Computers in Human Behavior 55 (2016) 87–99

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Full length article

The efficiency of worked examples compared to erroneous examples, tutored problem solving, and problem solving in computer-based learning environments

Bruce M. McLaren [a, 1, *], Tamara van Gog [b, c, 1], Craig Ganoe [a], Michael Karabinos [d], David Yaron [d]

[a] Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, USA
[b] Department of Pedagogical and Educational Sciences–Education, Utrecht University, The Netherlands
[c] Institute of Psychology, Erasmus University Rotterdam, The Netherlands
[d] Chemistry Department, Carnegie Mellon University, Pittsburgh, USA

ARTICLE INFO

Article history:
Received 1 June 2015
Received in revised form
19 August 2015
Accepted 24 August 2015
Available online xxx

Keywords:
Worked examples
Erroneous examples
Conventional problem solving
Intelligent tutoring

ABSTRACT

How much instructional assistance to provide to students as they learn, and what kind of assistance to provide, is a much-debated problem in research on learning and instruction. This study presents two multi-session classroom experiments in the domain of chemistry, comparing the effectiveness and efficiency of three high-assistance (worked examples, tutored problems, and erroneous examples) and one low-assistance (untutored problem solving) instructional approach, with error feedback consisting of either elaborate worked examples (Experiment 1) or basic correctness feedback (Experiment 2). Neither experiment showed differences in learning outcomes among conditions, but both showed clear efficiency benefits of worked example study: equal levels of test performance were achieved with significantly less investment of time and effort during learning. Interestingly for both theory and practice, the time efficiency benefit was substantial: worked example study required 46–68% less time in Experiment 1 and 48–69% in Experiment 2 than the other instructional approaches.

1. Introduction

A major and recurring question for teachers and developers of instructional software is how much guidance or assistance they should provide in order to lead to the best learning outcomes for students (see debates and research on high versus low or 'minimal' guidance instruction: e.g., Alfieri, Brooks, Aldrich, & Tenenbaum, 2011; Hmelo-Silver, Duncan, & Chinn, 2007; Kapur & Rummel, 2012; Kirschner, Sweller, & Clark, 2006; Mayer, 2004; Schmidt, Loyens, van Gog, & Paas, 2007; Tobias & Duffy, 2009; Wijnia, Loyens, Van Gog, Derous, & Schmidt, 2014). On the one hand, some educational researchers conjecture that too much instructional assistance can lead to lower learning outcomes and feelings of boredom and demotivation, as students have little to do on their own. On the other hand, other researchers have argued that too little assistance may lead to lower learning outcomes or inefficient and frustrating learning processes when students do not know what to do. The decision of how much assistance to provide students learning with instructional software, balancing between making instructional materials supportive and challenging, has been called the 'assistance dilemma' (Koedinger & Aleven, 2007). When it comes to teaching problem-solving skills, for instance, worked examples are on the high guidance side of the assistance continuum. Worked examples present students with a fully worked-out problem solution to study and (possibly) explain. On the low (or rather: no) guidance side of the continuum are problems that students attempt to solve themselves without any instructional guidance whatsoever.

It is well-established that for novices, studying *worked examples only* (Nievelstein, Van Gog, Van Dijck, & Boshuizen, 2013; Van Gerven, Paas, Van Merriënboer, & Schmidt, 2002; Van Gog, Paas, & Van Merriënboer, 2006) or *example-problem pairs* (Carroll, 1994; Cooper & Sweller, 1987; Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Mwangi & Sweller, 1998; Rourke & Sweller, 2009; Sweller & Cooper, 1985) is generally *more effective* for learning and transfer than practicing conventional problem solving (i.e.,

* Corresponding author. Human-Computer Interaction Institute, 5000 Forbes Avenue, Carnegie Mellon University, Pittsburgh, PA, 15213-3891 United States.
E-mail address: bmclaren@cs.cmu.edu (B.M. McLaren).
[1] First two authors should be considered as first author.

http://dx.doi.org/10.1016/j.chb.2015.08.038
0747-5632/

without any assistance). Moreover, worked examples or example–problem pairs have also been shown to be *more efficient* than conventional problem solving, in the sense that equal or higher test performance is reached in less study time and with less investment of mental effort (an indicator of cognitive load). This has become known as the 'worked example effect' (for reviews, see Atkinson, Derry, Renkl, & Wortham, 2000; Clark & Mayer, 2011; Renkl, 2014a, 2014b; Sweller, Ayres, & Kalyuga, 2011; Sweller, Van Merriënboer, & Paas, 1998; Van Gog & Rummel, 2010).

The efficiency of studying worked examples compared to problem solving makes sense when one looks at the cognitive processes involved. When novices, who lack knowledge of effective problem-solving procedures, have to practice solving problems without any assistance or instructional guidance, they are forced to resort to weak problem-solving strategies, such as means-ends analysis (Simon, 1981), in which learners search for operators to reduce the difference between the current problem state and the goal state (Sweller, 1988). This takes a lot of time and imposes a high load on working memory (i.e., is effortful) but is not effective for learning, that is, for building a cognitive schema of how such problems should be solved (Sweller, 1988; Sweller & Levine, 1982). Consequently, when learners are presented with a subsequent, similar practice problem, they again have to rely on the same, inefficient strategies. When studying worked examples, in contrast, learners do not have to spend time and effort on weak problem-solving strategies, but instead, can devote all of their attention to learning how such problems should be solved, that is, to constructing a cognitive schema that can guide future problem solving when instructional assistance is no longer available.

Worked example study, however, has been criticized as a relatively 'passive' form of instruction. Even though the cognitive schema of the solution procedure has to be actively constructed by a learner, it is constructed based on example study rather than production or generation of problem-solving steps. It has been argued that there is a benefit to sometimes withholding assistance in favor of having learners actively produce or generate solutions (Koedinger & Aleven, 2007). Koedinger and Aleven even suggest that it is unlikely that instruction consisting of only studying worked examples would be better than interleaving worked examples and problem solving (i.e., in which the learner studies *and* engages in problem solving), although they added, "We do not know of such a direct comparison …" (p. 243).

Indeed, in 2007, when their article appeared, no such direct comparisons had been conducted yet. More recent studies, however, have shown that there were no differences in learning outcomes or effort investment between examples only and example–problem pairs and that both were more effective than conventional problem solving only on an immediate test (Leppink, Paas, Van Gog, Van der Vleuten, & Van Merriënboer, 2014; Van Gog, Kester, & Paas, 2011). One might argue in light of research on the testing effect, though, that the benefits of alternating example study with problem solving would only arise on a delayed test. That is, research has shown that after initial study, testing is more effective for long-term learning than restudying, even though on an immediate test there may be no differences or restudy might even be more effective (Roediger & Karpicke, 2006; Rowland, 2014). Given that example–problem pairs resemble a study-test situation whereas example study only resembles restudy, one might expect that example–problem pairs would lead to better performance on a delayed test. However, several studies have shown that this is not the case, and example–problem pairs and example study are equally effective even when learning outcomes are measured one week later (Leahy, Hanham, & Sweller, 2015; Van Gog & Kester, 2012; Van Gog et al., 2015; potentially, this finding can be explained by the complexity of the learning material; see Van Gog & Sweller, 2015).

So in contrast to the suggestion that it is sometimes better to withhold assistance (Koedinger & Aleven, 2007), these findings suggest that giving novice learners full support (i.e., only having them study examples), is neither better nor worse for learning than first providing and then withholding support (i.e. example–problem pairs). However, it should be noted that the above studies were single-session experiments, conducted either in a lab setting or in a single classroom period, involving relatively short sequences of learning tasks. In other words, ecological validity was low and it cannot be ruled out that withholding assistance would have beneficial effects in real classroom settings. Yet, in at least one classroom study, conducted over a period of up to 6 class periods, McLaren and Isotani (2011) found that a condition consisting of only worked examples led to students learning just as much, in significantly less time, than both an alternating examples/tutored problems condition and an all tutored problems condition. This study was different from the aforementioned lab studies in that the worked examples contained some "active" elements (i.e., answering self-explanation questions after viewing worked example videos). Nevertheless, this study provides further evidence that exclusively studying worked examples may be more efficient – although not necessarily more effective – for learning than has been assumed.

Moreover, the McLaren and Isotani (2011) study is important because the effectiveness of worked examples was not compared to "conventional" problem solving, but rather to another 'high assistance' condition, namely tutored problems in which students are supported by hints and feedback on each step when needed. Koedinger and Aleven (2007) have suggested that the worked example effect arises mainly because no guidance whatsoever is given in conventional problem solving: "In the context of tutored practice as opposed to untutored practice, the information-giving benefits of worked examples may essentially be redundant. In essence, the tutor dynamically converts a problem-solving experience into an annotated worked example when the student is having enough trouble such that they request the final 'bottom-out' level of hint that tells them what to do next". (p. 257).

Koedinger and Aleven (2007) subsequently initiated several studies to investigate this assumption, and found instead that interleaving example study and tutored problem solving proved to be more efficient than tutored problem solving alone (McLaren, Lim, & Koedinger, 2008) and that faded examples with increasingly more steps for the learner to complete with tutor support were more efficient than tutored problem solving alone (Schwonke et al., 2009; for a review of effects of [faded] worked examples in tutoring systems, see Salden, Koedinger, Renkl, Aleven, & McLaren, 2010). The McLaren and Isotani (2011) study goes beyond the prior studies by comparing worked examples only to interleaved example-tutored problem pairs and tutored problems only. Their data show that if students did use the tutored problems to "dynamically convert problem-solving experience(s) into annotated worked example(s)" in this study, it did not help them learn more or learn more efficiently.

This lack of learning benefit might be explained as follows. Whereas it is true that a tutored problem can essentially amount to a worked example when the student gets to the bottom-out hints, getting there is an inefficient process. It is likely to take much more time and effort to work through to the bottom-out hints of many individual problem solving steps than to study a full example. It is questionable whether this is time and effort well spent (especially for low prior knowledge learners), that is, whether it would contribute much to learning compared to studying a fully worked-out solution presented as a whole. Instead, why not give learners an example of a correct solution procedure immediately, rather than

requiring them to make repeated attempts at solving each step or to click through a series of hints at each step in order to obtain that same information?

Next to tutored problem solving, another way to avoid the "passiveness" of worked example study, would be to present learners with erroneous examples and instruct them to find, explain, and fix the errors (e.g., Adams et al., 2014; McLaren, Adams, & Mayer, in press; Durkin & Rittle-Johnson, 2012; Grosse & Renkl, 2007). Not only would this prompt them to study the examples more carefully, it might also help students remember and avoid making those errors in the future. Findings regarding the effectiveness of erroneous examples compared to worked examples, have been somewhat mixed, though. Students with low prior knowledge have been found to benefit more from studying correct worked examples than from a mix of correct and erroneous examples, even if errors were highlighted. Students with more prior knowledge, in contrast, benefited from a mix of correct and erroneous examples (Grosse & Renkl, 2007). However, it seems that even novices can benefit from a mix of correct and erroneous examples compared to correct examples only, when they are explicitly instructed to compare (or contrast) the correct and incorrect examples (e.g., Durkin & Rittle-Johnson, 2012) or when elaborate feedback is given (e.g., Stark, Kopp, & Fischer, 2011). The effectiveness of erroneous examples has also been established compared to supported problem solving (i.e., computer-based problems that provide correctness feedback on student steps, but no hints or error messages; Adams et al., 2014; McLaren et al., in press). In these two studies, the erroneous examples condition led to better performance on a delayed posttest than the supported problem solving condition. Furthermore, this finding was the same for both low and high prior knowledge subjects.

In sum, whereas worked examples, erroneous examples, and tutored problems, all provide a high level of assistance, they differ in terms of whether learners have to actively construct or generate answers at problem-solving steps. The theorized cognitive processes involved in creating problem-solving schemas and the potential learning benefits of each instructional approach, are summarized in Table 1. Worked examples provide the highest degree of assistance of the three, but are also the most passive. An open question in light of the assistance dilemma, is whether these high-assistance forms of instruction are all equally effective and efficient for novice learners compared to no assistance, or whether the forms that require more (inter)activity on part of the learner (i.e., erroneous examples, tutored problems) would be more effective. Despite the fact that all of these instructional formats have been investigated in various empirical studies, such a direct comparison within a single study has never been made before. Moreover, while some of the prior studies have involved classroom work over multiple instructional sessions, a context very close to genuine educational practice, the majority of the evidence in support of the learning benefits of worked examples has come from single-session lab studies.

Therefore, the present study compared the effectiveness and efficiency (both in terms of time and mental effort investment) of worked examples, erroneous examples, tutored problems, and untutored problem solving (i.e. no hints or feedback on steps *during* problem solving but feedback shown after problem solving) over multiple classroom sessions. It is hypothesized that the three high-assistance forms of instruction will be more effective and efficient for learning than untutored problem solving. Based on the findings regarding the efficiency of worked examples versus tutored problems (McLaren & Isotani, 2011), we expect worked example study to be more efficient (though not necessarily more effective) for learning than tutored problem solving. With regard to differences between erroneous examples on the one hand and worked examples or tutored problems on the other hand, it is less clear what to expect. Although Adams et al. (2014) and McLaren et al., (in press) found erroneous examples to be more effective than 'supported problem solving' (i.e., correctness feedback on individual steps, but no hints or error messages), tutored problem solving provides more assistance than their supported problem solving condition. As mentioned above, findings on the effectiveness of erroneous

Table 1
The four instructional approaches compared in this study, with a summary of the theorized cognitive processes to create problem solving schemas, the active learning steps taken by students, and the potential learning benefits of each instructional approach.

Instructional approach	Process to create problem solving schema	Active learning	Potential learning benefits
Worked Examples (WE)	Observe problem-solving steps and create a reusable problem-solving schema to guide future (analogous) problem solving.	The learner actively and directly constructs the problem-solving schema, with no search to interfere with the active schema construction, since all steps are given.	No cognitive resources are wasted on search; Learner uses all of his/her resources to construct the problem solving schema based on given steps. The process of reviewing problem solving steps is much faster than generating those steps.
Erroneous Examples (ErrEx)	Like worked examples, observe problem solving steps and create a reusable problem solving schema to guide future (analogous) problem solving. The schema is tested and revised during construction by identifying steps that are incorrect and correcting them.	The learner actively and directly constructs the problem-solving schema, with relatively little search to interfere with the active schema construction, since all steps are given. However, since some steps are incorrect, the student must then actively – and resource intensively – review and revise the schema.	The learner initially uses all of his/her cognitive resources to construct the problem solving schema, but also has to actively test each given step and if necessary (i.e., if it is incorrect) revise the schema to include the correct step. This is expected to reinforce both the learner's understanding of the problem-solving schema and help in avoiding (typically common) errors in future problem solving.
Tutored Problems (TPS)	Means-ends search to solve problem; consequently, limited cognitive resources are available to construct problem-solving schema. Feedback and hints help student solve problem and construct the schema when learner is stuck.	Learner actively solves the problem and constructs the problem solving schema, but since search is required, it is challenging and resource consuming to construct. Student uses feedback and hints to help in constructing the schema when impasses are reached.	The learner has to create the problem-solving schema based on self-generated steps. The process is aided by the feedback and hints, which prevent the learner from getting stuck on impasses and individual steps.
Problem Solving (PS)	Means-ends search to solve problem; consequently, limited cognitive resources available to construct problem solving schema. With little or no feedback, correctness checking of schema is limited to what learner can do him or her-self.	Learner actively solves the problem and attempts to construct the problem solving schema, but since little/no feedback is given, it is very difficult to construct an accurate schema (i.e., unclear for learner which moves actually brought solution closer).	The learner has to create the problem-solving schema based on self-generated steps, without any instructional support or guidance during problem solving.

examples compared to worked examples have been mixed and seem to depend on opportunities for comparison and elaborate feedback (i.e., explaining not only what was wrong but also why it was wrong), so it is also difficult to formulate a clear hypothesis about how erroneous examples and worked examples will compare to one another.

We used an identical user interface for all conditions, so that the problems looked the same, with only the (inter)actions of students with the interface differing (i.e., more passively observing animated step-by-step examples vs. actively solving problems with or without hints). Next to effectiveness and efficiency, we also explored whether the instructional conditions differentially affected how students liked the instructional materials and their preference for using similar materials in the future, as well as their confidence in their posttest performance. The different degrees of assistance might have an effect on students' confidence; for instance, it has been shown that students often overestimate how much they have learned from worked examples, but actively engaging in solving an entire problem or completing missing steps in a partially worked-out problem, might reduce their overconfidence (Baars, Van Gog, De Bruin, & Paas, 2014; Baars, Visser, Van Gog, De Bruin, & Paas, 2013). If varying degrees of assistance would have an effect on students' enjoyment of working with the instructional materials, this would be useful to take into account when revising the instructional materials for further classroom use.

2. Experiment 1[2]

2.1. Method

2.1.1. Participants

Participants were 179 students from the 10th and 11th grade of two high schools in the U.S. All participants were taking an introductory chemistry course, had covered the basics of the topic of this study, stoichiometry, earlier in the course, and were told that their test scores would be used for a class grade. Twenty-four participants had to be excluded because they did not fully complete all phases of the study. The remaining 155 students had a mean age of 15.4 ($SD = .59$); 75 were male, 80 female.

2.1.2. Design

Participants were randomly assigned to one of the four instructional conditions: (1) Worked Examples (*WE*; $n = 39$ after exclusion), (2) Erroneous Examples (*ErrEx;* $n = 43$ after exclusion), (3) Tutored Problems to Solve (*TPS*; $n = 36$ after exclusion), or (4) Problems to Solve (*PS*; $n = 37$ after exclusion).

2.1.3. Materials

A web-based stoichiometry-learning environment, developed for and used in earlier experiments (McLaren & Isotani, 2011; McLaren et al., 2008), was updated and revised for this experiment. Stoichiometry is a subdomain of chemistry in which basic mathematics (i.e., multiplication of ratios) is applied to chemistry concepts. A detailed description of the study materials is described below and the ordering of the materials is summarized in Table 2.

[2] Some of the data from Experiment 1 have previously been reported in a conference proceedings paper (McLaren, van Gog, Ganoe, Yaron, & Karabinos, 2014). In that proceedings paper we also analyzed whether there were differential effects of instructional condition for high and low prior knowledge learners within our sample, and there were not. For the sake of brevity, those data are not reported here.

2.1.3.1. Pre-questionnaire. Prior to participating in the study, students were presented with an online questionnaire (i.e., "Pre-Questionnaire" in Table 2) containing standard demographic questions (e.g., age and gender), as well as questions about their use of computers (e.g., "How many hours a week do you normally use a computer?" <1 h, 1–5 h, 5–10 h, 10–15 h, >15 h) and their prior knowledge of chemistry and stoichiometry (e.g., "Check all that apply: I know what the '2' stands for in H_2O; I know what a mol is; I know what Na stands for; " etc.).

2.1.3.2. Pretest and posttest. The pretest and posttest consisted of four stoichiometry problems to solve (isomorphic to the Intervention problems, described below) and four conceptual knowledge questions to answer. The conceptual questions probed either understanding of representations used in the problem solving (e.g. molecular formulas – see Fig. 1 for an example) or transfer from the macroscopic to the microscopic level (see Fig. 2 for an example). Reliability (Cronbach's alpha) of the pretest was .448; of the posttest it was .571. Note that some questions had more than one part, such as the question in Fig. 1, where there are multiple check boxes to select. There was an A and B form of the test (see "Pretest (A/B)" and "Posttest (A/B)" in Table 2), which were isomorphic to one another and which were counter-balanced within condition (i.e., approximately ½ of the students in each condition received Test A as the pretest and Test B as the posttest, the other ½ received Test B as the pretest and Test A as the posttest).

2.1.3.3. Context-setting videos. Two brief narrated videos introduced students to the different stages of the study. The first ("Video: Intro to Study" in Table 2) was an introduction to the overall study, describing to students what they will see (i.e., the pretest, the intervention, the posttest). This initial video also provided some simple notational guidance, such as that chemical formulas like "H_2O" are rendered as "H2O" on the computer screen. The second context-setting video ("Intro to Posttest" in Table 2) alerted students to the start of the posttest. These videos were the same across all conditions.

2.1.3.4. Interface-usage videos. Two videos described the use of the computer-based interfaces the student used throughout the study. The first ("Video: Intro to the Test Interface" in Table 2) presented an example of a test problem and use of the test interface to solve the problem. This video was the same in all conditions. The second was a condition-specific video that provided a narrated example to explain how problems in this condition would be presented and how the student should use the interface (i.e., "*Video: Intro to WE/ErrEx/TPS/PS Interface*" in Table 1).

2.1.3.5. Instructional videos. Six instructional videos introduced new stoichiometry concepts and procedures used in the problems. These videos were the same in all conditions and were presented immediately before the relevant concepts and/or procedures were exercised. Videos included: presentation of dimensional analysis (prior to intervention problem 1 – "Video: Dim. Analysis"), a description of stoichiometry problem solving (prior to intervention problem 1 – "Video: Intro to Stoich Problem Solving"), a review of significant figures (prior to intervention problem 1 – "Video: Significant Figures"), presentation of molecular weight (prior to intervention problem 3 – "Video: Molecular Wt."), a presentation of composition stoichiometry (prior to intervention problem 5 – "Video: Composition Stoichiometry"), and a presentation of solution stoichiometry (prior to intervention problem 7 – "Video: Solution Concentration").

Table 2
Conditions and Materials used in the study, which was conducted over 6 class periods of 40 min each on different days. *Bold-italicized* items varied across conditions. Note that only the videos used to introduce the specific conditions and the instructional format of the 10 intervention items in each condition varied (i.e., *Video: Intro to WE/ErrEx/TPS/PS Interface*). The problem content of the 10 intervention items was the same across the conditions.

WE	*ErrEx*	*TPS*	*PS*
Pre-Questionnaire	Pre-Questionnaire	Pre-Questionnaire	Pre-Questionnaire
Video: Intro to Study	Video: Intro to Study	Video: Intro to Study	Video: Intro to Study
Video: Intro to the Test Interface	Video: Intro to the Test Interface	Video: Intro to the Test Interface	Video: Intro to the Test Interface
Pretest (A/B)	Pretest (A/B)	Pretest (A/B)	Pretest (A/B)
Video: Dim. Analysis	Video: Dim. Analysis	Video: Dim. Analysis	Video: Dim. Analysis
Video: Intro to Stoich Problem Solving	Video: Intro to Stoich Problem Solving	Video: Intro to Stoich Problem Solving	Video: Intro to Stoich Problem Solving
Video: Significant Figs	Video: Significant Figs	Video: Significant Figs	Video: Significant Figs
Video: Intro to WE Interface	***Video: Intro to ErrEx Interface***	***Video: Intro to TPS Interface***	***Video: Intro to PS Interface***
WE-1	***ErrEx-1***	***TPS-1***	***PS-1***
WE-2	***ErrEx-2***	***TPS-2***	***PS-2***
Embedded Test Ques. 1	Embedded Test Ques. 1	Embedded Test Ques. 1	Embedded Test Ques. 1
Video: Molecular Wt.	Video: Molecular Wt.	Video: Molecular Wt.	Video: Molecular Wt.
WE-3	***ErrEx-3***	***TPS-3***	***ErrEx-3***
WE-4	***ErrEx-4***	***TPS-4***	***ErrEx-4***
Embedded Test Ques. 2	Embedded Test Ques. 2	Embedded Test Ques. 2	Embedded Test Ques. 2
Video: Composition Stoichiometry	Video: Composition Stoichiometry	Video: Composition Stoichiometry.	Video: Composition Stoichiometry
WE-5	***ErrEx-5***	***TPS-5***	***ErrEx-5***
WE-6	***ErrEx-6***	***TPS-6***	***ErrEx-6***
Embedded Test Ques. 3	Embedded Test Ques. 3	Embedded Test Ques. 3	Embedded Test Ques. 3
Video: Solution Concentration	Video: Solution Concentration	Video: Solution Concentration	Video: Solution Concentration
WE-7	***ErrEx-7***	***TPS-7***	***PS-7***
WE-8	***ErrEx -8***	***TPS-8***	***PS -8***
Embedded Test Ques. 4	Embedded Test Ques. 4	Embedded Test Ques. 4	Embedded Test Ques. 4
WE-9	***ErrEx -9***	***TPS-9***	***PS -9***
WE-10	***ErrEx 10***	***TPS-10***	***PS 10***
Embedded Test Ques. 5	Embedded Test Ques. 5	Embedded Test Ques. 5	Embedded Test Ques. 5
Post-Questionnaire	Post-Questionnaire	Post-Questionnaire	Post-Questionnaire
Video: Intro to Posttest	Video: Intro to Posttest	Video: Intro to Posttest	Video: Intro to Posttest
Posttest (A/B)	Posttest (A/B)	Posttest (A/B)	Posttest (A/B)

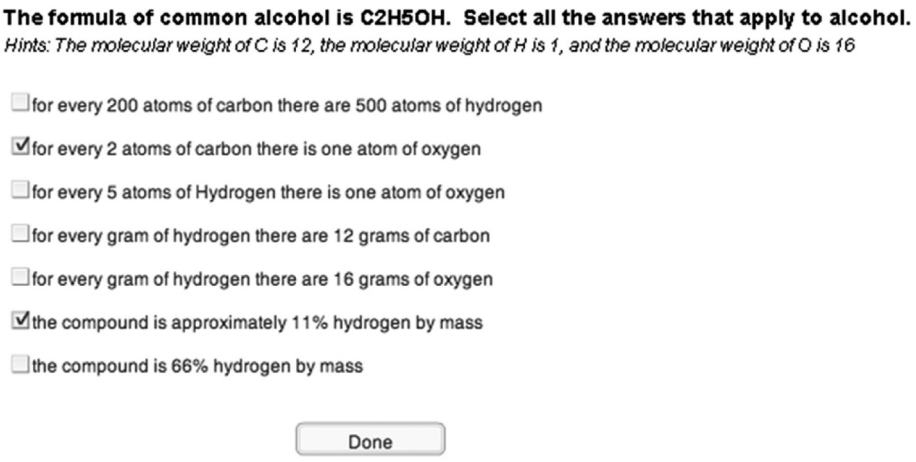

Fig. 1. Example conceptual question from the posttest that probes understanding of a chemical representation.

2.1.3.6. Intervention problems and feedback. Students were presented with a total of 10 intervention problems, in an instructional format specific to their condition. The problems were grouped in isomorphic pairs, as shown in Table 2 (e.g., WE-1 and WE-2 are an isomorphic pair, WE-3 and WE-4 are a second isomorphic pair, et cetera). The complexity of the problems presented in the intervention gradually increased.

The worked examples (*WE*) consisted of problem statements and screen-recorded animations of how to solve the problem, step-by-step. The animated examples had duration of between 30 and 70 s, could not be stopped or self-paced, and did not include any narration or explanation of why steps were taken; students only saw the steps being completed. When the animated example finished, students had to indicate the "reason" for each individual step by selecting an item from a drop-down menu. There were six options in each menu – Given Value, Unit Conversion, Avogadro's Number, Molecular Weight, Composition Stoichiometry, and Solution Concentration – corresponding to all of the possible reasons for a step. After entering all the reasons, they could click the "Done" button and feedback appeared. If they selected all reasons correctly, all steps in their problem would turn green and students were encouraged to study the final correct problem state: "Well done! You have correctly solved this problem. You might want to review the problem for a while. Select the 'Next' button when you are ready to proceed." If they did not select all reasons correctly, the correct steps turned green and incorrect steps turned red, and feedback

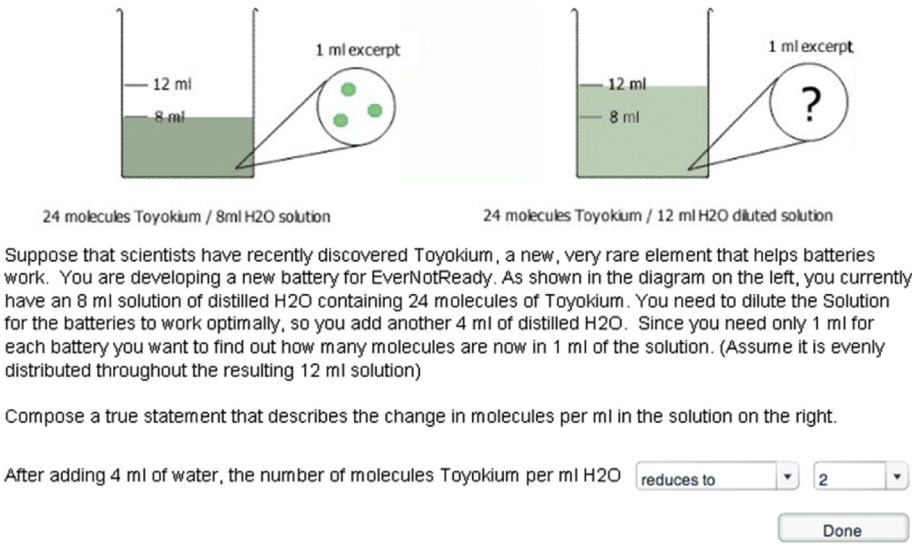

Fig. 2. Example conceptual question from the posttest that probes students understanding of transfer from the macroscopic to the microscopic scale.

appeared below the problem in the form of a fully worked-out final correct problem state (i.e., a static worked-out example). The students could study the correct solution as long as they wished, preceded by the message "You have some errors in your solution. The correct solution is below. You might want to review and compare your work to the correct solution. Select the 'Next' button when you are ready to proceed." Fig. 3 shows an example of an incorrectly completed worked example, with the correct, fully worked-out final solution below it.

The erroneous examples (*ErrEx*), consisted of screen-recorded animations of 30–70 s that could not be stopped or self-paced and demonstrated how to solve the problem step-by-step (i.e., dynamically), except the items contained 1 to 4 errors that students were instructed to find and fix. Part of the demonstrated solution included the "reasons" for the individual steps; these reasons could also be in error and could be corrected by the student. The inserted errors were those that most frequently occurred, as determined by examining data from a prior study with the stoichiometry materials (McLaren & Isotani, 2011). The students had to fill out at least one step before they could click the 'Done' button, at which point feedback appeared. When they managed to find and fix all errors correctly, all steps turned green and students were encouraged to study the final correct problem state for as long as they wanted (cf. message in the *WE* condition). When they did not manage to find and fix all errors correctly, the correct steps turned green and incorrect steps turned red, and feedback appeared below the problem (cf. *WE* condition and Fig. 3).

The tutored problems to solve (*TPS*) consisted of a problem statement and fields to fill in (similar to what is shown at the top of

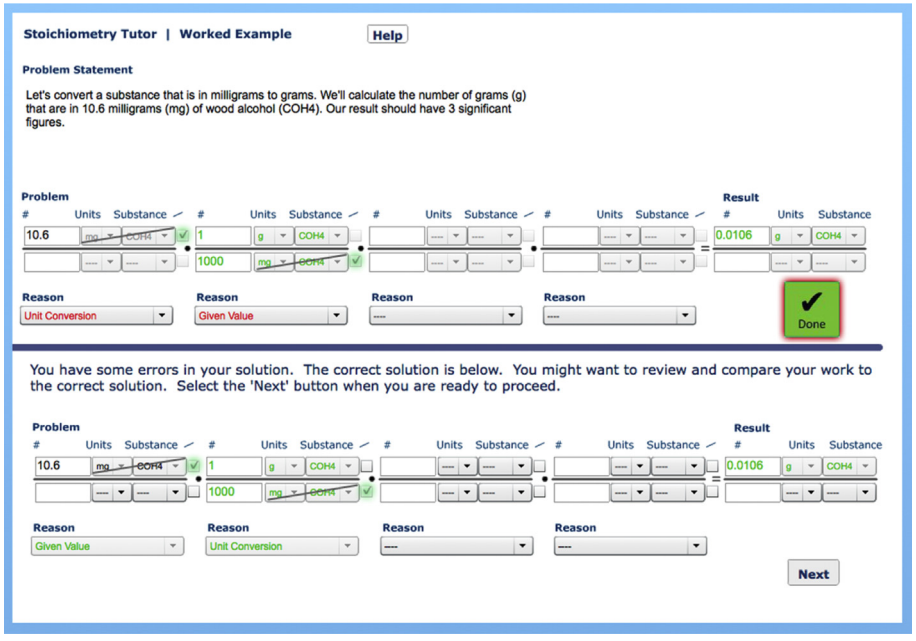

Fig. 3. Worked Example from Experiment 1, with feedback indicating incorrect reasons selected and the correct worked example shown below the student's work.

Fig. 3) and students had to attempt to solve the problem themselves, but with assistance received in the form of on-demand hints and error feedback. There were up to 5 levels of hints per step, with the bottom-out hint being both a message giving the answer to that step and a worked example of the problem solved to that point, shown below the interface (cf. position of the feedback example in Fig. 3). Because the tutored problems always ended in a correct final problem state, due to the given hints and the fact that students had to correctly solve every step in order to move on, an additional correct solution never appeared at the bottom of the screen in this condition, as it did when students made errors in the other conditions. Instead, students were encouraged to study their own correct problem state (with all steps turned green, for correct), prompted by the same message as the *WE* condition, but with no further feedback.

The problems to solve (*PS*) consisted of a problem statement and fields to fill in (similar to the top of Fig. 3) and students had to attempt to solve the problem themselves, without any assistance. They had to fill out at least one step before they could click the 'Done' button. When they clicked 'Done', feedback appeared. When they had solved the problem correctly, all steps turned green and students were encouraged to study the final correct problem state (cf. message in *WE* condition). When they did not manage to solve the problem correctly, the correct steps turned green and incorrect steps turned red, and feedback appeared below the problem in the form of a fully completed and correct solution (cf. message in *WE* condition and Fig. 3).

2.1.3.7. Embedded test problems. After every two intervention problems, students had to complete an embedded test problem. This problem was identical in content to the first item of the pair (i.e., embedded test problems 1 through 5 corresponded to the intervention problems 1, 3, 5, 7, and 9), but in the form of a conventional problem that students had to attempt to solve without any guidance or feedback. The embedded test items did not vary across conditions. Reliability (Cronbach's alpha) of the embedded test problems was .774.

2.1.3.8. Mental effort rating scale. The 9-point mental effort rating scale developed by Paas (1992) was administered after each intervention problem to assess how much effort students invested in completing the tasks in the intervention.

2.1.3.9. Post-questionnaire. After completing the intervention, students were presented with a second questionnaire (i.e., "Post-Questionnaire" in Table 2) that probed their confidence in tackling the posttest (i.e., "How confident are you that you will be able to solve …" 1 of the 8 problems, 2 of the 8 problems, 3 of the 8 problems, etc.), asked how much they liked working with the materials (i.e., "I liked working with the instructional materials", (1) Strongly disagree to (5) Strongly agree), and queried whether they would like to work with these materials again (i.e., "I would like to work again with these instructional materials", (1) Strongly disagree to (5) Strongly agree).

2.1.4. Procedure

The experiment was conducted at students' schools within their regular science classrooms. In total, the study took 6 class periods of 40 min to complete. Students received a login for the web-based environment and could work at their own pace on the materials they encountered in the learning phase (see Table 2). When they had finished with the intervention phase, however, they could not progress to the posttest; this test took place on the sixth and final period for all students.

2.1.5. Data analysis

The maximum total score on the pretest and posttest was 101 points: The four stoichiometry problems to solve consisted of a total of 94 steps, with one point gained for each step correctly performed, and the four conceptual questions had a total of 7 possible answers with one point per correct answer. The maximum score on the embedded test problems was 122 points, as the five problems consisted of a total of 122 steps and one point could be gained for each correctly solved step. Performance on each step was automatically scored as correct or incorrect via built-in rules in the learning environment, which also logged students' responses on the questionnaire items and mental effort rating scales.

2.2. Results

Data are presented in Table 3 and were analyzed with ANOVA and Bonferroni post-hoc tests (in case of unequal variances the Welch test is additionally reported along with Games-Howell post-hoc tests).

2.2.1. Pre-questionnaire and pretest

There were no significant differences among conditions in students' self-reported computer use, $\chi^2(12, N = 155) = 5.825$, $p = .925$ (on average, ca. 5% of students used computers less than 1 h per week; 32% between 1 and 5 h a week; 35% between 5 and 10 h a week; 20% 10–15 h a week and 8% more than 15 h a week) or self-reported prior knowledge of concepts and terms used in the problems, $F(3, 151) = 2.053$, $p = .109$ (on average, students indicated they knew 6.8 [SD = 1.3] out of 9 concepts). Analysis of the pretest scores confirmed that there were no significant differences among conditions in prior knowledge,[3] $F(3,151) = .359$, $p = .783$. Pretest scores correlated significantly with embedded test and posttest performance (embedded: $r = .511$, $p < .001$; post: $r = .482$, $p < .001$), but self-reported computer use (embedded: $r = -.027$, $p = .739$; post: $r = .049$, $p = .542$) and self-reported prior knowledge (embedded: $r = .077$, $p = .339$; post: $r = .008$, $p = .919$) did not.

2.2.2. Embedded and posttest performance

Overall, students' performance significantly improved from pretest to posttest, $F(1,154) = 255.319$, $p < .001$, $\eta_p^3 = .624$. However, there were no significant differences in performance among conditions, either on the embedded test problems, $F(3,151) = 1.163$, $p = .326$, or on the posttest, $F(3,151) = .485$, $p = .693$ (with pretest score as covariate: embedded: $F(3,150) = .934$, $p = .426$; posttest: $F(3,150) = .276$, $p = .843$).

2.2.3. Mental effort

There was a significant difference among conditions in the average mental effort invested in the intervention problems, $F(3,151) = 9.994$, $p < .001$, $\eta_p^2 = .166$. Bonferroni post-hoc tests showed that students in the *WE* condition invested significantly less mental effort in the intervention problems than students in all other conditions (*ErrEx*: $p < .001$, $d = .891$; *TPS*: $p < .001$, $d = .954$; *PS*: $p < .001$, $d = 1.04$). No other comparisons were significant.

2.2.4. Study time

Regarding the time students spent on the intervention problems, significant differences among conditions were found $F(3,151) = 51.005$, $p < .001$, $\eta_p^2 = .503$ (Welch: $F(3, 74.275) = 96.345$, $p < .001$). Bonferroni post hoc tests showing that the time spent in

[3] Note that total pretest and posttest performance is reported (i.e., score on stoichiometry problems and conceptual questions combined). Analyzed separately, results do not differ.

Table 3
Performance, mental effort, time on task, and post-questionnaire ratings per condition in Experiment 1.

	Condition:			
	WE (n = 39)	*ErrEx* (n = 36)	*TPS* (n = 43)	*PS*(n = 37)
Pretest (max = 101)	48.6 (12.8)	48.8 (15.4)	49.4 (13.5)	46.3 (14.3)
Posttest (max = 101)	68.5 (17.3)	68.3 (18.4)	71.1 (13.4)	66.4 (17.1)
Embedded test (max = 122)	89.4 (23.7)	88.3 (27.0)	95.3 (23.3)	84.8 (23.1)
Mental effort on intervention problems (1–9)	4.4 (1.8)	5.8 (1.4)	6.1 (1.7)	6.1 (1.3)
Time on intervention problems (min.)	19.8 (5.8)	37.2 (9.6)	62.4 (17.2)	52.1 (25.2)
Reflection time on feedback (min.)	1.7 (1.1)	4.3 (2.6)	1.3 (1.0)	6.5 (3.9)
Posttest confidence (correct out of 8; $N = 135$)	4.9 (2.3)	4.7 (1.9)	3.9 (2.4)	4.3 (2.0)
Liked materials (1–5; $N = 135$)	2.7 (1.2)	2.7 (1.3)	2.6 (1.3)	3.3 (1.1)
Want to work again with materials (1–5; $N = 135$)	2.2 (1.0)	2.6 (1.2)	2.3 (1.2)	2.8 (1.1)

Note. Time on task concerns only the intervention problems which differed among conditions; it does not include the time spent on the pre-questionnaire, pretest, instruction videos, effort ratings, embedded test problems, post-questionnaire, and posttest.

the intervention for the *WE* condition was significantly lower than in all the other conditions, *WE* vs. *ErrEx*: $p < .001$, $d = 2.195$, *WE* vs. *TPS*: $p < .001$, $d = 3.312$, *WE* vs. *PS*: $p < .001$, $d = 1.762$ (Games-Howell: same results); that it was significantly lower in the *ErrEx* condition than in both the *TPS* and *PS* conditions, *ErrEx* vs. *TPS*: $p < .001$, $d = 1.812$, *ErrEx* vs. *PS*: $p < .001$, $d = .782$ (Games-Howell: same results except *ErrEx* vs. *PS* at $p = .008$), and that the time was significantly lower in *PS* than in *TPS*, $p = .038$, $d = .478$ (but the Games-Howell post-hoc test was not significant, so this seems an artifact of the unequal variances).

If we look at the time students spent reflecting on the worked example given as feedback, there are also significant differences among conditions, $F(3,151) = 36.204$, $p < .001$, $\eta_p^2 = .418$ (Welch: $F(3, 79.098) = 33.632$, $p < .001$). As can be seen in Table 3, reflection time in the *WE* and *TPS* conditions was very low, and Bonferroni post hoc tests showed it was significantly lower than in the *ErrEx* condition, *WE* vs. *ErrEx*: $p < .001$, $d = 1.253$, *TPS* vs. *ErrEx*: $p < .001$, $d = 1.507$ (Games-Howell: same result) and *PS* condition, *WE* vs. *PS*: $p < .001$, $d = 1.670$, *TPS* vs. *PS*: $p < .001$, $d = 1.848$ (Games-Howell: same result), while there were no significant differences between the *WE* and *TPS* condition, $p = 1.000$ (Games-Howell: $p = .273$). Time spent reflecting in the *PS* condition was highest, and significantly higher than in all other conditions, *WE* vs. *PS*: $p < .001$, $d = 1.670$; *ErrEx* vs. *PS*: $p < .001$, $d = .672$; *TPS* vs. *PS*: $p < .001$, $d = 1.848$ (Games-Howell: same results, except *ErrEx* vs. *PS* at $p = .022$).

2.2.5. Post-questionnaire

Twenty participants did not answer the questions on the post-questionnaire, so these data are based on $N = 135$ (*WE*: 34; *ErrEx*: 35; *TPS*: 32; *PS*: 34). There were no significant differences in students' confidence in their posttest performance (i.e., how many problems out of 8 they were confident they could solve correctly) among conditions, $F(3,131) = 1.348$, $p = .262$. As can be seen in Table 3, students in the untutored problem-solving condition were somewhat more positive when answering the questions "I liked working with the instructional materials" and "I would like to work again with these instructional materials" than students in the other conditions, but the difference in average ratings on the 5 point scale was not statistically significant (liked: $F(3, 131) = 2.227$, $p = .088$; again: $F(3, 131) = 1.797$, $p = .151$).

2.3. Discussion

Results from Experiment 1 show a large efficiency benefit of worked examples compared to all other conditions. Equal learning outcomes were attained while less study time (between 46 and 68% less than the other conditions) and effort were spent on the intervention problems. The time and effort efficiency benefit is in line with prior studies comparing worked example study to conventional problem solving (e.g., Nievelstein et al., 2013; Van Gog et al., 2006), and the time efficiency benefit was also found in studies comparing worked example study to tutored problem solving (McLaren & Isotani, 2011). However, the four conditions of this experiment – worked examples, erroneous examples, tutored problem solving, and untutored problem solving – have never before been compared to each other directly, in a single experiment, on efficiency in terms of both time and effort.

The finding that worked example study is more efficient in terms of both time and effort compared to other instructional formats that provide a high degree of assistance but require more active involvement from learners (i.e., erroneous examples and tutored problems) is interesting in light of the assistance dilemma (Koedinger & Aleven, 2007) and debates about direct instruction (e.g., Hmelo-Silver et al., 2007; Kapur & Rummel, 2012; Kirschner et al., 2006; Schmidt et al., 2007; Tobias & Duffy, 2009). Our results show that the additional time and effort spent on finding and fixing errors in erroneous examples, or reviewing hints in tutored problems, does not improve learning outcomes compared to the more passive worked example study or the effort intensive conventional problem solving.

The latter is quite remarkable; whereas the finding that the high-assistance instructional formats all lead to comparable levels of learning outcomes is perhaps not entirely surprising, it is mystifying why they did not outperform the conventional problem solving condition. One possibility is that the feedback in the form of a worked example made the conventional problem solving condition too similar to the other conditions. As can be seen from Table 3, students in the problem-solving condition (as well as in the erroneous example condition, for that matter) spent quite a lot of time reflecting on the fully worked-out solution, especially when one realizes that the animated examples in the worked example condition were between 30 and 70 s. In other words, students in the problem-solving condition spent almost as much time reflecting on the fully worked-out solution as students in the example condition studied examples. In the tutored problem-solving condition, this reflection time was much lower, which makes sense because in this condition, students generate the solution, but they also effectively get help during problem solving by asking for, reading, and reflecting upon hints and by seeing worked examples as part of the bottom-out hints, so the intervention time in this condition already includes some feedback-processing time.

In the few other studies in which examples were provided as feedback after conventional problem solving (Paas, 1992; Paas & Van Merriënboer, 1994), a worked example effect was established; however, in those studies students could only review the feedback for a restricted amount of time that was less than the amount of time available for worked example study. To determine

whether the correct worked example feedback contributed to the equal performance across conditions in Experiment 1, a second experiment was conducted. Instead of receiving the correct worked example as feedback, students only saw feedback highlighting correct and incorrect steps completed in Experiment 2. This also allowed for a replication of the direct comparison of these four instructional conditions, under different feedback circumstances.

3. Experiment 2[4]

3.1. Method

3.1.1. Participants

Participants were 131 students from the 10th and 11th grade of two high schools in the U.S. (different from the schools in Experiment 1, but, like Experiment 1, the students were taking an introductory chemistry class, had covered the basics of the topic of this study, stoichiometry, earlier in the course, and were told that their test scores would be used for a class grade.). Fifteen participants had to be excluded because they did not fully complete all phases of the study. The remaining 116 students had a mean age of 16.45 (SD = .76); 48 were male, 66 female.

3.1.2. Materials, procedure, and data analysis

The same materials, experimental procedure and data scoring methods were used as in Experiment 1, with the exception that the final feedback, which no longer consisted of worked examples to study, but only of highlighting the steps that were correctly (green) and incorrectly (red) completed. In case the solution was correct, all steps turned green and a feedback message appeared stating: "Well done! You have correctly solved this problem. You might want to review the problem for a while. Select the 'Next' button when you are ready to proceed." In case errors were made, the correct steps turned green and incorrect steps turned red, and students received the message: "There are some errors in the solution. The steps in red are incorrect. Please take some time to review your work. When you are ready, select the 'Next' button to move on." (see Fig. 4). Reliability (Cronbach's alpha) of the pretest was .592, of the posttest was .692, and of the embedded test problems was .811.

3.2. Results

Data are presented in Table 4 and were analyzed with ANOVA (in case of unequal variances the Welch test is additionally reported along with Games-Howell post-hoc tests).

3.2.1. Pre-questionnaire and pretest

One student from the *WE* condition failed to fill out the self-rated computer use and prior knowledge questions on the pre-questionnaire so these data are based on N = 115. There were no significant differences among conditions in students' self-reported computer use, $\chi^2(12, N = 115) = 12.022$, $p = .444$ (on average, ca. 13% of students used computers less than 1 h per week; 29% between 1 and 5 h a week; 23% between 5 and 10 h a week; 18% 10–15 h a week and 17% more than 15 h a week) or self-reported prior knowledge of concepts and terms used in the problems, $F(3, 111) = .099$, $p = .960$ (on average, students indicated they knew 7.3 [SD = 1.7] out of 9 concepts).

Analysis of the pretest scores confirmed that there were no significant differences among conditions in prior knowledge, $F(3,112) < 1$, $p = .500$. Pretest scores correlated significantly with embedded and posttest performance (embedded: $r = .495$, $p < .001$; post: $r = .499$, $p < .001$; $N = 116$), and – in contrast to Experiment 1 - so did self-reported prior knowledge (embedded: $r = .238$, $p = .010$; post: $r = .365$, $p < .001$; $N = 115$). Self-reported computer use did not correlate with embedded or posttest performance (embedded: $r = .019$, $p = .837$; post: $r = .130$, $p = .168$; $N = 115$).

3.2.2. Embedded and posttest performance

As in Experiment 1, students' performance improved from pretest to posttest overall, $F(1,115) = 157.965$, $p < .001$, $\eta_p^2 = .579$, but there were no significant differences in performance among conditions, either on the embedded test problems, $F(3,112) = 1.031$, $p = .382$, or on the posttest, $F(3,112) < 1$, $p = .883$ (with pretest score as covariate: embedded: $F(3,111) = 1.289$, $p = .282$; posttest: $F(3,151) = .683$, $p = .564$).

3.2.3. Mental effort

There was a significant difference among conditions in the average mental effort invested in the intervention problems, $F(3,112) = 9.709$, $p < .001$, $\eta_p^2 = .206$. Bonferroni post-hoc tests showed that students in the *WE* condition invested significantly less effort in the intervention problems than students in the *TPS* ($p < .001$) and *PS* ($p = .002$) condition, but in contrast to Experiment 1, not compared to students in the *ErrEx* condition ($p = 1.000$). Moreover, in contrast to Experiment 1, students in the *ErrEx* condition invested significantly less effort than students in the *TPS* ($p = .003$) condition. There was no significant difference between the *ErrEx* and *PS* conditions ($p = .069$) or between the *TPS* and *PS* conditions ($p = 1.000$). No other comparisons were significant.

3.2.4. Study time

Time spent on the intervention problems differed significantly among conditions, $F(3,112) = 72.93$, $p < .001$, $\eta_p^2 = .661$ (Welch: $F(3, 56.164) = 120.265$, $p < .001$). Bonferroni post-hoc tests showed that students in the *WE* condition took significantly less time than students in all other conditions to complete the intervention problems, all $p < .001$ (Games-Howell: same result), that students in the *ErrEx* condition took less time than students in the *TPS* and *PS* conditions, both $p < .001$ (Games-Howell: same result), and that students in the *TPS* condition took more time than students in the *PS* condition, $p = .014$ (although the Games-Howell post-hoc test was not significant, so this seems an artifact of the unequal variances).

Looking at the time students spent reflecting on the feedback that highlighted the steps they had and had not performed correctly, there was a significant difference among conditions, $F(3,112) = 3.295$, $p = .023$, $\eta_p^2 = .081$ (Welch: $F(3, 61.401) = 3.370$, $p = .024$). Bonferroni post-hoc tests showed that students in the *ErrEx* condition spent significantly less time on the correctness feedback than students in the *WE* condition, $p = .027$ (Games-Howell: same result except $p = .019$); no other differences were significant.

3.2.5. Post-questionnaire

One participant from the *WE* condition did not answer the questions on the post-questionnaire, so these data are based on $N = 115$. There was a significant difference among conditions in students' confidence in their posttest performance (i.e., how many problems out of 8 they were confident they could solve correctly), $F(3,111) = 3.823$, $p = .012$, $\eta_p^2 = .094$. As can be seen in Table 4, students in the *PS* (i.e., no assistance) condition were less confident in their posttest performance than students in the other conditions, but Bonferroni post-hoc tests showed that their confidence was

[4] Some of the data from Experiment 2 have previously been reported in a short conference proceedings paper (McLaren, van Gog, Ganoe, Yaron, & Karabinos, 2015).

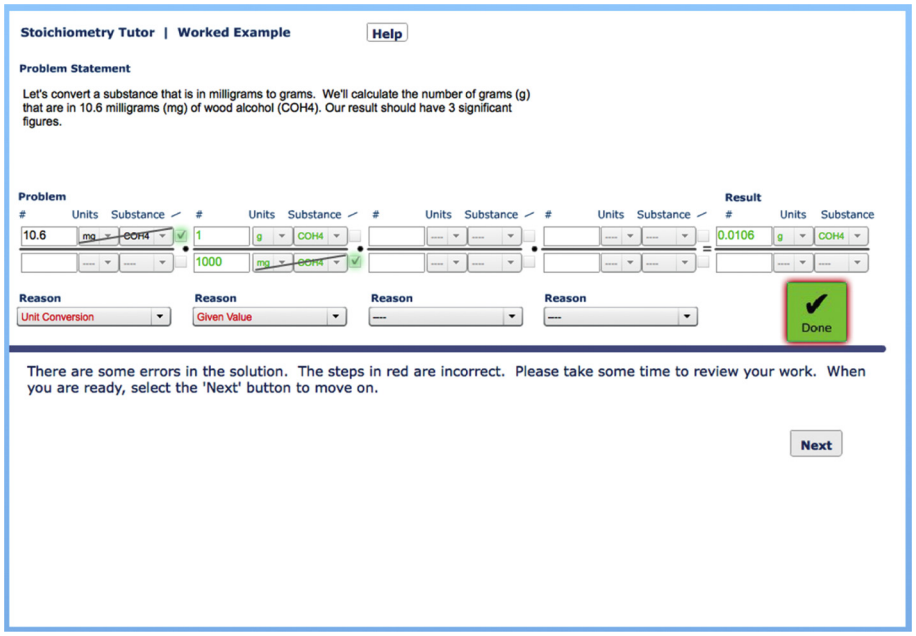

Fig. 4. Worked Example from Experiment 2, with feedback indicating incorrect reasons selected.

Table 4
Performance, mental effort, time on task, and post-questionnaire ratings per condition in Experiment 2.

	Condition			
	WE (n = 29)	*ErrEx* (n = 28)	*TPS* (n = 27)	*PS*(n = 32)
Pretest (max. = 101)	48.7 (17.6)	47.5 (20.3)	41.9 (16.8)	45.3 (16.3)
Posttest (max. = 101)	68.2 (18.2)	65.7 (23.1)	67.8 (20.0)	69.9 (19.2)
Embedded test (max. = 122)	92.2 (25.0)	79.8 (33.3)	85.3 (30.9)	80.7 (31.1)
Mental effort on intervention problems (1–9)	4.9 (1.4)	5.3 (1.7)	6.7 (1.3)	6.3 (1.3)
Time on intervention problems (min.)	20.9 (5.5)	40.5 (11.3)	67.1 (18.9)	56.8 (11.8)
Reflection time on feedback (min.)	2.3 (1.6)	1.3 (.9)	1.5 (1.0)	1.8 (1.7)
Posttest confidence (correct out of 8; $N = 135$)	4.8 (2.1)	3.8 (2.0)	4.9 (1.7)	3.5 (2.2)
Liked materials (1–5; $N = 135$)	2.9 (1.2)	2.8 (1.2)	3.2 (1.2)	2.6 (1.3)
Want to work again with materials (1–5; $N = 135$)	2.6 (1.3)	2.4 (1.3)	2.9 (1.1)	2.2 (1.1)

Note. Time on task concerns only the intervention problems which differed among conditions; it does not include the time spent on the pre-questionnaire, pretest, instruction videos, effort ratings, embedded test problems, post-questionnaire, and posttest.

only significantly lower compared to the *TPS* condition ($p = .038$; compared to *WE*: $p = .062$, to *ErrEx*: $p = 1.000$). Students in the *PS* condition seemed to be slightly more negative when answering the questions "I liked working with the instructional materials" and "I would like to work again with these instructional materials" than students in the other conditions (Table 4), but this was not statistically significant (liked: $F(3, 111) = 1.355$, $p = .261$; again: $F(3, 111) = 1.841$, $p = .144$).

3.3. Discussion

The results of Experiment 2 generally replicate the findings from Experiment 1: there were no differences in learning outcomes across conditions, but worked example study was much more efficient in terms of time (i.e., between 48 and 69% study time reduction compared to the other conditions) and effort spent on the intervention problems. Interestingly, with regard to effort investment, there was a difference compared to Experiment 1: the effort invested in correct and erroneous example study did not differ significantly and students in the erroneous example study condition invested less effort than students in the tutored and conventional problem-solving conditions. Possibly, this is related to the difference in feedback received; the time on task data also show that students in the erroneous examples condition spent less time reflecting on the feedback in which correct and incorrect steps were highlighted than students in the worked examples condition. In fact, compared to Experiment 1, students in the erroneous examples, and untutored problem solving condition spent a lot less time on the feedback. This suggests that students either may find it easier or more useful to study a worked example as feedback than to see only their errors highlighted; they do not seem to actively search for what the correct answer should have been.

With regard to the exploratory data (i.e., confidence; how much students liked the materials), in contrast to Experiment 1, there was a significant difference among conditions in confidence in posttest performance, with students in the problem-solving condition being least confident, and significantly less confident than students in the tutored problem-solving condition. This difference with Experiment 1 might well be due to the change from elaborate worked example feedback to correct/incorrect feedback in Experiment 2: students in the problem-solving condition were made aware of their errors but without showing them the correct answer, they may not have been confident about being able to avoid such errors in the future. A possibly related finding is that in Experiment 1, students in the problem-solving condition seemed to be slightly (though not significantly) more positive when answering the

questions "I liked working with the instructional materials" and "I would like to work again with these instructional materials" than students in the other conditions, but this was the other way around in Experiment 2 were they seemed to be slightly more negative than students in the other conditions, although again, this was not statistically significant.

4. General discussion

This study is the first to directly compare the effects of four instructional methods that vary in the type and amount of assistance: studying worked examples, studying and correcting erroneous examples, working on tutored problems, and engaging in untutored problem solving. While pairs of these approaches have been compared in prior work, no prior study has compared all four approaches at once. Direct comparisons of multiple methods in a single study are best for determining their relative effectiveness and efficiency. For instance, rather than comparing two methods, A and B, in one study, two methods, A and C, in a second study, and the final combination, B and C, in a third study, and then come to conclusions about how the three methods compare based on separate results, it is simpler and more reliable to compare the three methods directly in a single study. The naturally varying aspects of separate studies, such as population, time of the day and year the study is conducted, etc., even if identical instructional materials are used across studies, makes it difficult to combine and compare results and come to solid conclusions. On the other hand, a direct comparison of A, B, and C in a single study is more likely to lead to a better, more accurate comparison.

Across two experiments, the results showed clear efficiency benefits of worked example study, both in terms of time and effort investment, compared to the other conditions.[5] This is in line with prior research on the worked example effect, in which examples only or example problem pairs showed efficiency benefits compared to attempting to solve problems without any support (e.g., Cooper & Sweller, 1987; Nievelstein et al., 2013; Van Gog et al., 2006) and tutored problems (McLaren & Isotani, 2011). However, most prior research, especially on comparisons of examples and conventional problem solving, concerned single-session experiments in the lab or at schools. What our study shows is that this efficiency benefit holds when learning and testing is spread out over multiple classroom periods, a more ecologically valid context than the single-session lab and school studies of most prior research. Students in the worked example condition achieved the same level of embedded and posttest performance with substantially less study time than students in all other conditions.

Especially the time efficiency benefit achieved with worked examples is not only interesting in light of the assistance dilemma (Koedinger & Aleven, 2007) and debates about direct instruction (e.g., Kapur & Rummel, 2012; Hmelo-Silver et al., 2007; Kirschner et al., 2006; Schmidt et al., 2007; Tobias & Duffy, 2009), but with increasing strains on curricula, it is also highly relevant for educational practice. Note that the size of this time efficiency benefit (i.e., a reduction of 46 to 68% in Experiment 1 and 48 to 69% in Experiment 2 compared to the other conditions) resonates well with Zhu and Simon (1987), where it was reported that a redesigned Chinese mathematics curriculum, relying on self-guided study of worked examples alternated with problem solving practice, could be completed by students in two years while achieving the same level of test performance as attained by students who followed the regular three year lectures-plus-practice curriculum. Although Zhu and Simon were appropriately careful in interpreting this 'finding' (which they only mentioned, but did not support with any data or details), our data also suggest that it is not unthinkable that a curriculum redesign with a central role for worked example study could be completed substantially faster without compromising learning outcomes.

Nevertheless, an interesting question that remains is why we did not find any differences in learning outcomes among conditions. One would expect the high-assistance approaches, and especially the worked examples condition, given the robustness of the worked example effect, to outperform the conventional problem-solving condition on the embedded test problems and the posttest. Following Experiment 1, we hypothesized that the worked examples we provided as feedback might have contributed to this lack of difference in learning outcomes. The few prior studies in which students in the conventional problem-solving condition were provided with feedback in the form of a worked example still showed a worked example effect (e.g., Paas, 1992; Paas & Van Merriënboer, 1994). However, in those studies students could only review the feedback for a limited amount of time, less than the amount of time available for worked example study. Students in our problem-solving condition studied the examples for a substantial amount of time (about as long as the worked examples condition). So we conducted Experiment 2 to rule out that this lack of effect on learning outcomes was caused by the worked example feedback. Yet, Experiment 2, in which only correctness feedback was given, still showed no benefits of worked examples in terms of learning outcomes, only – and importantly – on efficiency.

We can only speculate about other potential explanations for the lack of performance differences and in particular the lack of worked example effect. One potential explanation is that for students in the worked examples condition, there was a longer time lag until the posttest; their condition was much more time efficient and the posttest was on the same day for all students. So it is possible that they would have outperformed the students in the other conditions if the time interval between study and test had been similar. Another possibility, given that these were multi-session classroom experiments, is that other factors not under our control eliminated effects on learning outcomes. For instance, student behavior is no longer under experimenter control in multi-session studies; students from different instructional conditions may talk to each other in between sessions about the learning materials or might look up further information. In addition, there might be memory interference from other classes students take in between the experimental sessions. Note though, that it is unlikely that these would be the sole explanations for the lack of a worked example effect on learning outcomes, as all these concerns would apply to the posttest, but not – or at least to a much lesser extent – to the embedded test problems, on which we also did not find performance differences among conditions.

A third and perhaps more likely potential explanation is that the instructional videos on stoichiometry, which also included an example of how to apply a concept during problem solving, may have reduced differences among instructional conditions. Other studies on the worked example effect have also provided students with general theoretical instructions prior to the learning phase in which the experimental intervention took place (e.g., Paas, 1992; Paas & Van Merriënboer, 1994) and sometimes even with worked examples (e.g., Cooper & Sweller, 1987; Sweller & Cooper, 1985). However, because the instructional videos were interspersed throughout the learning phase in our experiments, they may have provided sufficient support for students in the problem-solving

[5] Although the erroneous examples were also more efficient than problem solving in terms of effort investment when no elaborate worked example feedback was provided.

condition to benefit from practice, although it was slower and more effortful. In classroom settings though, it is quite common for students to receive explanations about new concepts either from the teacher or incorporated in textbooks or e-learning environments when these are necessary for solving practice problems. Therefore, it would be interesting for future research to determine whether interspersed instruction indeed attenuates the effects of worked examples on learning outcomes in stoichiometry as well as other domains.

A potential limitation of this study is that we did not assess whether learning benefits across the conditions would have varied on a delayed posttest. There is reason to believe a delayed posttest might have yielded different results, given the aforementioned findings of Adams et al. (2014) and McLaren et al. (in press), in which students in an erroneous example condition exhibited more learning on a delayed posttest than a supported problem solving condition, even though the erroneous example students did not show more learning on an immediate posttest. Another potential limitation is that we did not assess far transfer, such as that assessed in other studies involving worked examples (e.g., Eysink et al. 2009). Exploring delayed effects on learning and transfer across the four conditions of the present study remains for future research. We considered that the examples might have been more effective had they also included narration and process information (i.e., not only showing the steps but also including an explanation of why those steps; Van Gog, Paas, & Van Merriënboer, 2004), but in this case the origin of potential benefits of example study would no longer be clear (i.e., building schema of steps versus deeper understanding from explanation). We therefore chose a more basic example type. Finally, it may have been useful to test for students' ability to identify errors, since this was a skill that was directly promoted in the erroneous examples condition. In other words, we don't know whether the erroneous examples instruction might have fostered error recognition skills that the other conditions did not promote. On the other hand, the ability to recognize errors also relies on the quality of the schema that is acquired, which we did test with the pre- and posttest tasks.

To conclude, there are several potential explanations for the absence of differences in learning outcomes among conditions. Despite the lack of learning benefits, however, we did find a clear advantage of worked examples in terms of time and effort reduction. That this efficiency benefit was still apparent in a study conducted in the classroom, over multiple lesson periods, with short theory videos interspersed, is extremely valuable for education. Of course, our study was conducted in a single domain, and therefore does not provide all that is needed to assess the relative efficiency and learning impact of the various instructional approaches. Therefore, future research should continue to compare the effects of varying degrees of instructional assistance in the classroom, given that direct comparisons are informative and multi-session experiments are scarce but much closer to the reality of educational practice. As discussed above, such studies, like the one reported in this paper, give more ecologically valid information about the impact of various instructional formats on learning processes and outcomes.

Acknowledgements

Funding for this research was provided by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation award No. SBE-0836012. During the realization of these experiments, Tamara van Gog was supported by a Vidi grant from the Netherlands Organization for Scientific Research (NWO, # 452-11-006).

References

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

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*, 1–18. http://dx.doi.org/10.1037/a0021017.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: instructional principles from the worked examples research. *Review of Educational Research, 70*, 181–214. http://dx.doi.org/10.3102/00346543070002181.

Baars, M., Van Gog, T., De Bruin, A. B. H., & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology, 28*, 382–391. http://dx.doi.org/10.1002/acp.3008.

Baars, M., Visser, S., Van Gog, T., De Bruin, A. B. H., & Paas, F. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemporary Educational Psychology, 38*, 395–406. http://dx.doi.org/10.1016/j.cedpsych.2013.09.001.

Carroll, W. M. (1994). Using worked out examples as an instructional support in the algebra classroom. *Journal of Educational Psychology, 86*, 360–367. http://dx.doi.org/10.1037/0022-0663.86.3.360.

Clark, R., & Mayer, R. E. (2011). *e-Learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. San Francisco: Pfeiffer.

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79*, 347–362. http://dx.doi.org/10.1037/0022-0663.79.4.347.

Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction, 22*, 206–214. http://dx.doi.org/10.1016/j.learninstruc.2011.11.001.

Eysink, T. H. S., de Jong, T., Berthold, K., Kolloffel, B., Opfermann, M., & Wouters, P. (2009). Learner performance in multimedia learning arrangements: an analysis across instructional approaches. *American Educational Research Journal, 46*, 1107–1149.

Grosse, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: can this foster learning outcomes? *Learning and Instruction, 17*, 612–634. http://dx.doi.org/10.1016/j.learninstruc.2007.09.008.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: a response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*, 99–107. http://dx.doi.org/10.1080/00461520701263368.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93*, 579–588. http://dx.doi.org/10.1037/0022-0663.93.3.579.

Kapur, M., & Rummel, N. (2012). Productive failure in learning from generation and invention activities. *Instructional Science, 40*, 645–650. http://dx.doi.org/10.1007/s11251-012-9235-4.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*, 75–86. http://dx.doi.org/10.1207/s15326985ep4102_1.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review, 19*, 239–264. http://dx.doi.org/10.1007/s10648-007-9049-0.

Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review, 27*, 291–304. http://dx.doi.org/10.1007/s10648-015-9296-4.

Leppink, J., Paas, F., Van Gog, T., Van der Vleuten, C. P. M., & Van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30*, 32–42. http://dx.doi.org/10.1016/j.learninstruc.2013.12.001.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? the case for guided methods of instruction. *American Psychologist, 59*, 14–19. http://dx.doi.org/10.1037/0003-066X.59.1.14.

McLaren, B. M., & Isotani, S. (2011). When is it best to learn with all worked examples? In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Lecture notes in computer science, 6738: Artificial intelligence in education* (pp. 222–229). Berlin: Springer. http://dx.doi.org/10.1007/978-3-642-21869-9_30.

McLaren, B. M., Lim, S., & Koedinger, K. R. (2008). When and how often should worked examples be given to students? new results and a summary of the current state of research. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 2176–2181). Austin, TX: Cognitive Science Society.

McLaren, B. M., van Gog, T., Ganoe, C., Yaron, D., & Karabinos, M. (2014). Exploring the assistance dilemma: Comparing instructional support in examples and problems. In S. Trausan-Matu, et al. (Eds.), *Proceedings of the twelfth international conference on intelligent tutoring systems (ITS-2014)* (pp. 354–361). Switzerland: Springer International Publishing. LNCS 8474.

McLaren, B. M., van Gog, T., Ganoe, C., Yaron, D., & Karabinos, M. (2015). Worked examples are more efficient for learning than high-assistance instructional software. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.),

Proceedings of the 17th international conference on artificial intelligence in education (AIED 2015) (pp. 710–713). LNAI 9112.

McLaren, B. M., Adams, D. M., & Mayer, R.E. (in press). Delayed learning effects with erroneous examples: a study of learning decimals with a web-based tutor. *International Journal of Artificial Intelligence in Education*

Mwangi, W., & Sweller, J. (1998). Learning to solve compare word problems: the effect of example format and generating self-explanations. *Cognition and Instruction, 16*, 173–199. http://dx.doi.org/10.1207/s1532690xci1602_2.

Nievelstein, F., Van Gog, T., Van Dijck, G., & Boshuizen, H. P. A. (2013). The worked example and expertise reversal effect in less structured tasks: learning to reason about legal cases. *Contemporary Educational Psychology, 38*, 118–125. http://dx.doi.org/10.1016/j.cedpsych.2012.12.004.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive load approach. *Journal of Educational Psychology, 84*, 429–434. http://dx.doi.org/10.1037/0022-0663.84.4.429.

Paas, F., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: a cognitive load approach. *Journal of Educational Psychology, 86*, 122–133. http://dx.doi.org/10.1037/0022-0663.86.1.122.

Renkl, A. (2014a). Towards an instructionally-oriented theory of example-based learning. *Cognitive Science, 38*, 1–37. http://dx.doi.org/10.1111/cogs.12086.

Renkl, A. (2014b). The worked examples principle in multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (2nd rev. ed., pp. 391–412). New York: Cambridge University Press.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210. http://dx.doi.org/10.1111/j.1745-6916.2006.00012.x.

Rourke, A., & Sweller, J. (2009). The worked-example effect using ill-defined problems: learning to recognize designers' styles. *Learning and Instruction, 19*, 185–199. http://dx.doi.org/10.1016/j.learninstruc.2008.03.006.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432–1463. http://dx.doi.org/10.1037/a0037559.

Salden, R., Koedinger, K. R., Renkl, A., Aleven, V., & McLaren, B. M. (2010). Accounting for beneficial effects of worked examples in tutored problem solving. *Educational Psychology Review, 22*, 379–392. http://dx.doi.org/10.1007/s10648-010-9143-6.

Schmidt, H. G., Loyens, S. M. M., van Gog, T., & Paas, F. (2007). Problem-based learning is compatible with human cognitive architecture: commentary on Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*, 91–97. http://dx.doi.org/10.1080/00461520701263350.

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. J. C. M. (2009). The worked-example effect: not an artefact of lousy control conditions. *Computers in Human Behavior, 25*, 258–266. http://dx.doi.org/10.1016/j.chb.2008.12.011.

Simon, H. A. (1981). *The sciences of the artificial*. Cambridge, Mass: MIT Press.

Stark, R., Kopp, V., & Fischer, M. R. (2011). Case-based learning with worked examples in complex domains: two experimental studies in undergraduate medical education. *Learning and Instruction, 21*, 22–33. http://dx.doi.org/10.1016/j.learninstruc.2009.10.001.

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cognitive Science, 12*, 257–285. http://dx.doi.org/10.1207/s15516709cog1202_4.

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

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*, 59–89. http://dx.doi.org/10.1207/s1532690xci0201_3.

Sweller, J., & Levine, M. (1982). Effects of goal specificity on means-ends analysis and learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 8*, 463–474. http://dx.doi.org/10.1037/0278-7393.8.5.463.

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–295. http://dx.doi.org/10.1023/A:1022193728205.

Tobias, S., & Duffy, T. M. (Eds.). (2009). *Constructivist instruction: Success or failure?*. New York: Taylor & Francis.

Van Gerven, P. W. M., Paas, F., Van Merriënboer, J. J. G., & Schmidt, H. G. (2002). Cognitive load theory and aging: effects of worked examples on training efficiency. *Learning and Instruction, 12*, 87–105. http://dx.doi.org/10.1016/S0959-4752(01)00017-2.

Van Gog, T., & Kester, L. (2012). A test of the testing effect: acquiring problem-solving skills from worked examples. *Cognitive Science, 36*, 1532–1541. http://dx.doi.org/10.1111/cogs.12002.

Van Gog, T., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., & Verkoeijen, P. P. J. L. (2015). Testing after worked example study does not enhance delayed problem-solving performance compared to restudy. *Educational Psychology Review, 27*, 265–289. http://dx.doi.org/10.1007/s10648-015-9297-3.

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology, 36*, 212–218. http://dx.doi.org/10.1016/j.cedpsych.2010.10.004.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2004). Process-oriented worked examples: Improving transfer performance through enhanced understanding. *Instructional Science, 32*, 83–98.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction, 16*, 154–164. http://dx.doi.org/10.1016/j.learninstruc.2006.02.003.

Van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22*, 155–174. http://dx.doi.org/10.1007/s10648-010-9134-7.

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: the testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review, 27*, 247–264. http://dx.doi.org/10.1007/s10648-015-9310-x.

Wijnia, L., Loyens, S. M. M., Van Gog, T., Derous, E., & Schmidt, H. G. (2014). Is there a role for direct instruction in problem-based learning? comparing student-constructed versus integrated model answers. *Learning and Instruction, 34*, 22–31. http://dx.doi.org/10.1016/j.learninstruc.2014.07.006.

Zhu, X., & Simon, H. A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction, 4*, 137–166. http://dx.doi.org/10.1207/s1532690xci0403_1.