Erroneous examples: effects on learning fractions in a web-based setting

Dimitra Tsovaltzi*

Department of Educational Technology,
Saalrland University,
Building C5.4, D-66123 Saarbrücken, Germany
Email: Dimitra.tsovaltzi@mx.uni-saarland.de
*Corresponding author

Bruce M. McLaren

Center for e-Learning Technology (CeLTech),
Saalrland University,
Building C5.4, D-66123 Saarbrücken, Germany
Email: bmclaren@cs.cmu.edu

Erica Melis

Deceased; formerly of: German Research Centre for Artificial Intelligence

Ann-Kristin Meyer

DFKI GmbH,
German Research Centre for Artificial Intelligence,
Stuhlsatzenhausweg 3 (Building D3 2),
D-66123 Saarbrücken, Germany
Email: annkristinmeyer@gmx.de

Abstract: Learning from errors can be a key 21st century competence, especially for informal learning where such metacognitive skills are a prerequisite. We investigate whether, how and when web-based interactive erroneous examples promote such competence, and increase understanding of fractions and learning outcomes. *Erroneous examples* present students with common errors or misconceptions. Three studies were conducted with students of different grade levels. We compared the cognitive, metacognitive, conceptual, and transfer learning outcomes of three conditions: a control condition (problem solving), a condition that learned with erroneous examples without help, and a condition that learned with erroneous examples with error detection and correction support. Our results indicate significant metacognitive learning gains of erroneous examples with help for 6th-graders. They also show cognitive and conceptual learning gains for 9th and 10th-graders when additional help is provided. No effects were found for 7th-graders. We discuss the implications of our findings for instructional design.

Keywords: erroneous examples; learning from errors; empirical studies; fractions misconceptions; adaptive learning; conceptual learning; metacognition; learner support.

Reference to this paper should be made as follows: Tsovaltzi, D., McLaren, B.M., Melis, E. and Meyer, A-K. (2012) 'Erroneous examples: effects on learning fractions in a web-based setting', *Int. J. Technology Enhanced Learning*, Vol. 4, Nos. 3/4, pp.191–230.

Biographical notes: Dimitra Tsovaltzi is faculty member at Saarland University, Department of Educational Technology and an affiliated researcher of the DFKI in Saarbrücken, Germany. She received her PhD in Educational Technology at Saarland University, Department of Informatics. The focus of her research is the design, empirical investigation and automation of learner support for technology enhanced learning. She has investigated learning from errors in online environments, natural language hinting for mathematics, computer-supported collaborative scripting, and is currently working on learning in social networks. She has published over 30 publications in the main international conferences in e-learning (ITS, EC-TEL) and journals (*IJTEL*).

Bruce McLaren is a senior systems scientist at Carnegie Mellon University, USA, and has an adjunct position as a principal researcher within the Center for e-Learning Technology (CeLTech) at Saarland University, Germany. He has interests in intelligent tutoring systems, e-learning principles, and collaborative learning. For instance, he has done work on developing educational technology using AI techniques to help teachers moderate collaborative e-Discussions and online arguments. He has over 100 publications spanning peer-reviewed journals, conferences, workshops, symposiums and book chapters. According to Microsoft Academic search, he is one of the top-ranked authors in the world in computers and education.

Erica Melis established the Intelligent Learning Environments Group as a principal researcher at DFKI. She served as a member of numerous international conference committees including 'Intelligent Tutoring Systems' and 'AI in Education'. She has pioneered research on analogy, automated proof planning and intelligent, web-based learning environments. She received her PhD from Humboldt-University, Berlin and was a research fellow at Carnegie Mellon University (Pittsburgh, USA) and at University of Edinburgh (UK) later. Her research interests include knowledge representation, adaptive systems, user modelling, support of meta-reasoning, data mining and interdisciplinary work with cognitive psychologists on effective support of human learning.

Ann-Kristin Meyer received her BA in Psychology from Saarland University in 2012. Between 2008 and 2012, she worked as student assistant at the DFKI Saarbrücken in the CCeL (department of e-learning) and in the department of Educational Technology, Saarland University, in the area of learning from erroneous examples. She contributed to several papers and posters in the main international conferences in e-learning. She is now working as a clinical psychologist in Eichenberg Institut Koblenz, Gesundheitsmanagement und Personalentwicklung GmbH. She also has a post as therapist trainee in St. Elisabeth Krankenhaus Lahnstein.

1 Introduction

There is a growing interest and a body of knowledge regarding worked examples (correct solutions) and a lot of evidence of their effectiveness as an instructional method in learning mathematics and in science education (Catrambone, 1994; Catrambone, 1998; McLaren, et al., 2008; Paas, 1992; Renkl, 1997; Sweller and Cooper, 1985; Trafton and Reiser, 1993; Van Gog, et al., 2006). The benefits of worked examples are especially discussed in connection to cognitive load theory (Pass and Merrienboer, 1994; Sweller, 1988; Sweller et al, 1998), which emphasises their ability to reduce cognitive load in comparison to standard problem solving. Moreover, in the context of informal learning that is rapidly gaining ground, learning from errors with its inherent metacognitive skills of spotting and correcting errors may be an important competence to warrant the validity of informally acquired knowledge. Therefore, erroneous examples are a potential teaching strategy for promoting such skills. *Erroneous examples* are counterparts of worked examples that include one or more errors. Although there has been some interest in investigating the use of erroneous examples in conjunction with worked examples, erroneous examples have been scarcely investigated in their own right. Moreover, erroneous examples are rarely used in mathematics teaching, because many mathematics teachers are sceptical about discussing errors in the classroom (Tsamir and Tirosh, 2003). Teachers are cautious of exposing students to errors in fear that it could lead to incorrect solutions being assimilated by students, in behaviourist fashion (Skinner, 1938). As a consequence, it remains open (1) if and when erroneous examples are beneficial for learning and (2) what form of erroneous examples is more beneficial.

In particular, the question of what form or what type of erroneous examples presentation is beneficial can be carefully explored in the context of learning technologies, where erroneous examples can be implemented in an interactive fashion, thus opening new possibilities for adaptive instruction. The presentation of erroneous examples can vary by the kind and amount of feedback provided, diverse tutorial strategies can be used, and the choice and sequencing of the learning material can be decided on the fly (e.g. erroneous examples provided in conjunction with, for instance, standard problem-solving exercises, or worked examples). Adaptation to the needs of individual students has two main advantages. First, it can shed light on learning research, as it facilitates testing how students learn under different manipulations. Second, it may contribute to better learning outcomes in formal education (in or after the classroom).

We focus on fractions as a core topic in middle school math curricula around the world. Fractions are a good target for adaptive, web-based instruction. There is evidence that students, and even preservice teachers, do not have the expected level of understanding of fractions (Newton, 2008). Persistent misconceptions lead to poor performance in solving fraction problems (Stafylidou and Vosniadou, 2004). Since fractions are also essential to other key subjects, such as physics and chemistry problems, they represent a "gateway" topic to success for any student of science and mathematics. Thus, new, successful forms of teaching fractions could have a profound impact on science and math learning.

Theoretical and empirical work provides some support for studying errors that can promote student learning of mathematics (Borasi, 1994; Oser and Hascher, 1997; Strecker, 1999; Müller, 2003; Seidel and Prenzel, 2003). For example, Borasi argues that mathematics education could benefit from the discussion of errors by encouraging critical thinking about mathematical concepts, by providing new problem solving opportunities, and by motivating reflection and inquiry.

Siegler and Chen (2002; 2008) conducted a controlled comparison of correct and incorrect examples for mathematical equality problems. They found that when students studied and self-explained *both* correct and incorrect examples they learned better than when students studied and self-explained *only* correct examples. They hypothesised that self-explanation of correct and erroneous examples strengthened correct strategies and weakened incorrect problem solving strategies, respectively.

Grosse and Renkl (2007) studied whether explaining both correct and incorrect examples of probability problems makes a difference to learning and whether highlighting errors helps students learn from those errors. Their empirical studies (in which no help or feedback was provided) showed some learning benefit of erroneous examples, but unlike the results of Siegler and colleagues (2002; 2008), the benefit they uncovered was only for learners with strong prior knowledge and for far transfer.

Both Siegler (2002) and Grosse and Renkl (2007) concluded that in order for students to benefit from incorrect solutions, they have to be able to explain "why" the solutions are incorrect. In particular, a later study by Grosse and Renkl (2007) analysed think-alouds on self-explanation strategies. The analysis revealed that spontaneous self-explanations of errors are very important for learning, but that they inhibit principle-based explanations (explanations based on principles of the domain) that are normally produced when self-explaining worked examples, for instance. However, such principle-based self-explanations are crucial to learning.

Durkin and Rittle-Johnson (2008, 2012) and Rittle-Johnson and Wagner Alibali (2001) tested whether comparing incorrect and correct examples of decimal problems promotes greater learning than comparing two correct decimals examples. They hypothesised that comparing incorrect examples to correct examples may be particularly effective for emphasising the critical attributes of correct examples as suggested by Grosse and Renkl (2007). They found that students in the incorrect condition had higher procedural posttest scores, as well as higher conceptual posttest scores on a delayed posttest two weeks later, than students in the correct condition.

In the domain of medical education, research on erroneous examples has demonstrated the benefits of erroneous examples in combination with elaborate feedback in the acquisition of problem-solving schemata. This was compared to the use of erroneous examples without feedback (Kopp et al, 2008) and with knowledge of correct solution feedback (Stark et al., 2011). The diagnostic knowledge, which included conceptual, strategic and teleological knowledge, increased more for students who worked with erroneous examples and elaborate feedback on "why" the step was wrong and "which" step would be correct. The effects of elaborate feedback were replicated for a more complex domain that imposed additional cognitive load, but the effects of erroneous examples or their interaction were not replicated (Stark et al., 2011). Erroneous examples had a significantly better effect on cognitive skills in a delayed posttest. This effect was persistent regardless of prior knowledge.

Finally, in the domain of decimal numbers, internet-based interactive erroneous examples with feedback on correctness of solution and on error explanation were compared to a problem solving with feedback on correctness (McLaren et al., 2012).). They found that middle school students who worked with erroneous examples did better on a delayed posttest than the students who worked with standard problems and attributed this finding to "desirable difficulties" (Schmidt and Bjork, 1992). In particular, they hypothesised that challenging students with difficult problems, which erroneous examples could be described as, did not lead to immediate learning benefits, but *did* lead to delayed learning benefits.

This scientific findings are also supported by the results of the highly-publicised TIMSS studies (OECD, 2001) showed that Japanese math students outperformed their counterparts in most of the western world. The key curriculum difference cited was that Japanese educators present and discuss incorrect solutions and ask students to locate and correct errors.

1.1 Contribution of our studies

We take the earlier controlled studies further by investigating erroneous examples decoupled from worked examples in the context of technology enhanced learning with ActiveMath, a web-based system for mathematics (Melis et al., 2006). Our ultimate goal is to develop micro and macroadaptation for the presentation of erroneous examples for individual students since the benefit of erroneous examples may depend on individual skills, grade level, etc. By microadaptation we mean the teaching strategy, or step-by-step feedback, inside an erroneous example based on the student's performance. By macroadaptation we mean the choice of task for the student, as well as the frequency and sequence of the presentation of erroneous examples.

We focus on the empirical results that inform our work on the adaptive technology. In contrast to the Siegler (Siegler, 2002; Siegler and Chen, 2008) studies, we are interested in the interaction of students' with erroneous examples and how situational and learner characteristics impact that interaction. Extending the work of Grosse and Renkl (Grosse and Renkl, 2007; Renkl, 1997), we investigate interactive erroneous examples with adaptive error-detection and error-correction help. This novel design relies on the intelligent technology of ActiveMath. Our primary rationale for including error detection and correction help in the empirical studies is that students are not accustomed to working with and learning from erroneous examples in mathematics. Thus, they may not have the required skills to review, analyse, and reflect upon such examples, as Grosse and Renkl (Grosse and Renkl, 2007) have hypothesised based on their results, thus additional help may be necessary. Taking this strand and providing additional elaborate help, we also extend the work of Kopp and colleagues (Kopp et al., 2008) in medical education to the domain of mathematics education. Moreover, we include feedback that emphasises conceptual principle-based knowledge in order to counter-balance the effect reported by Grosse and Renkl (2007). They found that such reflections were missing in the students' spontaneous self-explanations of errors and hypothesised that, due to this lack of more conceptual explanations, learning opportunities created by errors were not exploited. Providing such help in an adaptive fashion to students of different knowledge levels might eliminate the aptitude-treatment effect for transfer, which was one of their main findings. Additionally, we did studies with school kids of lower and higher levels, to test if the benefits reported by Grosse and Renkl (2007) transfer to different the school level and for which grades in particular.

With regard to the possible drawbacks of erroneous examples, we hypothesise that a student is less likely to exhibit the feared 'conditioned response' of behaviourist theory (Skinner, 1938) when studying errors that the student has not made him/herself and thus has not (necessarily) internalised. On the contrary, students may benefit from erroneous examples when they encounter them at the right time and in the right way. For example, rewarding a student for error detection may lead to memory annotation such that errors will be avoided in subsequent retrieval. At the same time, a student is unlikely to be demotivated by studying common errors in the domain, made by others, as when emphasising errors the student has made him/herself. In fact, some of our own work has already demonstrated the motivational potential of erroneous examples (Melis, 2004).

In summary, we believe that learning from errors can help students develop (or enhance) their critical thinking, error detection, and error awareness skills, something that is not possible with correct examples and difficult with unsupported problem solving (Borasi, 1994). Moreover, erroneous examples may weaken students' incorrect strategies, as opposed to worked examples that strengthen correct strategies (Siegler, 2002). Additionally, similar to worked examples, erroneous examples do not ask students to perform as in problem solving, but instead provide a worked-out solution that includes one or more errors. Thus, they could, reduce extraneous cognitive load in comparison to problem solving (Paas et al., 2003), while increasing germane cognitive load in the sense of creating cognitive conflict situations. Adaptive help, in particular, might support deeper reflection on errors and help induce such cognitive conflict. Especially the kind of adaptive help that elaborates on conceptual understanding of errors may catalyse the creation and exploitation of such learning opportunities. Furthermore, erroneous examples may guide learners toward learning orientation rather than performance orientation; specifically in combination with help that increases student's involvement in the learning process and in more conceptual understanding (Siegler, 2002).

In the course of our investigation of erroneous examples, we aim to answer the following research questions:

When

1 Do advanced students, in terms of grade level, gain more from erroneous examples than less advanced students?

How

1 Can students' cognitive skills, conceptual understanding, and transfer abilities improve through the study of erroneous examples?

2 Does work with erroneous examples help to improve the metacognitive competencies of error detection, error awareness and error correction?

3 Does adaptive help play a role in whether and how students learn from erroneous examples?

Based on these considerations and research questions, our primary hypotheses are:

Hypothesis 1: Presenting erroneous examples to students will improve:

H1a: their cognitive skills,

H1b: conceptual knowledge,

H1c: transfer skills, and

H1d: metacognitive skills

Cognitive skills refer to solving standard fraction addition and subtraction exercises. Conceptual knowledge refers to understanding the domain concepts necessary for solving each specific problem, for instance "addition as increasing". Transfer refers to solving more difficult problems using the same concept, e.g. three-fraction addition as opposed to two-fraction addition, or solving problems using a theoretically related concept. Metacognitive skills refer to error detection and error correction.

A control group learning through partially supported problem solving is compared to the erroneous examples groups on the dependent variables, cognitive skills, metacognitive skills, conceptual learning, and transfer.

*Hypothesis 2: The learning effect of erroneous examples is stronger when students are supported in finding and correcting the error with additional help.*Two experimental groups were used, one with help and one without help, to test this hypothesis.

Hypothesis 3: The effect of erroneous examples with adaptive help will be independent of grade level. Three levels of students are tested spanning five grade levels.

Moreover, we explore the following supplementary conjectures:

The learning effect of erroneous examples depends on when they are presented to the students. The order of presentation of erroneous examples is varied between studies, to allow drawing some conclusions.

The cognitive load of students will be reduced through working with erroneous examples, as opposed to standard problem solving, and that they will be more motivated to learn and understand the materials, which results from a shift to learning orientation. Self-reports were analysed to test these conjectures.

To assess the learning effects of erroneous examples at different grade levels and settings, we conducted lab studies with 6th, 7th and 8th-graders and classroom studies with 9th and 10th-graders. The participants came from both urban and suburban German schools from two states. In a previous article (Tsovaltzi et al., 2010), we presented results of the first two studies and preliminary results of the third study. Here we present the analysis of the third study with additional data that we collected to account for group size differences. We also present the new analysis of the questionnaires of all three studies and discuss the relevance of these results with regard to the learning gains analysis. In view of the new analyses, we further present implications that can be drawn from our results.

2 Study 1: 6th-grade lab study

2.1 Methods

2.1.1 Design

One control group and two experimental groups were used. The control condition, No-Erroneous-Examples (NOEE), trained with partially supported standard fraction exercises (Figure 1), but no erroneous examples. The experimental condition Erroneous-Examples-With-Help (EEWH) trained with standard exercises, but also with erroneous examples (Figure 2) and provision of additional help within the erroneous examples for explaining the error. The condition Erroneous-Examples-Without-Help (EEWOH) trained with standard exercises, and erroneous examples but without additional help. The participants completed the experiment on a single day in approximately 2 hours and 40 minutes with three breaks of between five and ten minutes. Breaks were not obligatory, so participants could choose to skip them. Participants sat together in a computer room, but all parts of the study were completed individually on separate computers. All sessions were completed over the course of three weeks and were supervised by the experimenter (first author) and her assistant (fourth author).

Figure 1 A standard exercise in ActiveMath (with English translations in the legends)

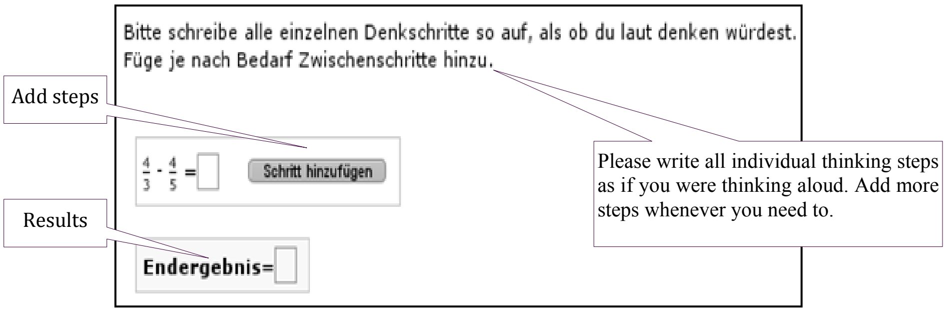
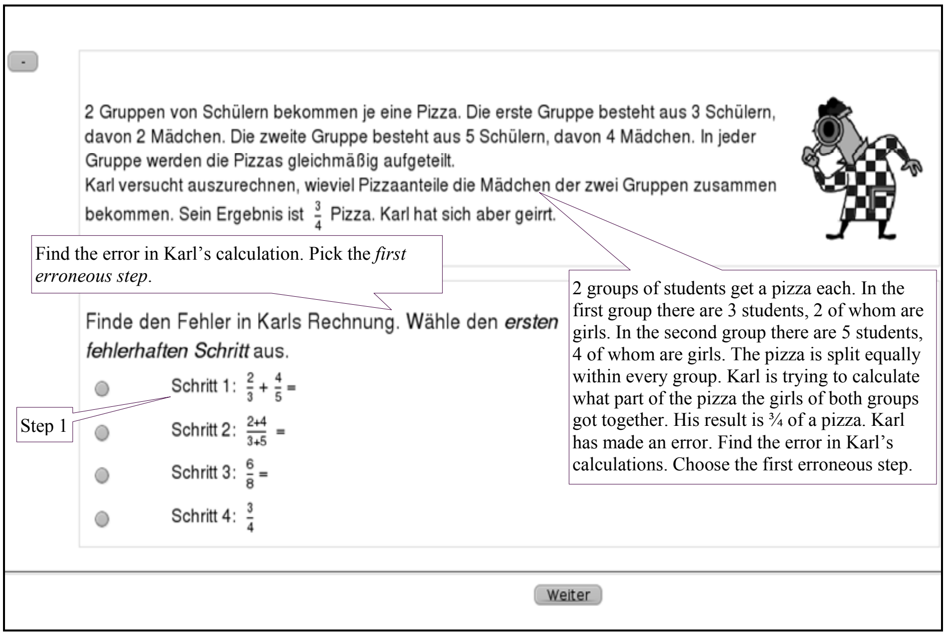


Figure 2 Interactive erroneous example in ActiveMath on the typical error of adding numerators and denominators of fractions with unlike denominators



2.1.2 Participants

Twenty-three volunteers from the 6th-grade at German schools participated in this study, which took place in a lab at the DFKI (German Research Center for Artificial Intelligence) in Saarbrücken, Germany. The participants were recruited through a press release announcing the study that was described as software testing that gives students a possibility to practice mathematics. All students who expressed interest were accepted for participation based on availability criteria during the time planned for the studies. Their parents signed a letter of consent informing them that the participants were free to drop out at any point during the study. Participants came from different urban and suburban schools in Germany (Saarland). They received a payment of ten Euro at the end of the session, irrespective of whether they completed all parts. They were randomly distributed to the groups by the experimenter and her assistant as follows: NOEE=8, EEWH=8, EEWOH=7. The experimenter's assistant was also mainly responsible for the communication with the participants prior to the experiment. All participants had just completed a course on fractions at school. The mean of their term-grade in mathematics across conditions was 2.04 (SD=.88) (best=1 vs. fail=6), so the participants were generally good students. There was no significant difference in the means of the pretest among conditions ($F(2,20)=0.23$, $p=.79$, $n^2=0.02$).

2.1.3 Materials

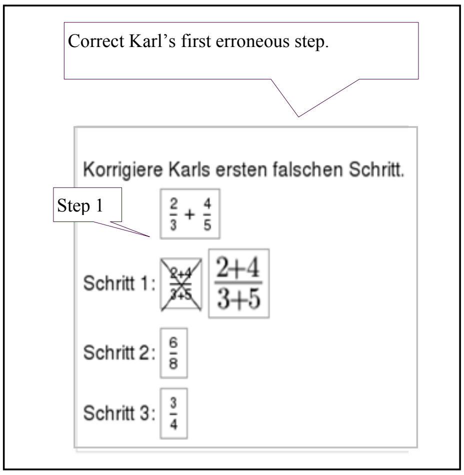
The design included a pre-questionnaire, a familiarisation, a pretest, an intervention, a posttest and a post-questionnaire, which were presented in this order to all students in the ActiveMath software environment.

Familiarisation. The familiarisation in ActiveMath allowed students to train with the system. All conditions trained in writing fractions in the system using a specialised input editor and in interacting with the system in general. The exercises used in this phase asked students to order the following fractions from smallest to largest: 1, 1/6, 7/6. This skill was not trained during the intervention or tested in the pre and posttest. Correct and incorrect feedback as well as the correct worked out solution were presented to all conditions. The EEWH condition received additional help to get familiar with how help is presented in ActiveMath. No erroneous examples were used during the familiarisation.

Standard Fraction Exercises. Standard fraction exercises included addition and subtraction of fractions represented in ActiveMath. A simple exercise of fraction subtraction with unlike denominators is shown in Figure 1. We asked the students to write all thinking steps, as if they were thinking aloud, so that the system could more accurately assess the students' performance on an exercise. After entering their result, students got feedback from ActiveMath to indicate whether their result was correct or wrong and the correct worked out solution was presented.

Interactive Erroneous Examples. The presentation of erroneous examples in ActiveMath is done through a tutorial strategy, which defines when and how to provide help, signal correct and incorrect answers, give answers away, show previous steps of the students, etc. Previous steps are folded and hidden automatically, to allow students to concentrate on the current step. Students can choose to unfold previous steps if they want to refer back to them. Erroneous examples include instances of typical errors students made in rule-application and errors that address common fractions misconceptions. Figure 2 displays the task presented in the first phase. Each step of the erroneous solutions is presented as choices in a multiple-choice question (MCQ) and students have to select the erroneous step. After completing this phase, students are prompted to correct the error, as shown in Figure 3.

Figure 3 Error-correction phase

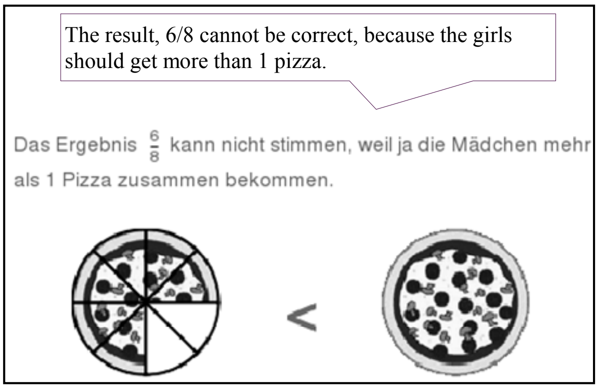


Feedback Design. Based on pilot studies (Tsovaltzi et al., 2009), we designed feedback for helping students understand and correct the errors. There are four types of unsolicited feedback: standard feedback, error-awareness and error detection (EAD) feedback, self-explanation feedback and error-correction scaffolds.

Standard feedback consists of flag feedback (checks for correct and crosses for incorrect answers) along with a text indication. It also consists of the correct answer or correct worked solution, which is presented to the student at the conclusion of an attempt.

EAD feedback (Figure 4) focuses on supporting the metacognitive skills of error detection and awareness that may trigger cognitive conflict. It appears on the screen after the student has indicated having read the problem statement.

Figure 4 EAD feedback with additional visual example

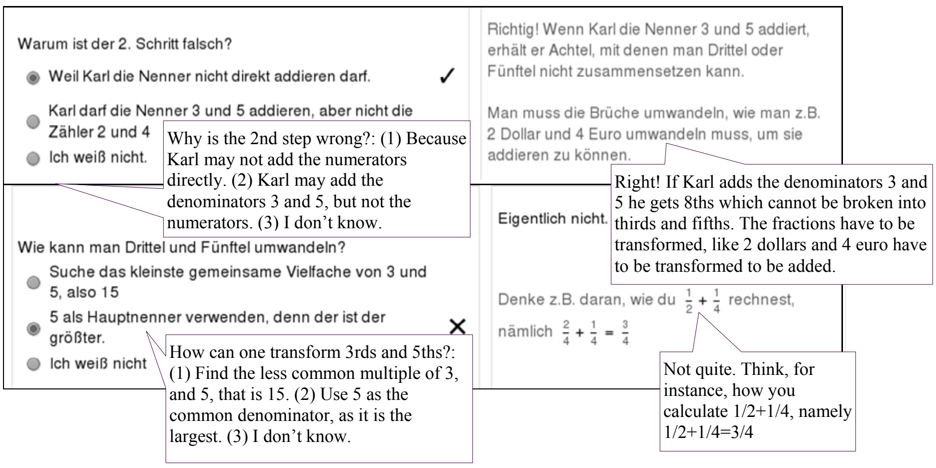


Self-explanation feedback (Figure 5) is presented in the form of MCQs. It aims to help students understand and reason about the error through "why" questions (Figure 5, top).

"Why" questions are asked to further prompt reflection that can lead to cognitive conflict, elaboration on errors, and conceptual understanding of errors. After a choice, the system indicates whether the response was correct or not and provides additional conceptual explanation of the error and of what the right thing to do would be (Figure 5, top right)

Error-correction scaffolds prepare the student for correcting the error in the second phase and also have the form of MCQs. They start with "how" questions that concentrate rather on procedural skills and attempt to facilitate the acquisition of practical knowledge. Additional conceptual explanations are provided depending on the student's response. The incorrect choices in the MCQs correspond to typical misconceptions or performance errors. For example, the second choice at the top part of Figure 5, "Karl may add the numerators but not the denominators", tries to see if the students understand that both numerators and denominators have to be transformed when making fractions like. By addressing such misconceptions and errors, MCQs are meant to prepare the students for correcting the error in Phase 2. Students receive correct and incorrect feedback on their choices, and eventually the correct answer. The "how" question at the bottom of Figure 5, which follows the "why" question, asks the student: The second choice, "By using 5 as the common denominator, because it is larger." is an over-generalisation error that students make by analogy to when adding e.g. 1/5+1/15. The student in this case gets the feedback that the answer is wrong together with additional help (Figure 5, bottom right).

Figure 5 "Why" and "How" MCQs with choices and conceptual explanations



MCQs are nested (2 to 5 layers). If a student chooses the right answer at the two top-level MCQs (the "why" and "how" questions), then the next levels, the error-correction MCQs, are skipped, under the assumption that the student probably knows how to correct the error and to avoid providing unneeded help which might frustrate students or interfere with existing problem-solving schemata that would have to be extended (Kalyuga et al., 2003).

In the second, error-correction, phase the chosen step is crossed out, and an additional editable box is provided for correcting the error (cf. Figure 2). After that error-specific feedback is provided, e.g. "*You forgot to expand the numerators*", along with the correct solution. Here, we allow students one attempt to correct the mistake. Only one attempt is allowed so that this process is not too much like problem solving

In the intervention, all groups solved six sequences of three exercises. The control group solved only standard exercises. The sequences for the experimental groups included: standard exercise - standard exercise - erroneous example. In the EEWH group, erroneous examples were presented with additional help (EAD, error detection/correction MCQs, and error-specific help). The condition Erroneous-Examples-Without-Help (EEWOH) included standard exercises, and erroneous examples but without additional help.

These sequences trained skills that are typical fraction topics taught at school, e.g. fraction addition/subtraction with like denominators and with unlike denominators, addition of whole numbers with fractions, as well as word problems, that did not include complex modelling tasks, which would require students to use fraction operators to represent the word problems.

Pretest and Posttest. The pretest and posttest were the same for all three conditions and were counter-balanced and consisted of similar problems to those used in the intervention and a transfer problem (a four-fraction addition, as opposed to the maximum of three in the intervention). However, there was no feedback or additional help provided in the pretest and postest. Finally, three erroneous examples were part of the posttest only, as we did not want the control group to see any erroneous examples before the intervention. The posttest erroneous examples consisted of two phases, similar to the intervention erroneous examples, but instead of feedback they included three open conceptual questions on error detection and awareness. The questions were of the kind "Why cannot Oliver's solution be correct?", "What mistake did Oliver make?", "Why did Oliver make this mistake? What does he not understand about fractions?" These questions were designed to test students' error detection skills as well as their understanding of basic fraction principles. For example, the mistake Oliver made was that he added the denominators 6 and 8 in the exercise 7/6+5/8. The answer to the question about what Oliver did not understand would be "That if one adds the denominators 6 and 8, one gets 14ths, which one cannot break in neither 6ths nor 8ths.", which refers to the basic concept of common denominators.

Questionnaires. The pre- and post-questionnaires used in all studies were based on MSLQ[1] (Pintrich et al., 1991) and on CAQ[2] (Knezek and Rhonda, 1996), which contain six-point Likert scale questions for self-report. The items were adjusted and translated into German. The questionnaires consisted of six constructs each: motivation, error-awareness, critical thinking, cognitive load, learning orientation, and self-efficacy. There were eighteen items in total per questionnaire. The greatest number of items were dedicated to motivation (5) and the least to self-efficacy, error-awareness and critical thinking (2). The pre- and post-questionnaires were designed to have equivalent constructs and items. For example, a pre-motivation item was: "I know that computers give me the opportunity to learn many new things" (German: "Ich weiss, dass Computer mir die Möglichkeit geben, viele neue Dinge zu lernen."). The equivalent post-motivation item was: "I learned many new things through the learning software" (German: "Durch das Lernprogramm habe ich viele neue Sachen gelernt").

3.1.1 Results: 6th-Grade Lab Study

ANOVA Results. The results for the erroneous examples scores follow our hypotheses, although they were mostly insignificant (cf. Table 1). The EEWH condition scored highest in almost all scores. For all these scores, EEWOH came second, followed by NOEE. The big variances between conditions (cf. Figure 6) were only significant for correcting the error (EE-correct) in the erroneous examples. Nevertheless, we ran an ANOVA for that score, since the group size is almost the same across conditions. The condition showed no significant effect in the ANOVA, there was a significant difference when comparing EEWH and NOEE for finding the error in the planned contrasts (Helmert) ($t(20)=2.14$, $p<.05$, $d=1.29$, $r=.54$). Another quite big difference was between EEWH and NOEE for the total erroneous example score ($t(20)=1.95$, $p=.065$, d=1.02, $r=.46$), which includes correcting the error and answering conceptual questions. These learning gains related to erroneous examples did not transfer to the cognitive skills where the differences between pretest and posttest are minimal in either direction for all conditions and there was a ceiling effect both in the pretest ($M=84.1$, $SD=19.3$) and the posttest ($M=84.4$, $SD=15.8$). This was probably due to the high prior knowledge level of the participants.

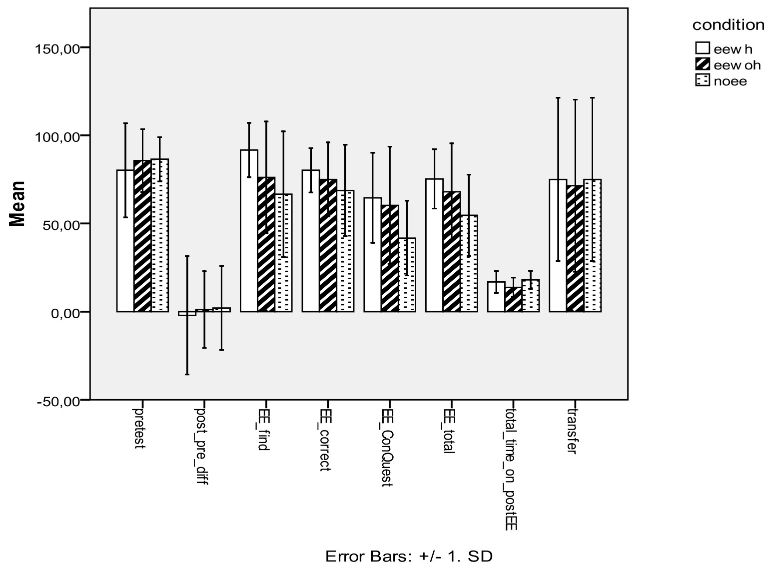
ANCOVA Results. As we did not have access to the term grades of the participants before the experiment, the conditions were not balanced in that respect. Therefore, we analysed the data with the term-grade but also with the pretest score as covariates, to capture the possible influence of previous math and fraction knowledge, respectively, on the learning effects. With this analysis, there is a main effect for erroneous examples in answering conceptual questions ($t(20)=2.25$, $p<.05$, $d=1.01$, $r=.45$) , and in the total erroneous examples score ($t(20)=2.34$, $p<.05$, $d=1.04$, $r=.46$), when comparing the two erroneous example conditions with the control. The same scores were also significantly higher for EEWH vs. NOEE (conceptual questions: t(20)=2.48, $p<.05$, $d=1.11$, $r=.49$/ erroneous examples: $t(20)=2.96$, $p<.05$, $d=1.32$, $r=.55$) respectively). Additionally, the difference for finding the error was significantly higher for EEWH vs. NOEE ($t(20)=2.37$, $p<.05$, $d=1.06$, $r=.47$).

Table 1 Descriptive Statistics: Lab Study 6th-Grade

	Condition	*EEWH N=8*	*EEWOH N=7*	*NOEE N=8*
Score	*Subscore*	*mean(sd)%*	*mean(sd)%*	*mean(sd)%*
Cognitive Skills	Pretest	80.2(26.7)	85.7(17.8)	86.5(12.5)
	Post-pre-diff	–2.1(33.6)	1.2(21.7)^	2.1(23.9)+
Metacognitive Skills (EE)	EE-find	91.7(15.4)+	76.2(31.7)^	66.5(35.6)
	EE-correct	80.2(12.5)+	75.0(21.0)^	68.7(25.9)
	EE-ConQuest*	64.6(25.5)+	60.2(33.3)^	41.7(21.2)
	EE-total	75.3(16.8)+	67.9(27.5)^	54.7(23.0)
	Total-time-on-postEE	16.9(6.2)^	13.8(5.5)+	18.0(5.1)
Transfer	Transfer	75.0(46.2)+	71.4(48.8)	75.0(46.3)^

Note: +=best, ^=middle learning gains, *= also conceptual skill.

Figure 6 Descriptive Statistics for 6th grade



Questionnaires' Results. The questionnaires of sixteen participants were evaluated: EEWH=5, EEWOH=5, NOEE=6. Due to technical reasons some pre and post questionnaires' data was lost. Paired sample t-test revealed that most self-reports were worse in the post-questionnaires than in the pre-questionnaires (cf. Table 2), however, these results were significant only for two constructs: motivation (t(14)=2.66; p<.05, d=0.92, r=0.42) and error-awareness (t(14)= 2.95; p<.05 d=1.05, r=0.47). Exceptions were the self-reports on cognitive load (for EEWOH and NOEE), learning orientation (for EEWOH and NOEE), and self-efficacy (for EEWOH), which were better in the post-questionnaire.

Table 2 Self-report in pre and post-questionnaires for 6th-grade

	Condition	*EEWH N=5*	*EEWOH N=6*	*NOEE N=5*
Construct	pre vs. post	mean(sd)%	mean(sd)%	mean(sd)%
motivation	Pre	78.00(10.95)	83.33(6.24)+	80.56(13.24)^
	Post	67.33(13.21)	75.33(7.67)+	72.22(5.44)+
Err-awareness	Pre	76.67(19.00)	78.33(22.52)+	77.78(13.61)^
	Post	63.33(9.50)+	61.67(18.26)^	54.17(22.82)
Crit-thinking	Pre	71.67(12.64)+	68.33(12.36)	70.83(14.67)^
	Post	68.33(10.87)	48.33(25.95)+	62.50(21.57)^
Cognitive-load	Pre	42.22(15.01)+	27.78(8.78)	38.89(17.57)^
	Post	47.78(9.30)	25.56(8.43)^	35.19(9.07)+
Learn-orient.	Pre	73.33(7.57)+	65.83(6.18)	71.53(13.29)^
	Post	70.83(13.82)^	69.17(12.36)	72.22(9.00)+
Self-efficacy	Pre	80.00(12.64)^	80.00(12.64)	83.33(14.91)+
	Post	75.00(10.21)	93.33(10.87)+	76.39(17.01)^

Note: +=best, ^=middle.

When comparing the conditions with ANOVA and planned contrasts, the difference in the reported cognitive load in the post-questionnaire is significantly better for NOEE than the two experimental conditions ($F(2,13)=7.76$, $p=.006$, $n^2=0.54$). The individual group differences were also significant: EEWH vs. NOEE ($t(8)=2.32$, $p<.05$, $d=1.29$, $r=.57$) and EEWH vs. EEWOH ($t(9)=3.93$, $p<.05$, $d=2.18$, $r=.78$). The ANCOVA and planned contrasts with covariates the pretest score and the term grade also revealed that EEWOH reported significantly more self-efficacy than EEWH ($t(9)=3.05$, $p<.05$, $d=2.15$, $r=.73$).

3.2 Discussion: 6th Grade

We found significant differences in the scores for erroneous examples, which show that erroneous examples, in general, and the additional help, in particular, supported better the metacognitive skills of error detection and error correction. The higher performance in the conceptual questions related to understanding the error also indicates better conceptual understanding for the erroneous examples conditions and for the help condition. To illustrate this, the erroneous example "Oliver must calculate how much 7/6+5/8 is. His results is 6/7." was followed by the conceptual question „Why cannot Oliver's result be correct?". An example of a good answer in the NOEE condition is "Because the common denominator is not 7 and it cannot be reduced to 7." This is correct but it does not explain the reason why this is7 cannot be the denominator why the denominator cannot be reduced to 7, therefore it does not get to the necessary reasoning for spotting the error. An answer from the EEWH conditions is "Because the first summand is greater than his result", which gets to the point of the error recognition, indicating that the sum in Oliver's addition is even smaller than one of the added fractions. Recognising that, which was trained in the erroneous example conditions, is the skill necessary for spotting errors.

The better performance found on metacognitive skills is not in line with the self-reports on self-efficacy. This scale focused on understanding complex fraction problems and basic concepts of fractions. EEWH reported more self-efficacy in comparison to EEWOH, who performed better. Furthermore, we had no evidence that studying erroneous examples had an effect on standard cognitive skills, where the level was very high to begin with. Interestingly, the term grade was not a significant covariate of the cognitive load self-reports. However, our hypothesis that erroneous examples and the additional help would cause less cognitive load does not seem to be supported by the comparison of the conditions reports on post cognitive load.

4 Study 2: lab study 7th and 8th grade

4.1 Methods

4.1.1 Design

The design in this study was the same as in Study 1.

4.1.2 Participants

Twenty-four paid volunteers in the 7th and 8th-grade participated in the study, eight in each of the three conditions. They were recruited and assigned to groups in the same way as participants in Study 1. 7th and 8th-graders are similarly advanced beyond 6th-graders in their understanding of fractions, according to our expert teachers. They have had more opportunity to practice, but often retain their misconceptions in fractions. The mean of their term grade in mathematics was again at the upper-level of the grading scale and a little higher compared to the 6th grade (*M*=2.8, *SD*=1.2) (best=1 vs. fail=6). The pretest mean difference was not significant between conditions ($F(2,21)=0.23$, $p=.80$, $n^2=0.02$). Consistent with the judgments of the expert teachers, there was no significant difference in the scores of the 7th compared to the 8th grade ($t(22)=0.71$, $p>.05$, $n^2=0.02$, $d=0.29$, $r=.14$).

4.1.3 Materials

The materials overlapped to a large degree with those of Study 1, but participants in Study 2 also solved world problems that were not used in the 6th-grade, since such exercises are not typically encountered in German schools in this grade, so teachers advised us against using them. An example of a world problem is: "Eva invited her friends to her birthday party. They drank 8 3/7 bottles of apple juice as well as 1 5/6 bottles of lemonade. How many bottles did they drink all together?"[3] The expected transformation into a mathematical expression in this exercise is: 8 3/7 + 1 5/6. In total, there were seven sequences of exercises in this study. A world problem also testing transfer was added to the posttest. By including such fraction modelling, we aimed to induce and measure conceptual understanding.

4.1.4 Results: 7th-8th-grade lab study

As a whole, the results do not support our hypotheses for the 7th and 8th-grade (cf. Table 3), although differences in scores are small and not significant. NOEE scored better in almost all scores, apart from the conceptual questions, where EEWOH did best. EEWOH was also second best in finding the error and in the total erroneous examples score. EEWH came second in the cognitive skills, correcting the error, transfer exercises, and modelling. The standard deviation for all scores except for improvement on cognitive skills was highest for EEWH (cf. Figure 8).

ANOVA Results. Since the group size is the same across conditions, the results of the ANOVA can be considered robust although Levene's test was significant for finding the error (p=.018), conceptual questions (p=.000) and for the total score on erroneous examples (p=.000). The only statistically significant score in the ANOVA test was the time spent on the posttest erroneous examples ($F(2,21)=5.59$, $p=.011$, $n^2=.35$), where NOEE spent significantly more time than the erroneous-examples conditions together ($t(22)=2.88$, $p<.05$, $d=1.23$, $r=.52$) and EEWH alone ($t(22)=3.04$, $p<.05$, $d=1.63$, $r=.63$).

ANCOVA Results. The ANCOVA with covariates the term grade and the pretest score showed that only the term-grade is a significant covariate for answering conceptual questions ($F(1,21)=4.49$, $p=.047$, $n^2=.18$) and also has quite a big covariating effect for the total erroneous examples score ($F(1,21)=4.03$, $p=.059$, $n^2=.17$). In both cases, considering the covariating effect decreases the difference between the control and the erroneous example conditions that originally scored worse. There is also a significant effect of the condition for the time spent on erroneous examples ($F(2,21)=5.28$, $p=.014$, $n^2=.59$) when term-grade is considered as a covariate.

Other Results. An important result in this study is the significant difference in the scores for finding and correcting the error (t(23)=4.89, $p<.001$, $d=0.59$, $r=.28$). The standard deviation for the two metacognitive competencies is comparable, but the mean for correcting is more than 0.5 point lower than for finding the error ($M=3.12$, $SD=.95$ for finding, $M=2.54$, $SD=.99$ for correcting), which means that a significant number of participants were able to find the error but not to correct it. This is also true when comparing separate conditions. Where the difference for EEWOH and for NOEE between finding and correcting the error is significant (EEWOH: $t(7)=4.33$, $p<.05$, d=1.15, $r=.49$; NOEE: $t(7)=4.32$, $p<.05$, $d=1.44$, $r=.58$), but not significant for EEWH ($t(7)=2.19$, $p>.05$, $d=1.64$, $r=.63$). The same phenomenon occurred even with students who could solve exercises. Most students could add fractions with unlike denominators, but could not correct related errors. For example, they could solve solve the addition 1/6 + 3/8 = 4/24 + 9/24 = (4+9)/24 = 13/24 correctly, in the erroneous example Oliver (Step 1: 7/6 + 5/8, Step 2: (7+5)/(6+8), Step 3: 12/14, Step 4: 6/7) they identified Step 2 as wrong, but when asked to correct it, they often forgot to extend the numerators after calculating the common denominator, probably because they concentrated on extending the denominators. The MCQs to the conceptual questions after spotting the error are shown in Figure 7 and the correct answer is marked.

In other words, the problem of not finding the less common multiple was accepted as the first occurring problem without further mentioning that they also had to extend the numerators. This gave the following erroneous solution: 7/6+5/8=12/24.

Questionnaires' Results. The questionnaires of fifteen participants were evaluated: EEWH=6, EEWOH=6, NOEE=3. Unfortunately, some data was lost due to technical reasons, which led to a very small *N* in the NOEE condition. Therefore, the results reported can only be considered indicative. As, in the 6th-grade, most self-reports were worse in the post-questionnaires than in the pre-questionnaires (cf. Table 4), as measured in a paired sample t-test. However, none of the differences were significant. Self-reports that improved in the post-questionnaire include the ones on cognitive load (for NOEE), on learning orientation (for EEWOH and NOEE), and on self-efficacy (EEWOH and NOEE).

Figure 7 MCQs for the posttest erroneous example "Oliver"

What did Oliver do wrong in the step?

1. All steps are actually correct.
2. He added numerator with numerator and denominator with denominator. (Correct)
3. He simplified wrongly.
4. His common denominator is wrong.
5. I don't know.

Why did Oliver make this error? What did he not understand about fractions?

1. He actually understood everyhing.
2. That he cannot add whole numbers direct with fractions.
3. That he has to extend the numerators because he now has one denominator.
4. That he has to find the less common multiple of 6 and 8, because he cannot make 8ths or 6ths out of 14ths (6+8=14) (Correct)
5. I don't know.

Table 3 Descriptive Statistics: Lab Study 7th-8th-Grade

	Condition	*EEWH N=8*	*EEWOH N=8*	*NOEE N=8*
Score	*Subscore*	*mean(sd)%*	*mean(sd)%*	*mean(sd)%*
Cognitive Skills	Pretest	73.7(26.7)	71.2(19.7)	77.9(12.4)
	Post-pre-diff	2.4(24.4)^	-4.3(26.6)	6.9 (17.9)+
Metacognitive Skills (EE)	EE-find	68.7(34.7)	75.0(13.4)^	90.6(12.9)+
	EE-correct	57.8(26.7)^	54.7(21.1)	65.6(20.8)+
	EE-ConQuest*	55.2(46.5)	62.5(12.6)+	61.5(19.4)^
	EE-total	59.3(37.1)	63.7(11.9)^	69.8(15.0)+
	Total-time-on-postEE	8.1(4.3)+	11.5(4.2)^	15.5(4.8)
Transfer	Transfer	45.2(45.8)^	38.0(36.0)	67.3(28.5)+
Conc. Underst.	Modelling	36.4(42.2)^	19.8(35.0)	40.8(48.6)+

Note: +=best, ^=middle learning gains, *= also conceptual skill.

There were some significant differences when comparing self-reports from the pre- and post-questionnaires. The reports on self-efficacy were significantly better for NOEE vs. EEWH ($t(7)=2.69$, $p<.05$, $d=2.03$, $r=.71$). In ANCOVA contrasts with the covariates pretest and term-grade, the difference reported on cognitive load also became significantly better for NOEE vs. EEWOH ($t(13)=2.52$, $p<.05$, $d=1.9$, $r=.69$).

Figure 8 Descriptive Statistics 7^{th}-8^{th}-Grade with standard deviations

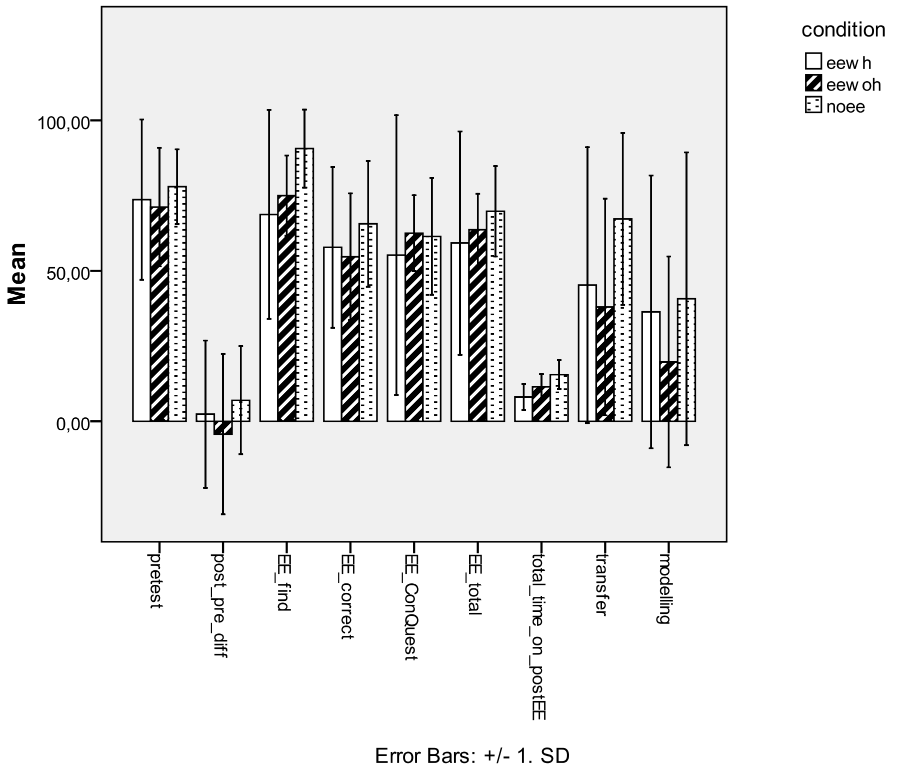


Table 4 Self-report in pre and post-questionnaires for 7th-8th-grade

	Condition	*EEWH N=6*	*EEWOH N=6*	*NOEE N=3*
Construct	*pre vs. post*	*mean(sd)%*	*mean(sd)%*	*mean(sd)%*
motivation	Pre	69.45(13.07)	72.78(12.72)^	84.45(8.39)+
	Post	58.33(6.12)	63.33(26.25)^	83.33(17.64)+
Err-awareness	Pre	68.06(18.57)	73.61(19.31) ^	80.56(12.73)+
	Post	52.78(14.59)	62.50(25.14) ^	69.45(26.79)+
Crit-thinking	Pre	70.83(15.59)	70.83(13.69) ^	75.00(16.67)+
	Post	62.50(21.57)+	59.72(20.69) ^	58.33(8.33)
Cognitive-load	Pre	48.15(25.50)+	45.37(14.24) ^	42.59(22.45)
	Post	53.70(14.34) ^	58.33(29.76)+	22.22(5.56)
Learn-orient.	Pre	65.97(12.48)	66.67(7.45) ^	72.22(2.41)+
	Post	61.11(7.76)	68.06(19.84) ^	87.50(18.16)+
Self-efficacy	Pre	73.61(12.27) ^	76.39(16.17)+	69.45(9.62)
	Post	68.06(22.00)	80.56(13.61) ^	94.45(9.62)+

Note: +=best, ^=middle.

4.2 Discussion: 7th-8th-grade

An explanation for the fact that the erroneous examples conditions, and especially the EEWH condition, did not perform better in the metacognitive skills tested through erroneous, is the little time students spent on erroneous examples in the posttest. Moreover, the long session might have overloaded the students and especially the ones in the EEWH condition whose sessions last long (over two and a half hours) because of the help provided. The possible resulting fatigue might be the reason why they did not spent more time on erroneous examples in the posttest. The self-reports on cognitive load are consistent with this hypothesis. Moreover, the high self-reports of NOEE on self-efficacy especially in comparison to EEWH might also mean that NOEE was more motivated in the posttest.

A plausible interpretation for the fact that the term grade is a significant covariate for answering conceptual questions, but not for cognitive skills is that a higher level of prior math knowledge is required to process new conceptual knowledge. This high-level knowledge is not necessary to deal with trained (almost automated) cognitive skills, which can be mastered by using well-practiced solutions steps (algorithmically). The difference between finding and correcting the error may mean that although students know the correct rules for performing operations on fractions and can recognise errors that violate these rules, they still have knowledge gaps that surface when asked to correct the error. A simpler explanation that is easier to find the error (recognise it) than to correct it is plausible, but elucidate the reasons behind this difference. Moreover, the inability to correct erroneous steps, for example, not extending the numerators when adding unlike fractions, which we observed with the same students who otherwise solve standard exercises with unlike fractions reveals that students do not understand the principle behind extending numerators. Rather they extend numerators automatically (algorithmically) and easily forget to when they don't put their algorithmic procedure in action from the beginning. One can think of this phenomenon as analogous to reciting a whole poem when the first line is provided, but not without the first line.

5 Study 3: classroom study 9th-10th-grade

To test the use of erroneous examples outside the lab we conducted classroom studies. Apart from the general ecological validity, this decision lab was also motivated by an attempt to avoid another ceiling effect, which is unlikely to occur in standard mixed-level classes. We previously reported results from our classroom studies for this level (Tsovaltzi et al., 2010) which, were not reliable due to a combination of big variances and unequal group sizes that the dropout of participants resulted to. In order to raise the reliability of our results, we collected additional data. Moreover, the data come from a different school, making the sample more representative. The results reported here and the corresponding discussion refers to a new analysis with the additional data. Moreover, we report and discuss the questionnaires analysis, which was not included in the past report.

5.1 Methods

5.1.1 Design

The design was similar to that of Study 1 and Study 2. Differences include that the students were not strict volunteers, but they agreed to take part in the studies in coordination with their mathematics teacher and their parents signed a consent form. They did not receive payment. Participants were informed that the study was not going to be assessed as part of their course-work. Another important difference is that in this study we were able to run the experiments on two different days, which was not possible in the lab studies. We were thus hoping to reduce the possibility of fatigue. This difference adds to the ecological validity of the results, in terms of the time students spent working with mathematics. Each session lasted two classroom hours with standard school breaks. The sessions took place in the computer labs of the schools, where students often work as part of their mathematics course.

5.1.2 Participants

Seventy-seven students in the 9th and 10th-grade participated in the study. Fifty-seven students completed the study successfully, fourteen did not attend school on the second day of the experiment and 6 either did not complete the intervention or entered values that showed non-attempts to more than 50% of the exercises (for instance, only “1” and “2” instead of fractions) and were screened. These classroom studies tested students from two different schools, one urban and one suburban, of yet a higher level (9th and 10th-grade). Our expert teachers advised that students of these levels typically still exhibit common fractions misconceptions. Moreover, 9th and 10th-graders have, on average, higher math knowledge. Since we found that the level of math knowledge has a covariating effect on conceptual understanding, we wanted to test if erroneous examples would have a stronger effect with these higher grade students.

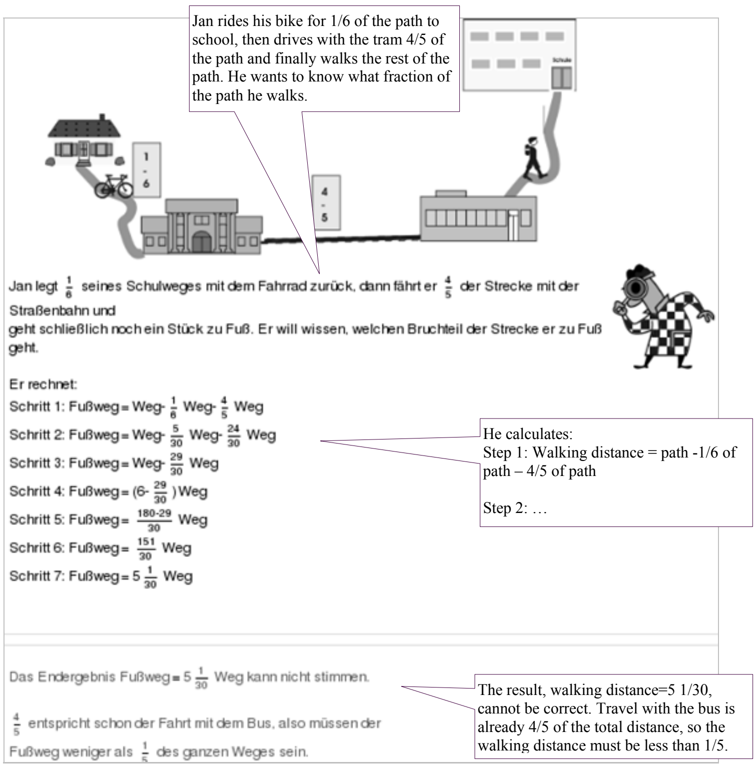
Participants were semi-randomly distributed to conditions, but the conditions were balanced so that the mean term-grade was about the same in each condition. The final distribution to conditions of the participants who completed all sessions was as follows: EEWH=18, EEWOH=20, NOEE=19. The difference in the pretest was not significant either between 9th and 10th grade ($F(2,54)=3.03$, $p=.057$, $n^2=.33$), or between conditions ($F(2,54)=1.24$, $p=.29$, $n^2=.053$).

5.1.3 Materials

Taking into account teachers' emphasis on fractions misconceptions as the common problem at this level we shifted from the traditional school fraction curriculum and included more conceptual exercises to address the basic principles of fractions, and common misconceptions. For instance, the exercises used the principles of "addition as increasing", "subtraction as decreasing", and "part of a whole" (Malle, 2004). In effect, we reorganised our sequences to reflect this shift. Capturing this structure in the presentation of sequences (although it was not explicitly indicated) intended to raise the awareness of these underlying principles. We added one sequence to train the basic concept "part of a whole", to explicitly include conceptual errors on top of the rule-application errors, which were the focus of the previous lab studies. In total, there were seven sequences. Figure 9 displays a task that trained the concept "part of a whole". The EAD feedback for this task is at the bottom of Figure 9.

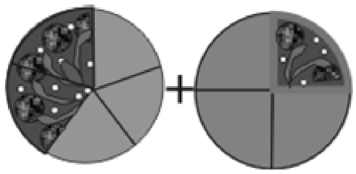
Moreover, we changed the order of presentation of the erroneous examples in the intervention; a sequence here consisted of standard exercise – erroneous examples – standard exercise, to test whether allowing students to train a bit after the erroneous examples would make a difference in learning outcomes. Furthermore, we adjusted the pretest and posttest exercises to test these concepts by adding world problems on them and also added two transfer exercises: one for fraction subtraction and one for the basic concept "relative part of" (Malle, 2004).

Figure 9 Interactive Erroneous Example on the Concept "part of a whole" with Error-Awareness and Error Detection (EAD) Feedback (bottom)



Two more new exercises asked students to transform a fraction operation represented by pizzas into a numerical fraction representation. For example, the task in Figure 10 had to be represented as 3/5+1/4. This type of exercise is commonly used at schools and was meant to give us a better assessment of the students' standard fraction competencies.

Figure 10 Pizza Representation of the Fraction problem 3/5+1/4



5.1.4 Results: Classroom Study 9th- and 10th-Grade

The results of the classroom studies supported our hypothesis. The participants in the EEWH condition scored higher in all four scores for learning (cognitive skills: Diff-post-pre-total, metacognitive skills: EE-total, transfer: transfer-total, and conceptual understanding: modelling-total), and in all subscores except for modelling the concept "relative part of". NOEE comes second for the four main scores, but this varies for individual subscores. The variances tend to be high for all variables (cf. Figures 11–14), but they are comparable between conditions, that allows an analysis of variance, except from transformation (p=.002) and "relative part of" (p=.003), for which we report contrasts assuming unequal variance.

Table 5 Descriptive statistics classroom studies 9th and 10th-Grade

	Condition	*EEWH N=18*	*EEWOH N=20*	*NOEE N=19*
Type of score	Type of Subscore	mean(sd)%	mean(sd)%	mean(sd)%
Time-on-task	Total-interv-duration	32.5(8.8)	26.4(6.9)	21.7(6.2)
	EE-or-equiv-duration	16.2(4.5)	10.6(3.8)	6.0(2.4)
Cognitive Skills	Pretest	74.5(14.2)	66.4(21.1)	64.9(17.2)
	Transform	16.2(23.0)+	4.9(33.2)^	-10.2(45.4)
	Diff-post-pre-total	8.9(12.8)+	1.4(23.5)	4.9(18.8)^
Metacognitive Skills (EE)	EE-find	61.1(28.7)+	50.0(28.1)	60.5(28.0)^
	EE-correct	40.3(28.0)+	21.3(30.6)	30.3(33.9)^
	EE-ConQuest*	50.9(20.7)+	50.4(24.9)^	47.8(25.1)
	EE-total	50.8(22.1)+	44.5(24.0)	46.8(24.7)^
	Total-time-on-EE	5.9(3.2)+	4.1(3.1)	5.9(3.9)+
Transfer	Add-subtr-total (cog. transfer)	32.0(30.1)+	20.0(34.3)	29.0(34.6)^
	Conc-transf-total*	46.8(34.7)+	30.4(29.3)^	29.5(30.30)
	Transfer-total	39.4(20.3)+	25.2(25.8)	29.2(26.8)^
Conceptual Understanding	Part-of-whole	11.1(47.3)+	-5.0(59.4)^	-9.9(44.6)
	Addition-as-incr	65.3(44.7)+	56.3(48.6)^	30.5(46.4)
	Subtr-as-decreas	52.9(49.9)+	27.5(44.4)	34.2(47.3)^
	Rel-part-of	22.2(42.8)^	7.5(24.5)	23.7(42.1)+
	Modelling-total	54.5(30.5)+	33.1(24.6)	35.6(27.4)^

Note: +=best, ^=middle learning gains, *=also conceptual skill.

Figure 11 Descriptive statistics with standard deviation for cognitive skills (9th-10th-grade)

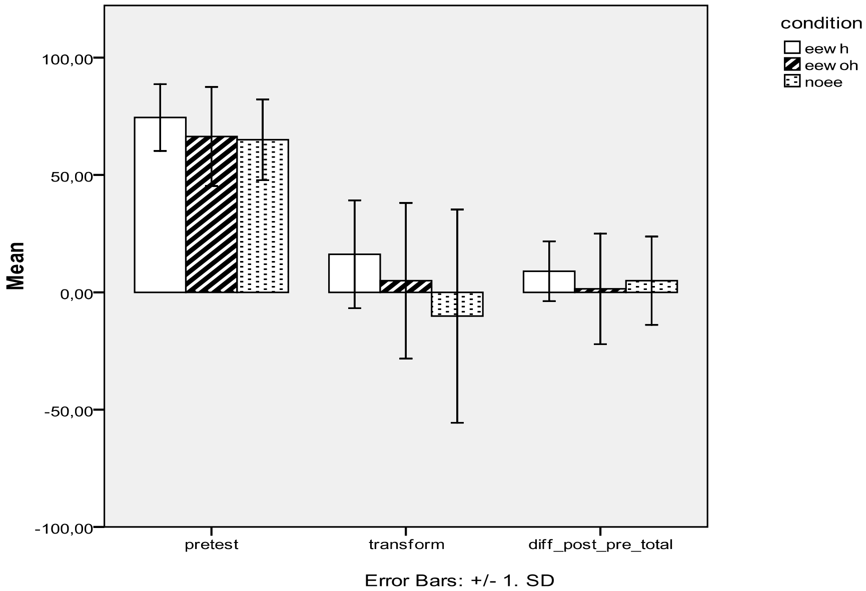


Figure 12 Descriptive statistics with standard deviation for metacognitive skills (9th-10th-grade)

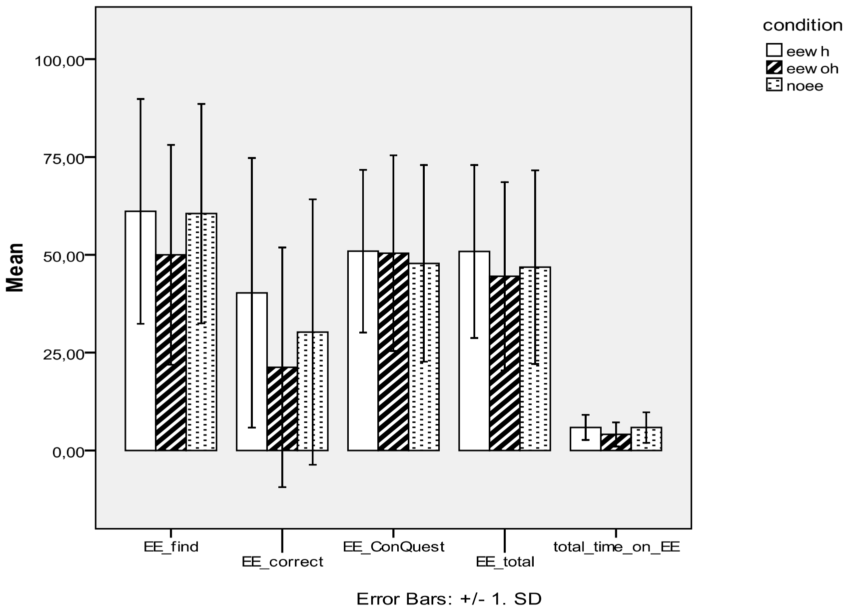


Figure 13 Descriptive statistics with standard deviation for transfer (9th-10th-grade)

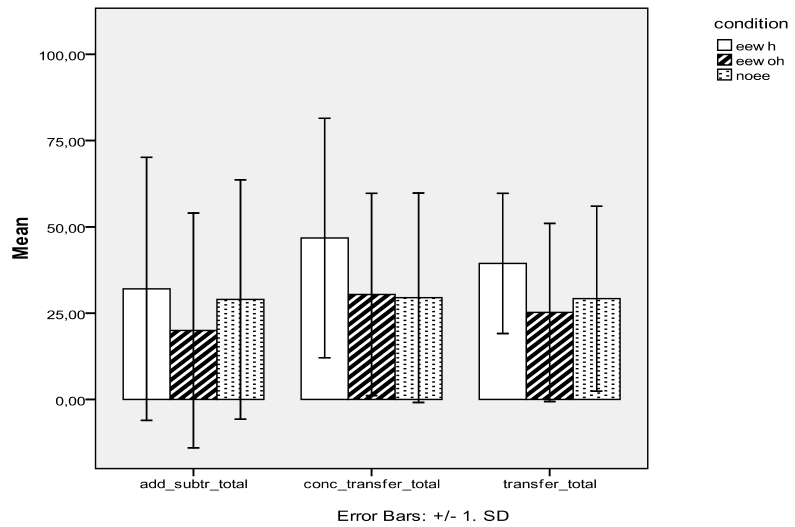
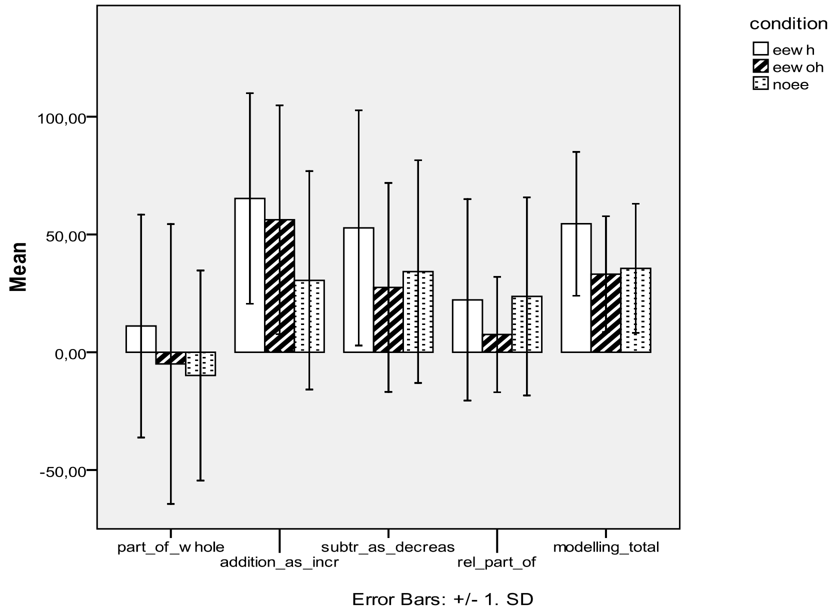


Figure 14 Descriptive statistics with standard deviation for conceptual understanding (9th-10th-grade)



ANOVA Results

The difference in favour of EEWH for the time-on-task were significant, both for the total intervention duration ($F(2,54)=10.1$, $p=.000$, $n^2=.29$) and for the time spent on erroneous examples or equivalent standard exercises, which applies for NOEE, (F(2,54)=35.45, p=.000, n^2=.57). The biggest non-significant differences were also in favour of EEWH and for the variables conceptual knowledge (world problem of basic concepts) in total (F(2,54)=3.03, p=.057, n^2=.11), and for modelling the basic concept “addition as increasing” (F(2,54)= 2.81, p=.067, n^2=.09) (cf. also Table 3).

Moreover, the cognitive skills in the exercises increased more for EEWH who also had a lower variance than for the other two conditions, although the difference was not significant in the analysis of variance. EEWH reached the mean of 83.4 (SD=14.1) in the posttest and surpassed the other two conditions by about 15% (EEWOH: M=67.9, SD=21.1 and NOEE: M=69.9, SD=17.2) although they started with a higher pretest (cf. Table 3). This difference in the posttest was also significant (F(2,56)=3.49, p=.038, n^2=.13).

ANOVA Planned Contrasts

Main Effects. In ANOVA planned contrasts there were main effects for erroneous examples for *time-on-task* (intervention duration: t(53)=4.03, p<.001, d=0.86, r=.40 / EE or equivalent: t(49.72)=8.45, p<.001, d=2.4, r=.77, unequal variance assumed), for the subscore transformation (t(54)=2.09, p<.05, d=0.57, r=.27), but not for cognitive skills in general as well as for the subscore “addition as increasing” (t(54)=2.31, p<.05, d=0.63, r=.30), but not for conceptual understanding as a whole. However, NOEE spent significantly more time on the standard exercises common to all conditions in comparison to EEWH and EEWOH together (t(53)=3.22, p<.05, d=0.88, r=.40).

EEWH vs. NOEE. EEWH had more time-on-task (intervention duration: t(23)=4.67, p<.001, d=1.28, r=.45 / EE or equivalent: t(23)=8.43, p<.001, d=3.46, r=.86, unequal variance assumed) than NOEE. EEWH was better in transformation (subscore for cognitive skills) (t(23)=2.24, p<.05, d=0.86, r=.40, unequal variance assumed), in conceptual understanding (t(23)=2.09, p<.05, d=0.57, r=.27) and its subscore “addition as increasing” (t(23)=2.27, p<.05, d=0.62, r=.30).

EEWH vs. EEWOH. EEWH was better than EEWOH in conceptual understanding in general (t(30)=2.54, p<.05, d=0.69, r=.33), in *transfer* (t(30)=2.54, p<.05, d=0.69, r=.33), and they also had more time-on-task (intervention duration: t(30)=2.54, p<.05, d=0.7, r=.33 / EE or equivalent: t(30)=4.06, p<.001, d=1.43, r=.58, unequal variance assumed).

ANCOVA Results

We tested the possible covariating effect of the pretest score. The pretest score was meant to indicate significant differences based on the prior fraction knowledge. The results show that it has a covariating effect on learning for the cognitive skills (F(1,54)=12.88 p=.001, n^2=.50) and separately for the transformation subscore (F(1,54)=6.60, p=.013, n^2= .34), as well as for the cognitive transfer (F(1,54)=5.16, p=.027, n^2= .30). It also had a covariating effect on the metacognitive scores (total score on erroneous examples) (F(1,54)=4.36, p=.042, n^2=.28) as well as for correcting the error separately (F(1,54)=5.09, p=.028, n^2=.29). Taking these into account, the time-on-task remains significantly longer for EEWH (F(2,54)=9.64, p=.000, n^2=.49), the effect for transformation subscore becomes significant (F(2,54)=3.52, p=.037, n^2=.34) whereas the effect for the conceptual knowledge subscore “addition as increasing” is stronger but remains insignificant (F(2,54)=2.89, p=*064*, n^2=.32), always in favour of EEWH.

ANCOVA Planned Contrasts

Main Effects The ANCOVA planned contrasts showed the same main effects of erroneous examples as the ANOVA contrasts. More specifically, there are main effects of erroneous examples for the *time-on-task* ($t(54)=3.56$, $p=.001$, $d=0.97$, $r=.43$), for the transformation subscore ($t(54)=2.42$, $p<.05$, $d=0.66$, $r=.31$), and for the concept "addition as increasing" ($t(54)=2.32$, $p<.05$, $d=0.63$, $r=.30$).

EEWH vs. NOEE Differences between EEWH and NOEE are significant for time-on-task ($t(23)=4.23$, $p<.001$, $d=0.78$, $r=.36$), for transformation ($t(23)=2.87$, $p<.05$, $d=0.97$, $r=.43$), and for the subscore "addition as increasing" ($t(23)=2.35$, $p<.05$, $d=0.64$, $r=.30$), but not for conceptual understanding as a whole ($t(23)=1.74$, $p=.09$, $d=0.47$, $r=.25$).

EEWH vs. EEWOH The significant differences between EEWH and EEWOH include the scores for cognitive skills (t(30)=2.13, p<.05, d=0.58, r=.27), and for conceptual understanding (t(30)=2.10, p<.05, d=0.58, r=.27).

*Other Results A*lthough we did not find any significant difference between conditions in metacognitive skills, we again found that significantly more students across conditions could find the error in the posttest erroneous examples than could correct it ($t(56)=8.94$, $p<.001$, $d=0.87$, $r=.397$). This difference was also significant for individual conditions, when comparing finding vs. correcting the error (EEWH : $t(17)=3.83$, $p<.05$, $d=0.66$, $r=.31$; EEWOH: $t(19)=5.88$, $p<.001$, $d=0.98$, $r=.44$; NOEE: $t(18)=5.75$, $p<.001$, $d=0.97$, $r=.44$). However, the effect is less strong for EEWH.

Questionnaires' Results. Forty-eight participants completed both the pre- and the post-questionnaire: EEWH=18, EEWOH=16, NOEE=14. Some students from the EEWOH and the NOEE conditions chose not to fill in the post-questionnaire. The students who did not fill in the questionnaires were students who struggled throughout the experimental sessions, which is what probably led to their lack of motivation to fill in the post-questionnaire. This probably makes the results much harsher on the EEWH condition whose participants, including the ones who struggled, all filled in the questionnaires.

In paired sample t-test, all self-reports were significantly worse in the post-questionnaire, apart from cognitive load, which was better, but not significantly. However, there were no significant differences between conditions when comparing the drop between pre and posttest.

There were no interesting results in the analysis of variance, however, as expected, the term-grade had a covariating effect on the cognitive load ($F(1,45)=8.15$, $p=.007$, $n^2=0.16$), unlike in the 6th-grade. This makes the difference in the reported for cognitive load drop significantly higher for EEWH than for NOEE ($t(30)=2.22$, $p<05$, $d=0.24$, $r=.012$), whereas the difference between EEWH and EEWOH just missed significance ($t(28)=2.05$, $p=.05$, $d=0.14$, $r=.07$). However, the effect sizes are small in both cases.

Another interesting result is that there is a significant negative correlation between the reports of self-efficacy in the pre-questionnaire with both the amount of help ($r(46)=-.71$, $p=.001$) and the amount of time spent on erroneous examples ($r(46)=-.49$, $p=.045$) during intervention. This possibly means that the more students felt able to tackle fractions, the less help they received and the less time they needed to work through erroneous example, thus confirming their self-reports.

With regard to the students' self-reports on motivation, they did not correlate with the time they spent on the erroneous examples ($r(46)=-.21$, $p=.43$). This means that they did not apply themselves as expected from their self-reports, which is also reflected on the rather low learning effects. The motivation (*b*=.13, *t*(45)=.65, *p*>.05) and self-efficacy (*b*=-.08, *t*(45)=-.43, *p*>.05) reported in the posttest were also not good predictors of the time spent in the posttest.

Table 6 Descriptive Statistics of Questionnaires 9th, 10th-Grade

	Condition	*EEWH N=18*	*EEWOH N=16*	*NOEE N=14*
Construct	*pre vs. post*	*Mean (sd)%*	*Mean (sd)%*	*Mean (sd)%*
motivation	Pre	52.93(14.84)^	49.38(15.59)	61.43(14.73)+
	Post	35.00(18.26)^	29.69(17.37)	42.14(18.26)+
Err-awareness	Pre	57.89(34.57)^	66.25(32.43)+	52.86(24.32)
	Post	37.89(27.40)^	26.25(21.56)	45.71(35.46)+
Crit-thinking	Pre	50.53(22.23)+	45.63(24.21)^	39.29(12.69)
	Post	33.16(17.34)^	32.50(22.06)	42.86(27.01)+
Cognitive-load	Pre	36.49(20.05)+	39.58(16.77)	38.57(21.59)^
	Post	30.53(16.67)+	36.67(20.37)^	37.62(19.67)
Learn-orient.	Pre	50.00(14.81)^	50.94(11.72)+	49.64(12.93)
	Post	42.63(20.51)^	31.56(20.79)	43.93(17.34)+
Self-efficacy	Pre	71.05(16.29)+	61.88(14.71)	67.14(18.58)^
	Post	52.63(24.00)^	50.00(25.29)	57.14(29.20)+

Note: +=best, ^=middle.

Additionally, we found that students self-report on error-awareness (*b*=.17, *t*(45)=1.21, *p*>.05) and critical-thinking (*b*=.002, *t*(45)=.012, *p*>.05) in the pre-questionnaire was probably not an accurate estimation as it could not predict the performance on the relevant metacognitive skills in the posttest: finding the error, correcting it and answering conceptual questions.

5.2 Discussion: 9th and 10th-Grade Classroom Study

The most striking result is that erroneous examples with help had a significant effect on the cognitive skills as compared to erroneous examples without help. This was not the case in the comparison to no erroneous examples. The reason for that might be that the NOEE condition spent significantly more time on standard exercises practicing cognitive skills unlike the erroneous examples conditions as evidenced by the ANOVA contrasts (main effect for NOEE for standard-exercises duration; *t*(53)=3.22, *p*<.05, *d*=0.88, *r*=.430). Despite of that, there were main effects of erroneous examples on the transformation subscore of cognitive skills. One should be careful with the interpretation of that finding, as the EEWH condition saw a few pizza representations as part of some EAD feedback (cf. Figure 4), which bore similarities to the representations they were later asked to transform in order to make the calculations. Still, two facts make this finding interesting. First, that EEWOH who did not see any such representations also scored significantly higher in this kind of exercise than NOEE. Second, that there was no significant difference on transformation skills between EEWH and EEWOH.

Moreover, the main effect on the conceptual understanding subscore "addition as increasing" shows, at least partially, that the erroneous example conditions benefitted indeed from the conceptual focus of the erroneous examples. This focus was even stronger in the error-detection and error-correction help (see Sections 2.1.3 and 5.1.3), which is also reflected in the significant differences in conceptual understanding between EEWH and EEWOH and big between EEWH and NOEE.

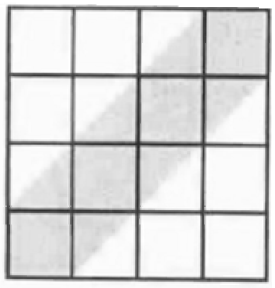
The effects of erroneous examples, especially in combination with help, become more interesting if one considers that EEWH also reported more reduced cognitive load in the post-questionnaire in comparison to the pre-questionnaire. Although the effect size is small, this is a good indication that for students of higher grade working with erroneous examples makes it easier to understand and deal with fraction problems, including erroneous examples. This is not true for erroneous examples without help.

A puzzling result at first sight is the high variances and very low means observed in modelling the basic concepts tested in this experiment. This is an indication that some students could understand the principle behind them and had no problems applying them, whereas others were just confused. This effect is particularly high for the EEWOH in modelling "part of a whole", as well as for modelling "relative part of" that was not taught at all during intervention, but was meant to test transfer from the more general concept "part of a whole". Both of these concepts seem to have been particularly confusing for EEWOH and NOEE. The explanation for the NOEE seems to be obvious, namely that they did not receive training with erroneous examples which, based on our hypothesis, would increase their conceptual understanding. On the contrary, the cause of the higher variance and the negative learning effect in modelling "part of a whole" for EEWOH is not that clear. It may mean that this condition was confused by being asked to represent the difficult concept "part of a whole" explicitly and conceptually, as opposed to the standard school algorithmic approach. Since they received no help, they could not recover from the confusion at all, unlike EEWH, and scored badly both in this trained concept ("part of a whole"), and in the transfer concept ("relative part of").

On the contrary, the somewhat higher learning effect of EEWH can be attributed to the extra help they had in dealing with the new approach to this concept. This resulted in scoring better at the relevant exercise, as well as in transferring from the concept "part of a whole" to "relative part of". The high variances in the EEWH condition are an indication that some students remained confused and did not grasp the underlying concept. Looking at the data, students who did not solve the exercise correctly often did not make an attempt at the first step, which supports that they did not grasp the underlying concept necessary for the first modelling step. These might be students who rely on purely procedural/algorithmic solutions and would need more practice than the one exercise they trained with. Another supportive evidence for the students' confusion is mirrored in the fact that many students in the NOEE used the standard algorithmic solution learned at school to solve modelling problems. For example, in the posttest they had to calculate the part of the square that is not shaded in Figure 15. The expected conceptually adequate answer was 1-7/16, indicating that the students have understood that they have to find the part of the whole and that the whole is represented through 1. The solution a lot of students in the NOEE condition provided was 16/16-7/16. This solution is correct and was counted as correct, but does not make it clear that the students have understood the underlying concept. Similarly, NOEE managed to score better than the erroneous example conditions in modelling the untaught "relative part of" (although not significantly) by simply using the standard algorithmic strategy taught at school.

A simple explanation for the lack of the expected transfer between the concepts "part of a whole" and the "relative part of" is that the participants never mastered the taught concept in order to be able to transfer from it, but, at the same time, their original algorithmic strategy had been destabilised through the experimental intervention. However, one cannot exclude the possibility that the theoretically subordinate category of "relative part of" is actually not cognitively subordinate, which is prerequisite for transfer to occur.

Figure 15 Posttest exercise on the concept "part of a whole"



It is intriguing that there were no effects for erroneous examples with regard to metacognitive skills. Although there is no clear explanation for that, it is possible that students, and especially the more competent ones, did not spend the necessary time on erroneous examples in the posttest, which measured these competencies. The fact that the students' reports on self-efficacy did not correlate with the time spent on erroneous examples during intervention, and the negative correlation between the reports on self-efficacy and the steps taken during intervention imply that possibly the more competent students who could spot the error and directly choose the right explanation might have actually needed more help on correcting the error to improve their metacognitive skills.

The students' inability to assess their error awareness and critical-thinking, which did not predict their performance on finding and correcting the error in the posttest, could be an indication that in fact erroneous examples fine-tuned their self-assessment. That is, students who worked with erroneous examples during intervention were made aware of their lack of error-awareness and critical thinking, which they reported in the posttest. It is quite interesting, that these self-reports in the post-questionnaires are actually closer to their scores in correcting the error. Especially for NOEE, the students' perception did not change as they did not get any feedback on their relevant abilities. This interpretation would explain the unexpected, although not significant, results in the error-awareness and critical thinking constructs (cf. Table 6).

Moreover, the fact that the term-grade has a covariating effect on the cognitive load reported by the students of 9th and 10th grade in the questionnaires could mean that the erroneous examples with help imposed less cognitive load on the more competent students in mathematics. That is in line with work on how automated schemata can explain differences between novices and experts (Chi et al., 1981; Reimann and Chi, 1989), as well as with the findings of Gross and Renkl (Gross and Renkl, 2007; Renkl, 1997). In fact it could be the explanation behind why more competent students benefit more from erroneous examples.

6 General discussion and implications for cognitive modelling

In general, we had some results that supported the use of erroneous examples with additional help in teaching fractions and some that reveal different tendencies depending on the class level. In the following, we discuss these results thematically based on our hypothesis, while in every section we also review the influence of grade level. We compare the different grades although the two lower grades-levels (6th and 7th-8th) were tested in the lab, because the two differences in the setting arguably counter-balance each other. These differences are, the presence of the teacher in the class studies, which could add motivation for grades 9th and 10th that were tested in the classroom, and the payment received by grades 6th, 7th and 8th for their participation in the lab study, which could also motivate students to work harder. Other differences, for example in the materials used, are taken into consideration in the relevant discussion sections. Still, when comparing the results between grades 6th, 7th and 8th with those of grades 9th and 10th one must keep in mind that their ecological validity is lower as they were lab studies.

6.1 Hypothesis 1

6.1.1 Cognitive Skills (H1a), Conceptual Understanding (H1b), and Transfer (H1c)

In our studies, we found that more advanced students (9th and 10th-grade) benefit from erroneous examples with help in terms of cognitive skills (including standard problem solving) in general, as opposed to erroneous examples without help, and partially as opposed to no use of erroneous examples. Although this was not the case for either of the two less advanced levels that we tested, it might have been an artefact of the very high prior fraction knowledge of the particular participants (6th, 7th, and 8th-grade). In particular for the middle grade level (7th, and 8th-grade), it is possible that the problems they face with fractions are also more conceptual rather than procedural and that they might benefit rather from the conceptual material. Moreover, we had some evidence that deep conceptual understanding is supported by erroneous examples with additional error-detection and error-correction help. Such evidence includes the better performance of the EEWH over the NOEE condition at the conceptual questions in the 6th-grade, as well as the main effects in modelling "addition as increasing" for EEWH vs. NOEE and in modelling in general for EEWH vs. NOEE (big but not significant) and EEWH vs. EEWOH (significant) for the 9th and 10th grade. The higher grades (9th, 10th) are the ones that received more intervention materials aiming at conceptual understanding. The difference in conceptual understanding between EEWH and EEWOH for the same grade levels might have also instigated the respective difference in cognitive skills.

Our results do not show a benefit using erroneous examples, with or without help, for increasing cognitive skills or conceptual knowledge in the 7th and 8th-grade. For this grade levels, prior knowledge seems to play a crucial role. A reason for that might be the combination of the high grade level but also the high competence (term grade and pretest scores) which the participants had. Students of the 9th and 10th grade shared the high grade level, but not the level of competency. They were an average school class, and hence a more representative sample.

The higher transfer scores of EEWH in the 9th and 10th grades are promising, but little transfer occurred in 7th and 8th grades. Moreover, there were no significant differences in any of the grades 6th, 7th, or 8th. The transfer scores for 6th graders are high across conditions, which is probably the result of the corresponding high metacognitive learning gains that were observed in this level. Similarly, the low cognitive and metacognitive gains in the 7th and 8th grade explain the low transfer scores. On the contrary, 9th and 10th grade scored rather low because transfer was also measured on modelling exercises that were far more demanding than standard fraction exercises. These results together probably mean the conceptual categorisation of problems inside a sequence, which was done for the 9th and 10th grade is in the right direction for transfer to occur and for the learning potential of erroneous examples to unfold. Grades 6th, like grades 7th and 8th did not receive concept-related sequences during intervention, which might be one reason behind the lack of differences in transfer scores between conditions. A more explicit representation of the concept dealt with in the sequences that were used for grades 9th and 10th might be necessary for students to assign a problem-solving schema to a concept, like suggested by Catrambone and Holyoak (1989) and be able to retrieve it later for application. Research on conceptual chunks by Koedinger and Anderson (1990), points at the same direction for improving transfer skills.

As a whole, our results from the more advanced 9th and 10th grade show clear indications that fostering conceptual understanding through the use of erroneous examples with additional help can result in significant learning effects for conceptual knowledge, but also for cognitive skills. Moreover, although standard cognitive skills are also fostered through extensive practice with standard exercises, such practice does not suffice to improve all kinds of cognitive skills, or conceptual knowledge. In our results this is especially true for the well-practiced fraction addition, were students learned or reminded themselves of the algorithmic steps, but they could not improve significantly either in transformation skills that also addressed fraction addition, or in conceptual understanding of fraction addition. We consider the results from the 9th and 10th grades particularly important, first, because the turn to the more conceptual learning material was made in this study and, second, because there was no ceiling effect, and third, because the setting was more ecological.

6.1.2 Metacognitive competencies (H1d): error detection vs. error correction

We had evidence that erroneous examples can influence the metacognitive skill of error detection for lower-grade (6th-grade) but highly competent students. There is a possible twofold explanation for this. First, these students, who have just learned fractions can handle the demanding erroneous examples because the cognitive skills and domain knowledge that erroneous examples presuppose is readily available to them. Second, there is room for improving their error-detection significantly as they have not yet applied much of what they have learned to make errors themselves and, hence, to practice error-detection on their own errors.

There were no significant differences in students' metacognitive skills for the other class levels. Nonetheless, it was interesting to find out that students of the higher grade level (9th-10th-grade) often did not judge correctly their ability for critical thinking and error awareness. The results indicated that dealing with erroneous examples made their judgement more accurate.

An interesting mismatch between the competencies of finding and correcting the error across conditions is evident in our results. This mismatch persisted in all our studies independent of student level or material design, and it was significant in our studies with the two groups of higher-grade students (7th-8th and 9th-10th-grades). Ohlsson (1996) has described this phenomenon as dissociation between declarative and practical knowledge. Here declarative knowledge here means rule definitions, which relates to recognising the violation of rules and hence spotting the resulting errors. Practical knowledge means rule applications, which relates to applying the correct rule after spotting the error in order to correct it. It is intriguing that in our classroom studies with 9th and 10th-graders, students' cognitive skills did improve through erroneous examples, despite the fact that their ability to find errors developed significantly more than that of correcting errors. This might show that the competence of correcting typical errors is not necessary for monitoring, correcting, or avoiding one's own errors. That is consistent with Ohlsson's (1996) argument, that when the competency for finding errors is active, it functions as a self-correction mechanism that, given enough learning opportunities, can lead to a reduction of performance errors. However, it is a new finding in comparison to previous research in erroneous examples that has not differentiated between the competencies of finding and correcting errors.

6.2 Hypothesis 2: erroneous examples with or without help

The choice between help or no help pertains to the microadaptation of erroneous examples. Although we found some main effects for erroneous examples for the less advanced 6th-grade (metacognitive skills) and the more advanced 9th and 10th-grade (conceptual understanding), most effects were for erroneous examples with help. This is consistent with the results of Kopp and colleagues (2008) in the medical domain in terms of the benefit of erroneous examples with help, although the domains differ a lot and therefore a comparison is tenuous.

We also found that the use of erroneous examples without help might be worse than no use of erroneous examples for conceptual and transfer skills, which is not reliably true for metacognitive skills. As a whole, the inconsistent performance observed in the classroom study with regard to the modelling might mean that there was a conflict between the standard procedural way that teachers normally apply to teach fractions at school and the conceptual way our erroneous examples deal with fractions. This effect might be stronger for EEWOH who are left confused, due to the lack of guidance. However, more familiarity with erroneous examples and the conceptual strategy might counter-balance this confusion, especially when combined with provision of help. Siegler (2002) suggested that requests for explanation of correct and incorrect strategies lead to a period of "cognitive ferment" (p. 51), following cognitive conflict, and only later do they cause the development of correct strategies and the ability to self-explain. He attributes this delay to a state of increased uncertainty and variability.

For medium-advanced students (7th and 8th-grade), no difference was found between erroneous examples with or without help.

In general, to continue on Ohlsson's (1996) argument, it seems like erroneous examples with error-detection and error-correction help that specifically train finding errors and explaining them might offer the required learning opportunities without the need to develop error-correction skills, which was very moderately observed in our data. The help we provided assisted students to explain errors conceptually, but also to understand the practical/procedural implications of these conceptual explanations in terms of problem solving. The contribution of such help is also in line with the theoretical work by van Gog and her colleagues (Van Gog et al., 2004) who have advocated its use in the context of worked examples as a way of promoting conceptual understanding.

6.3 Hypothesis 3:grade level

We already discussed differences in grade level in the previous sections. As a summary, we have found more support for the use of erroneous examples as an instructional method for the more advanced students of the 9th and 10th grades who have had fraction courses in previous years.

For students just learning fractions, namely 6th-grade students, we found that their metacognitive abilities were enhanced. These metacognitive gains for erroneous examples with help did not give rise to enhanced cognitive skills. One could suspect that the cognitive load might have been too large to allow the pass from metacognitive skills to schema creation and hence cognitive skills. In fact, cognitive load was experienced as high by students of this level independent of their previous mathematical knowledge, as we found no significant covariating effect of the term grade on the cognitive load self-reports, contrary to what we expected. The possibility, however, still remains that the existing high level of cognitive skills (ceiling effect) did not allow learning effects to occur.

We did not find supportive evidence for the use of erroneous examples with students of medium level (7th and 8th grade). As mentioned above, the reason for that might be that the materials used were not appropriate to induce learning at this level.

Moreover, contrary to what we expected due to the use of adaptive help, the grade level appears to play a role in whether students learn from erroneous examples with help. This can be an indication that the more conceptual adaptive help triggered germane cognitive load for students of a higher grade level (and hence higher prior knowledge). For students of lower grade level, for which the material was less conceptual the adaptive help was not enough to cause the required germane cognitive load in the form of cognitive conflict. This difference could have led to the comparatively higher learning gains.

6.4 Supplementary Conjectures

6.4.1 Presentation of erroneous examples

Regarding the presentation of erroneous examples, which relates to macroadaptation, we have at least a first indication that they are more beneficial when presented after the students have been confronted with standard exercises and followed again by standard exercises, since we only found a significant improvement at tasks other than erroneous examples when this order of presentation was used. A potential explanation is that this gives students the opportunity to review the material before working with erroneous examples that might also increase the perceived relevance of erroneous examples, as well as to practice what they have learned after the presentation of the erroneous examples. However, this might be different for students who are just learning fraction operations, or for students of lowers competency and self-regulation skills who might need more practice with standard fraction problems before confronting them with erroneous examples. This could allow them, first, to practice with the problems at all and, second, to become more aware of the difficulties involved before they can understand and work with erroneous examples.

6.4.2 Motivation, cognitive load, and learning orientation

There were no significant differences between conditions for measures of motivation, however, neither motivation nor self-efficacy seem to be a good criterion for whether students learn on not from erroneous examples and for whether help is effective.

Self-reports of higher-grade students (9th and 10th grade) show that working with erroneous examples and additional adaptive help reduce the perceived cognitive load that is caused by solving fraction problems together with erroneous examples. This would be consistent with our hypothesis. However, the results were not the same for the other grades. Since we used more conceptual materials for the 9th and 10th-grade and we intended to induce germane cognitive load through the use of conceptual help ("why" and "how" questions), this might be an indication that students experienced the required cognitive conflict but also were assisted by the help for resolving it. On the contrary, the materials for the other levels were possibly too easy for cognitive conflict to occur, so that the additional help was perceived as extraneous cognitive load.

We did not have any clear indications that learning orientation is fostered through the use of erroneous examples.

6.5 Open questions

Two main questions remain open: first, how interactive erroneous examples can be improved in general; second, if and how medium advanced students (7th and 8th-grade) can be assisted in learning with erroneous examples to profit from them. In the following, we discuss these questions from different perspectives and suggest possible solutions.

6.5.1 Design of interactive erroneous examples

A practical measure, in terms of the design of interactive erroneous examples, may be to allow students to explicitly request more help, which would amount to more help on procedural "how" knowledge. It is likely that they will use this extra feature if they feel uncertain about their answer, thus overcoming a possible shortcoming of our design of interactive erroneous examples, which assumes that if students can answer the basic "why" and "how" process-oriented MCQs they do not need error detection and correction help. Currently, MCQs providing such additional error detection and correction help are skipped once the student has answered the first two MCQs correctly, in an attempt to avoid a possible "expertise reversal effect" (Kalyuga et al, 2003). Following Kalyuga and his colleagues, we tried to track the existence of knowledge and avoid providing students with redundant help. For that reason, we considered answering the top-level self-explanation MCQs as evidence that the students would also possess the knowledge dealt with by the following MCQs. However, this might be too coarse an indicator for when and how much help is needed. Moreover, it underestimates the difficulty students have with applying rules (practical knowledge), as opposed to recognising (declarative knowledge) and the respective benefit of explaining detailed "how" questions in combination with "why" questions. Support for this reasoning is the fact that the students of the 9th and 10th grade who felt able to cope with fractions, based on their self-reports, and received less help did not score as well as one expected. Had they received some additional help on the errors, they might have learned more.

6.5.2 Materials and instructional design

The materials and instructional design might also need modifications. For instance, the results might be clearer if we enrich our conceptual exercises and test rather whether errors that reveal lack of conceptual understanding are committed. We want to elaborate more on such conceptual exercises since the standard fraction exercises practiced at school might be too simple to influence students' performance alone through process-oriented ("how") help, as we have observed in our studies with the less-advanced and medium-advanced students. This is hypothesised from a theoretical perspective by Ohlsson (1996) and van Gog and colleagues (Van Gog et al., 2004) and was empirically tested in the medical domain with positive results for erroneous examples with help (Stark et al., 2011). A good start would be to try to replicate our results for the advanced students (9th and 10th-grade) using the new, more conceptual materials with the other grade levels, and especially with the 7th and 8th grades. A more representative sample in terms of prior math and fraction knowledge is also a prerequisite for this test. Furthermore, the replication of the results would help rule out the possibility that the materials alone and not the level made the difference in our results.

6.5.3 Presentation

We plan to test whether the order of presentation really plays a significant role, by using the more conceptual material and varying the order of presentation between different conditions. Moreover, it could be the case that explicitly making students aware of the basic concept handled in each sequence would further increase awareness of such concepts and the related errors that indicate lack of awareness of these principles. This might also contribute to better transfer as students would be trained in categorising problem types based on their basic concepts.

7 Conclusions and implications for instructional design

As a whole, our studies reveal a good potential for erroneous examples as an instructional method that can help students in the demanding domain of fractions, although they show room for further improvement. The overall finding that working with erroneous examples with help produces better learning effects than working without help replicates the results of Kopp and colleagues (Kopp et al., 2008; Stark et al., 2011). The studies also indicate that previous results on the benefits of self-explaining correct and incorrect examples by Siegler and colleagues in water displacement and mathematical equality problems (Siegler, 2002; Siegler et al., 2008) and Grosse and Renkl (2007) in probability problems are transferable, first, to using Interactive erroneous examples alone, and second, to the fraction domain. Analogous to the aptitude-treatment effect that Grosse and Renkl (2007) observed with regard to transfer, and despite our expectation that help might counter-balance such an effect, we found that the students' grade level may be important for potential benefit from erroneous examples in general. However, we did not find transfer effects for erroneous examples. Overall, the fact that erroneous examples with help caused less cognitive load to students of higher grade levels who received conceptual materials suggests a potential similar effect to worked examples (correct solutions), as often discussed in the relevant work (Pass and Merrienboerg, 1994; Renkl, 1997; Trafton and Reiser,1993). Stark et al. (2011) have looked at cognitive load as a covariate of learning from erroneous examples. A more detailed investigation of cognitive load to differentiate between the kinds of cognitive load induced through erroneous examples with help would be even more interesting in view of the desired cognitive conflict, which would constitute germane cognitive load in the case of erroneous examples.

The work presented, generated interesting research questions that remain to be answered. As an outcome of this work, first implications for instructional design can be formulated.

In general, erroneous examples are recommended rather as an instructional method for higher-level grades if the aim is to enhance both cognitive skills and conceptual knowledge. They should, however, be used with additional help which should be elaborate when erroneous examples first start being presented for learning, which is consistent with the findings of Stark et al. (2011).

Our current results indicate that erroneous examples should concentrate on finding the error and explaining it, rather than on correcting it. The competency of correcting common errors or misconceptions in the domain does not seem to be necessary for avoiding making errors. However, it has the disadvantage of being time consuming. This is particularly important for educational technologies as it reduces the costs of developing software, including domain reasoners that are necessary to provide error-specific feedback and feedback modules or authoring tools for designing or authoring this feedback.

Moreover, erroneous examples seem to be more effective when addressing conceptual knowledge directly, as compared to only dealing with practical errors commonly committed by students. This is true, even though practical errors are often indications of missing knowledge, or misconceptions. In our next steps, we will be testing this finding and the influence of grade level further.

Furthermore, when basic concepts are addressed by erroneous examples, caution should be taken that the inconsistencies with the standard algorithmic approaches are addressed and resolved. The aim of such caution is not just to avoid confusion, but rather to take advantage of the cognitive conflict induced by the erroneous examples and reveal the common underlying principle of both approaches. Specifically, in relation to the cognitive conflict caused by erroneous examples, the delayed effects of erroneous examples should also be tested to replicate effects from previous studies (Mclaren et al, 2012; Stark et al., 2011).

Self-efficacy seems to be a decisive learner characteristic that influences whether students learn from erroneous examples or not.

In conclusion, these first directions for instructional design must be further tested and elaborated. In addition, a cognitive model of how erroneous examples with help advance learning should be sketched based on empirical results and relevant theoretical viewpoints. This will allow the formulation and testing of hypotheses combined in a coordinated attempt. Such testing should also involve, for instance, the examination of cognitive processes through the collection and analysis of think-alouds.

Beyond learning in the classroom, learning from errors in general and acquiring metacognitive skills of detecting and fixing errors can prove to be a key 21st century competence especially in the context of informal learning. For instance, it can be a crucial supplement of information validation.

Acknowledgements

In memory of Erica Melis who pioneered the research of erroneous examples in the context of technology enhanced learning. This work was supported by the DFG - Deutsche Forschungsgemeinschaft under the ALoE project ME 1136/7.

References

Borasi, R. (1994) 'Capitalising on errors as "springboards for inquiry": a teaching experiment', *Journal for Research in Mathematics Education*, Vol. 25, No. 2, pp.166–208.

Catrambone, R. and Holyoak, K.J. (1989) Overcoming contextual limitations on problem-solving transfer', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 15, No. 6, pp.1147–1156.

Catrambone, R. (1994) 'Improving examples to improve transfer to novel problems', *Memory & Cognition*, Vol. 22, pp.606–615.

Catrambone, R. (1998) 'The subgoal learning model: creating better examples so that students can solve novel problems', *Journal of Experimental Psychology: General*, Vol. 127, No. 4, pp.355–376.

Chi M.T.H., Feltovich, P. and Glaser, R. (1981) 'Categorization and representation of physics problems by experts and novices', *Cognitive Science*, Vol. 5, pp.121–152.

Durkin, K.L. and Rittle-Johnson, B. (2008) 'Comparison of incorrect examples in math learning', Poster presented at *The IES annual research conference*, Washington, D.C.

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

Grosse, C.S. and Renkl, A. (2008) 'Finding and fixing errors in worked examples: Can this foster learning outcomes?', *Learning and Instruction*, Vol. 17, pp.612–634.

Newton, J.K. (2008) 'An extensive analysis of preservice teachers' knowledge of fractions', *American Educational Research Journal*, Vol. 45, No. 4, pp.1080–1110.

Kalyuga, S., Ayres, P., Chandler, P. and Sweller, J. (2003) 'The expertise reversal effect', *Educational Psychologist*, Vol. 38, No. 1, pp.23–31.

Knezek, G. and Rhonda, C. (1996) 'Validating the Computer Attitude Questionnaire (CAQ)', Paper presented at *The Annual Meeting of the Southwest Educational Research Association*, New Orleans, LA.

Koedinger, K.R. and Anderson, J.R. (1990) 'Abstract planning and perceptual chunks: elements of expertise in geometry', *Cognitive Science*, Vol. 14, pp.511–550.

Kopp, V., Stark, R. and Fischer, M.R. (2008) 'Fostering diagnostic knowledge through computer-supported, case-based worked examples: effects of erroneous examples and feedback', *Medical Education*, Vol. 42, pp.823–829.

Malle, G. (2004) 'Grundvorstellungen zu Bruchzahlen', *Mathematik Lehren*, Vol. 123, pp.4–8.

McLaren, B.M., Lim, S.J. and Koedinger, K.R.(2008) 'When and how often should worked examples be given to students?', in Love, B.7.C., McRae, K., V.M.S., (Eds): *New results and a summary of the current state of research*', *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, Cognitive Science Society, pp.2176–2181.

McLaren, B.M., Adams, D., Durkin, K., Goguadze, G. Mayer, R.E., Rittle-Johnson, B., Sosnovsky, S., Isotani, S. and Van Velsen, M. (2012) 'To err is human, to explain and correct is divine: A study of interactive erroneous examples with middle school math students', in Ravenscroft, A., Lindstaedt, S., Delgado Kloos, C. and Hernández-Leo, D. (Eds): *Proceedings of ECTEL 2012: Seventh European Conference on Technology Enhanced Learning*, LNCS 7563, Springer, Berlin, pp.222–235.

Melis, E. (2005) 'Design of erroneous examples for ActiveMath', in Bredeweg, B., Looi, C-K., McCalla, G. and Breuker, J. (Eds): *Artificial Intelligence in Education. Supporting Learning Through Intelligent and Socially Informed Technology. 12th International Conference (AIED 2005)*, IOS Press pp. 125, 451–458.

Melis, E., Goguadze, G., Homik, M., Libbrecht, P., Ullrich, C. and Winterstein, S. (2006) 'Semantic-aware components and services in ActiveMath', *British Journal of Educational Technology*, *Semantic Web for E-learning*, Vol. 37, No. 3, pp.405–423.

Müller, A. (2003) 'Aus eignen und fremden Fehlern lernen', *Praxis der Naturwissenschaften*, Vol. 52, No. 1, pp.18–21.

OECD (2001) *International report PISA plus*.

Ohlsson, S. (1996) 'Learning from performance errors', *Psychological Review*, Vol. 103, No. 2, pp.241–262.

Oser, F. and Hascher, T. (1997) Lernen aus Fehlern - Zur Psychologie des negativen Wissens. *Schriftenreihe zum Projekt: Lernen Menschen aus Fehlern? Zur Entwicklung einer Fehlerkultur in der Schule, Pädagogisches Institut der Universität Freiburg*.

Paas, F.G., Renkl, A. and Sweller, J. (2003) 'Cognitive load theory and instructional design: recent developments', *Educational Psychologist*, Vol. 38, pp.1–4.

Paas, F. (1992) 'Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive load approach', *Journal of Educational Psychology*, Vol. 84, pp.429–434.

Paas, F.G. and van Merrienboerg, J.J.G. (1994) 'Variability of worked examples and transfer of geometrical problem-solving skills: a cognitive-load approach', *Journal of Educational Psychology*, Vol. 86, No.1, pp.122–133.

Pintrich, P.R., Smith, D.A.F., Garcia, T. and McKeachie, W.J. (1991) 'A Manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)', Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning, University of Michigan.

Reimann, P. and Chi M.T.H. (1989) 'Human expertise', in Gillhooly, K.J. (Ed.): *Human and machine problem solving*, Plenum, New York, pp.161–191.

Rittle-Johnson, B. and Wagner Alibali, M. (2001) 'Conceptual and procedural knowledge of mathematics: does one lead to the other?', *Journal of Educational Psychology*, Vol. 91, No. 1, pp.175–189.

Renkl, A. (1997) 'Learning from worked-out examples: a study on individual differences', *Cognitive Science*, Vol. 21, pp.1–29.

Schmidt, R.A. and Bjork, R.A. (1992) 'New conceptualizations of practice: common principles in three paradigms suggest new concepts for training', *Psychological Science*, Vol. 3, No. 4, pp.207–217.

Seidel, T. and Prenzel, M. (2003) 'Mit Fehlern umgehen – Zum Lernen motivieren', *Praxis der Naturwissenschaften*, Vol. 52, No. 1, pp.30–34.

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

Siegler, R.S. and Chen, Z. (2008) 'Differentiation and integration: guiding principles for analyzing cognitive change', *Developmental Science*, Vol. 11, pp.433–448.

Skinner, B.F. (1938) 'The behavior of organisms: an experimental analysis', Appleton-Century, New York, US.

Stafylidou S. and Vosniadou, S. (2004) ''The development of students' understanding of the numerical values of fractions', *Learning and Instruction*, Vol. 14, pp.503–518.

Strecker, C. (1999) *Aus Fehlern lernen und verwandte Themen*. Available online at http://www.blk.mat.uni-bayreuth.de/material/db/33/fehler.pdf (accessed on 20 September 2010)

Sweller, J. (1988) 'Cognitive load during problem solving: effects on learning', *Cognitive Science*, Vol. 12, pp.257–285.

Sweller, J., Van Merriënboer, J.J.G. and Paas, F. (1998) 'Cognitive architecture and instructional design', *Educational Psychology Review*, Vol. 10, pp.251–295.

Sweller, J. and Cooper, G.A. (1985) 'The use of worked examples as a substitute for problem solving in learning algebra', *Cognition and Instruction*, Vol. 2, pp.59–89.

Stark, R., Kopp, V. and Fischer, M.R. (2011) 'Case-based learning with worked examples in complex domains: two experimental studies in undergraduate medical education', *Learning and Instruction*, Vol. 21, pp.22–33.

Trafton, J.G. and Reiser, B.J. (1993) 'The contribution of studying examples and solving problems', in *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*.

Tsamir, P. and Tirosh, D. (2003) 'In-service mathematics teachers' views or errors in the classroom', in *International Symposium: Elementary Mathematics Teaching*, Prague.

Tsovaltzi, D., Melis, E., McLaren, B.M., Dietrich, M., Goguadze, G. and Meyer, A-K. (2009) 'Erroneous examples: a preliminary investigation into learning benefits', in Cress, U., Dimitrova, V., Specht, M. (Eds): *Proceedings of the Fourth EC-TEL 2009*, LNCS 5794, Springer-Verlag, Heidelberg, Berlin, pp.688–693.

Tsovaltzi, D., Melis, E., McLaren, B.M., Meyer, A-K., Dietrich, M. and Goguadze, G. (2010) 'Learning from erroneous examples: when and how do students benefit from them?', in Wolpers, M., Kirschner, P.A., Scheffel, M., Lindstaedt, S. and Dimitrova, V. (Eds): *Proceedings of the 5th European Conference on Technology Enhanced Learning (EC-TEL 2010), Sustaining TEL: From Innovation to Learning and Practice*, LNCS 6383, Barcelona, Spain, Springer-Verlag, Heidelberg, Berlin, pp.357–373.

Van Gog, T., Pass, F. and van Merrienboerg, J.J.G. (2004) 'Process-oriented worked examples: improving transfer performance through enhanced understanding', *Instructional Science*, Vol. 32, pp.83–98.

Van Gog, T., Paas, F. and Van Merriënboer, J.J.G. (2006) 'Effects of process-oriented worked examples on troubleshooting transfer performance', *Learning and Instruction*, Vol. 16, pp.154–164.

Notes

1 Motivated Strategies for Learning Questionnaire

2 Computer Attitude Questionnaire

3 The problem entails the usual assumption that apple juice and lemonade bottles have the same volume. There was no evidence that students did not understand this assumption