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

Martine Baars, Tamara van Gog, Anique de Bruin & Fred Paas

Published online: 09 Mar 2016.

Submit your article to this journal ☐

Article views: 2087

View related articles ☐

View Crossmark data ☐

Citing articles: 8 View citing articles ☐

Routledge
Taylor & Francis Group

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

Martine Baars[a], Tamara van Gog[a], Anique de Bruin[b] and Fred Paas[a,c]

[a]Institute of Psychology, Erasmus University Rotterdam, Rotterdam, The Netherlands; [b]Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands; [c]Interdisciplinary Educational Research Institute, University of Wollongong, Wollongong, Australia

ABSTRACT

Monitoring accuracy, measured by judgements of learning (JOLs), has generally been found to be low to moderate, with students often displaying overconfidence, and JOLs of problem solving are no exception. Recently, primary school children's overconfidence was shown to diminish when they practised problem solving after studying worked examples. The current study aimed to extend this research by investigating whether practising problem solving after worked example study would also improve JOL accuracy in secondary education. Adolescents of 14–15 years old ($N = 143$) were randomly assigned to one of five conditions that differed in timing of JOLs, whether practice problems were provided, and timing of the practice problems provided: (1) worked examples – JOL, (2) worked examples – delay – JOL, (3) worked examples – practice problems – JOL, (4) worked examples – practice problems – delay – JOL or (5) worked examples – delay – practice problems – JOLs. Results showed that practice problems improved absolute accuracy of JOLs as well as regulation accuracy. No differences in final test performance were found.

Introduction

Students can only learn effectively in a self-regulated way if they have accurate knowledge about their own learning process. Monitoring, that is, keeping track of one's own performance during the learning process (e.g. by making judgements of learning [JOLs]), provides the learner with information about the quality of the learning process, which can subsequently be used to regulate further study (Serra & Metcalfe, 2009). Models of self-regulated learning imply that with accurate monitoring of the learning process, subsequent regulation choices can be made based on better information, and consequently, the process of self-regulated learning can become more effective and lead to better learning outcomes (Thiede, Anderson, & Therriault, 2003; Winne & Hadwin, 1998).

Generally, monitoring is not very accurate, but it improves with age (Bryce & Whitebread, 2012). Moreover, it can be improved through addition of instructional strategies (Dunlosky & Lipko, 2007; Maki, 1998; Thiede, Griffin, Wiley, & Redford, 2009). More accurate monitoring during a learning task can in turn be used to regulate further learning more accurately (e.g. Dunlosky & Lipko, 2007; Thiede et al., 2003). Much less is known, however, about ways to improve monitoring and regulation accuracy when learning to solve problems, despite the prominent role of problem solving in subjects such as

---

CONTACT Martine Baars ✉ baars@fsw.eur.nl

math, science or biology. Problem solving tasks encountered in these subjects in secondary education are usually well-structured problems that consist of a well-defined initial state, a known goal state, and can be solved using a constrained set of logical operators (Jonassen, 2011). An effective way to learn to solve such problems is by studying worked-out examples of the solution procedure (Atkinson, Derry, Renkl, & Wortham, 2000; Renkl, 2013; Sweller, Van Merriënboer, & Paas, 1998; van Gog & Rummel, 2010). Given that self-regulated learning is also very important in subjects involving problem solving tasks, the present study investigated whether solving a practice problem after studying a worked example (on the biology topic of heredity) would be an effective generation strategy to improve secondary education students' monitoring and regulation accuracy.

Since most of the research on self-monitoring and improving self-monitoring was conducted with word pairs and expository texts, we will first shortly describe the findings on improving monitoring accuracy when learning from word pairs and expository texts as these findings form an important background for the current study.

Improving monitoring accuracy through the timing of JOLs

In a study by Nelson and Dunlosky (1991), it was found that *relative* JOL accuracy was higher when students gave their JOLs after they had studied the whole list of word pairs than when they gave the JOLs directly after each word pair. This so-called 'delayed-JOL effect' was explained by the monitoring-dual-memories principle which states that participants use their long-term memory (LTM) to make a delayed JOL which is more indicative for future test performance, whereas to make an immediate JOLs, both LTM and short-term memory are used which causes noise in the immediate JOL. This delayed-JOL effect has been replicated in many studies of which most use similar designs with delays varying from less than 1 min up to 10 min. The delayed-JOL effect was found to be robust with paired associates, category exemplars, sentences and single words – at least for adults, but to a much lesser extent for young children (Rhodes & Tauber, 2011). Moreover, Scheck and Nelson (2005) found that for difficult word pairs, after practice (i.e. on second trials), *absolute* accuracy was higher for immediate JOLs compared to delayed JOLs. In the study by Scheck and Nelson, two study–test cycles were used in which students studied easy and difficult English–Swahili word pairs, gave an immediate or delayed JOL and took a self-paced recall test. In the second study–test cycle, they found that absolute accuracy was higher for immediate JOLs compared to delayed JOLs on difficult items. Possibly, participants use a psychological anchor to base their JOLs on. This anchor is assumed to be the mid-point in the range of performance and is based on earlier learning experiences. Such an anchor could have caused overconfidence with difficult items as performance drops below the anchor for difficult items (Scheck & Nelson, 2005). So, timing of JOLs is important but the difficulty of the materials should also be considered.

Indeed, a negative relationship was found between item difficulty and monitoring accuracy when studying word pairs (e.g. Griffin & Tversky, 1992; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein & Fischhoff, 1977). For instance, in a series of experiments conducted by Lichtenstein and Fischhoff (1977), participants had to judge the probability of the correctness of their answers to general knowledge questions and they found that judgements were less accurate when item difficulty was higher. According to Lichtenstein and Fischhoff (1977), this shows the insensitivity of participants to what they know or do not know. An alternative explanation could be that complexity affects monitoring accuracy because monitoring requires WM resources (e.g. Griffin, Wiley, & Thiede, 2008; van Gog, Kester, & Paas, 2011a) which are limited (Baddeley, 1986; Cowan, 2001; Miller, 1956). Studying more complex materials requires more WM resources (Sweller et al., 1998), and consequently, leaves less WM resources for monitoring the learning process. This could affect the cues available for making monitoring judgements (i.e. JOLs).

Improving monitoring accuracy through generation strategies

For more complex learning materials, which require learners to remember and understand the materials, such as learning from text, no delayed-JOL effect was found (Maki, 1998; Thiede et al., 2009). Yet, when

learning from text was combined with strategies that help learners judge their understanding of the text, the relative accuracy of JOLs was improved. For example, for adults making summaries (Thiede & Anderson, 2003) or generating keywords (Thiede et al., 2003; Thiede, Dunlosky, Griffin, & Wiley, 2005) at a delay after reading all the six texts in the experiment improved relative JOL accuracy. Generating keywords after reading a text was found to improve relative JOL accuracy for children as well (De Bruin, Thiede, Camp, & Redford, 2011). Moreover, it has been found that this improved JOL accuracy also enhanced regulation accuracy (a relative measure consisting of the gamma correlation between JOL and whether a text was selected for restudy) for both adults (Thiede et al., 2003) and children (De Bruin et al., 2011). Furthermore, immediate generation strategies, such as self-explaining (Griffin et al., 2008) and making concept maps for both adults (Thiede, Griffin, Wiley, & Anderson, 2010) and children (Redford, Thiede, Wiley, & Griffin, 2012) directly after reading a text, were also found to improve JOL accuracy. Thus, generation strategies seem to help adults and children to judge their own performance more accurately when working on more complex learning materials like texts.

Thiede et al. (2009) explained the effect of both immediate and delayed generation strategies by the situation model approach. A situation model can be defined as the mental representation of the situation described in a text that is created by the reader in conjunction with a text-based representation of the text. In the situation model, the reader integrates already existing knowledge from LTM with information from the text to understand the representation of the described situation (Zwaan, 1999; Zwaan & Radvansky, 1998). According to the situation model approach, using information from the situation model helps students to make more accurate JOLs because the deeper level of understanding about the text, which is also measured on the future test, resides in the situation model. After a delay, generation strategies help students to access the situation model as it is stored in LTM. Yet, immediate generation strategies help students to access their situation model while constructing it. So in different ways, both immediate and delayed generation strategies help students to access their situation model, which provides them with more valid cues to base JOLs on, which in turn enhances JOL accuracy.

The majority of research investigating how to improve JOL accuracy has focused on adults. As mentioned above, however, some research has shown that JOL accuracy of children and adolescents is generally low, but improvable by generation strategies (De Bruin et al., 2011; Redford et al., 2012; Thiede, Redford, Wiley, & Griffin, 2012). However, there might be a difference in how this information acquired through monitoring is used for regulation of the learning process. Older children (11–12 years old) were found to be better able to regulate their learning process (i.e. indicating restudy choices) compared to younger children (8–10 years old; De Bruin et al., 2011; Destan, Hembacher, Ghetti, & Roebers, 2014; Krebs & Roebers, 2010; Roebers, Schmid, & Roderer, 2009). So, although most research on monitoring and regulation was focused on adults, it seems that monitoring accuracy in children can be improved and especially older children are capable of using monitoring to regulate their learning process.

Improving monitoring accuracy through practice problems when learning to solve problems

Only very few studies have investigated JOL accuracy when learning to solve problems and these have mostly focused on adults (e.g. De Bruin, Rikers, & Schmidt, 2005, 2007). De Bruin et al. (2005) asked university students to provide a JOL and select a move for restudy or to only select a move for restudy (i.e. without providing a JOL first) when learning to play a chess endgame. They found that students who gave JOLs showed better self-regulatory behaviour. However, this did not affect their learning outcomes. Furthermore, De Bruin et al. (2007) showed that JOL accuracy and performance were higher for more experienced chess players when learning to play a chess endgame. So, when learning to perform a more procedural task like learning to play a chess endgame, making JOLs can be beneficial for (self-regulated) learning. However, De Bruin et al. (2005, 2007) used chess problems, which are very different from the well-structured problems encountered in primary and secondary education school subjects. When monitoring how well one has learned to solve such well-structured problems, students have to monitor whether they *understand* the problem solving *procedure*.

Table 1. Overview of design (WE = worked example; JOLs = judgements of learning). After the pretest, the WE or WE – practice problem and JOL sequence was repeated three times with problems which increased in complexity.

Worked examples only		Worked example – Practice problems		
Pretest (5 min)				
(1) Immediate JOLs	(2) Delayed JOLs	(3) Immediate practice problem and immediate JOLs	(4) Immediate practice problem and delayed JOLs	(5) Delayed practice problem and immediate JOLs
WE (3 min)	WE (3 min)	WE (3 min)	WE (3 min)	WE (3 min)
JOL	*Filler task*	Problem (3 min)	Problem (3 min)	*Filler task*
Filler task	JOL	JOL	*Filler task*	Problem (3 min)
Filler task	*Filler task*	*Filler task*	JOL	JOL
Restudy choices				
Criterion test (9 min)				
Restudy phase (9 min)				
Final test (9 min)				

Some recent studies have begun to focus on JOL accuracy when acquiring such problem solving skills. For primary school children, JOL accuracy when solving problems was quite low and relative and absolute accuracy of immediate JOLs tended to be higher than delayed JOLs although this difference did not reach statistical significance (Baars, van Gog, de Bruin, & Paas, 2015). When learning to solve problems from worked examples, which is a very effective instructional method when students have little or no prior knowledge of the problem (for reviews see, Atkinson et al., 2000; Renkl, 2011; van Gog & Rummel, 2010), there was no difference in primary school children's absolute accuracy of immediate and delayed JOLs following worked example study (Baars, van Gog, de Bruin, & Paas, 2014). It was demonstrated though that solving a practice problem after studying the worked example significantly improved primary school children's JOL accuracy, although this improved monitoring accuracy did not affect their regulation. These primary school children did not become better at determining which items they should restudy (Baars et al., 2014).

In analogy to learning from expository text, solving a practice problem after studying a worked example is a generation strategy that presumably gives learners the opportunity to access and test the quality of the mental model they have built during worked example study. According to Griffin, Jee, and Wiley (2009), one of the routes towards making a JOL is making a judgement of test performance based on cues from performance on a task that was just completed. They therefore call this the 'postdiction route'. Solving a practice problem after worked example study gives students the opportunity to make a postdiction judgement about their performance on the practice problem and use this to predict future test performance when making a JOL.

In order to determine whether solving a practice problem after worked example study is an effective generation strategy to improve JOL accuracy for other learners as well, it is important to perform a conceptual replication of the results found by Baars et al. (2014) and test this hypothesis with adolescent secondary education students. It is expected that practice problems will also provide secondary education students with the opportunity to use a postdiction judgement about their performance, which helps them to judge their future performance more accurately. Moreover, as mentioned above, there seems to be a developmental component in whether learners use the information they gain from monitoring in regulating further study (Roebers et al., 2009), and it might be that secondary education students would not only benefit in terms of monitoring accuracy, but also in terms of regulation accuracy. Therefore, the present study investigated the effects of practice problems after worked example study on 14–15 year old secondary education students' monitoring and regulation accuracy when learning to solve problems by studying worked examples.

The present study

In the present study, in secondary education, five instructional conditions were compared in terms of their effects on JOL and regulation accuracy, and these conditions differed in timing of JOLs, whether practice problems were provided, and timing of practice problems provided: (1) worked example – JOL, (2) worked example – delay – JOL, (3) worked example – practice problem – JOL, (4) worked example – practice problem – delay – JOL and (5) worked example – delay – practice problem – JOL (see Table 1).

As for the effects of timing of JOLs on JOL accuracy (Nelson & Dunlosky, 1991; Rhodes & Tauber, 2011), we hypothesise that immediate JOLs will be more accurate than delayed JOLs after problem solving because problem solving inherently provides feedback about performance, such as whether a step could be completed, how easily it could be completed, et cetera, that is present immediately after solving the problem but not at a delay (Hypothesis 1a: condition 3 > condition 4). Moreover, we hypothesise that immediate JOLs after studying a worked example would also be more accurate than delayed JOLs because learners would be better able to judge whether they have understood the procedure demonstrated in the example and how easily they could understand it right after studying it than at a delay (Hypothesis 1b: condition 1 > condition 2); it should be noted that this is in line with the hypothesis of (Baars et al., 2014) but not with their findings; nevertheless, secondary education students might be better able to monitor cues about their understanding during example study than primary education students.

Regarding the effects of practice problems on JOL accuracy, we hypothesise, in line with the findings by Baars et al. (2014), that solving a problem after worked example study will be an effective generation strategy, as it provides students with the opportunity to test the quality of the schema they acquired by studying a worked example, which would enhance JOL accuracy compared to only studying worked examples (i.e. Hypothesis 1c: conditions 3, 4, and 5 > conditions 1 and 2).

Regarding timing of practice problems, studies on learning from expository text found that delayed keyword or summary generation (De Bruin et al., 2011; Thiede & Anderson, 2003; Thiede et al., 2003) led to more accurate JOLs compared to immediate keyword or summary generation. Therefore, it was hypothesised that delayed practice problems would enhance JOL accuracy more than immediate practice problems (i.e. Hypothesis 1d: condition 5 > conditions 3 and 4). It should be noticed that this hypothesis was not confirmed for primary school children (Baars et al., 2014) but, again, secondary school students might be better able to use the cues from delayed practice problems to monitor their learning process.

According to models of self-regulation (e.g. Winne & Hadwin, 1998; Zimmerman, 2008), monitoring influences regulation of further study behaviour. Students use the information from monitoring processes to decide what to study next. So, improved JOL accuracy should result in improved regulation of study, which should result in improved test performance (Son & Metcalfe, 2000; Thiede, 1999; Thiede et al., 2003; Thiede & Dunlosky, 1999). For adults, Thiede et al. (2003) found that generating keywords at a delay after reading an expository text improved monitoring accuracy, led to more effective regulation and improved test performance. For children, De Bruin et al. (2011) found that after generation of keywords, sixth graders were able to use their comprehension ratings to decide which text to study again. So, we expected a similar pattern of results on regulation accuracy (Hypothesis 2a–2d) and final test performance after the restudy phase (Hypothesis 3a–3d) as for JOL accuracy.

The effects of task complexity on monitoring accuracy were explored. Studying or solving more complex problems requires more WM resources (Sweller et al., 1998), and consequently leaves less WM resources for monitoring the learning process. This could affect the cues available for making monitoring judgements (i.e. JOLs) after the task is completed (cf. Kostons, van Gog, & Paas, 2009). Therefore, we explored JOL accuracy over tasks at three levels of complexity. Finally, it was explored whether practice problems had an effect on initial learning (i.e. on the criterion test: conditions 1 and 2 vs. conditions 3–5) and whether restudy had a positive effect on learning by analysing whether students' performance improved from criterion to final test.

Method

Participants and design

Participants were 143 Dutch 11th-grade students (which would be USA 9th-grade) from six different classrooms of two secondary schools (M_{age} = 14.63 years, SD = .58; 79 boys and 64 girls). Within each classroom, participants were distributed among the five conditions via random assignment: (1) worked example – JOL (n = 29), (2) worked example – delay – JOL (n = 29), (3) worked example – practice problem – JOL (n = 29), (4) worked example – practice problem – delay – JOL (n = 28) and (5) worked example – delay – practice problem – JOL (n = 28). As such, condition was a between-subjects factor; task complexity was manipulated within-subjects (see materials section). Table 1 provides an overview of the design.

Materials

All materials were paper-based and each worked example, problem solving task or rating scale was presented on a new page. In this experiment, students had to learn to solve heredity problems (laws of Mendel). The problems could be solved in six steps: (1) translating the phenotype of the father (i.e. expressions of genetic traits) described in the cover story into genotypes (i.e. a pair of upper and lower case letters representing genetic information), (2) translating the phenotype of the mother described in the cover story into genotypes, (3) making a genealogical tree, (4) putting the genotypes in a Punnett square, (5) extracting the genotype of the child from a Punnett square, (6) determining the phenotype of the child. All materials were developed by the researchers based on problem solving tasks from the study of Kostons, van Gog, and Paas (2012).

Pretest
The pretest consisted of nine open-ended questions measuring conceptual knowledge about heredity. For example, one of the questions was: 'What is a genotype in reference to a hereditary trait?' Pretest performance was scored by the experimenter using a standard of the correct answers. For each correct answer, one point was assigned, except for question 9 for which 2 points could be obtained, adding up to a maximum score of 10 points for the whole pretest.

Worked examples
Three worked examples with gradually increasing complexity were used which provided a step-by-step demonstration of how to solve heredity problems (laws of Mendel). The problems were at three different complexity levels, from lowest to highest: (1) one generation with an unknown child, (2) one generation with an unknown mother and (3) two generations with an unknown child (see also Kostons et al., 2012). Task complexity was manipulated within subjects. An example of a worked example can be found in Appendix 1.

Practice problems
Practice problems that students had to solve after studying a worked example consisted of heredity problems that were isomorphic to the ones that were explained in the worked examples (i.e. the same solution procedure but different surface features). The fact that these practice problems were isomorphic prevented students from filling out the steps in the practice problems from memory only. An example of a practice problem can be found in Appendix 2.

JOL rating
Specific JOLs about each step in the problem solving tasks were used (cf. term-specific JOLs in studies with text: Dunlosky, Rawson, & McDonald, 2002; Dunlosky, Rawson, & Middleton, 2005; Rawson & Dunlosky, 2007). JOLs were provided on a seven-point rating scale, which asked students to indicate how well they expected they could perform the *step* that was shown, in a comparable problem on a future

test, ranging from (0) *not at all* to (6) *very well* (see Appendix 3 for an example). JOLs were either asked directly after studying the worked example (condition 1), after a 3-min delay (condition 2), immediately after the practice problem (condition 3), after a 3-min delay after the practice problem (condition 4), or directly after a delayed practice problem (condition 5).

Indication of restudy
At the end of the study phase, before the criterion test was taken, participants were asked to indicate which worked examples they should study again to perform as well as possible on a future test (and they got the opportunity to do so after the criterion test; see Procedure section).

Criterion test problems
The criterion test (see Table 1) consisted of three problem solving tasks, one at each of the three complexity levels, and these tasks were identical in both solution procedure and content of the problems that were explained in the worked examples.

Final test problems
The final test (see Table 1) also consisted of three problem solving tasks, one at each of the three complexity levels, which were isomorphic (i.e. the same solution procedure but different surface features) to the ones explained in the worked examples, to the ones practiced and to the ones used in the criterion test.

Filler task
A paper-based number puzzle (Sudoku puzzle), which was unrelated to the other tasks used in this experiment, was used as a filler task (see Table 1).

Procedure

The study was run in group sessions in students' classrooms, which lasted approximately 70 min. Students were randomly assigned to one of the conditions and received a set of numbered booklets which the experiment leader used to structure the procedure. In the first booklet, all students completed the pretest (5 min). In booklets 2–12, all participants studied a worked example (3 min), and students in the conditions with practice problems solved a practice problem (for 3 min) either immediately after studying the worked example or after a filler task at a 3-min delay. After the worked example (immediately in condition 1 or at a delay in condition 2) or after the practice problem (immediately in condition 3 and 5 or at a delay in condition 4), students gave a JOL (Appendix 3). This study-JOL or study-practice-JOL cycle was repeated three times with problems of increasing complexity, after which students indicated if they needed to study a specific worked example again (booklet 13). Then, in booklet 14, all students completed the criterion test (9 min), after which they were instructed to restudy the worked examples they had chosen for restudy which were provided in a separate booklet (9 min). In this booklet, the page with the title of the example was stapled to the page with the example and students had to rip open the examples they wanted to restudy which made it possible to check which examples were restudied. Finally, in the last booklet, all students completed the final test (9 min). All students worked individually on their own booklet under supervision of the experiment leader and assistants.

Data analysis

Performance scores
Test performance on the criterion and final test was scored by the experimenter by assigning 1 point for each step correctly performed, resulting in a maximum score of 6 points per test problem and a maximum total score of 18 points on each test.

Table 2. Scoring of absolute monitoring accuracy per step and scoring of absolute regulation accuracy per problem.

Criterion test performance/Restudy choice JOL rating	Correct performance (1)	Incorrect performance(0)	No restudy (0)	Restudy (1)
0	0	1	0	1
1	.17	.83	.17	.83
2	.33	.67	.33	.67
3	.50	.50	.50	.50
4	.67	.33	.67	.33
5	.83	.17	.83	.17
6	1	0	1	0

Relative monitoring accuracy

Relative monitoring accuracy was measured with the Goodman–Kruskal gamma correlation between JOLs and performance on the criterion test problem steps. Gamma correlations between JOLs and performance on the criterion test problem steps were calculated for each individual participant, and the closer to 1, the higher the monitoring accuracy. Thirteen participants had indeterminate gamma correlations due to invariance in either JOLs or performance on the criterion test. For seven participants, no gamma correlation could be calculated because they did not fill out all JOLs. The mean of the intra-individual gamma correlations was calculated based on the following numbers of participants (1) worked example – JOL: $n = 24$, (2) worked example – delay – JOL: $n = 25$, (3) worked example – practice problem – JOL: $n = 26$, (4) worked example – practice problem – delay – JOL: $n = 22$ and (5) worked example – delay – practice problem – JOL: $n = 26$.

Absolute monitoring accuracy

We developed a gradual measure of absolute accuracy that varies between 0 and 1, based on each possible combination of JOL (0–6) and criterion test performance per step of the problem (0 or 1). The scoring system is shown in Table 2. As can be inferred from the table, lower JOLs combined with a criterion test performance of 0 resulted in higher absolute accuracy, whereas lower JOLs combined with a criterion test performance of 1 resulted in lower absolute accuracy; similarly, higher JOLs combined with a criterion test performance of 0 resulted in lower accuracy, whereas higher JOLs combined with a criterion test performance of 1 resulted in higher accuracy. Mean absolute accuracy over the three problem solving tasks from the criterion test was calculated. The higher this absolute accuracy score was, the better the absolute monitoring accuracy was. We could not calculate absolute accuracy for seven participants because they did not fill out all JOLs. The mean absolute accuracy was calculated based on the following numbers of participants (1) worked example – JOL: $n = 28$, (2) worked example – delay – JOL: $n = 29$, (3) worked example – practice problem – JOL: $n = 29$, (4) worked example – practice problem – delay – JOL: $n = 23$ and (5) worked example – delay – practice problem – JOL: $n = 26$.

Regulation accuracy

We expected students to make restudy choices based on their JOLs and expected them to choose the tasks that received a lower JOL for restudy (cf. the Discrepancy Reduction model of self-regulated study, Dunsloky & Thiede, 1998; Thiede & Dunlosky, 1999). To calculate regulation accuracy, we used a similar gradual measure as was used to calculate absolute accuracy for monitoring, which varies between 0 and 1, based on each possible combination of mean JOL for a whole problem (0–6) and restudy choice for a whole worked example (0 or 1). As can be inferred from Table 2, lower JOLs combined with the choice not to restudy the task resulted in lower regulation accuracy, whereas lower JOLs combined with the choice to restudy the task resulted in higher regulation accuracy; similarly, higher JOLs combined with the choice not to restudy the task resulted in higher regulation accuracy, whereas higher JOLs combined with the choice to restudy the task resulted in lower regulation accuracy. Mean regulation accuracy over the three problem solving tasks was calculated. The higher this regulation accuracy score

Table 3. The mean practice problem performance (range: 0–6), JOLs (range: 0–6), criterion test performance (range: 0–6), absolute accuracy JOLs (range: 0–6), relative accuracy JOLs (range: −1–1), regulation accuracy (range: 0–1) and final test performance are presented.

	Immediate JOLs (1)	Delayed JOLs (2)	Immediate practice problem and JOLs (3)	Immediate practice problem and delayed JOLs (4)	Delayed practice problem and JOLs (5)
Mean pretest performance	.18 (.22)	.22 (.20)	.29 (.23)	.25 (.17)	.23 (.18)
Practice problem performance	–	–	4.09 (1.35)	4.07 (1.07)	3.56 (1.35)
JOLs	3.76 (1.35)	3.43 (1.08)	4.09 (1.55)	3.97 (1.21)	3.37 (1.62)
Complexity 1	3.68	3.29	3.77	3.86	3.00
Complexity 2	3.52	3.32	4.12	3.62	3.33
Complexity 3	4.06	3.69	4.38	3.92	3.77
Criterion test performance	3.60 (1.58)	3.75 (1.55)	4.20 (1.56)	4.08 (1.37)	3.83 (1.37)
Complexity 1	4.69	4.66	4.86	4.89	4.54
Complexity 2	2.62	3.24	3.28	3.04	3.14
Complexity 3	3.48	3.34	4.45	4.32	3.82
Absolute accuracy JOLs	.60 (.15)	.58 (.12)	.71 (.16)	.64 (.16)	.61 (.15)
Complexity 1	.60	.56	.67	.64	.60
Complexity 2	.59	.57	.68	.58	.60
Complexity 3	.60	.61	.76	.66	.63
Relative accuracy JOLs	.33 (.48)	.20 (.45)	.39 (.53)	.34 (.50)	.17 (.38)
Regulation accuracy	.50 (.15)	.54 (.13)	.64 (.20)	.60 (.16)	.60 (.19)
Final test performance	4.15 (1.68)	4.03 (1.49)	4.49 (1.69)	4.21 (1.41)	4.50 (1.42)
Complexity 1	4.97 (1.72)	4.72 (1.44)	4.97 (1.90)	5.00 (1.39)	5.14 (1.48)
Complexity 2	3.48 (1.96)	3.14 (2.13)	3.86 (2.03)	3.25 (1.86)	3.86 (1.65)
Complexity 3	4.00 (2.24)	4.24 (1.90)	4.83 (1.69)	4.39 (2.14)	4.39 (2.15)

Note: Standard deviations are provided within parentheses.

was the better JOLs and restudy choices corresponded. We could not calculate regulation accuracy for seven participants because they did not fill out all JOLs. The mean regulation accuracy was calculated based on the following numbers of participants per condition (1) worked example – JOL: $n = 28$, (2) worked example – delay – JOL: $n = 29$, (3) worked example – practice problem – JOL: $n = 29$, (4) worked example – practice problem – delay – JOL: $n = 23$ and (5) worked example – delay – practice problem – JOL: $n = 27$.

Results

As a check on randomisation, the pretest performance scores were compared, which showed no differences between conditions, $F(4, 138) = 1.25$, $p = .294$. The mean pretest performance, mean practice problem performance, JOLs, criterion test performance, absolute accuracy, relative accuracy, regulation accuracy and final test performance per condition are presented in Table 3.

Monitoring accuracy

Relative accuracy
Planned comparisons were conducted to test our hypotheses. The planned comparisons showed that there was no significant difference in relative accuracy between conditions that gave an immediate vs. delayed JOL after practice problems (Hypothesis 1a: condition 3 vs. 4), $t(118) < 1$, $p = .747$, between conditions that gave an immediate vs. delayed JOL after worked example study (Hypothesis 1b: condition
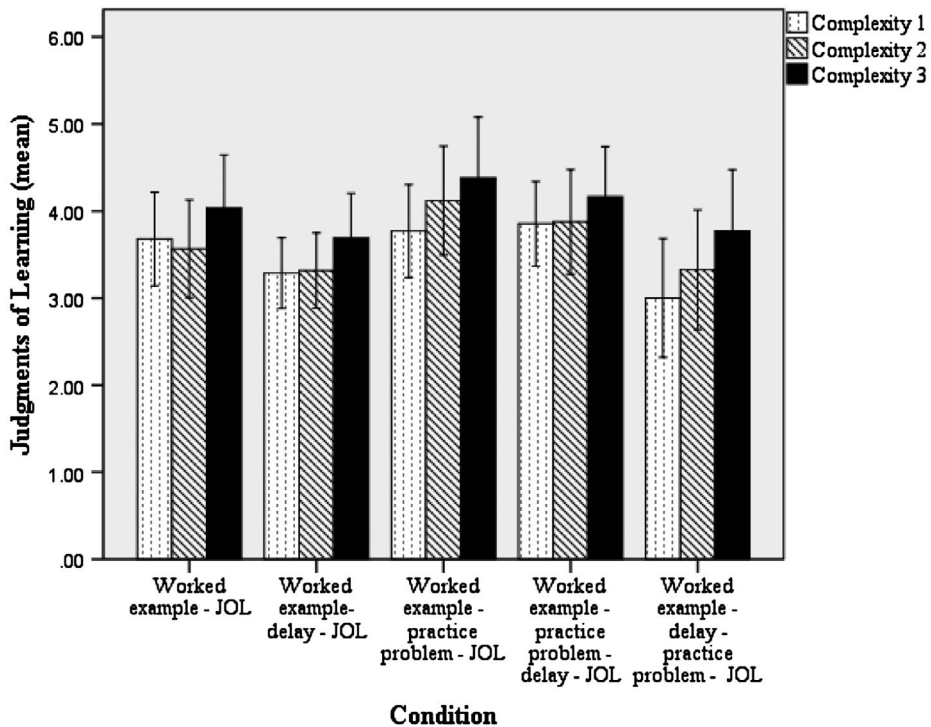
Figure 1. Mean JOLs for the three complexity levels per condition. Note: Standard errors of the mean are represented by the error bars.

1 vs. 2), $t(118) = 1.02, p = .308$ and between conditions in which students solved problems after worked example study and conditions in which students did not solve problems (Hypothesis 1c: condition 1 and 2 vs. condition 3–5), $t(118) < 1, p = .700$. The fourth planned comparison (Hypothesis 1d: condition 3 and 4 vs. condition 5) showed a trend that suggested a difference in mean relative accuracy between delayed and immediate problem solving, but this was not statistically significant, $t(118)=-1.71, p = .090$. The conditions in which students could solve a practice problem directly after studying the worked example showed a higher mean relative accuracy.

Absolute accuracy
To test our hypotheses in terms of absolute accuracy between JOLs and performance, we conducted the same planned comparisons as for absolute accuracy. The first planned comparison (Hypothesis 1a: condition 3 vs. condition 4) showed that although numerical absolute accuracy of immediate JOLs was higher, there was no significant difference between conditions that gave an immediate vs. delayed JOL after practice problems, $t(131) = 1.64, p = .104$. The second planned comparison (Hypothesis 1b: condition 1 vs. condition 2) showed that there was no significant difference between conditions in which an immediate vs. delayed JOL was given after worked example study, $t(131) < 1, p = .567$. The third planned comparison (Hypothesis 1c: condition 1 and 2 vs. condition 3–5) showed that absolute accuracy scores of students who solved practice problems after worked example study differed significantly from students who did not solve problems after worked example study, $t(131) = 2.39, p = .018$, Cohen's $d = -.42$. Conditions in which students could solve practice problems after studying worked examples showed higher absolute accuracy. Note that after a Bonferroni adjustment of the alpha level ($a = .013$), this result was not significant, but may reflect a trend in the data. The fourth planned comparison (Hypothesis 1d: conditions 3 and 4 vs. condition 5) revealed that there may be a trend, indicating that the conditions in which students could solve practice problems directly after worked example study showed a higher

absolute accuracy than the conditions in which students solved practice problems at a delay after worked example study, although again, this was not statistically significant, $t(131) = -1.71$, $p = .092$.

Furthermore, we analysed absolute accuracy of the JOLs as a function of complexity and the condition. In Figure 1, the mean JOL per complexity level and per condition is shown. A repeated measures ANOVA with complexity (three levels) as within-subjects factor and condition as between-subjects factor showed that absolute accuracy significantly changed over the levels of complexity, $F(2, 262) = 6.04$, $p = .003$, $\eta_p^2 = .04$, and that the complexity effect did not vary across conditions (i.e. there was no Complexity × Condition interaction), $F(8, 262) < 1$, $p = .840$. Absolute accuracy was higher for the third and most complex task. Post hoc comparisons revealed that absolute accuracy was significantly higher for the third complexity level compared to the first, $p = .017$, and second complexity level, $p = .018$. Furthermore, there was a significant effect of condition on absolute accuracy, $F(1, 262) = 3.12$, $p = .017$, $\eta_p^2 = .09$. Post hoc comparisons showed that absolute accuracy was significantly higher in the condition with immediate practice and immediate JOL compared to the condition with delayed JOL only, $p = .014$.

Regulation accuracy

To test our hypotheses about regulation accuracy, we conducted planned comparisons. The first hypothesis (Hypothesis 2a: condition 3 vs. condition 4) showed that there was no significant difference in regulation accuracy between conditions that gave an immediate vs. delayed JOL after practice problems, $t(131) = 1.06$, $p = .292$. The second planned comparison (Hypothesis 2b: condition 1 vs. condition 2) showed that there was no significant difference between conditions in which an immediate vs. delayed JOL was given after worked example study, $t(131) < 1$, $p = .395$. The third planned comparison (Hypothesis 2c: condition 1 and 2 vs. condition 3–5) showed a significant difference between conditions in which practice problems were provided and conditions in which no practice problems were provided, $t(131) = 2.77$, $p = .007$, Cohen's $d = .48$. Regulation accuracy was higher for the conditions with practice problems. This result remains significant after a Bonferroni adjustment of the alpha level ($\alpha = .013$). The fourth planned comparison (Hypothesis 2d: conditions 3 and 4 vs. condition 5) showed no significant difference between delayed and immediate practice problem solving, $t(131) = -1.20$, $p = .351$.

Yet, not all actual restudy choices made after the criterion test were the same as restudy indications made before the first test. To get an idea about the amount of students that restudied different problems than indicated, actual restudy choices (0 or 1) were subtracted from indicated restudy indications (0 or 1). 81.1% of the students restudied as they indicated, 11.2% of the students restudied more than they indicated and 7.7% of the students restudied less than they indicated. Planned comparisons on regulation accuracy when using student's actual restudy behaviour showed the same pattern of results as the planned comparisons on regulation accuracy as described above (Condition 3 vs. condition 4, $t(131) = 1.13$, $p = .259$, Condition 1 vs. condition 2, $t(131) < 1$, $p = .578$, Conditions 1 and 2 vs. conditions 3–5, $t(131) = 2.31$, $p = .023$ and conditions 3 and 4 vs. condition 5, $t(131) = 1.60$, $p = .113$.).

Test performance

To test our hypotheses that improved monitoring would lead to improved regulation and therefore to improved *final* test performance, the same planned comparisons were conducted. Not surprisingly given the lack of findings on monitoring and regulation accuracy, the first planned comparison (Hypothesis 3a: condition 3 vs. condition 4) showed that there was no significant difference in final test performance between conditions that gave an immediate vs. delayed JOL after practice problems, $t(138) < 1$, $p = .495$. The second planned comparison (Hypothesis 3b: condition 1 vs. condition 2) showed that there was no significant difference in mean final test performance between conditions in which an immediate vs. delayed JOL was given after worked example study, $t(138) < 1$, $p = .777$. What was surprising, given the results on monitoring and regulation accuracy, is that the third planned comparison (Hypothesis 3c: conditions 1 and 2 vs. conditions 3–5) showed no significant difference between the conditions in which practice problems were provided and the conditions in which no practice problems were

provided, $t(138) = 1.38$, $p = .239$. And again, not surprisingly given the lack of findings on monitoring and regulation accuracy, the fourth planned comparison (Hypothesis 3d: conditions 3 and 4 vs. condition 5) showed no significant difference in final test performance between conditions with immediate vs. delayed practice problem solving, $t(138) < 1$, $p = .683$.

The explorative analysis of whether practice had a positive effect on *criterion* test performance (Conditions 1 and 2 vs. conditions 3–5) showed that this was not the case, $t(138) = 1.44$, $p = .153$.

The explorative analysis of the effect of restudy on learning was conducted with a repeated measures ANOVA with test moment (Criterion test vs. Final test) as within-subjects factor and condition as between-subjects factor which showed that test performance significantly increased from criterion test to final test, $F(1, 138) = 29.58$, $p < .001$, $\eta_p^2 = .18$, but there was no significant difference among conditions, $F(4, 138) < 1$, $p = .702$ and no interaction between test moment and conditions, $F(4, 138) = 1.84$ 1, $p = .125$.

Discussion

The present study investigated the effect of immediate and delayed practice problems after worked example study on the accuracy of JOLs, regulation accuracy and test performance. No significant difference was found between immediate and delayed JOL accuracy, both relative and absolute accuracy, after practice problems (Hypothesis 1a). Note that the immediate JOL condition (3) seemed to show higher absolute accuracy than the delayed JOL condition (4) in line with our expectation, but this was not statistically significant. Neither was there an effect of delaying JOLs after worked example study (Hypothesis 1b). Despite possible trends in mean scores suggesting that immediate JOLs would be more accurate in problem solving tasks, this does not seem to be a significant or reliable effect. Note that these findings are in line with findings regarding JOLs about expository texts, where delaying JOLs did not affect accuracy, unless a generation strategy was added (Maki, 1998; Thiede et al., 2009).

The current study replicates and extends the findings from our previous study in primary education, which showed that practice problems diminished overconfidence in JOLs (Baars et al., 2014). In line with our expectation (Hypothesis 1c), practice problems helped students to make more accurate JOLs. Absolute accuracy was higher for students who worked on practice problems after worked example study than for students who did not solve problems after worked example study. Similar to the generation strategies that have been found to improve JOL accuracy when learning from expository text (i.e. keywords, Thiede et al., 2003; summaries, Anderson & Thiede, 2008; self-explanations, Griffin et al., 2008; concept maps, Thiede et al., 2010), practice problems seem an effective generation strategy to improve JOL accuracy when learning to solve problems. Also, in line with our hypothesis, but in contrast to the findings with primary school children, regulation accuracy was higher for adolescents who were provided with practice problems compared to students who did not receive practice problems after worked example study. However, in contrast to our expectation that delayed practice problems would lead to the highest JOL accuracy (Hypothesis 1d), both relative and absolute accuracy showed a trend suggesting that accuracy was higher for students who solved practice problems *immediately* after worked example study. Regulation accuracy and final test performance did not differ between conditions with immediate or delayed practice problems.

In analogy to the explanation offered in the studies on generation strategies when monitoring learning from text (i.e. the situation model approach, Thiede et al., 2009), the effect of practice problems can be explained by the opportunity they provide students to test their mental model of how to solve this type of problem should be solved, and to use this information to make a JOL and regulate further study. Another explanation, which is not mutually exclusive with the mental model explanation, is that the practice problems allowed students to use mnemonic cues like encoding or retrieval fluency to base their JOLs on (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Roediger & Karpicke, 2006). Both explain why practice problems could provide students with more valid cues for making JOLs compared to studying worked examples alone. Moreover, mental effort may have been a cue: solving isomorphic practice problems after studying worked examples could have caused a high load on WM because

students had to replicate the procedural steps with different surface information, requiring students to invest more mental effort. Prior research has shown a negative relationship between JOLs and effort ratings (e.g. Baars, Visser, van Gog, de Bruin, & Paas, 2013). The effect of mental effort investment on metacognitive monitoring during problem solving would be an interesting subject for future research.

Interestingly, timing of practice problems does seem to be important, but in contrast to our expectation, we found that absolute accuracy was higher for immediate practice problems compared to delayed practice problems, even though students have to fully rely on information from LTM when solving delayed practice problems. Also, we found a trend suggesting that relative accuracy tended to be better for immediate practice problems. The delay in the current study was 3 min, but perhaps with a longer delay, different results would be found. Future research could investigate the effects of longer delays on monitoring accuracy. Furthermore, one might argue that on immediate practice problems, students recall the example better, but if anything, one would expect this to affect problem solving performance, not necessarily monitoring accuracy (since the test is also taken a substantial amount of time after example study and JOLs prompted the students to predict future test performance). Surprisingly, however (though in line with the study in primary education), no effects of practice problems on criterion test performance were found. Prior studies comparing worked example study only with example-problem pairs did not find differences in performance on an immediate test (van Gog & Kester, 2012; van Gog, Kester, & Paas, 2011b) either; however, in those studies, solving a problem meant getting one example less to study. In our study, the practice problems were additional; students in the worked examples conditions simply got fewer learning tasks. It is therefore quite surprising that the additional opportunity to practise a problem did not lead to better outcomes on the criterion test. Possibly, the opportunity to solve a problem only allows for learning when performance on that problem is high; that is, when learners have a high level of prior knowledge or have acquired a lot of knowledge from example study, but that would probably require studying multiple examples. This is an interesting issue for future research on example-based learning to address.

It was also quite remarkable that we did not find differences among conditions in performance on the final test, given that practice problems led to higher regulation accuracy. After the criterion test, students were able to actually restudy the worked examples they had indicated they should study again at the end of the learning phase. According to models of self-regulation (e.g. Winne & Hadwin, 1998) and earlier findings on learning from expository text (Thiede et al., 2003), better monitoring accuracy should lead to better regulation accuracy which should lead to better test performance if students have the opportunity to control their study time allocation. Possibly, solving a problem after studying an example made students aware of what they knew and did not know, but might not have been sufficient to improve their final test performance. This would have allowed them to make more accurate restudy choices, but again, restudying the example only once might not have been sufficient to boost their final test performance. In addition, the restudy option was to study the worked example again which could have been experienced by students as being too limited to actually mend their suboptimal understanding. Therefore, it would be interesting for future research to use multiple examples or tasks for restudy to investigate the effect of regulation on learning. Also, future studies could include other possible restudy options in a classroom, such as asking a teacher for feedback or asking a peer for help. Furthermore, because JOLs were not perfectly accurate, regulation choices might have been more accurate, but still suboptimal. Also, some students (although only 13 students) restudied other examples than the ones they indicated during the learning phase, which might have interfered with the relation between regulation accuracy and test performance.

Because some of the findings in the current study are in line with findings from studies on learning from expository text (e.g. effect of generation strategies) but other findings are not (e.g. no effect on relative monitoring accuracy), the cognitive and metacognitive processes when learning to solve problems compared to other materials like word pairs or texts, and their consequences for (measuring) monitoring accuracy, would be an interesting topic of future research. With regard to the monitoring process and consequences for monitoring, procedural problem solving tasks such as the ones used in the current study are cumulative, in the sense that the answers to previous steps

in the problem are needed to generate answers to the next steps in the problem solving task. This is an important difference with learning from text, where students may not understand a part in the middle, but may still be able to answer questions about a later part. This may also explain why regulation was affected even though relative accuracy was not: Students gave JOLs for each step in the worked example or problem solving task and relative accuracy only shows whether students can discriminate between the different steps when predicting performance on a future test (i.e. criterion test); but the value of being able to discriminate will be lower towards the end of the problems, as inability to solve one step affects all subsequent steps. Moreover, because regulation choices were made per whole problem solving task, absolute accuracy might be more important for regulation of these procedural problem solving tasks. Even though it was also based on judgements per step in our study, improved absolute accuracy shows that students' awareness of their understanding of the specific steps has improved and this may help them to decide whether to study a whole problem solving task again or not.

Limitations of this study are the small number of problem solving tasks used and the fact that tasks were only available at three complexity levels which had to be presented sequentially because of the difficulty of the tasks. With more problem solving tasks, students might become more experienced with making JOLs about the tasks, which could lead to better JOL accuracy. Also, JOLs were found to be most accurate for the most complex problem solving task, yet this was also the third problem solving task students had to judge, which points out a possible confound. It is not clear whether complexity or experience caused JOLs to be most accurate at the third and most complex problem solving task. In addition, Moos and Azevedo (2008) have shown that prior knowledge is also related to monitoring and could therefore also be of importance when investigating complexity and experience when monitoring problem solving tasks. Future research should try to disentangle these possible causes. Furthermore, in the current study, monitoring of complex tasks was far from perfect (gamma correlation between .20 and .39). Research has shown that scaffolds can enhance metacognitive activities of triads working together on complex learning tasks (Molenaar, Van Boxtel, & Sleegers, 2010). The triads showed more metacognitive activities when provided with scaffolds, which in turn stimulated individual metacognitive skills. So, moving from individual study to small group learning, and scaffolding the process of monitoring during small group learning, might further enhance monitoring accuracy when learning to solve biology problems. Finally, we only measured JOLs to investigate monitoring and restudy choices to measure regulation. Other measures might give additional insight into the processes of monitoring and regulation and could therefore be used in future research.

Finally, perhaps the number of participants (N = 143) in the present study might have prevented the detection of some effects we expected. A post hoc power analysis revealed that on the basis of the standard deviation in absolute accuracy in the current study and the expected difference in absolute accuracy between conditions of .50, at the recommended .80 level (Cohen, 1988), the number of participants needed to be able to reject the H0 hypothesis in a comparison should be 64 participants. In the current study, 143 secondary school students participated but they were divided among five conditions, which means that for some comparisons, this number was not reached.

In sum, the current study showed that providing secondary education students with practice problems after worked example study led to improved JOL and regulation accuracy. To the best of our knowledge, the current study was the first to influence regulation accuracy using a generation strategy when learning to solve problems in the classroom. Next to the theoretical implications of this study, this study has practical relevance, in the sense that practice problems could be implemented relatively easily in educational practice. However, despite better regulation, final test performance was not affected. Therefore, future research should follow up on these findings and should attempt to gain more insight into the relationships among JOLs, regulation of study and performance.

Acknowledgements

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & 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

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

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. doi:10.1002/acp.3008

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2015). *Accuracy of primary school children's immediate and delayed judgments of learning about problem solving tasks*. Manuscript submitted for publication.

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

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

Bryce, D., & Whitebread, D. (2012). The development of metacognitive skills: Evidence from observational analysis of young children's behavior during problem solving. *Metacognition and Learning, 7*, 197–217. doi:10.1007/s11409-012-9091-2

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. San Diego, CA: Academic Press.

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

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 in cognitive skill acquisition: The effect of learner expertise. *European Journal of Cognitive Psychology, 19*, 671–688. doi:10.1080/09541440701326204

De Bruin, A. B. H., Thiede, K. W., 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

Destan, N., Hembacher, E., Ghetti, S., & Roebers, C. M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5-to 7-year-old children. *Journal of Experimental Child Psychology, 126*, 213–228. doi:10.1016/j.jecp.2014.04.001

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*, 228–232. doi:10.1111/j.1467-8721.2007.00509

Dunlosky, J., Rawson, K., & McDonald, S. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In T. Perfect & B. Schwartz (Eds.), *Applied metacognition* (pp. 68–92). Cambridge: Cambridge University Press.

Dunlosky, J., Rawson, K., & Middleton, E. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*, 551–565. doi:10.1016/j.jml.2005.01.011

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

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*, 411–435. doi:10.1016/0010-0285/92

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37*, 1001–1013. doi:10.3758/MC.37.7.1001

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 & 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, NY: Routlegde.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 107. doi:10.1037/0278-7393.6.2.107

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

Kostons, D., van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction, 22*, 121–132. doi:10.1016/j.learninstruc.2011.08.004

Krebs, S. S., & Roebers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology, 80*, 325–340. doi:10.1348/000709910X485719

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.

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

Molenaar, I., Van Boxtel, C. A., & Sleegers, P. J. (2010). The effects of scaffolding metacognitive activities in small groups. *Computers in Human Behavior, 26*, 1727–1738. doi:10.1016/j.chb.2010.06.022

Moos, D. C., & Azevedo, R. (2008). Self-regulated learning with hypermedia: The role of prior domain knowledge. *Contemporary Educational Psychology, 33*, 270–298. doi:10.1016/j.cedpsych.2007.03.001

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

Rawson, K., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*, 559–579. doi:10.1080/09541440701326022

Redford, J. S., Thiede, K. W., Wiley, J., & Griffin, T. (2012). Concept mapping improves metacomprehension accuracy among 7th graders. *Learning and Instruction, 22*, 262–270. doi:10.1016/j.learninginstruc.2011.10.007

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.

Renkl, A. (2013). Towards an instructionally-oriented theory of example-based learning. *Cognitive Science*, 38(1), 1–37. doi:10.1111/cogs.12086

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*, 131–148. doi:10.1037/a0021705

Roebers, C. M., Schmid, C., & Roderer, T. (2009). Metacognitive monitoring and control processes involved in primary school children's test performance. *British Journal of Educational Psychology, 79*, 749–767. doi:10.1348/978185409X429842

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

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

Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition and education* (pp. 278–298). New York, NY: Routledge.

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., 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., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*, 331–362. doi:10.1080/01638530902959927

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 of metacognition and self-regulated learning*, 85–106. Mahwah, NJ: Erlbaum.

Thiede, K., Redford, J. S., Wiley, J., & Griffin, T. D. (2012). Elementary school experience with comprehension testing may influence metacomprehension accuracy among seventh and eighth graders. *Journal of Educational Psychology, 104*, 554–564. doi:10.1037/a002860

van Gog, T., & Kester, L. (2012). A test of the testing effect: Acquiring problem solving skills from worked examples. *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., & 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 D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Hillsdale, NJ: LEA.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166–183.

Zwaan, R. A. (1999). Situation models: The mental leap into imagined worlds. *Current Directions in Psychological Science, 8*, 15–18. doi:10.1111/1467-8721.00004

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162–185. doi:10.1037/0033-2909.123.2.162

Appendix 1.

<div align="center">

Worked example

1 generation with a homozygote parent and a heterozygote parent

</div>

Given:

1. The gene for curly hair (H) dominates the gene for straight hair (h).
2. The father Josh has curly hair.
3. The mother Annie has curly hair too.
4. Josh has a homozygote genotype and Annie has a heterozygote genotype.

Question: What could the genotypes (genes) and phenotypes (looks) for hair of Josh's and Annie's children be?

| Step	Answer
Step 1. Determine the genotype for father's hair We know that the father (Josh) has curly hair. Also, we know that the gene for curly hair is *dominant* and that it is depicted with a capital letter *H*. When a dominant feature is visible in the way somebody looks (phenotype), then it could be the case that both genes in the genotype are different (*Hh*) or the same (*HH*). We also know that Josh is *homozygote* for hair. If a person is homozygote for a feature then both genes in the genotype are the same. In this example it means that the father has genotype *HH*.	HH
Step 2. Determine the genotype for mother's hair We know that the mother (Annie) has curly hair. Also, we know that the gene for curly hair is *dominant* and that it is depicted with a capital letter *H*. We also know that the mother is heterozygote for hair. When a person is heterozygote for a feature then both genes in the	Answer Hh

genotype are different. In this example it means that the mother has the genotype *Hh.*

Step 3. Make a family tree

A family tree is a graphical representation of the genotypes. The parents are in the top and below them are the children.

Answer
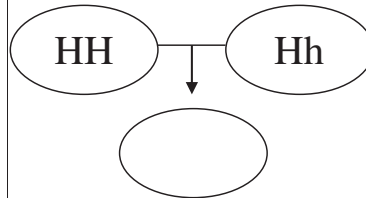
Step 4. Make a crosstable to mix the genotypes of the parents and put down the possible genotypes for their children

a. Make a crosstable and divide the genes of the genotypes of the mother in the two cells of the upper row and the genes of the genotypes of the father in the left column.

b. Fill out the crosstable by combining the genes of the father and the mother.

Answer

		Annie Hh	
		H	**h**
Josh	**H**	**HH**	**Hh**
HH	**H**	**HH**	**Hh**

Step 5. Determine the possible genotypes for hair for the children and the chance to get those genotypes

GENOTYPE = GENES

		Annie Hh	
		H	**h**
Josh	**H**	**HH**	**Hh**
HH	**H**	**HH**	**Hh**

You can get this information from the crosstable you just made. In the four cells of the crosstable you find the four possible genotypes for a child. If this genotype is in one cell that means there is a 25% chance for a child to get this genotype.

Answer
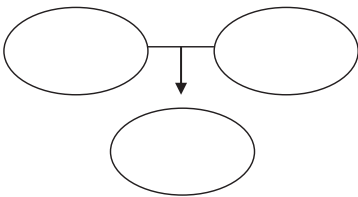
50% HH and 50% Hh

In this example: two cells have HH = 50% and two cells have Hh = 50%.	
Step 6. Determine the possible phenotypes for hair for the children and the chance to get those phenotypes	Answer
PHENOTYPE = LOOKS	
Genotype HH means that the dominant feature will show (H = curly hair). Genotype Hh means that the dominant feature will show (H = curly hair). Genotype hh mean that the recessive feature will show (h = straight hair) In this example we know that a child would have a 50% chance to get genotype HH or genotype Hh. This means that the child will have a 100% chance to have curly hair.	100% curly hair

Practice problem

1 generation with a homozygote parent and a heterozygote parent

Given:

1. The gene for freckles (F) dominates the gene for no freckles (f).

2. The father Josh has freckles.

3. The mother Annie has freckles too.

4. Josh has a homozygote genotype and Annie has a heterozygote genotype.

Question: What could the genotypes (genes) and phenotypes (looks) for freckles of Josh's and Annie's children be?

Step	Answer
Step 1. Determine the genotype for father's freckles	
Step 2. Determine the genotype for mother's freckles	Answer
Step 3. Make a family tree	Answer

Step 4. Make a crosstable to mix the genotypes of the parents and put down the possible genotypes for their children

Answer

		Annie	
Josh			

Step 5. Determine the possible genotypes for freckles for the children and the chance to get those genotypes

GENOTYPE = GENES

		Annie
Josh		

Answer

Step 6. Determine the possible phenotypes for freckles for the children and the chance to get those phenotypes

Answer

Appendix 2.

Example: 1 generation with a homozygote parent and a heterozygote parent
Question: What could the genotypes (genes) and phenotypes (looks) for hair of the children be?

Step 1. Determine the genotype for father's hair

How well would you be able to solve the step in which <u>the genotype of the father has to be determined</u> during a future test?

Indicate on a scale from 'Not at all' to 'Very well'.

Not at all						Very well

Step 2. Determine the genotype for mother's hair

How well would you be able to solve the step in which <u>the genotype of the mother has to be determined</u> during a future test?

Indicate on a scale from 'Not at all' to 'Very well'.

Not at all						Very well

Example: 1 generation with a homozygote parent and a heterozygote parent
Question: What could the genotypes (genes) and phenotypes (looks) for hair of the children be?

Step 3. Make a family tree

How well would you be able to solve the step in which <u>a family tree has to be filled out</u> during a future test?

Indicate on a scale from 'Not at all' to 'Very well'.

Not at all						Very well

Step 4. Make a crosstable to mix the genotypes of the parents and put down the possible genotypes for their children

How well would you be able to solve the step in which <u>genotypes are mixed in a crosstable</u> during a future test?

Indicate on a scale from 'Not at all' to 'Very well'.

Not at all						Very well

Appendix 3.

Example: 1 generation with a homozygote parent and a heterozygote parent
Question: What could the genotypes (genes) and phenotypes (looks) for hair of the children be?

Step 5. Determine the possible genotypes for hair for the children and the chance to get those genotypes

How well would you be able to solve the step in which <u>possible genotypes of the child and the chance to get those has to be determined</u> during a future test?

Indicate on a scale from 'Not at all' to 'Very well'.

Not at all						Very well
←						→

Step 6. Determine the possible phenotypes for hair for the children and the chance to get those phenotypes

How well would you be able to solve the step in which <u>possible phenotypes of the child and the chance to get those has to be determined</u> during a future test?

Indicate on a scale from 'Not at all' to 'Very well'.

Not at all						Very well
←						→