel1.p3numline ButtonPressed	InCorrect	134.866
15	0.5	middlepanel1.p3numline ButtonPressed	InCorrect	134.893
16	1	middlepanel1.p3numline ButtonPressed	InCorrect	134.85
17	1	middlepanel1.p3numline ButtonPressed	InCorrect	134.832
18	0.5	middlepanel1.p3numline ButtonPressed	InCorrect	134.807
19	0.5	middlepanel1.p3numline ButtonPressed	InCorrect	134.774
20	1	middlepanel1.p3numline ButtonPressed	Correct	134.731
21	1	_root showMiddle2	Correct	-1
22	8.5	middleQChoices UpdateRadioButton	Correct	middlepanel2.p3Answer1: R
23	8.5	_root showAdviceQ	Correct	0
24	11	p4group1 UpdateRadioButton	Correct	adviceQ.p4Answer1: R
25	11	_root showDone	Correct	-1
26	2	done ButtonPressed	Correct	-1

Fig. 3. Part of an example of a clip used in coding and detecting confrustion, taken from a student's attempt to solve the number line problem in the ErrEx condition shown in Fig. 1. All data is in DataShop format (http://pslcdatashop.org; Koedinger et al., 2010; 2013).

computing. We delineated clips (coding units) by treating each problem as its own clip. Each problem in the ErrEx condition comprised four to five steps: self-explanation of the erroneous example, providing the correct answer, self-explanation of the correct answer, and answering one or two advice questions. Each problem in the PS condition comprised two to three steps: providing the correct answer and answering one or two advice questions. As such, each clip was comprised of multiple steps. Fig. 3 shows an example clip.

In the clip shown, a student is first presented with the erroneous example (line 2), with a hypothetical student's incorrect solution to a question about placing a decimal number on a number line. After reading the erroneous example, the student advanced by prompting the tutor to give them the first-step question (line 3), which asks the student to explain the error in the erroneous example. The student in this clip selected the correct explanation from the multiple-choice options on the first attempt (line 4). The student was then presented with the second step, the number line with the incorrectly placed decimal number (line 5), and was asked to place the decimal number in the correct position on the number line. The student in this clip made fifteen attempts at answering this step (lines 6–20) before getting the correct answer: 134.731. The student's first response was 134.798 (line 6), and the student's guesses continued until getting the correct answer in line 20. Changes in the student's guesses included increasing values (i.e., from 134.798 to 134.947), decreasing values (i.e., from 134.947 to 134.542), and several large jumps in value. The third step (line 21) asked the student a multiple choice self-explanation question, which the student answered correctly on the first attempt (line 22). The final step (line 23) asked the student another multiple-choice self-explanation question, which student again answered correctly on the first try (line 24). Finally, the “Done” button appeared (line 25), the student clicked it (line 26), and they progressed to the next problem.

Coders recognized confrustion based on their overall judgment regarding a clip, based on evidence that holistic reasoning produces richer representations of complex constructs than attempts to produce coding rules by hand (see Paquette, de Carvalho, & Baker, 2014 for an example of the complex reasoning used in holistic judgments). For example, a simple set of rules might contain items like “takes a long time to respond,” which could by itself represent many cognitive and affective states. Using a holistic approach, a student who paused for a substantial amount of time before responding would not be coded as confrusted based solely on that feature, but a student who paused for a substantial amount of time before responding, gave an incorrect response, and then went

on to make more incorrect attempts would be coded as confrusted. In discussing their approach, the coders came to the following consensus, which covered most of the common cases where confrustion was seen across conditions. For multiple-choice questions, a student who spent a substantial amount of time on a first, incorrect attempt and then went on to make at least one additional incorrect attempt was labeled confrusted. On number line problems, students were labeled as confrusted if either of the following conditions were met: 1) the student made multiple incorrect attempts in both directions on the number line (e.g., first attempts 0.7, then 0.81, then 0.55, such as what is shown in Fig. 3), or 2) the student made more than two attempts where the current attempt was substantially distant from the previous attempt (e.g., first attempt 0.3, then 1.1, then 1.8). On ordering problems, students who made at least two incorrect attempts were labeled as confrusted. On problems where students were asked to complete a sequence, a student was labeled as confrusted if they made at least two incorrect attempts on each empty slot of the sequence. Finally, in decimal addition problems, students were labeled as confrusted if either of the following conditions were met: 1) the student triggered different errors within the same step, or 2) the student made at least two incorrect attempts on a single step before triggering a different error.

In order to establish ground truth in this data set, the two coders first discussed a small number of clips together to establish that they were thinking about confrustion similarly ($n = 50$). They then labeled the same set of clips independently and checked for inter-rater reliability ($n = 130$), achieving high agreement between the two coders ($\kappa = 0.82$, $p < .001$). After that, the two coders coded the rest of the clips independently ($n = 1420$), splitting the remaining clips between them. The 1600 clips were stratified to equally represent all four problem types (i.e., ordering of decimals, placement on the number line, completing the sequence, and decimal addition), all student cohorts present in the data set, and both conditions in the original study. Of the 1600 clips, 512 clips (32%) were coded as confrusted. This confrustion proportion is at the upper end, but still in range, of what has been seen in past studies of confusion and frustration where other coding methods were used (cf. Andres & Rodrigo, 2014; Baker et al., 2010).

This labeled sample of clips was then used to build a confrustion detector that predicted confrustion at the problem level. The detector was built using the Extreme Gradient Boosting (XGBoost) classifier (Chen & Guestrin, 2016), which uses an ensemble technique in which an initial, weak decision tree is trained, and its prediction errors are calculated. Subsequent decision trees are then trained iteratively to predict the error of the decision tree before them. The final prediction is the sum of the predictions of all the trees in the set (Chen & Guestrin, 2016). We used 10-fold student-level cross-validation to validate this model, repeatedly building the model on some students' data and testing it on other students' data, and we determined that it was effective at inferring confrustion in unseen students, e.g., in the testing data ($\kappa = 0.84$, AUC = 0.97, precision (0) = 0.95, recall (0) = 0.93, precision (1) = 0.87, recall (1) = 0.9). The detector was applied to the rest of the dataset, comprising a total of 27,439 clips across 598 students.

To predict confrustion, the detector used 37 features that were representative of the students' interaction with the decimal tutor. These features can be divided into four main categories:

1. Total amount of time spent, including time on the entire problem attempt; on each of the steps of the problem; between starting the problem and the first attempt on step 1; between getting the correct answer on the final step and proceeding to the next problem; and the minimum and maximum amount of time spent on any one step
2. Amount of time spent on the first attempt of each step
3. Number of attempts, including total number of attempts on the problem; attempts per step; and total incorrect attempts on the problem

7.4. Average amount of time spent per attempt per step

For the most important features, see Table 3. The XGBoost algorithm weighed the importance of each feature as its proportion of contribution to the final prediction model, which ranged from zero and one. The contributions of all the features thus added up to one.

8. Results

Main effects of the intervention on students' test performance, survey responses, and confidence have been reported previously (Adams et al., 2014; McLaren et al., 2015). As reported in previous papers, there was no significant difference in posttest performance

Table 3
Feature descriptions and importance in predicting confrustion.

Feature	Importance
Total number of incorrect problem attempts	0.105
Total reading time	0.077
Minimum amount of time spent within any step	0.069
Total time spent on a problem attempt	0.061
Total time spent on the “Providing the Correct Answer” step	0.052
Total reflection time	0.05

Note. Only the features with contribution value of 0.05 or greater, of the 37 total features, are included in the table.

Table 4
Learning gains by condition and misconception type.

Type	Pre-posttest gains *M* (*SD*)		Pre-delayed test gains *M* (*SD*)		Pre-post *t*-test	Pre-delayed *t*-test
	PS	ErrEx	PS	ErrEx		
Megz	.14 (.26)	.14 (.26)	.17 (.27)	.20 (.27)	t (596) = 0.25, p = .80	t (596) = 1.31, p = .19
Segz	.11 (.29)	.13 (.29)	.13 (.28)	.18 (.29)	t (596) = 0.48, p = .63	t (596) = 1.87, p = .062
Pegz	.07 (.18)	.08 (.18)	.11 (.18)	.11 (.19)	t (596) = 0.85, p = .39	t (596) = 0.42, p = .67
Negz	.05 (.23)	.09 (.26)	.08 (.23)	.15 (.24)	t (596) = 2.27, p = .024*	t (596) = 4.05, p < .001*

Note: Levene's test for equality of variances was rejected for pre-post Negz problem gain, F = 5.83, p = .016; therefore, equal variances were not assumed for *t*-tests on this misconception type. * indicates a significant difference. Results originally reported in Adams et al. (2014) and McLaren et al. (2015).

between students in the PS (M = 0.64, SD = 0.22) and ErrEx conditions (M = 0.67, SD = 0.21) when controlling for pretest, F (2, 595) = 2.24, p = .14, d = 0.14. There was, however, a significant effect of condition predicting delayed posttest when controlling for pretest, F (2, 595) = 15.83, p < .001, d = 0.27, with students in the ErrEx condition (M = 0.73, SD = 0.19) performing better than students in the PS condition (M = 0.68, SD = 0.21). Table 4 reports gain means, standard deviations, and *t*-test results comparing conditions by misconception.

In this paper, we re-analyze this data to infer students' affective states while completing the intervention materials, based on their behaviors in the tutor. Students' overall confrustion levels were calculated by taking the probability that they were confrusted on each individual intervention problem, assessed by the detector, and then averaging across those probabilities. Averaging probabilities retains more information than treating each problem as involving either confrustion (1) or non-confrustion (0); a student with a 45% probability of confrustion across 10 problems should be treated as confrusted 45% of the time rather than 0% of the time. All results reported below relate to our automated detector measure of confrustion and its relation to other variables.

8.1. Does confrustion predict learning outcomes?

Confrustion was significantly, negatively correlated with performance on the pretest (r = −0.74, p < .001), posttest (r = −0.74, p < .001) and delayed posttest (r = −0.73, p < .001). Given that the pretest was a significant predictor of confrustion, we tested multiple regression models using confrustion to predict students' performance on the posttest and delayed posttest while controlling for pretest. The model predicting posttest performance was significant, F (2, 595) = 561.90, p < .001, as were both pretest (β = 0.49, p < .001) and confrustion (β = −0.37, p < .001) within the model. Similarly, the model predicting delayed posttest performance was significant, F (2, 595) = 575.49, p < .001, as were both pretest (β = 0.52, p < .001) and confrustion (β = −0.35, p < .001). In other words, confrustion was associated with lower posttest and delayed posttest performance even after controlling for pretest.

8.2. Do the groups differ in their levels of confrustion while working through the instructional materials?

A one-way analysis of variance (ANOVA) revealed a large effect of condition on confrustion, F (1, 596) = 43.00, p < .001, d = 0.54, with students in the ErrEx condition (M = 0.34, SD = 0.16) experiencing a significantly higher level of confrustion than students in the PS condition (M = 0.25, SD = 0.16). We also conducted t-tests to compare the longest period of time, measured in seconds and in number of problems, that students experienced confrustion across the two conditions. Students in the ErrEx condition tended to have longer episodes of confrustion, whether measured in number of problems F (1, 596) = 10.67, p < .005, d = 0.27, or in seconds of time, F (1, 596) = 23.41, p < .001, d = 0.40. These measures indicate that although students in the ErrEx condition performed better on the delayed posttest, they also experienced more confrustion, which was negatively correlated with posttest and delayed posttest performance. To investigate these seemingly contradictory results, we examined interactions between confrustion and condition.

Condition was tested as a moderator of the relation between confrustion and posttest performance using a PROCESS 1 moderation model to predict test scores (Hayes, 2013). PROCESS is an SPSS macro that tests mediation and moderation using 5000 bootstrap estimates to create confidence intervals for indirect effects. Results indicated that the interaction between condition and confrustion was a significant predictor of posttest performance, B = .88, 95% CI [0.05, 0.31], and the inclusion of the interaction term explained significantly more variance in the model, ΔR^2 = 0.005, F (1, 594) = 7.36, p = .007. As shown in Fig. 4, while confrustion was negatively related to performance in both conditions, an increase in confrustion had less of a negative impact on posttest performance for students in the ErrEx condition than students in the PS condition.

A similar interaction effect between condition and confrustion occurred when predicting delayed posttest using the same moderation model. Results indicated that the interaction between condition and confrustion was significant, B = 0.17, 95% CI [0.05, 0.29], and the inclusion of the interaction term explained significantly more variance in the model, ΔR^2 = 0.011, F (1, 594) = 8.24, p = .004. As shown in Fig. 5, an increase in confrustion again had less of a negative impact on delayed posttest performance for students in the ErrEx condition than students in the PS condition.

A similar picture emerged when examining the mean duration of confrustion. We calculated duration by first identifying all

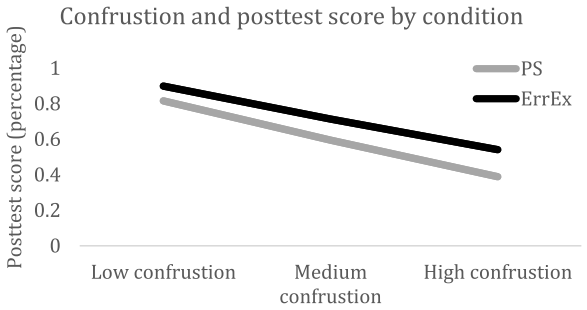

Fig. 4. Interaction of confrustion and condition predicting posttest score. Test scores are calculated across conditions using the regression equation for low (16th percentile), medium (50th percentile), and high (84th percentile) values of confrustion.

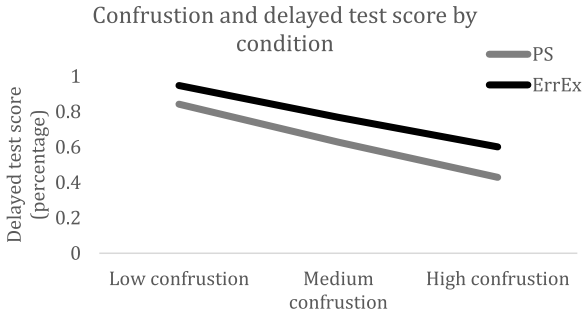

Fig. 5. Interaction of confrustion and condition predicting posttest score. Test scores are calculated across conditions using the regression equation for low (16th percentile), medium (50th percentile), and high (84th percentile) values of confrustion.

occasions where students were confrusted on one or more problems in a row, based on a confrustion probability rate equal to or greater than 0.5. Sequences of confrustion were broken up by any problems on which students were not confrusted, i.e., probability rate less than .5. We then calculated times for each confrustion sequence by summing the time spent on each problem within the sequence. Confrustion sequence times were averaged for each student across all their sequences of confrustion.We removed 19 students who had average confrustion durations greater than two standard deviations above the mean ($M = 166.78$, $SD = 160.21$). Students in the erroneous examples condition ($M = 163.88$, $SD = 82.01$) had significantly longer confrustion durations than students in the problem solving condition ($M = 128.89$, $SD = 92.73$), F (1, 576) = 22.92, $p < .001$. Across both conditions, confrustion duration was negatively correlated with pretest, $r = -0.52$, posttest, $r = -0.52$, and delayed posttest, $r = -0.48$. Unlike with probability of confrustion, however, a moderation analysis showed no interaction between confrustion duration and condition when predicting posttest, $b = -0.0001$, $p = .71$, 95% CI [-0.0004, 0.0003], or delayed posttest, $b = -0.0001$, $p = .66$, 95% CI [-0.0004, 0.0002].

There was a significant effect of gender on confrustion, with female students ($M = 0.31$, $SD = 0.17$) experiencing higher levels of confrustion than male students ($M = 0.27$, $SD = 0.17$), F (1, 596) = 5.37, $p = .021$, $d = 0.219$. Given that there was also a significant effect of gender on pretest scores, F (1, 596) = 13.44, $p < .001$, $d = 0.30$, with female students ($M = 0.54$, $SD = 0.21$) receiving lower pretest scores than male students ($M = 0.61$, $SD = 0.22$), we conducted an ANCOVA to assess the effect of gender on confrustion when controlling for pretest scores. While the pretest covariate was significant, F (1, 595) = 822.33, $p < .001$, $\eta_p^2 = 0.55$, gender was not, F (1, 595) = 0.34, $p = .56$, $\eta_p^2 = 0.001$, suggesting the apparent effect of gender on confrustion was a product of female students' lower pretest scores. Moderation analyses in PROCESS (Hayes, 2013) showed no significant interaction between gender and condition when predicting confrustion, $b = 0.012$, $p = .63$, 95% CI [-0.04, 0.06].

8.3. Do confrustion levels differ based on the misconception targeted?

To test the hypothesis that students' levels of confrustion would differ between conditions based on the misconceptions targeted by different items, we conducted a mixed ANOVA that included the between-subjects variable of condition (PS or ErrEx) and the within-subjects variable of targeted misconception (Megz, Segz, Pegz, and Negz, as described in Table 1). A violation of the sphericity assumption is considered a serious problem that increases Type 1 error rate in mixed ANOVAs. Mauchly's sphericity test indicated that the main effect of problem type did not violate the sphericity assumption, $W = 0.98$, χ^2 (5) = 10.98, $p = .052$, Greenhouse-Geisser $\varepsilon = 0.99$, meaning a mixed ANOVA was appropriate for these data. Results revealed a significant effect of misconception type, F (3, 1788) = 41.68, $p < .001$, $\eta_p^2 = 0.065$, and a significant interaction between misconception type and condition, F (3, 1788) = 20.45, $p < .001$, $\eta_p^2 = 0.033$. This indicates that students in the ErrEx and PS conditions differed in how their confrustion varied across misconception types (Fig. 6).

To understand this interaction, we conducted pairwise comparisons between conditions on each misconception type, as well as

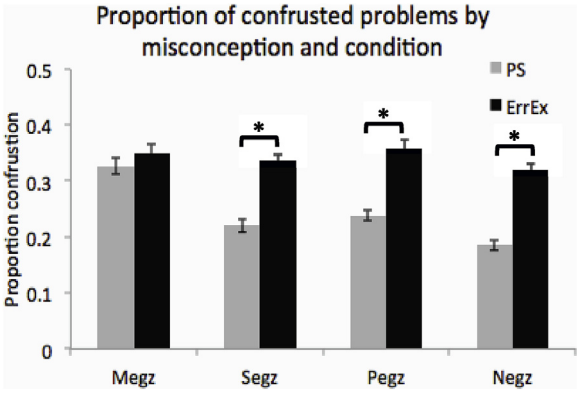

Fig. 6. Proportion of instructional problems on which students experienced confrustion, by misconception type. * denotes a significant difference between instructional conditions for each misconception.

pairwise comparisons between misconception types for each condition separately (Table 5). We applied Benjamini and Hochberg's (1995) false discovery rate post-hoc procedure, using a false discovery rate of 0.05. Benjamini and Hochberg's method controls for false positives due to conducting multiple comparisons (Type I error) while avoiding the considerable over-conservatism seen for familywise error rate methods like the Bonferonni correction (Type II error). Pairwise comparisons revealed significant differences in levels of confrustion between conditions on Segz, Pegz, and Negz misconceptions, with students in the ErrEx condition experiencing more confrustion than students in the PS condition. However, there was no difference in confrustion between conditions on Megz problems.

8.4. Do confrustion levels decline as students work through the materials?

To test the hypothesis that students' levels of confrustion would change across the course of the intervention in different ways between conditions, we conducted an ANCOVA that included the between-subjects variable of condition (PS or ErrEx) and the within-subjects variable of problem number. Results indicated a significant effect of problem number, F (3, 27334) = 210.85, $p < .001$, $\eta_p^2 = 0.0075$, condition F (3, 27334) = 37.64, $p < .001$, $\eta_p^2 = 0.0013$, and a significant interaction, F (3, 27334) = 7.97, $p = .0047$, $\eta_p^2 = 0.0003$. The significant effect of problem number indicated that students tended to experience less confrustion later in the materials, and the significant interaction effect suggests that students in the ErrEx and PS conditions differed in how their confrustion varied across the course of the intervention (Fig. 7).

9. Discussion

9.1. Main findings

This paper examined confrustion, a combined measure of confusion and frustration inferred through students' behaviors in a computer-based tutoring system, as a possible mechanism to explain how students learn from erroneous examples. Prior research has shown that studying incorrect worked examples can be more beneficial than problem-solving practice or studying correct worked examples, possibly by highlighting common student errors and prompting students to understand why the demonstrated errors are incorrect (Siegler, 2002, pp. 31–58). This research aimed to examine confrustion as a possible mechanism to explain the greater learning gains experienced by students who studies erroneous examples.

As predicted, students in the erroneous example condition experienced greater confrustion than students in the problem-solving condition and, as previously reported, also showed greater performance gains on the delayed posttest (Adams et al., 2014; McLaren et al., 2015). Contrary to predictions, however, confrustion was associated with worse test performance. These two results appear to

Table 5
Condition effects on average confrustion levels by misconception problem type.

Type	PS *M* (*SD*)	ErrEx *M* (*SD*)	*t*-test
Megz	.33 (.24)	.35 (.21)	t (595.58) = 1.10, $p = .27$
Segz	.24 (.21)	.35 (.19)	t (594.42) = 6.69, $p < .001$*
Pegz	.23 (.17)	.33 (.17)	t (596) = 7.34, $p < .001$*
Negz	.20 (.17)	.32 (.20)	t (567.10) = 8.31, $p < .001$*

Note: Levene's test for equality of variances was rejected for Megz, $F = 7.48$, $p = .006$, Segz, $F = 6.41$, $p = .012$ and Negz, $F = 5.72$, $p = .017$. Therefore equal variances were not assumed for *t*-tests on these misconception types. * denotes significant effect based on Benjamini & Hochberg's post-hoc control.

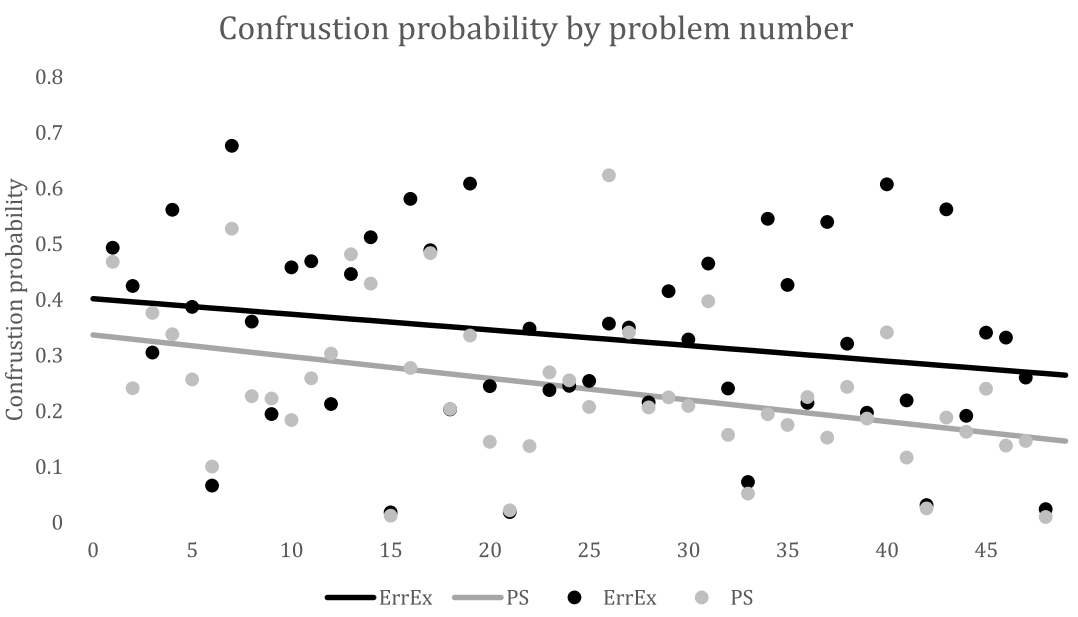

Fig. 7. Probability of detected confrustion across the problem set, divided by condition.

be contradictory: students in the ErrEx condition experienced more confrustion, which was associated with worse posttest and delayed posttest performance, but also performed better on the delayed posttest. The negative relation between confrustion and test performance remained significant even when controlling for pretest, suggesting that confrustion was not simply capturing the degree to which students already understood decimal number concepts at the beginning of the experiment. Consistent with prior research, students who experienced longer durations of confrustion had worse outcomes and students who experienced shorter durations of confrustion had better outcomes (Liu et al., 2013). Students in the erroneous example condition also experienced longer durations of confrustion, which contradicted our prediction that confrustion would be resolved more quickly in the erroneous example condition. However, given that confrustion was assessed at the problem-level, duration had to be calculated at that level as well, in this case as the average time across consecutive problems in which students experienced confrustion. If confrustion could be calculated at a more fine-grained level, we could assess whether students in the erroneous example condition resolved confrustion more quickly within individual problems.

The absence of condition effects on immediate posttest is consistent with distinctions between performance and learning (Kapur, 2016; Soderstrom & Bjork, 2015) and between near and far transfer, which can be distinguished based on content or context, including temporal delays (Barnett & Ceci, 2002). Barnett and Ceci (2002) identify nine dimensions of transfer, including time. Consistent with our results, Kapur (2016) argues that rote practice, which is most similar to the PS condition, is best suited for improving performance but often falls short on fostering learning. On the other hand, he suggests that tasks that encourage students to struggle with what they do not know or understand can, over the long run, promote greater learning. While our use of erroneous examples differs from the "productive failure" paradigm Kapur has studied (Kapur, 2014; Kapur & Bielaczyc, 2012), both approaches predict long-term learning gains from interventions aimed at helping students to recognize and wrestle with what they do not understand. Research on desirable difficulties has also suggested that students can reap long-term learning benefits from more challenging materials, as long as the student has sufficient knowledge or support to overcome the challenge (Bjork & Bjork, 2011; Schmidt & Bjork, 1992). Confrustion may be viewed as an intrinsic cost of engaging in a more challenging learning activity, but one that eventually pays dividends.

A moderation analysis showed that greater confrustion was associated with a *smaller* drop in performance in the ErrEx condition, compared to the PS condition. In other words, the link between confrustion and learning outcomes was weaker in the ErrEx condition, particularly when students were experiencing high levels of confrustion; the intentionally engineered confrustion in the ErrEx condition (cf. Lehman et al., 2013) may have been more beneficial than the less intentional confrustion in the PS condition. The different relations between learning and confrustion across conditions could be explained by confrustion coming from different sources in the different conditions. Specifically, in the PS condition, confrustion most likely came from a student not knowing how to solve problem. A student's main focus in this condition was on solving the problem, and so confrustion was likely to increase with each attempt until they reached the correct solution. In contrast, students in the ErrEx condition may have experienced confrustion primarily through the process of making sense of errors, rather than from seeking the correct answer. Making sense of errors might lead to more confrustion than problem solving, but prior research suggests it is also a potentially more productive process and thus, perhaps, a more productive form of confrustion (Booth et al., 2013; Groβe & Renkl, 2007; Siegler, 2002, pp. 31–58). Better understanding the sources of confrustion is an essential step toward understanding and eventually developing methods for optimizing it.

The past literature on these constructs within learning has not considered its sources, treating confusion from one source as representative of all confusion (e.g. Lehman et al., 2013; D'Mello et al., 2014), or considering the overall proportion, incidence, or duration of confusion or frustration without differentiating based on its source (Liu et al., 2013; Rodrigo et al., 2009; Schneider et al., 2015).

The pattern of students experiencing more confrustion on ErrEx materials compared to PS materials held for learning items targeting the Segz, Pegz, and Negz misconceptions, but not the Megz misconception. While confrustion levels for students in the ErrEx condition were relatively consistent across all four misconception types, students in the PS condition experienced notably more confrustion on the Megz problems than they did on other problems. Megz ("longer decimals are larger") is identified in prior work as the most common misconception for students similar to those in our sample (Sackur-Grisvard & Léonard, 1985; see review in; Isotani et al., 2011), which could explain why students in the PS condition encountered greater levels of confrustion when trying to solve Megz problems. However, Megz is also one of the first decimal number misconceptions that students resolve, typically around the age of our sample (Resnick et al., 1989; Steinle, 2004). Perhaps for this reason, it is also the one for which all students on average showed the greatest learning gains from pre-to posttest and pre-to delayed posttest. The unusual nature of the Megz misconception—common, yet easier to resolve—may explain why students' experiences of confrustion were more similar across conditions on the Megz problems than others. Students in the PS condition may have been more likely to experience confrustion as a result of the high frequency of Megz misconcepions, but they may have also been better equipped to reach a correct solution and avoid a long duration of confrustion, which is when confusion and frustration become harmful to learning (D'Mello & Graesser, 2012; Liu et al., 2013). Similarly, students in the ErrEx condition may have experienced confrustion as they engaged deeply in understanding and generalizing the common misconception, but this process may have been particularly beneficial for Megz items, as supported by the greater learning gains.

9.2. Limitations and future directions

Students were assigned to conditions at the individual level, creating the potential limitation that individuals sitting next to one another could receive different materials and potentially notice the different activities on neighboring computer screens. However, we think it is unlikely that students would experience any learning benefits from a neighbor's activities, and the interfaces were similar enough across conditions that neither set of materials appeared particularly more motivating than the other. We considered any potential risk of exposure to the other condition to be outweighed by the benefits of randomizing at the individual level, which allowed us to control for cohort factors that might arise from different classes (e.g., ability level, time of day).

Another potential limitation to the interpretation of our findings stems from the multiple sources of potential confrustion in the ErrEx condition. Participants who held the misconception demonstrated in the erroneous example might have experienced confrustion because they believed the answer to be correct, while students who did not hold the misconception might have experienced confrustion because they did not understand how anyone could think the erroneous example was correct. Pretest scores indicated that nearly all students made at least some errors consistent with all misconceptions, so it is unlikely that many students viewed all erroneous examples as obviously—and perhaps confrustingly—incorrect. Nevertheless, future research might actively track the misconceptions students hold at a given point in the intervention and either customize the erroneous examples to match their misconceptions or empirically test the effects of seeing erroneous examples that illustrate misconceptions the student either does or does not hold.

Finally, while previous research has shown the value in predicting performance through affective states, some studies suggest that transitions between affective states can be even better predictors of student learning (Lee et al., 2011; Liu et al., 2013). The nature of the materials (e.g., no hints) and current evidence regarding affect detection using text-replay coding led us to focus in the current study on confrustion, which we considered to be the most relevant affective state for understanding learning from erroneous examples. However, additional analyses using inferences of other learner-centered affective states, such as engaged concentration and boredom, might also shed light on the apparent contradiction of confrustion. It may also be relevant to replicate prior research regarding the length of time students spend in a state of confrustion within individual problems (D'Mello & Graesser, 2012; Liu et al., 2013). Additionally, the tutor could be modified to intervene with hints when students demonstrate persistent levels of confrustion, potentially reducing any extraneous confrustion caused by the task of understanding the erroneous examples. In general, confrustion seems to accompany successful learning within erroneous examples. Explicitly regulating it through a combination of inducing it (through erroneous examples and other strategies – e.g. Lehman et al., 2013) to produce deep cognitive engagement with complex learning material, and providing assistance to help students get past their confrustion when it persists for too long, may help to optimize student learning over time. For example, by embedding automated detectors of confrustion and knowledge in a learning system, it could be possible to find students who are not making progress, yet are also not experiencing confrustion, and give them an erroneous example that induces confrustion. Correspondingly, it would also be possible to detect confrustion that persists for more than a couple minutes, and automatically pop up a hint message or other learning support.

An important step for revising materials is to identify the reasons that students in the ErrEx condition experienced more confrustion. One reason might be that the process of studying an incorrect example before correctly solving a problem is not intuitive to students and they were confrusted by the activity itself. To the degree that confrustion arose from the confusing interface or engagement in an unexpected activity, we would expect to see confrustion decline more rapidly in the more novel ErrEx condition. Although students in both the ErrEx and PS conditions experienced less confrustion as they worked through the materials, students in the ErrEx condition, in addition to experiencing more confrustion initially, saw their confrustion decline more slowly. This suggests that the greater levels of confrustion in the ErrEx condition were not simply a product of the novelty of the task or interface, providing

tentative support for alternative explanations like those mentioned above.

Future work should examine whether confrustion can be reduced in the ErrEx condition without eliminating the learning benefits of studying erroneous examples. Research on cognitive load has identified "extraneous load," which is not essential to the task itself but consumes cognitive resources and negatively affects learning (e.g., having to switch between multiple representations in different locations; Paas, Renkl, & Sweller, 2003). This is contrasted with "germane load," which is also not an intrinsic part of the task but which promotes greater learning (e.g., having to retrieve prior knowledge and connect it to a new problem). It may be that there are analogous sources of extraneous and germane confrustion. If the higher levels of confrustion experienced by ErrEx students included significant levels of extraneous confrustion, then reducing the sources of that extraneous confrustion could further enhance learning. For example, if some of the greater confrustion caused by erroneous examples is a result of the novelty of the task, a brief tutorial administered before the learning materials might reduce differences in confrustion between conditions.

10. Conclusion

This research investigated underlying mechanisms that might explain why erroneous examples lead to greater learning. We leveraged student data collected in an educational technology platform and educational data mining to examine confrustion—a combination of confusion and frustration—as a potential factor in understanding differences in learning outcomes between students who studied erroneous examples and students who completed more traditional problem-solving practice. We hypothesized that students in the erroneous example condition would experience greater confrustion, and that confrustion in the erroneous example condition would be beneficial to learning.

While our results do not support the conclusion that students learn more from erroneous examples *because* of greater confrustion, they indicate that affect detectors have predictive value when examining learning from erroneous examples. This paper reports measurements of confrustion at the problem level, and therefore cannot identify the precise sources of confrustion within individual problems. Future research may be able to apply a more fine-grained analysis for understanding which components of erroneous examples are most responsible for students' higher levels of confrustion, which might also help distinguish between extraneous confrustion (e.g., confrustion caused by the particular format of the erroneous examples) and the confrustion that necessarily results from wrestling with incorrect examples that represent common student misconceptions. These steps would further advance theoretical understanding of learning from erroneous examples while also demonstrating the potential for affect detection to shape researchers' and teachers' efforts to create effective, responsible learning materials.

One general recommendation that can be drawn from this paper is that creating the type of data logging available in the system studied here can be a powerful tool for understanding learning better. By logging every student action in a fine-grained fashion, it was possible not only to study performance on specific skills over time, but also to conduct retrospective analyses on affect that were not envisioned at the initial time of data collection. It is especially useful if – as seen here – specific items are tagged with the skills relevant to them, enabling skill-level analyses of performance over time. Developers building systems of this nature should also incorporate careful step-level tagging, recording of actual student responses, and retention of timing and multiple attempt data. These types of data, still sometimes deleted in the interest of saving bandwidth, are essential to the type of analysis conducted here.

Overall, by better understanding the role of confrustion, and affect in general, in learning from erroneous examples and problem-solving, we can develop next-generation, affect-appropriate learning technologies that use these methods to improve learning outcomes and create more positive affective experiences for learners.

References

Adams, D., McLaren, B. M., Durkin, K., Mayer, R. E., Rittle-Johnson, B., Isotani, S., et al. (2014). Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior, 36*, 401–411. https://doi.org/10.1016/j.chb.2014.03.053.

Aleven, V., McLaren, B. M., & Sewall, J. (2009). Scaling up programming by demonstration for intelligent tutoring systems development: An open-access website for middle school mathematics learning. *IEEE Transactions on Learning Technologies, 2*(2), 64–78.

Aleven, V., McLaren, B. M., Sewall, J., van Velsen, M., Popescu, O., Demi, S., et al. (2016). Example-tracing tutors: Intelligent tutor development for non-programmers. *International Journal of Artificial Intelligence in Education, 26*(1), 224–269. https://doi.org/10.1007/s40593-015-0088-2.

Andres, J. M. L., & Rodrigo, M. M. T. (2014). The Incidence and persistence of affective states while playing Newton's playground. *7th IEEE international conference on humanoid, nanotechnology, information technology, communication and control, environment, and management*.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70*(2), 181–214.

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology, 95*(4), 774–783.

Baker, R. S. J.d., Corbett, A. T., & Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. *Proceedings of the educational data mining workshop at the 8th international conference on intelligent tutoring systems* (pp. 29–36). .

Baker, R. S. J.d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to Be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*(4), 223–241.

Baker, R. S. J.d., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., et al. (2012). Sensor-free automated detection of affect in a cognitive tutor for algebra. *Proceedings of the 5th international conference on educational data mining* (pp. 126–133). .

Baker, R. S. J.d., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson, & B. White (Eds.). *Learning analytics: From research to practice*. Berlin, Germany: Springer.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612.

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, 57*, 289–300.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W.

Pew, L. M. Hough, & J. R. Pomerantz (Eds.). *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York: Worth Publishers.
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. (1987). Exploring mathematics through the analysis of errors. *For the Learning of Mathematics, 7*(3), 2–8.
Borasi, R. (1994). Capitalizing on errors as "springboards for inquiry": A teaching experiment. *Journal for Research in Mathematics Education,* 166–208.
Botelho, A. F., Baker, R., Ocumpaugh, J., & Heffernan, N. (2018). Studying affect dynamics and chronometry using sensor-free detectors. *Proceedings of the 11th international conference on educational data mining* (pp. 157–166). .
Brown, D. E. (1992). Using examples and analogies to remediate misconceptions in physics: Factors influencing conceptual change. *Journal of Research in Science Teaching, 29*(1), 17–34.
Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18–37.
Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). ACM.
Chi, M. T. H. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.). *Handbook of research on conceptual change* (pp. 61–82). Hillsdale, NJ: Erlbaum.
Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73–105.
Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(2), 145–182.
Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477.
DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., et al. (2018). Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education, 28*(2), 152–193.
Desmet, L., Gregoire, J., & Mussolin, C. (2010). Developmental changes in the comparison of decimal fractions. *Learning and Instruction, 20*, 521–532. https://doi.org/10.1016/j.learninstruc.2009.07.004.
Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction, 22*(3), 206–214.
D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology, 105*(4), 1082–1099.
D'Mello, S. K., Craig, S. D., Witherspoon, A. W., McDaniel, B. T., & Graesser, A. C. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction, 18*(1 – 2), 45–80.
D'Mello, S. K., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction, 20*(2), 147–187. https://doi.org/10.1007/s11257-010-9074-4.
D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157.
D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction, 29*, 153–170.
Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist, 46*(1), 6–25.
Gadgil, S., Nokes-Malach, T. J., & Chi, M. T. H. (2012). Effectiveness of holistic mental model confrontation in driving conceptual change. *Learning and Instruction, 22*(1), 47–61.
Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2013). Automatically recognizing facial expression: Predicting engagement and frustration. *Proceedings of the international conference on educational data mining (EDM)* (pp. 43–50). .
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.
Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* New York, NY, US: Guilford Press.
Heitzmann, N., Fischer, F., & Fischer, M. R. (2018). Worked examples with errors: When self-explanation prompts hinder learning of teachers diagnostic competences on problem-based learning. *Instructional Science, 46*(2), 245–271.
Hiebert, J., & Wearne, D. (1985). A model of students' decimal computation procedures. *Cognition and Instruction, 2*, 175–205.
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? *The Proceedings of the sixth European Conference on technology enhanced learning: Towards Ubiquitous learning (EC-TEL 2011)* (pp. 181–195). .
Johnson, C. I., & Mayer, R. E. (2010). Adding the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior, 26*, 1246–1252.
Kai, S., Almeda, M. V., Baker, R., Heffernan, C., & Heffernan, N. (2018). Decision tree modeling of wheel-spinning and productive persistence in skill builders. *JEDM | Journal of Educational Data Mining, 10*(1), 36–71. Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM/article/view/210.
Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93*, 579–588.
Kapur, M. (2014). Productive failure in learning math. *Cognitive Science, 38*(5), 1008–1022.
Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist, 51*(2), 289–299.
Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *The Journal of the Learning Sciences, 21*(1), 45–83.
Koedinger, K. R., Baker, R. S. J.d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J.d. Baker (Eds.). *Handbook of educational data mining.* Boca Raton, FL: CRC Press.
Koedinger, K. R., Stamper, J. C., Leber, B., & Skogsholm, A. (2013). LearnLab's DataShop: A data repository and analytics tool set for cognitive science. *Topics in Cognitive Science, 5*(3), 668–669.
Kostyuk, V., Almeda, M. V., & Baker, R. S. (2018). Correlating affect and behavior in reasoning mind with state test achievement. *Proceedings of the international conference on learning analytics and knowledge* (pp. 26–30). .
Lee, D. M. C., Rodrigo, M. M. T., Baker, R. S. J.d., Sugay, J. O., & Coronel, A. (2011). Exploring the relationship between novice programmer confusion and achievement. *Proceeds of the 4th bi-annual Conference on affective Computing and intelligent interaction, 2011*.
Lehman, B., D'Mello, S., Strain, A., Mills, C., Gross, M., Dobbins, A., ... Graesser, A. (2013). Inducing and tracking confusion with contradictions during complex learning. *International Journal of Artificial Intelligence in Education, 22*(1–2), 85–105.
Limón, M. (2001). On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learning and Instruction, 11*(4–5), 357–380.
Liu, Z., Pataranutaporn, V., Ocumpaugh, J., & Baker, R. S. J.d. (2013). Sequences of frustration and confusion, and learning. *Proceedings of the 6th international conference on educational data mining* (pp. 114–120). .
Mayer, R. E., & Johnson, C. I. (2010). Adding instructional features that promote learning in a game-like environment. *Journal of Educational Computing Research, 42*, 241–265.
McLaren, B. M., Adams, D. M., & Mayer, R. E. (2015). Delayed learning effects with erroneous examples: A study of learning decimals with a web-based tutor. *International Journal of Artificial Intelligence in Education, 25*(4), 520–542.
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* (pp. 2176–2181). Austin, TX: Cognitive Science Society.
McLaren, B. M., van Gog, T., Ganoe, C., Karabinos, M., & Yaron, D. (2016). The efficiency of worked examples compared to erroneous examples, tutored problem solving, and problem solving in classroom experiments. *Computers in Human Behavior, 55*, 87–99. https://doi.org/10.1016/j.chb.2015.08.038.
McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38*(1), 1–30.

Melis, E. (2004). Erroneous examples as a source of learning in mathematics. *CELDA, 2004*, 311–318.
Muldner, K., Burleson, B., & VanLehn, K. (2010). "Yes!": Using tutor and sensor data to predict moments of delight during instructional activities. *Proceedings of the international conference on user modeling and adaptive presentation (UMAP'10)* (pp. 159–170). .
Ocumpaugh, J., Baker, R. S., & Rodrigo, M. M. T. (2015). *Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual*Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4.
Paquette, L., Baker, R. S., Sao Pedro, M. A., Gobert, J. D., Rossi, L., Nakama, A., et al. (2014, June). Sensor-free affect detection for a simulation-based science inquiry learning environment. *International conference on intelligent tutoring systems* (pp. 1–10). Cham: Springer.
Paquette, L., de Carvalho, A. M. J. A., & Baker, R. S. (2014). Towards understanding expert coding of student disengagement in online learning. *Proceedings of the 36th annual cognitive science conference* (pp. 1126–1131). .
Pardos, Z. A., Baker, R. S., San Pedro, M., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics, 1*(1), 107–128.
Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36*(1), 36–48.
Pekrun, R., Goetz, T., Titz, W., & Perry, R. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37*(2), 91–105. https://doi.org/10.1207/s15326985ep3702_4.
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. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*(1), 1–29.
Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanations. *Learning and Instruction, 12*(5), 529–556.
Renkl, A., & Atkinson, R. K. (2002). Learning from examples: Fostering self-explanations in computer-based learning environments. *Interactive Learning Environments, 10*(2), 105–119. https://doi.org/10.1076/ilee.10.2.105.7441.
Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist, 38*(1), 15–22. https://doi.org/10.1207/S15326985EP3801_3.
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,* 8–27.
Rittle-Johnson, B., & Star, J. R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology, 101*(3), 529.
Rodrigo, M. M. T., Baker, R. S., Jadud, M. C., Amarra, A. C. M., Dy, T., Espejo-Lahoz, M. B. V., ... Tabanao, E. S. (2009). Affective and behavioral predictors of novice programmer achievement. *Proceedings of the ACM SIGCSE annual conference on innovation and technology in computer science education: Vol. 41*, (pp. 156–160). New York, NY: ACM Press.
Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education,* (1–2), 115–133. https://doi.org/10.3233/JAI-2011-019.
Rushton, S. J. (2018). Teaching and learning mathematics through error analysis. *Fields Mathematics Education Journal, 3*(1), 4.
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.
San Pedro, M. O. Z., Baker, R. S. J.d., Bowers, A. J., & Heffernan, N. T. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. *Proceedings of the 6th international conference on educational data mining* (pp. 177–184). .
Sao Pedro, M. A., de Baker, R. S., Gobert, J. D., Montalvo, O., & Nakama, A. (2013). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction, 23*(1), 1–39.
Schmidt, R. A., & Bjork, R. A. (1992). New conceptualization of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207–217.
Schneider, B., Krajcik, J., Lavonen, J., Salmela-Aro, K., Broda, M., Spicer, J., et al. (2015). Investigating optimal learning moments in U.S. and Finnish science classes. *Journal of Research in Science Teaching, 53*(3), 400–421. https://doi.org/10.1002/tea.21306.
Shute, V. J., D'Mello, S., Baker, R., Cho, K., Bosch, N., Ocumpaugh, J., et al. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education, 86*, 224–235. https://doi.org/10.1016/j.compedu.2015.08.001.
Siegler, R. S. (2002). *Microgenetic studies of self-explanation. Microdevelopment: Transition processes in development and learning*.
Siegler, R. S., & Chen, Z. (2008). Differentiation and integration: Guiding principles for analyzing cognitive change. *Developmental Science, 11*(4), 433–448.
Sinatra, G. M., & Pintrich, P. R. (Eds.). (2003). *Intentional conceptual change*. Routledge.
Sinha, T., Bai, Z., & Cassell, J. (2017). A new theoretical framework for curiosity for learning in social contexts. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.). *Proceedings of 12th European Conference on technology enhanced learning (EC-TEL '17)* (pp. 254–269). Cham: Springer.
Smith, J. P., III, DiSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115–163.
Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199.
Stacey, K. (2005, July). Travelling the road to expertise: A longitudinal study of learning. *Proceedings of the 29th conference of the international group for the psychology of mathematics education: Vol. 1*, (pp. 19–36). Melbourne: PME.
Stacey, K., Helme, S., & Steinle, V. (2001). Confusions between decimals, fractions and negative numbers: A consequence of the mirror as a conceptual metaphor in three different ways. In M. v. d. Heuvel-Panhuizen (Vol. Ed.), *Proceedings of the 25th conference of the international group for the psychology of mathematics education: Vol. 4*, (pp. 217–224). Utrecht: PME.
Stacey, K., & Steinle, V. (1998). Refining the classification of students' interpretations of decimal notation. *Hiroshima Journal of Mathematics Education, 6*, 49–69.
Steinle, V. (2004). *Changes with age in students' misconceptions of decimal numbers*PhD thesis. Department of Science and Mathematics Education, The University of Melbourne.
Steinle, V., & Stacey, K. (2004). Persistence of decimal misconceptions and readiness to move to expertise. In M. J. Hoines, & A. B. Fuglestad (Vol. Eds.), *Proceedings of the 28th conference of the international group for the psychology of mathematics education: Vol. 1*, (pp. 225–232). Bergen-Norway: Bergen University College.
Trafton, J. G., & Reiser, B. J. (1993). The contribution of studying examples and solving problems to skill acquisition. *Proceedings of the 15th annual conference of the cognitive science society* (pp. 1017–1022). Hillsdale, NJ: Erlbaum.
Tsovaltzi, D., Melis, E., & McLaren, B. M. (2012). Erroneous examples: Effects on learning fractions in a web-based setting. *International Journal of Technology Enhanced Learning, 4*(3–4), 191–230.
Van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology, 36*(3), 212–218.
Vosniadou, S. (Ed.). (2009). *International handbook of research on conceptual change*. Routledge.
Vosniadou, S. (2012). Reframing the classical approach to conceptual change: Preconceptions, misconceptions and synthetic models. *Second international handbook of science education*. Dordrecht: Springer.
Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7*(1), 1–39.
Wittwer, J., & Renkl, A. (2010). How effective are instructional explanations in example-based learning? A meta-analytic review. *Educational Psychology Review, 22*(4), 393–409.
Wu, C. H., Huang, Y. M., & Hwang, J. P. (2015). Review of affective computing in education/learning: Trends and challenges. *British Journal of Educational Technology, 47*(6), 1304–1323. https://doi.org/10.1111/bjet.12324.