2014

Effects of problem solving after worked example study on primary school children's monitoring accuracy

Martine Baars
Erasmus University

Tamara Van Gog
Erasmus University

Anique de Bruin
Maastricht University

Fred Paas
University of Wollongong, fredp@uow.edu.au

Effects of problem solving after worked example study on primary school children's monitoring accuracy

Abstract

Research on expository text has shown that the accuracy of students' judgments of learning (JOLs) can be improved by instructional interventions that allow students to test their knowledge of the text. The present study extends this research, investigating whether allowing students to test the knowledge they acquired from studying a worked example by means of solving an identical problem, either immediately or delayed, would enhance JOL accuracy. Fifth grade children (i) gave an immediate JOL, (ii) a delayed JOL, (iii) solved a problem immediately and then gave a JOL, (iv) solved a problem immediately and gave a delayed JOL, or (v) solved a problem at a delay and then gave a JOL. Results show that problem solving after example study improved children's JOL accuracy (i.e., overestimation decreased). However, no differences in the accuracy of restudy indications were found. Results are discussed in relation to cue utilization when making JOLs.

Keywords

problem, accuracy, effects, monitoring, children, school, primary, study, example, worked, after, solving

Disciplines

Education | Social and Behavioral Sciences

Publication Details

Baars, M., Van Gog, T., de Bruin, A. & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. Applied Cognitive Psychology, 28 (3), 382-391.

Running head: IMPROVING JOL ACCURACY

Published as: Baars, M., Van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology, 28*, 382-391.

Effects of Problem Solving after Worked Example Study on Primary School Children's Monitoring Accuracy

Martine Baars[1], Tamara van Gog[1], Anique de Bruin[2], and Fred Paas[1,3]

[1]Institute of Psychology, Erasmus University Rotterdam, The Netherlands

[2]Department of Educational Development and Research, Maastricht University, The Netherlands

[3]Early Start Research Institute, University of Wollongong, Australia

Author Note:

Correspondence concerning this manuscript should be addressed to Martine Baars, Institute of Psychology, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. T: +31 10 408 9021; F: +31 10 4089009; E: baars@fsw.eur.nl

Abstract

Research on expository text has shown that the accuracy of students' Judgments of Learning (JOLs) can be improved by instructional interventions that allow students to test their knowledge of the text. The present study extends this research, investigating whether allowing students to test the knowledge they acquired from studying a worked example by means of solving an identical problem, either immediately or delayed, would enhance JOL accuracy. Fifth grade children 1) gave an immediate JOL, 2) a delayed JOL, 3) solved a problem immediately and then gave a JOL, 4) solved a problem immediately and gave a delayed JOL, or 5) solved a problem at a delay and then gave a JOL. Results show that problem solving after example study improved children's JOL accuracy (i.e., overestimation decreased). However, no differences in the accuracy of restudy indications were found. Results are discussed in relation to cue utilization when making JOLs.

Keywords: Judgments of learning, Monitoring accuracy, Worked examples, Problem solving

Effects of Problem Solving after Worked Example Study on Primary School Children's Monitoring Accuracy

To effectively regulate their own learning process, students must be able to monitor their progress towards learning goals and use this information to regulate further study (Metcalfe, 2009; Winne & Hadwin, 1998). For example, if students are trying to solve a math problem, it is important for them to monitor whether they understand the problem and its solution procedure, or whether more problems should be studied or practiced in order to grasp the procedure for solving this type of problem. The quality of the monitoring process is frequently measured by asking students to provide a judgment of learning (JOL) in terms of a prediction of future test performance, and relating this to actual test performance (see e.g., Anderson & Thiede, 2008; Dunlosky & Lipko, 2007; Koriat, Ackerman, Lockl, & Schneider, 2009a; Koriat, Ackerman, Lockl, & Schneider, 2009b; Metcalfe & Finn, 2008; Nelson & Dunlosky, 1991; Thiede, Anderson, & Therriault, 2003; Thiede, Dunlosky, Griffin, & Wiley, 2005). Research suggests that JOL accuracy, used as an indicator of the quality of monitoring, may affect the quality of self-regulated learning. That is, if JOLs are more accurate, students are better able to regulate the time they spend or the restudy choices they make (Kornell & Metcalfe, 2006; Metcalfe, 2009; Thiede, Anderson, & Therriault, 2003). Even though studies on accuracy of JOLs about learning word pairs and about learning from expository texts have shown that accuracy is generally low, they also showed that it can be improved by certain instructional interventions (for reviews, see Dunlosky & Lipko, 2007; Rhodes & Tauber, 2011; Thiede, Griffin, Wiley, & Redford, 2009).

Very little is known, however, about JOL accuracy when acquiring problem-solving skills by means of worked example study, even though problem-solving tasks play an important role in education, for instance in subjects like science and math. Problem solving tasks can vary greatly, from insight problems to well-structured transformation problems to ill-structured problems. Problem solving tasks used in education, for example in math or biology, are generally well-structured. Well-structured problems consist of a well-defined initial state, a known goal state, and can be solved using a constrained set of logical operators (Jonassen, 2011). For effective self-regulated learning in domains in which problem solving tasks are used, it is as important that students are able to accurately monitor and regulate their learning. Therefore, this study extends the research on JOL accuracy and how to improve it, to learning from worked examples. Before describing our approach, we will first shortly describe the findings from previous research on improving JOL accuracy when learning from word pairs and expository texts.

Monitoring Accuracy when Learning Word Pairs and Texts

In a typical experiment in which monitoring accuracy is measured by JOLs, participants first study word pairs (e.g., Nelson & Dunlosky, 1991) or expository texts (e.g., Maki, 1998) and are then asked to judge their learning by predicting their future test performance for each word pair or text. After all materials have been studied and judged, participants take a test on which their performance is measured. The accuracy of the JOLs is established by comparing them to actual test performance. In studies investigating monitoring accuracy with lists of items (e.g., word pairs, single words, sentences), the timing of JOLs and item difficulty were all found to affect JOL accuracy.

In studies investigating monitoring accuracy with texts, generation strategies were shown to affect JOL accuracy.

Effects of timing and item difficulty on monitoring accuracy with items. Regarding timing, it was found that delaying JOLs, that is, making JOLs only after studying a list of word pairs, improved relative accuracy compared to immediate JOLs, that is, JOLs given directly after studying each word pair. This so-called delayed-JOL effect (Nelson & Dunlosky, 1991) was shown for young adults (e.g. Dunlosky & Nelson, 1994; Dunlosky & Nelson, 1997), and for primary school children (Schneider, Visé, Lockl, & Nelson, 2000). In their meta-analysis, Rhodes and Tauber (2011) showed that for adults, the delayed-JOL effect was robust with paired associates, category exemplars, sentences, and single words; whereas the effect was not so convincing for children. However, when taking into account item difficulty, immediate and delayed JOLs are affected differently (Scheck, Meeter, & Nelson, 2004; Scheck & Nelson, 2005) Scheck and Nelson (2005) found that with difficult English-Swahili word pairs, absolute accuracy was higher for immediate JOLs compared to delayed JOLs that showed significant overconfidence after practice (i.e., on second trials). On easy items the reverse was found for delayed JOLs, that is, both immediate and delayed JOLs showed underestimation of performance.

Next to effects of timing of JOLs, the item difficulty also seems to play a role in monitoring accuracy. Several studies have shown that the difficulty of items negatively affects the accuracy of judgments about the correctness of performance (e.g. Griffin & Tversky, 1992; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein & Fischhoff, 1977). For instance, Lichtenstein and Fischhoff (1977) conducted a series of experiments in which participants had to judge the probability of the correctness of their answers to general knowledge questions and found that judgments were less accurate when item difficulty was higher. Furthermore, it was found that difficult items yielded overconfidence and easy items yielded underconfidence (Lichtenstein & Fischhoff, 1977; Scheck & Nelson, 2005).

Effects of generation strategies on monitoring accuracy with texts. Studies on learning from expository text found that JOL accuracy was generally low, and could not be improved by delaying JOLs (Maki, 1998). It should be noted though, that making a JOL about text requires a judgment about text *comprehension*, which is much more complex than a judgment about whether or not a target word from a word pair can be recalled. Subsequent research has shown that JOL accuracy could be improved by focusing participants' attention on their comprehension of a text prior to making a JOL. This was done, for instance, by asking them to use generation strategies, such as summarizing the texts (Thiede & Anderson, 2003), or generating keywords about the texts (Thiede et al., 2003), prior to making delayed JOLs (for a review: Thiede et al., 2009). This positive effect of generating keywords and summaries at a delay on JOL accuracy is called the 'delayed-generation effect' (Thiede et al., 2009). Thiede, Dunlosky, Griffin, and Wiley (2005) explained the delayed generation effect in terms of the involvement of different memory systems. Because of the time lag between reading and generating keywords, superficial information about the text in working memory (WM) is no longer available when generating keywords. Instead, after this delay, information from long-term memory (LTM) has to be used to generate keywords, and it is this information that also needs to be activated in order to answer test questions.

According to the cue-utilization approach to judgments of learning (see Koriat, 1997) JOLs are inferential and can be based on different memory cues or contextual cues. From this perspective, generating keywords or summaries at a delay activates more valid cues about how well a text has been learned than immediate generation would, thereby enhancing the accuracy of JOLs after delayed keyword or summary generation (Thiede et al., 2009). Recently, De Bruin, Thiede, Camp, and Redford (2011) have replicated the delayed-keyword effect in a study with primary and middle school children.

In sum, research with expository texts has shown that delayed-generation strategies, which allow students to test their comprehension of a text, can enhance the accuracy of delayed JOLs. The question addressed here, is whether an equivalent instructional strategy can be found that would enhance JOL accuracy when acquiring problem-solving skills by studying worked examples.

Monitoring Accuracy when Learning to Solve Problems by Studying Worked Examples

Little is known thus far about JOL accuracy when learning to solve problems. There are several studies that investigated monitoring during problem solving by making other types of judgments such as feeling-of-knowing (e.g., Metcalfe, 1986; Metcalfe & Wiebe, 1987; Reder & Ritter, 1992), confidence judgments (e.g., Boekaerts & Rozendaal, 2010; Mitchum & Kelley, 2010), or feelings of difficulty (e.g., Efklides, Samara, & Petropoulou, 1999), but to the best of our knowledge, only a few studies investigated JOLs in problem solving tasks (De Bruin, Rikers, & Schmidt, 2005, 2007). Moreover, those studies used a type of problem (i.e., playing a chess endgame) that is very different from the kind of procedural problems encountered in math or science in schools. In a recent study, Authors (2013) investigated JOL accuracy in procedural arithmetic problem-solving tasks, in primary education. Although overall JOL accuracy was found to be low, relative accuracy of JOLs given immediately after solving a problem tended to be higher than delayed JOL accuracy, which is not in line with research on word pairs or texts. Possibly, this is the case because JOLs about problem-solving skills concern a judgment about comprehension of a *solution procedure*, which might be more difficult to make at a delay when the problem itself is no longer seen, only a description of the problem. In that study, however, students only solved practice problems; they were not taught how to solve problems. The present study investigates monitoring accuracy when learning to solve problems by means of worked example study.

Studying worked examples, which provide a step-by-step worked-out solution procedure to a problem, has proven to be an effective and efficient way of acquiring problem-solving skills for novices (for reviews see, Atkinson, Derry, Renkl, & Wortham, 2000; Renkl, 2011; Van Gog & Rummel, 2010). When solving problems, novice learners have to rely on weak strategies like trial-and-error or means-ends analysis, due to their lack of prior knowledge. Even though those strategies, which impose high cognitive load, may allow students to solve a problem eventually (i.e., good *performance*), they do not lead to the construction of adequate problem-solving schemas (i.e., *learning*; Sweller, Van Merriënboer, & Paas, 1998), that can guide the solving of similar problems after the learning phase. Because worked examples provide a step-by-step worked out solution to the problem for learners to study, they reduce ineffective cognitive load, and instead allow learners to devote all available working memory resources to studying the solution and constructing an adequate schema.

Research has shown that compared to problem-solving practice only, novices attain better test performance when studying examples (Nievelstein, Van Gog, Van Dijck, & Boshuizen, 2013; Van Gerven, Paas, Van Merriënboer, & Schmidt, 2002; Van Gog, Paas, & Van Merriënboer, 2006; Van Gog, Kester, & Paas, 2011b) or example-problem pairs in which example study is alternated with problem solving (Carroll, 1994; Cooper & Sweller, 1987; Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Mwangi & Sweller, 1998; Paas, 1992; Paas & Van Merrienboer, 1994; Rourke & Sweller, 2009; Sweller & Cooper, 1985; Van Gog et al., 2011b).

In terms of monitoring, there seems to be a parallel between learning from expository texts and acquiring problem-solving skills through worked example study. When making a JOL following example study, students also have to judge their comprehension rather than literal memory in order to predict their future test performance, that is, they have to judge the quality of the schema they constructed and how well they think they will be able to use that schema to solve a similar problem on a future test. In analogy to learning from expository text, then, a generation strategy that would allow participants to test the schema they constructed by studying a worked example might provide them with relevant cues that would enable them to make more accurate JOLs. Solving a problem after studying a worked example might be an appropriate generation strategy to enhance JOL accuracy, because it allows learners to test the quality of their schemas. As for the generation strategies when learning from expository text, it might be most effective if there is a delay between example study and problem solving, because to solve the problem at a delay, learners can only use information from LTM, which is what they have to rely on during the future test. In contrast, problem solving immediately after studying a worked example would probably lead to less valid cues about future performance because immediately after studying a worked example information from the worked example is still active in WM. Cues based on this information would be less informative about future test performance than cues solely based on LTM.

The Present Study

In this study, five instructional conditions will be compared in terms of their effects on JOL accuracy: 1) worked example – immediate JOL, 2) example – delayed JOL, 3) example – immediate problem – JOL, 4) example – immediate problem – delayed JOL, and 5) example – delayed problem – JOL (see Table 1). Most of the studies on using generation strategies to improve JOL accuracy when learning word pairs and expository texts, measured relative accuracy (e.g. Griffin, Wiley & Thiede, 2008; Maki, 1998; Nelson & Dunsloky, 1991; Thiede & Anderson, 2003; Thiede, Griffin, Wiley, & Anderson, 2010; Thiede et al., 2003). Relative accuracy (often measured by the Goodman-Kruskal gamma correlation) indicates whether students can discriminate among items, in such a way that items that received a higher JOL are indeed performed better on a test than items that received a lower JOL. Next to relative accuracy, absolute accuracy can also be used to analyze JOL accuracy. Absolute accuracy shows the precision of the judgments by comparing the JOL for an item with the performance on that item, and is often measured by bias scores (JOL – performance: negative values indicate underestimation, and positive values overestimation of performance) or absolute deviation (the absolute difference between JOL and test performance, regardless of the direction of the difference; Mengelkamp & Bannert, 2010; Schraw, 2009). In this study, we will focus on bias and absolute deviation, because this shows the precision of JOLs per problem solving task. While relative accuracy (i.e., the ability to distinguish between items) could also provide interesting information, it cannot be used here because research in the classroom allows only for a limited number of problem solving tasks but to calculate reliable gamma correlations many items are needed (Nelson, 1984; Schraw, Kuch, & Roberts, 2011).

Our first hypothesis is that the prior findings which show that immediate JOLs were more accurate than delayed JOLs for problem-solving tasks (Authors, 2013) and for difficult word pairs (Scheck & Nelson, 2005), also apply to worked examples (i.e., JOL accuracy in condition 1 > condition 2). That is, if judging comprehension of a procedure is more easily done immediately than at a delay, one would expect more accurate JOLs immediately after studying a worked example than after a delay.

Secondly, we hypothesize that being able to test the quality of the schema acquired by studying a worked example by means of solving the same problem that was demonstrated in the example, will enhance JOL accuracy compared to only studying worked examples (i.e., JOL accuracy in conditions 3, 4, and 5 > conditions 1 and 2).

Third, it is hypothesized that delayed problem solving will enhance JOL accuracy more than immediate problem solving, similar to the delayed-generation strategies for learning from expository text (i.e., JOL accuracy in condition 5 > condition 3 and 4).

Next to testing these hypotheses, effects of task complexity, effects on restudy choices, and effects on learning will be explored. Task complexity has been found to affect monitoring accuracy of items (e.g., Griffin & Tversky, 1992; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein & Fischhoff, 1977; Scheck & Nelson, 2005). For learning to solve problems it can be argued that monitoring requires working memory (WM) resources (e.g., Griffin et al., 2008; Van Gog, Kester, & Paas, 2011a) and WM resources are limited (Baddeley, 1986; Cowan, 2001; Miller, 1956). Therefore, the more complex a task is (i.e., the more WM resources would be needed to perform it), the less resources are available for monitoring performance during the task. This might affect the cues available for making JOLs after the task is completed (cf. Kostons, Van Gog, & Paas, 2009). Therefore, tasks at two levels of complexity are used in this study to explore whether task complexity affects JOL accuracy when learning to solve problems.

As for restudy choices, some studies have shown that improved JOL accuracy also resulted in improved regulation of study for adults (Son & Metcalfe, 2000; Thiede, 1999; Thiede & Dunlosky, 1999; Thiede et al., 2003) as well as for children (De Bruin et al., 2011). If this would also apply when acquiring problem-solving skills from worked examples, then the delayed problem-solving condition would not only show the most accurate JOLs, but also the most accurate restudy decisions.

Finally, regarding effects on learning, it is not entirely clear what to expect. Recent studies comparing a condition in which only examples were studied to a condition in which example-problem pairs were used showed that there was no difference between the conditions in performance on an immediate test (Van Gog & Kester, 2012; Van Gog et al., 2011b). In the present study, however, the problems are additional to the worked examples, not a replacement of the worked examples, and as such, it is possible that learning outcomes might be higher in the conditions with example-problem pairs.

Method

Participants and Design

Participants were 135 Dutch fifth grade students (M_{age}= 10.93 years, *SD* = 0.61, 67 boys and 68 girls) from five different classrooms in four different schools. Participants within each classroom were randomly assigned to one of the five conditions prior to the experiment: 1) example – immediate JOL (*n* = 26), 2) example – delayed JOL (*n* = 27), 3) example – immediate problem – JOL (*n* = 29), 4) example – immediate problem – delayed JOL (*n* = 28), and 5) example – delayed problem – JOL (*n* = 25) (see Table 1 for an overview of the design).

Materials

All materials were paper-based and each worked example, problem, and rating scale was presented on a new page.

Worked examples. Six worked examples were used that provided a step-by-step explanation of how to solve water jug problems. Three worked examples demonstrated the solution procedure to problems that could be solved by subtracting the volume(s) of available water jugs from the largest water jug. The other three worked examples demonstrated the solution procedure to more complex problems that could be solved by subtracting and adding the volume(s) of available water jugs from the largest water jug. An example of a worked example can be found in Appendix A.

Practice and posttest problems. The practice problems used during the learning phase consisted of six water jug problems that participants had to solve themselves. In each example and problem pair, the problem explained in the worked example, and the problem that had to be solved were identical. The worked example was not available while solving the practice problem. An example of a practice problem can be found in Appendix B. The six posttest problems were isomorphic to the problems explained in the worked examples (i.e., the same procedure could be used, but the numbers were different).

Rating scales. JOLs were provided on a 7-point rating scale, which asked students to predict how well they would be able to solve a similar problem on a future test (0 = *not at all* and 6 = *very well*). Above this question, the problem statement consisting of a picture of the water jugs and the goal amount of water was provided.

Filler task. Rebuses on paper were used as a filler task (see Table 1). The rebuses showed a Dutch proverb that children could find by changing or deleting letters from the names of the pictures that were shown in the puzzle picture.

Procedure

The experiment was run in group sessions in classrooms at participants' schools. All participants were told that they would learn to solve water jug problems by studying examples and that they would be asked to predict how well they would be able to solve similar problems on a test at the end of the session. It was explained that they had two minutes to study a worked example or solve a problem (which had been judged by the teachers to be sufficient time and this had been confirmed in a pilot test), that they should not progress before the experiment leader would tell them to move to the next page. During this general instruction, the experiment leader also showed participants a worked example about solving a water jug problem (one not used in the materials), the JOL rating scale, and an example of a test problem.

Then, the learning phase started, during which participants engaged in studying six worked examples. Depending on their assigned condition, they provided a JOL immediately after studying each example (example – immediate JOL condition), after a delay (2 min.) (example – delayed JOL condition), after solving a problem that followed each worked example directly (example – immediate problem – JOL condition), after a delay (2 min.) after immediate problem solving (example – immediate problem – delayed JOL condition), or after delayed (2 min.) problem solving (example – delayed problem – JOL condition). During problem solving, the worked examples were no longer available to the students. Subsequently, all participants indicated which worked examples they would like to study again (restudy: minimum: 0; maximum: 6). Finally, they completed the posttest. Note that participants did not actually get to restudy the examples prior to taking the posttest; they were asked to indicate this for the purpose of calculating a measure of the accuracy of restudy indications.

Data Analysis

Test performance. Posttest performance was scored by assigning 1 point for each correct step (i.e., maximally 6 points per test problem).

Monitoring accuracy. The accuracy of JOLs was analyzed by calculating bias and absolute deviation scores. Bias was calculated per test problem by subtracting test performance from the JOL that was given for that problem type. This resulted in a positive, negative, or zero deviation score, indicating an overestimation, underestimation, or correct estimation of performance, respectively. The mean bias over the test tasks was calculated for each student (min. = -6; max. = 6). Because negative and positive bias values can neutralize each other when the average bias per student or condition is calculated, this measure gives an indication of the direction of the difference, but not of the absolute magnitude of the difference between JOLs and test performance. Therefore, we also calculated this absolute deviation, that is, the square root of the squared bias for each item (min. = 0; max. = 6). The closer to zero bias or absolute deviation is, the more accurate monitoring was.

Regulation accuracy. We defined regulation accuracy in line with the discrepancy-reduction model of regulation (Dunsloky & Thiede, 1998; Thiede & Dunlosky, 1999), which states that more difficult items to learn are more often selected for restudy than more easy items to learn. Thus, we assumed students would choose to restudy worked examples of problem solving tasks that they gave a low JOL.

The accuracy of restudy indications is frequently analyzed using the Goodman-Kruskal Gamma correlation between JOLs and restudy choices (e.g., De Bruin et al., 2011; Thiede et al., 2003). We could not compute a reliable gamma correlation because we only used six tasks, which also limited the restudy choices to six. Therefore, we developed an absolute measure of regulation accuracy that varies between 0 and 1, based on each possible combination of JOL (0-6) and restudy choice (yes/no). The scoring system is shown in Table 2. As can be seen from the table, lower JOLs combined with a choice to restudy resulted in gradually higher accuracy, whereas lower JOLs combined with a choice not to restudy resulted in gradually lower accuracy; similarly, higher JOLs combined with a choice to restudy resulted in gradually lower accuracy, whereas higher JOLs combined with a choice not to restudy resulted in gradually higher accuracy. In total six restudy choices could be made, and therefore the total (summed) regulation accuracy score could lie between 0 and 6.

Results

The mean practice problem performance, JOL, mean bias, mean absolute deviation, regulation accuracy, number of restudy choices, and mean test performance are presented in Table 3.

Monitoring Accuracy

Bias. Planned comparisons were conducted to test our hypotheses. The first planned comparison (condition 1 vs. 2), showed that there was no significant difference in bias between conditions that gave an immediate vs. delayed JOL after worked example study, *t*(125) < 1, *p* = .810. The second planned comparison (condition 1 & 2 vs. condition 3, 4 & 5) showed that bias was significantly lower in the conditions in which children solved problems after worked example study (3, 4, & 5) than in the conditions in which children did not solve problems (1 & 2), *t*(125) = -2.32, *p* = .022, Cohen's *d* = 0.36. The third planned comparison (condition 3 & 4 vs. 5), showed that there was no difference between delayed and immediate problem solving, *t*(125) < 1, *p* = .418.

A closer look at the results concerning the second comparison, showed that children who made immediate or delayed JOLs, showed an average positive bias that was significantly different from zero (immediate: *t*(24) = 2.46, *p* = .021, Cohen's *d* = 0.26; delayed, *t*(25) = 2.43, *p* = .023, Cohen's *d* = 0.31), whereas the bias of children who engaged in problem solving was not significantly different from zero (immediate problem – JOL, *t*(27) = 1.03, *p* = .312; immediate problem – delayed JOL, *t*(26) < 1, *p* = .329; delayed problem – JOL, *t*(23) < 1, *p* = .894). This means that children who did not engage in problem solving after worked example study showed significant overestimation of their future test performance whereas children who did engage in problems solving after worked example study did not.

A paired t-test showed that bias changed significantly as the test problems increased in complexity (complexity level 1: *M* = -0.73, *SD* = 1.57, complexity level 2: *M* = 1.47, *SD* = 1.72), *t*(129) = -15.18, *p* < .001, Cohen's *d* = -1.34.

Absolute deviation. To test our hypotheses in terms of absolute deviations between JOLs and performance, we conducted the same planned comparisons as for bias. The first (condition 1 vs. 2), showed that there was no significant difference between conditions that gave an immediate vs. delayed JOL after worked example study, *t*(125) < 1, *p* = .766. The second planned comparison (condition 1 & 2 vs. condition 3, 4 & 5) showed that absolute deviation scores of children who solved problems after worked example study did not differ compared to children who did not solve problems after worked example study, *t*(125) < 1, *p* = .517. The third planned comparison (condition 3 & 4 vs. 5) showed that there was no difference between delayed and immediate problem solving, *t*(125) = -1.19, *p* = .237.

A paired t-test showed that absolute deviation increased significantly as the test problems increased in complexity (complexity level 1: *M* = 1.77, *SD* = 1.01, complexity level 2: *M* = 2.16, *SD* = 1.12), *t*(129) = -2.69, *p* = .008, Cohen's *d* = -0.36.

Practice problem performance and JOLs

To explore the relation between practice problem performance and JOLs (as requested by one of the reviewers), we calculated the absolute deviation between practice problem performance and JOLs (*range*: 0-6). The condition with delayed practice problems with immediate JOLs showed the lowest deviation (*M* = 1.40, *SD* = 0.69), compared to immediate practice problems with immediate JOLs (*M* = 1.71, *SD* = 0.73) and immediate practice problems with delayed JOLs (*M* = 1.67, *SD* = 0.60); however, there was no statistically significant difference among the three conditions, $F(2, 78) = 1.57, p = .214$.

Regulation Accuracy

A one way ANOVA showed no significant differences among conditions in regulation accuracy, $F(4, 125) < 1, p = .551$, or in the number of tasks selected for restudy, $F(4, 130) < 1, p = .533$.

Test Performance

A one way ANOVA showed that test performance did not differ among conditions, $F(4, 130) = 2.06, p = .089$.

Discussion

This study investigated the effects of immediate and delayed problem solving after studying worked examples as a strategy to improve JOL accuracy. In contrast to our first hypothesis that immediate JOLs would be more accurate than delayed JOLs, we did not find differences in bias or absolute deviation between participants who made immediate and delayed JOLs after worked example study. In other words, findings from a prior study on immediate vs. delayed JOLs about problem-solving tasks that suggested that immediate JOLs were more accurate (Authors, 2013), do not seem to apply to JOLs about worked examples. It should be noted though, that relative accuracy (i.e., gamma correlations) tended to be higher for immediate JOLs in the prior study, whereas the present study measured accuracy in terms of bias and absolute deviation, and used a different type of problem-solving task. So it is not entirely clear whether this difference is due to the format (problems vs. examples), the measures used (relative vs. absolute accuracy), or the content of the problem-solving tasks. It should, however, be noted that this lack of difference between immediate and delayed JOLs is in line with studies on learning from text, in which no differences in relative accuracy between immediate and delayed JOLs were found either (Maki, 1998) unless a generation strategy was added (Griffin et al., 2008; Thiede et al., 2003; Thiede et al., 2009; Thiede & Anderson, 2003). Future studies should use multiple measures of JOL accuracy to gain more insight in the accuracy of immediate and delayed JOLs about problems and worked examples.

In line with our second hypothesis, problem solving after worked example study was found to improve JOL accuracy, at least in terms of bias. Whereas the children in the examples only conditions showed significant overconfidence about their future performance, those who solved a problem after example study did not show significant overconfidence. This finding is in line with the findings from Agarwal et al. (2008) and Roediger and Karpicke (2006) who found that with studying prose passages, JOLs were less inflated after testing. In these studies it is suggested that after testing participants have access to mnemonic cues like encoding or retrieval fluency, which caused the JOLs to be less inflated. Although our study used a different design and different materials, the results do seem to imply that children got better cues about future test performance from problem solving after worked example study than from only studying worked examples, presumably because children who solved problems were able to test the knowledge they had acquired from the example about how to solve a certain problem. This opportunity probably gave them more valid cues when making a JOL.

It should be noted though that problem solving after worked example study had an effect on bias but not on absolute deviation. This might be the case because the range of bias is made up of negative and positive values whereas absolute deviation only reflects the magnitude of the difference between JOLs and test performance (no negative values). So, if students more often show negative bias values in one condition than in the other condition, average bias can differ between conditions whereas average absolute deviation does not. While the use of multiple measures of monitoring accuracy makes it more challenging to interpret findings, it has been advocated because it allows for analysing different aspects of monitoring accuracy (Schraw, 2009).

Regarding our third hypothesis that delayed problem solving would lead to the most accurate JOLs, there were no significant differences in accuracy of JOLs made after immediate or delayed problem solving. This contrasts with findings from studies with expository texts, in which both generating keywords and making summaries were found to enhance monitoring accuracy only at a delay (De Bruin et al., 2011; Thiede & Anderson, 2003; Thiede et al., 2003). In absolute deviation between *practice* problem performance and JOLs there were no significant differences among conditions either, which suggests that the cues students obtain from practice are not affected by the interval between study and practice. Possibly, our assumption that immediate problem solving would involve both retrieval from WM and LTM, rather than only from LTM, was unlikely in the current study design. That is, neither in the immediate nor in the delayed problem-solving conditions could learners go back to the worked example when solving the problem. Perhaps this meant that learners in the immediate problem solving condition already relied predominantly on the information available in LTM, generating similar cues as in delayed problem solving. However, this is an assumption that future research should test. Moreover, it might be interesting in future research to examine response times of practice problem performance and JOLs, which might provide insight into the extent to which students use retrieval fluency as a cue (see Metcalfe & Finn, 2008).

We also explored effects of task complexity on JOL accuracy, as well as effects of the different conditions on regulation and learning. In line with earlier findings for word pairs (Lichtenstein & Fischhoff, 1977; Scheck & Nelson, 2005), monitoring accuracy was lower for more complex tasks. We expected that task complexity could affect monitoring because more complex problem solving tasks, require more cognitive resources, leaving less cognitive resources for monitoring learning accurately (cf. Van Gog et al., 2011a); however, we did not measure cognitive load in this study. So, future research should follow up on this finding more thoroughly.

Regulation accuracy is an important aspect of self-regulated learning. Some studies have shown that enhanced monitoring accuracy also led to enhanced regulation accuracy (Kornell & Metcalfe, 2006; Metcalfe, 2009;Thiede et al., 2003). However, even though children in the conditions with problem solving showed less bias, their restudy choices were not more accurate than the restudy choices made in the conditions without problem solving. This finding suggests that children may not have been using their JOLs in deciding which worked examples they would need to study again. It should be noted though, that we defined our regulation accuracy measure based on the discrepancy-reduction model of regulation (Dunsloky & Thiede, 1998; Thiede & Dunlosky, 1999). That is, we assumed students would choose to restudy worked examples of problem solving tasks that they gave a low JOL. However, this measure of regulation accuracy does not take into account other possible ways of study time allocation, such as restudying items that are within the region of proximal learning (Ariel, Dunlosky, &

Bailey, 2009; Metcalfe & Kornell, 2005). Other models of study time allocation would lead to a different operationalization of regulation accuracy and could lead to different results on regulation accuracy. Future research should further investigate the relation between JOLs and restudy choices when learning problem-solving procedures from examples and take into account different models of study time allocation.

In terms of learning, our findings are quite surprising. Studies comparing example study only with example-problem pairs, showed that there was no difference between the conditions in performance on an immediate test (Van Gog & Kester, 2012; Van Gog et al., 2011b). However, in those studies, solving a problem meant getting one example less to study. In the present study, however, the problems were additional to the worked examples, not a replacement of the worked examples, but nevertheless, this additional problem-solving practice opportunity did not have a positive effect on learning.

This was the first study to investigate how to improve JOL accuracy when studying worked examples in primary education. However there are some limitations that should be mentioned. First, whereas many studies have used gamma correlations to measure JOL accuracy (e.g., Griffin, Wiley, & Thiede, 2008; Maki, 1998; Nelson & Dunsloky, 1991; Thiede & Anderson, 2003; Thiede, Griffin, Wiley, & Anderson, 2010; Thiede et al., 2003), the practical school context in this study did not allow for the use of enough problem solving tasks to calculated gamma correlations. Consequently, the results of the present study cannot easily be compared to those found in previous studies. In future research it would be interesting to use enough problem solving tasks to be able to draw conclusions on monitoring accuracy in a school context, based on gamma correlations Second, studying worked examples is an effective and efficient way of acquiring problem-solving skills for novices (Atkinson et al., 2000; Renkl, 2011; Van Gog & Rummel, 2010), however, when worked examples are studied in a passive or superficial way it can lead to an illusion of understanding (Renkl, 1999; Renkl, 2002; Stark, Mandl, Gruber, & Renkl, 1999). This drawback of worked example study is related to metacognitive processes like monitoring. Students studying worked examples might be prone to overestimation, because of the illusion of understanding that can be encountered when studying worked examples.

To summarize, this is the first study on primary school children's JOL accuracy when learning to solve problems by studying worked examples in the classroom. It showed that fifth grade children studying worked examples tend to overestimate their performance on a future problem–solving test. The opportunity to solve a problem after example study seems to decrease this bias regardless of the timing of problem solving or JOLs. Furthermore, children showed more accurate JOLs on the less complex tasks. Because this was the first study to investigate problem solving as a strategy for children to improve JOL accuracy when learning from worked examples, findings should be interpreted with caution and should be replicated in future studies, with other types of problems and with other student populations. It is very important but also challenging to conduct controlled experiments in an actual classroom, and such settings do not allow for process-tracing methods like verbal reports or eye-tracking to be used. Therefore, future research might complement classroom studies with lab studies in order to unravel the cues students use when monitoring and regulating their learning from worked examples.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger III, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22,* 861-876. DOI:10.1002/acp.1391

Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica, 128*, 110-118. DOI:10.1016/j.actpsy.2007.10.006

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General, 138,* 432-447. DOI: 10.1037/a0015928

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. DOI: 10.3102/00346543070002181

Authors (2013). *Accuracy of immediate and delayed comprehension judgments about problem solving tasks*. Manuscript submitted for publication.

Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.

Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction, 20,* 372-382. DOI:10.1016/j.learninstruc.2009.03.002

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

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

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences, 24*, 87–114. DOI: 10.1017/S0140525X01003922

De Bruin, A. B. H., Thiede, T., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology*, *109*, 294-310. DOI:10.1016/j.jecp.2011.02.005

De Bruin, A. B. H., Rikers, R. M. J. P., & Schmidt, H. G. (2005). Monitoring accuracy and self-regulation when learning to play a chess endgame. *Applied Cognitive Psychology, 19*, 167-181. DOI:10.1002/acp.1109

De Bruin, A. B. H., Rikers, R. M. J. P., & Schmidt, H. G. (2007). Improving metacomprehension accuracy and self-regulation when learning to play a chess endgame: The effect of learner expertise. *European Journal of Cognitive Psychology, 19*(4-5), 671-688. DOI:10.1080/09541440701326204

Dunlosky, J. & Nelson, T. O. (1994). Does sensitivity of Judgments of Learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33,* 545-565. DOI:10.1006.jmla.1994.1026

Dunlosky, J. & Nelson, T. O. (1997). Similarity between the cue for Judgments of Learning (JOL) and the cue for test is not the primary determinant of JOL accuracy. *Journal of Memory and Language, 36,* 34-49. DOI : 10.1006/jmla.1996.2476

Dunsloky, J., & Thiede, K. (1998). What makes people study more? An evaluation of factors that affect self-paced study. *Acta Psychologica, 98*, 37-52. DOI: 10.1016/S0001-6918(97)00051-6

Efklides, A., Samara, A., & Petropoulou, M. (1999). Feeling of difficulty: An aspect of monitoring that influences control. *European Journal of Psychology of Education, XIV*, 461-476. DOI: 10.1007/BF03172973

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24,* 411-435. DOI: 0010-0285/92

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory and Cognition, 36*, 93-103. DOI: 10.3758/MC.36.1.93

Jonassen, D. H. (2011). *Learning to solve problems: A handbook for designing problem-solving learning environments.* New York, Routlegde.

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. DOI: 10.1037/0022-0663.93.3.579

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349-370. DOI: 10.1037/0096-3445.126.4.349

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009a). The easily learned, easily remembered heuristic in children. *Cognitive Development, 24*, 169-182. DOI:10.1016/j.cogdev.2009.01.001

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 609–622. DOI: 10.1037/0278-7393.32.3.609

Kostons, D., Van Gog, T., & Paas, F. (2009). How do I do? Investigating effects of expertise and performance-process records on self-assessment. *Applied Cognitive Psychology, 23*, 1256-1265. DOI: 10.1002/acp.1528

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know. *Organizational Behavior and Human Performance, 20,* 159-183. DOI: 10.1016/0030-5073(77)90001-0

Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117-144). Mahwah, NJ: Erlbaum.

Mengelkamp, C., & Bannert, M. (2010). Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome. *Memory and Cognition, 38*, 441–451. DOI: 10.3758/MC.38.4.441

Metcalfe, J. (1986). Feeling of knowing in memory and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 288-294. DOI: 10.1037/0278-7393.12.2.288

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*, 159-163. DOI: 10.1111/j.1467-8721.2009.01628.x

Metcalfe, J., & Finn, B. (2008). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1087-1097. DOI: 10.1037/a0012580

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Jounral of Memory and Language, 52,* 463-477. DOI: 10.1016/j.jml.2004.12.001

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory and Cognition, 15*, 238-246. DOI: 10.3758/BF03197722

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97. DOI: 10.1037/h0043158

Mitchum, A. L., & Kelley, C. M. (2010). Solve the problem first: Constructive solution strategy can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 699-710. DOI: 10.1037/a0019182

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. DOI: 10.1207/s1532690xci1602_2

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109-133. DOI: 10.1037/0033-2909.95.1.109

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2*, 267-270. DOI: 10.1037/0033-2909.95.1.109

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. DOI: 10.1016/j.cedpsych.212.12.004

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. DOI: 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. DOI: 10.1037/0022-0663.86.1.122

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 435-451. DOI: 10.1037/0278-7393.18.3.435

Renkl, A. (1999). Learning mathematics from worked-out examples: Analyzing and fostering self-explanations. *European Journal of Psychology of Education, 14,* 477- 488. DOI: 10.1007/BF03172974

Renkl, A. (2002). Worked- out examples: Instructional explanations support learning by self-explanations. *Learning and Instruction, 12,* 529- 556. DOI: 10.1016/S0959-4752(01)00030-5

Renkl, A. (2011). Instruction based on examples. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 272-295). New York, NY: Routledge.

Rhodes, M. G., & Tauber, S. K. (2001). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137,* 131-148. DOI: 10.1037/a0021705

Roediger, H. L., & Karpicke, J. D. (2006). Test enhanced learning: Taking memory tests improves long term retention. *Psychological Review, 17,* 249-255. DOI: 10.1111/j.1467-9280.2006.01693.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. DOI: 10.1016/j.learninstruc.2008.03.006

Scheck, P., Meeter, M., & Nelson, T. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language, 51,* 71-79. DOI: 10.1016/j.jml.2004.03.004

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary condition and an explanation via anchoring. *Journal of Experimental Psychology: General, 134,* 124-128. DOI: 10.1037/0096-3445.134.1.124

Stark, R., Mandl, H., Gruber, H., & Renkl, A. (1999). Instructional means to overcome transfer problems in the domain of economics: Empirical studies. *International Journal of Educational Research, 31,* 591-609. DOI: 10.1016/S0883-0355(99)00026-9

Schneider, W., Visé, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring: Evidence from a judgment-of-learning task. *Cognitive Development, 15*, 115-134. DOI: 10.1016/S0885-2014(00)00024-1

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4,* 33-45. DOI: 10.1007/s11409-008-9031-3

Schraw, G., Kuch, F., & Roberts, R. (2011). Bias in the gamma coefficient: A Monte Carlo study. In P. Alexander (Chair), *Calibrating calibration: Conceptualization, measurement, calculation, and context.* Symposium presented at the annual meeting of the American Educational Research Association, New Orleans.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 204–221. DOI: 10.1037/0278-7393.26.1.204

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. DOI: 10.1207/s1532690xci0201_3

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296. DOI: 10.1023/A:1022193728205

Thiede, K. W. (1999). The importance of self-monitoring and self-regulation during multi-trial learning. *Psychonomic Bulletin & Review, 6,* 662-667. DOI: 10.3758/BF03212976

Thiede, K.W., & Anderson, M.C.M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*, 129–160. DOI: 10.1016/S0361-476X(02)00011-5

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66-73. DOI: 10.1037/0022-0663.95.1.66

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of selfpaced study: An analysis of selection of items for study and self paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1024–1037. DOI: 10.1037/0278-7393.25.4.1024

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1267-1280. DOI: 10.1037/0278-7393.31.6.1267

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J.S. (2009). Metacognitive monitoring during and after reading. In D.J. Hacker, J. Dunlosky, & A.C. Graesser, (Eds.) *Handbook ofMetacognition and Self-Regulated Learning*. Mahwah, NJ: Erlbaum.

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. DOI: 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 example study. *Cognitive Science, 36,* 1532-1541. DOI: 10.1111/cogs.12002

Van Gog, T., Kester, L., & Paas, F. (2011a). Effects of concurrent monitoring on cognitive load and performance as a function of task complexity. *Applied Cognitive Psychology, 25,* 584-587. DOI: 10.1002/acp.1726

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

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. DOI: 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. DOI: 10.1007/s10648-010-9134-7

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds). *Metacognition in Educational Theory and Practice.* (pp. 277-304). Hillsdale, NJ: LEA.

Table 1

Overview of design (WE = worked example; JOL = Judgment of Learning)

No self-test		**Self-test**		
Immediate JOL	**Delayed JOL**	**Immediate problem and immediate JOL**	**Immediate problem and delayed JOL**	**Delayed problem and immediate JOL**
WE	WE	WE	WE	WE
JOL	*Filler task*	Problem	Problem	*Filler task*
Filler task	JOL	JOL	*Filler task*	Problem
Filler task	*Filler task*	*Filler task*	JOL	JOL
		Restudy choices		
		Test		

Table 2

Scoring of regulation accuracy.

JOL scale/ Restudy choices	No (0)	Yes (1)
0	0	1
1	0.17	0.83
2	0.33	0.67
3	0.50	0.50
4	0.67	0.33
5	0.83	0.17
6	1	0