

Instructional Strategies for Improving Self-Monitoring of Learning to Solve Problems

The research reported here was carried out at the



In the context of the research school



(Interuniversity Center for Educational Research)

And was funded by



Netherlands Organisation for Scientific Research

(PROO project # 411-07-152)

ISBN: 978-90-5335-843-6

Copyright © Martine Baars, Rotterdam, The Netherlands, 2014

Cover design: Martine Baars

Printed by Ridderprint

All rights reserved.

**Instructional Strategies for Improving Self-Monitoring of Learning to
Solve Problems**

Instructiestrategieën ter verbetering van zelfbeoordeling tijdens het leren
probleem-oplossen

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam

op gezag van de rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor
Promoties.

De openbare verdediging zal plaatsvinden op
vrijdag 6 juni 2014 om 09:30 uur

door

Martine Baars
geboren te Woerden



Promotiecommissie

Promotoren Prof.dr. G. W. C. Paas
Prof.dr. T. A. J. M. Van Gog

Overige leden Prof.dr. J. J. G. Van Merriënboer
Prof.dr. J. T. C. Wirth
Dr. S. M. M. Loyens

Co-promotor Dr. A.B. H. De Bruin

Contents

Chapter 1 General Introduction.....	7
Chapter 2 Accuracy of Primary School Children's Immediate and Delayed Judgments of Learning about Problem-solving Tasks	17
Chapter 3 Effects of Problem Solving after Worked Example Study on Primary School Children's Monitoring Accuracy	37
Chapter 4 Effects of Problem Solving after Worked Example Study on Monitoring Accuracy in Secondary Education	59
Chapter 5 Completion of Partially Worked-out Examples as a Generation Strategy for Improving Monitoring Accuracy	89
Chapter 6 Effects of Training Self-assessment and Using Assessment Standards on Retrospective and Prospective Monitoring of Problem Solving	113
Chapter 7 Summary and General Discussion.....	149
References	163
Nederlandse samenvatting (Dutch summary).....	177
Publications.....	183
Dankwoord.....	185
Curriculum Vitae.....	187
ICO Dissertation Series.....	189

Chapter 1

General Introduction

General Introduction

Being able to regulate their own learning process is becoming increasingly important for students at all levels of education (OECD Programme for International Student Assessment, 2009). From early on in children's school careers, children are stimulated to be aware of what they are learning and to make choices about their own learning processes. Self-regulated learning can be defined as a self-directive process by which learners are able to improve their learning performance using the capabilities they already have (Zimmerman, 2008). According to the model of self-regulated learning by Winne and Hadwin (1998), monitoring and control are central processes to self-regulated learning. To effectively regulate their own learning process, students must be able to monitor their progress toward learning goals and use this information to regulate (i.e., control) further study (Metcalfe, 2009; Winne & Hadwin, 1998). For example, if students are trying to solve a math problem, it is important for them to keep track of their conceptual understanding of the problem and the steps of its solution procedure (i.e., monitoring), and to use this to determine whether more problems should be studied or practiced in order to grasp the procedure for solving this type of problem (i.e., control). Monitoring is assumed and has been shown to inform control (Kornell & Metcalfe, 2006; Metcalfe, 2009; Serra & Metcalfe, 2009; Thiede, Anderson, & Therriault, 2003; Winne & Hadwin, 1998), and can therefore be considered a crucial aspect of self-regulated learning.

Monitoring can be measured both *retrospectively*, by asking students to judge their performance on a task just completed, which is also known as self-assessment (in problem-solving tasks; Kostons, Van Gog, & Paas, 2012) or self-score judgment (in verbal tasks; Lipko et al., 2009; Rawson & Dunlosky, 2007) and *prospectively*, by asking students to predict their performance on that task on a future test, which is also known as a *Judgment of Learning* (JOL; e.g., Koriat, Ackerman, Lockl, & Schneider, 2009a, 2009b; Metcalfe & Finn, 2008; Nelson & Dunlosky, 1991).

The accuracy of monitoring can be measured both with relative and absolute measures. Relative accuracy shows whether students were able to discriminate between the different items that had to be monitored during a learning session. That is, relative accuracy measures to what extent students can distinguish between items that were learned versus items that were not learned. Relative accuracy is usually determined by calculating the Goodman-Kruskal gamma correlation, which correlates the JOLs and test performance pair

wise, and ranges between -1, indicating poor accuracy, and +1, indicating perfect accuracy. Absolute accuracy shows the absolute difference between monitoring judgments and actual test performance. The absolute accuracy of self-assessments or JOLs is usually determined by subtracting students' actual performance from their own judgment about their performance, with lower actual than judged performance resulting in overestimation, and higher actual than judged performance resulting in underestimation (this is called bias). One can also look at the difference between the two without considering the direction (i.e., absolute deviation; Mengelkamp & Bannert, 2010; Schraw, 2009).

Research has shown that the accuracy of both retrospective self-assessments and prospective JOLs are often low, but can be improved by instruction; however, this research has mainly focused on JOLs about learning items (e.g., word pairs) or learning from expository texts (for reviews, see Dunlosky & Lipko, 2007; Rhodes & Tauber, 2011; Thiede, Griffin, Wiley, & Redford, 2009). Only very few studies have investigated how to improve monitoring and regulation accuracy when learning to solve problems in an educational context. Yet, problem-solving tasks play an important role in learning school subjects like math, biology or science in which students are also increasingly expected to be(come) self-regulated learners.

In this dissertation, five empirical studies conducted in primary and secondary education are presented that investigated the question of whether and how JOLs and regulation accuracy can be improved when learning to solve problems. These studies will be outlined below, after a description of the findings on improving JOL accuracy when learning items or expository texts.

Improving Monitoring Accuracy When Learning Items or Texts

In a typical experiment in which monitoring accuracy is measured by JOLs, participants first study items such as word pairs (e.g., Nelson & Dunlosky, 1991) or study 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 (e.g., Nelson & Dunlosky, 1991) or by rating their comprehension for each text (e.g., Thiede et al., 2003). After all materials have been studied and judged, participants take a test on which their performance is measured. For word pairs it was found that delaying JOLs, that is, making JOLs per word pair only after studying multiple word pairs, improved 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 replicated and found to be a robust effect with other verbal items to be studied, such as paired associates, category exemplars, sentences, and single words for adults (Rhodes & Tauber, 2011).

The delayed-JOL effect can be explained by the memory systems involved in making JOLs (Nelson & Dunlosky, 1991). Immediate JOLs can be based on retrieval of information that is still available from working memory (WM). However, not all information from working memory is also learned, that is, stored in long-term memory (LTM). On a future test, students have to rely on information available in LTM, and when making delayed JOLs students rely on retrieval of information from LTM, which is a more valid source to base JOLs about future test performance on (Dunlosky & Nelson, 1997; Nelson & Dunlosky, 1991). Although there are some studies that show the delayed-JOL effect with children (Koriat et al., 2009a; 2009b; Schneider, Visé, Lockl, & Nelson, 2000), in their meta-analytic review Rhodes and Tauber (2011) conclude that it is not as robust as it is for adults.

Studies on learning from expository text, also found that JOL accuracy was generally low, but that it could *not* be improved by delaying JOLs (Maki, 1998a). Making a JOL about text is quite different from making a JOL about word pairs because it 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. Fortunately, other means to improve JOL accuracy for expository texts were found. For example, providing students with generation instructions that focused their attention on their comprehension of a text prior to making a JOL, improved the accuracy of JOLs.

Such instructional strategies (a.k.a. ‘generation strategies’), consisted for instance of summarizing texts (Thiede & Anderson, 2003), or generating keywords about texts (Thiede et al., 2003), at a delay after reading the text prior to making *delayed* JOLs (for a review, see Thiede et al., 2009). Generating keywords after reading a text was found to improve relative JOL accuracy for 9-10 and 12-13 year old children as well (De Bruin, Thiede, Camp, & Redford, 2011). This positive effect of generating keywords and summaries at a delay on JOL accuracy is called the ‘delayed-generation effect’ (Thiede et al., 2009). Similar to the delayed-JOL effect found for word pairs, the delayed-generation effect can be explained by the fact that delayed generation of keywords or summaries relies on

retrieving information from LTM, which is more indicative of future test performance than retrieving information from WM. That is, after a delay, generation strategies help students to access the situation model (i.e., mental representation) they constructed about the text as it is stored in LTM. Using information from the situation model helps students to make more accurate JOLs, because their deeper level of understanding about the text, which they also have to use during a future test, resides in the situation model (Thiede et al., 2009).

Interestingly, other generation strategies such as self-explaining expository text, which was investigated with adults (Griffin, Wiley, & Thiede, 2008), or making concept maps about the text, which was investigated with 12-13 year old children (Redford, Thiede, Wiley, & Griffin, 2012), were found to improve JOL accuracy even when implemented during or directly after studying expository text. Like delayed generation strategies, these immediate generation strategies help students to get access to their situation model. However, when using immediate generation strategies students can access cues about the quality of their situation model while constructing it (Thiede et al., 2009). Therefore, JOLs following immediate generation strategies can also help students to make more accurate JOLs.

For improving the accuracy of *retrospective* monitoring judgments about items or texts, providing students with a standard of the correct answers has proven effective. Students who had to self-score their own recall test performance on key concepts they studied, assigning themselves either no credit, full credit, or partial credit, tended to overestimate their own performance. However, when they were provided with a standard of the correct definitions of the key concepts, their self-scores judgments became more accurate (Lipko et al., 2009; Rawson & Dunlosky, 2007).

To summarize, monitoring accuracy when studying items such as word pairs or learning from expository text is generally low but prospective monitoring accuracy can be improved by means of delaying JOLs (items) or by means of generation strategies (texts) and retrospective monitoring accuracy can be improved by means of standards for self-scoring. As mentioned above, much less research has been conducted on how to improve monitoring accuracy when learning problem-solving tasks.

Monitoring Learning to Solve Problems from Worked Examples

Very little is known about JOLs when learning to solve problems, even though problem-solving tasks play an important role in education, for instance in primary school subjects like arithmetic and in secondary school subjects such as science, biology, economics, or math. In such domains, well-structured problems are commonly used 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 and efficient way of acquiring problem-solving skills when students have little or no prior knowledge of a task, is by studying worked examples, which provide a step-by-step worked-out solution procedure to a problem (for reviews see Atkinson, Derry, Renkl, & Wortham, 2000; Renkl, 2011; Sweller, Van Merriënboer, & Paas, 1998; Van Gog & Rummel, 2010). Studying worked examples is usually more effective for novice students than engaging in problem solving, because they do not have to resort to weak problem-solving strategies (e.g., trial and error). Novice students typically use weak problem-solving strategies, because they lack knowledge of effective problem-solving procedures (Sweller et al., 1998). Students who use these strategies may eventually solve the problems, but this problem-solving activity requires so much of their working memory capacity that no resources are left for learning, that is, acquiring the knowledge needed to solve similar problems in the future (Sweller et al., 1998). Because a worked example provides a step-by-step worked-out solution for learners to study, the high working memory load imposed by ineffective problem-solving strategies is eliminated. Instead, learners can devote all available working memory capacity to studying how to solve a problem, which fosters learning, that is, the construction of a schema of the solution procedure in long-term memory that can guide future attempts at solving the same type of problem.

Just like when learning word pairs and text, it is important to be able to monitor and regulate one's own learning process accurately when learning to solve problems. Nevertheless, there are only a few studies that have investigated monitoring judgments in problem-solving tasks. For example, feelings of knowing (FOKs judgments) have been investigated in insight and non-insight problem-solving tasks (Metcalfe, 1986; Metcalfe & Wiebe, 1987) and confidence judgments have been investigated before and after solving mathematical problems (Boekaerts & Rozendaal, 2010) and after multiple-choice problems (Mitchum & Kelley, 2010).

The only problems in which JOLs were investigated are chess problems (De Bruin, Rikers, & Schmidt, 2005; 2007), which differ from the well-structured problems that students typically encounter in primary and secondary education. University students learned to play a chess end game and students who had to provide JOLs showed better self-regulatory behavior in their move selections for restudy compared to students who did not provide JOLs. However, no differences in learning performance were found (De Bruin et al., 2005). Moreover, De Bruin et al. did not investigate effects of timing of JOLs (immediate vs. delayed) on monitoring accuracy.

As for retrospective monitoring, it has been found with educationally relevant problem-solving tasks (in biology) that monitoring accuracy could be improved by means of training self-assessment skills (Kostons et al., 2012) and that training both self-assessment (monitoring) and task selection (control) skills, resulted in better learning outcomes after self-regulated learning. Since prospective monitoring can be based on cues from past performance (Finn & Metcalfe, 2008; Koriat, 1997), retrospective monitoring judgments (i.e., self-assessment) about one's own performance could potentially inform prospective monitoring (i.e., JOLs). Therefore, improving self-assessment skills might also improve JOL accuracy.

In sum, it remains unclear: a) whether students are able to make accurate prospective judgments, like JOLs, and use these to control their learning process when solving or learning to solve well-structured problems, b) whether delaying JOLs about problem-solving tasks leads to higher accuracy, c) whether generation strategies are also effective for improving JOL accuracy when learning to solve problems by means of worked example study, and d) whether retrospective monitoring (self-assessment) of problem-solving can also be improved by using standards of the correct answers, and whether improvement in retrospective monitoring would also lead to improvement in prospective monitoring accuracy.

Overview of the Dissertation

This dissertation contains five empirical, experimental studies investigating monitoring and regulation accuracy when learning to solve problems in primary and secondary education. Analogous to the strategies that are known to improve monitoring and

regulation accuracy with word pairs and texts, a number of strategies to improve monitoring and regulation accuracy when learning to solve problems were investigated.

Chapter 2 presents a study exploring whether primary school children can monitor the complexity of arithmetic problem-solving tasks and investigating whether there are any differences in the accuracy of children's immediate and delayed JOLs about arithmetic problems. Chapters 3 and 4 include conditions in which the effect of delaying JOLs after worked example study is investigated.

The main focus in Chapters 3, 4, and 5, lies on whether monitoring when learning to solve problems by means of studying worked examples can be improved by generation strategies. As with learning from text, making JOLs about worked examples requires students to judge their *comprehension* rather than literal memory, if they are to correctly predict their future test performance. A major difference, however, is that they have to judge their comprehension of a *procedure*. That is, they have to evaluate the quality of the cognitive schema they constructed for how to solve this type of problem, in order to judge their ability to solve similar problems on a future test. When learning from text, generation strategies that focused participants' attention on the quality of their cognitive representation of the text were found to improve monitoring accuracy. As such, generation strategies 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.

The studies in Chapters 3 and 4 investigated whether problem solving after worked example study would be an effective generation strategy for children in primary school (Chapter 3) and adolescents in secondary education (Chapter 4). In both studies timing of JOLs and practice problems were varied in order to study whether this affected monitoring and regulation accuracy (cf. immediate vs. delayed generation strategies).

Chapter 5 describes an investigation into whether completing partially worked-out examples would be an effective immediate generation strategy to improve monitoring and regulation accuracy when learning to solve biology problems in secondary education.

In Chapter 6, two experiments are presented that investigated whether self-assessments (i.e., retrospective judgments) of practice problems could be improved by means of training or by means of providing students with standards of the correct solution procedures. Furthermore, it was investigated whether students used their self-assessments

of practice problem performance for making JOLs (i.e., prospective judgments), in which case potential improvements in self-assessment accuracy might also lead to improvements in JOL accuracy. The first experiment in Chapter 6 investigated whether training students in how to self-assess their performance on a practice problem after worked example study would improve self-assessment accuracy, and whether this, in turn, would lead to more accurate JOLs and regulation choices. The second experiment in Chapter 6 investigated whether self-assessment training and using standards of the correct solution, or both, would improve JOL and regulation accuracy.

Finally Chapter 7 provides a general discussion of the findings presented in this dissertation.

Chapter 2

Accuracy of Primary School Children's Immediate and Delayed Judgments of Learning about Problem-solving Tasks¹

¹ This chapter is submitted for publication as Baars, M., Van Gog, T., De Bruin, A., & Paas, F. (2013a). *Accuracy of primary school children's immediate and delayed judgments of learning about problem solving tasks*.

Accuracy of Primary School Children's Immediate and Delayed Judgments of Learning about Problem-solving Tasks

Research on monitoring learning and comprehension of study materials has mainly focused on word pairs and text. This study investigated whether children in grade 3 could differentiate in their Judgments of Learning (JOLs) between problem-solving tasks that varied in complexity, and whether those judgments made immediately or delayed would differ in accuracy. Participants engaged in solving four arithmetic problems, rated mental effort invested in each problem, gave either immediate or delayed JOLs, and completed a test containing isomorphic problems. The negative correlation that was found between invested mental effort and JOLs suggested that children's JOLs are sensitive to differences in complexity of the problem-solving tasks. Furthermore, results on the relative and absolute accuracy of JOLs showed that immediate JOLs were numerically higher than delayed JOLs, and relative accuracy of immediate JOLs was moderately accurate, whereas delayed JOLs were not accurate.

Research has shown that monitoring accuracy, that is, accuracy of students' judgments of what information they have or have not yet learned, plays an important role in self-regulated learning. When these monitoring judgments are not accurate, students will not be able to make optimal study choices, for example about how they should allocate their study time and what information they need to restudy (Dunlosky & Lipko, 2007; Metcalfe, 2009). Research on ways of enhancing the accuracy of students' monitoring judgments has mainly focused on study materials consisting of paired associates or short expository texts (for reviews, see Rhodes & Tauber, 2011; Thiede, Griffin, Wiley, & Redford, 2009). Much less is known about monitoring judgments regarding the kind of procedural problem-solving tasks typically seen in important school subject domains such as math or science (see Efkides, 2002).

There are many different kinds of problem-solving tasks; they vary from insight problems to well-structured transformation problems that have a clearly defined goal and solution procedure, to ill-structured problems that do not have a well-defined goal or solution procedure. Well-structured problems, such as math and biology problems encountered in primary and secondary education, consist of a well-defined initial state, a known goal state, and can be solved using a constrained set of logical operators (Jonassen, 2011). Even though monitoring one's own performance and being able to regulate further learning is just as important in this domain as it is in language learning, few studies have

investigated these issues in the domain of problem solving. Here, we take a first step toward investigating whether primary school children differentiate in their monitoring judgments between math problem-solving tasks that differ in complexity, and by exploring the accuracy of immediate and delayed monitoring judgments.

Metacognition and Self-Regulated Learning

Metacognition involves knowledge, monitoring, and control of a cognitive process, such as learning (Flavell, 1979; Serra & Metcalfe, 2009). Metacognition is held to play an important role in learning and especially self-regulated learning. Research has shown that when metacognitive knowledge, monitoring, and control are adequate, learning is enhanced (Azevedo & Cromley, 2004; Kornell & Metcalfe, 2006; Metcalfe, 2009; Thiede, Anderson, & Therriault, 2003). Monitoring involves judging how well information has been learned, and is especially important for self-regulated learning as it affects subsequent control (or regulation) of the learning process. For instance, research has shown that people tend to study longer on those items which they think they have not learned well (Metcalfe, 2009), and more accurate monitoring judgments have been found to lead to more accurate restudy choices and better final test performance (Thiede et al., 2003).

Judgments of Learning (JOLs) are probably the most widely used monitoring judgments. JOLs require participants to either predict their memory for items on a future test (e.g., Nelson & Dunlosky, 1991) or to rate their comprehension of items (e.g., Thiede et al., 2003) during or after the learning phase and prior to taking that test.

In typical studies on monitoring accuracy using JOLs (see e.g., Anderson & Thiede, 2008; Dunlosky & Lipko, 2007; Koriat, Ackerman, Lockl, & Schneider, 2009a; 2009b; Metcalfe & Finn, 2008; Nelson & Dunlosky, 1991; Thiede et al., 2003; Thiede, Dunlosky, Griffin, & Wiley, 2005), participants study items such as word pairs or texts, and provide JOLs by either predicting their future recall of each of the word pairs (e.g., Nelson & Dunlosky, 1991) or rating their comprehension of each of the texts (e.g., Maki, 1998b; Thiede et al., 2003). The relative accuracy of a judgment is then established by computing a Goodman-Kruskal gamma correlation between judgments and test performance, which can vary between -1 and +1; a gamma close to +1 would mean that criterion test performance on items that received higher recall/comprehension judgments was indeed better than performance on items that received lower judgments (Nelson, 1984). Relative accuracy

measured by the gamma correlation indicates whether students can discriminate among items (i.e., whether items that get a higher JOL are indeed performed better on a test than items getting a lower JOL). Next to relative accuracy, monitoring accuracy can also be determined using absolute measures (Mengelkamp & Bannert, 2010; Schraw, 2009), in which the judgment for an item is compared with the performance on that item.

Given the important role that accurate monitoring was considered (and later established; e.g., Thiede et al., 2003) to play for effective self-regulation, it was problematic that early studies on word pairs and text often found monitoring accuracy to be quite low (e.g., Glenberg, Sanocki, Epstein, & Morris, 1987; Maki, 1998a; Nelson, Gerler, & Narens, 1984; Vesonder & Voss, 1985), and consequently, a lot of subsequent research has focused on finding ways to improve monitoring accuracy.

Improving Monitoring Accuracy

In a well-known study with word pairs, Nelson and Dunlosky (1991) established that when asking JOLs not immediately after studying a word-pair but after studying all word pairs, relative accuracy was higher. They called this the delayed-JOL effect, which they explained based on memory systems involved in making JOLs. Note that with word pairs, participants study pairs of cue and target words and they are given the cue word when they are asked to provide a JOL, as well as on the test when they are asked to provide the appropriate target word. As a consequence, immediate JOLs might be less accurate because they are based on retrieval from both short-term memory (STM) and long-term memory (LTM), whereas delayed JOLs can only be based on LTM because STM traces of the item are no longer available. This more closely resembles the memory retrieval situation at the test and thus, delayed JOLs lead to more accurate judgments (Nelson & Dunlosky, 1991). Even though most of these studies focused on adults, Schneider, Visé, Lockl, and Nelson (2000) found the delayed-JOL effect also in primary school children from kindergarten, second grade, and fourth grade when learning word-picture pairs. Similar results were found by Koriat et al. (2009a; 2009b) with second and fourth grade primary school children learning word pairs. In their meta-analytic review, Rhodes and Tauber (2011) showed the robustness of this delayed-JOL effect on relative accuracy with paired associates, category exemplars, sentences, and single words. Effect sizes for prospective memory items and

information from videos were smaller. Also, the delayed-JOL effect was found to be less robust for children compared to other age groups.

Interestingly, however, early studies using texts suggested that this delayed-JOL effect would not apply to materials that are more complex than word pairs (Maki, 1998a). Maki (1998b) investigated text JOL accuracy under four conditions: (1) providing immediate JOLs and taking an immediate test after each text, (2) providing immediate JOLs but taking delayed tests after all texts were read and judged, (3) providing delayed JOLs and taking tests directly following the JOL about each text and (4) providing delayed JOLs and taking delayed tests after all JOLs were provided. The second condition is comparable to the immediate JOL and the fourth to the delayed JOL condition in the study by Nelson and Dunlosky (1991), but in contrast to their study with word pairs, Maki's (1998b) data showed no difference in accuracy between those conditions (see Thiede et al.'s, 2009, review for more studies that failed to find the delayed JOL effect with texts; e.g., Baker & Dunlosky, 2006; Dunlosky, Rawson, & Middleton, 2005).

However, as noted by Thiede et al. (2003), studies on monitoring accuracy with relatively simple tasks such as word pairs are different from studies on monitoring accuracy with more complex materials like texts. Task complexity can be defined in terms of element interactivity: the higher the number of interacting information elements that a learner has to relate and keep active in working memory when performing a task, the higher the complexity of that task and the higher the cognitive load it imposes (Sweller, 2010; Sweller, Van Merriënboer, & Paas, 1998). While learning word lists or word pairs requires memorization of isolated elements, learning texts requires building a mental representation consisting of multiple interacting elements. When providing a JOL about a text, then, learners have to judge the quality of their mental representation of the text, which differs markedly from JOLs about word pairs, which require learners to judge their ability to retrieve the learned information literally from memory. While simply delaying a judgment may provide a better cue for predicting memory of word pairs, it may not be sufficient for predicting the quality of the mental representation of a text.

Indeed, subsequent studies have shown that when participants are provided with instructions that focus their attention on the right cues (i.e., cues regarding the quality of their mental representation of the text) prior to making a comprehension judgment, their monitoring accuracy was enhanced. For example, generating keywords (Thiede et al., 2005)

or making a summary (Anderson & Thiede, 2008) at a delay (i.e., after studying several texts) improved the relative accuracy of subsequent JOLs (Thiede et al., 2009). Similarly, immediate instructional strategies (i.e., after each text) such as rereading or self-explaining the text (Griffin, Wiley, & Thiede, 2008) or making a concept map of the text (Thiede, Griffin, Wiley, & Anderson, 2010) enhanced relative accuracy of immediate JOLs. What these different instructional strategies have in common, is that they provide learners with better diagnostic cues to assess their understanding or predict their test performance, by focusing their attention on their situation model (i.e., mental representation) of the text (Thiede et al., 2009). Because the situation model is the result of learners' understanding of the text and influences their test performance (Kintsch, 1998), JOLs should be based on cues from the situation model in order to be accurate (Rawson, Dunlosky, & Thiede, 2000; Wiley, Griffin, & Thiede, 2005).

Again, most of the studies on improving accuracy of text comprehension judgments have focused on young adult learners, but De Bruin, Thiede, Camp, and Redford (2011) recently showed that generating keywords also enhanced monitoring accuracy for children in primary and secondary education. While there is a considerable amount of research on improving monitoring accuracy in language tasks such as word pairs or texts (in their review, Thiede et al., 2009, list 39 studies with such tasks), there is hardly any research with problem-solving tasks. As we will argue below, there are similarities between problem-solving tasks and texts in terms of monitoring, in that problem-solving tasks require students to judge their *comprehension* of a problem-solving *procedure* stored in a mental representation (i.e., cognitive schema). However, there are also important differences.

Judgments of Learning (JOLs) about Problem-solving Tasks

To the best of our knowledge, there are only two studies that have investigated JOLs in problem-solving tasks (De Bruin, Rikers, & Schmidt, 2005; 2007). However, De Bruin et al. used a type of problem (i.e., playing a chess endgame) that is very different from the kind of procedural problems encountered in educational domains such as math and science. There are some studies that have investigated confidence judgments made before and after completion of mathematical problems (Boekaerts & Rozendaal, 2010) and after multiple choice problems (Mitchum & Kelley, 2010). Interestingly, Boekaerts and

Rozendaal found that absolute accuracy of confidence judgments after solving a computational problem was higher than absolute accuracy of confidence judgments that were given before solving the problem. They suggest that children were able to pick up cues from the problem-solving process which allowed them to make a more accurate assessment of their performance. However, in these studies it was not investigated what the effects of timing of monitoring judgments (immediate vs. delayed) are on monitoring accuracy. Furthermore, it is unclear whether children are able to monitor changes in complexity of problem-solving tasks.

Therefore, the first goal of the present study was to explore whether children can monitor complexity of math problem-solving tasks. When confronted with problems that differ in complexity, are children able to monitor that they are more likely to comprehend problems they could solve easily and less likely to comprehend problems they found more difficult to solve? Task complexity can partly be determined objectively, by looking at the number of interacting elements a task contains. However, task complexity is also partially subjective. For instance, a learner's prior knowledge affects the number of interacting elements a task contains because elements that have been combined into a schema can be treated as a single element in working memory. Therefore, element interactivity is reduced for higher prior knowledge learners and therefore the cognitive load a task imposes will be lower for them than for learners with less prior knowledge (Kalyuga, 2007; Kalyuga & Sweller, 2004). Consequently, there will be individual differences in experienced cognitive load within objectively identified levels of task complexity. The question we will explore here is whether experienced cognitive load (as measured by ratings of invested mental effort; see Paas, Tuovinen, Tabbers, & Van Gerven, 2003) is negatively related to JOLs (i.e., the higher the experienced load, the lower the comprehension judgment).

The second goal of this study was to explore whether the timing of JOLs about problem-solving tasks (i.e., immediate vs. delayed) affects judgment accuracy. As mentioned above, like texts and unlike word pairs, a JOL about problem solving should not concern an evaluation of the ability to literally retrieve a piece of information from memory on the test (such as the number constituting the correct solution on a particular problem). Rather, it should be an evaluation of the ability to correctly perform a problem-solving *procedure* required to solve that *type* of problem. In other words, students have to judge

their cognitive schemas of solution procedures for certain problem types (see e.g., Sweller et al., 1998, for a discussion of problem-solving schemas).

As such, JOLs about problem solving seem more similar to JOLs about texts than JOLs about word pairs, as both require an assessment of the extent to which a mental representation (i.e., a problem schema or a situation model) has been acquired, in order to be an accurate predictor of test performance. In this case, one would expect that as with texts (Maki, 1998b), there should be no effect of timing on judgment accuracy. On the other hand, the need to monitor one's understanding of a step-by-step solution procedure and one's ability to actually generate a specific solution by applying that general procedure, makes monitoring of problem-solving tasks different from monitoring expository texts (where understanding the gist is sufficient). Thus, an important difference between problems and texts, is that the act of problem solving itself might provide important and immediate feedback to students regarding the quality of their problem schema (i.e., with a high quality schema, the solution procedure should be readily accessible from memory, easily implemented, and evoking feelings of success), and might thus focus their attention on accurate cues for making their judgment (Boekaerts & Rozendaal, 2010).

Griffin, Jee, and Wiley (2009) describe a model of different routes to making monitoring judgments about texts: 1) making a predictive judgment of test performance based on cues that are independent of the text representation (e.g., interest) and can be available before, during or after reading the text, which is called the '*heuristic route*'; 2) making a predictive judgment of test performance based on cues related to the representation of the text after reading it (e.g., ability to summarize), which is called the '*representation-based route*'; and 3) making a postdiction judgment of (future) test performance based on cues from performance on a test that was just completed, which is called the '*postdiction route*'. An example of the latter is the finding by Finn and Metcalfe (2007) that participants who learned word pairs and then took a test, used cues from their performance on that test (i.e., postdiction) to predict their future test performance (i.e., Memory for Past Test heuristic). In the case of problem solving, one may expect immediate JOLs about problem-solving tasks to be more accurate than delayed JOLs, because the act of solving (or attempting to solve) a problem provides participants with cues regarding their performance that will be most salient when making an immediate judgment.

Method

Participants and Design

Participants were 76 Dutch primary education students in grade three (8-10 years old, 39 boys and 37 girls). Only students with scores of B, C, or D on a standardized math test taken shortly before the study were included. This excludes the very very low [E] or very very high [A] math ability students because the learning materials used in this study presumably were too complex or too easy for these students to find sufficient variation in JOLs and test performance. Participants in each classroom were randomly assigned to one of the conditions prior to the experiment, resulting in 35 participants in the immediate JOLs condition (17 boys and 18 girls) and 41 participants in the delayed JOLs condition (22 boys and 19 girls).

Materials

All materials were paper-based.

Problems. In each phase of the experiment (pretest, practice, and posttest), four arithmetic problems were used which teachers considered to differ in complexity, one of each of the following types (in order of increasing complexity): addition without carrying (e.g., $414 + 135 + 250$), addition with carrying (e.g., $119 + 313 + 238$), subtraction with borrowing tens (e.g., $676 - 139$) and subtraction with borrowing tens and hundreds (e.g., $634 - 497$). The problems in the different phases were isomorphic (i.e., equivalent structure, but different numbers). According to the teachers, the children were familiar with the procedures used in addition and subtraction with borrowing tens with three digits numbers but had not yet practiced with addition with carrying and subtraction with borrowing tens and hundreds. Test performance was judged as either incorrect (0) or correct (1).

Mental effort ratings. Directly after each problem, students rated the amount of mental effort they invested in attempting to solve that problem on a 5-point rating scale, ranging from (1) very low mental effort, to (5) very high mental effort (cf. Paas, 1992). The original 9-point rating scale developed by Paas (1992) was adjusted to a 5-point rating scale to make it easier to understand and use for primary school children (cf. Van Loon-Hillen,

Van Gog, & Brand-Gruwel, 2012; for a review of other varieties of mental effort scales used, see Van Gog & Paas, 2008) and to make the use of this scale comparable to the JOL scale. The mental effort rating was prompted by the question: *How much effort did you invest in solving this problem?*

JOLs. JOLs were provided on a 5-point rating scale (cf. Thiede et al., 2003), asking students to rate their comprehension of this type of problem. The JOL was prompted with the title of the arithmetic problem in the question, for example: *How well do you think you understood the problem about subtracting with borrowing tens?* The answer scale that followed ranged from 1 (very poorly) to 5 (very well).

Procedure

This experiment was run in small group sessions ranging from 10 to 15 students in classrooms at participants' schools. All participants were told that they would have to solve arithmetic problems on paper and rate their invested mental effort and comprehension of the problems. Before the actual experiment started, both the mental effort and JOL rating scales were explained by the experimenter and practiced with one example problem. It was explained that they had two minutes to solve each 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 to the next one before this time had passed, and that the experiment leader would tell them when to start and stop working on solving each problem. Participants first completed the pretest. Then, in the practice phase, they engaged in solving four arithmetic problems, rating their invested mental effort after completing each problem. Depending on their assigned condition, they provided a JOL about each problem either immediately after each problem (immediate JOL condition) or after all four problems (delayed JOL condition). Then they completed the posttest.

Data Analysis

Relative accuracy. Relative monitoring accuracy was measured with the Goodman-Kruskal gamma correlation between JOLs and performance on the posttest, in line with previous studies (e.g., Dunlosky et al., 2005; Maki, 1998b; Nelson & Dunlosky, 1991; Thiede et al., 2003; 2005). The gamma correlation shows if participants are able to discriminate between problems on which they perform poorly and problems on which they perform well, that is, whether the problem types that were given a high JOL were also the

problem types participants performed well on the test (Maki, Shields, Wheeler, & Zaccilli, 2005). Gamma correlations between JOLs and performance on the posttest were calculated for each individual participant, and the closer to 1, the higher the monitoring accuracy. Twenty-four participants had indeterminate gamma correlations due to invariance in JOLs or performance. The mean of the intra-individual gamma correlations was calculated for each condition (immediate JOLs: $n = 26$; delayed JOLs $n = 26$).

Absolute accuracy. Because the JOL scores and test performance scores were not on the same scale, they cannot simply be subtracted in order to calculate absolute accuracy. We therefore developed a gradual measure of absolute accuracy that varies between 0 and 1, based on each possible combination of JOL (1-5) and test performance (0 or 1). The scoring system is shown in Table 1. As can be inferred from the Table, lower JOLs combined with a test performance of 0 resulted in higher absolute accuracy, whereas lower JOLs combined with a test performance of 1 resulted in lower absolute accuracy; similarly, higher JOLs combined with a test performance of 0 resulted in lower accuracy, whereas higher JOLs combined with a test performance of 1 resulted in higher accuracy. Mean absolute accuracy over the four problem-solving tasks was calculated. We could not calculate absolute accuracy for two participants because they did not fill out all JOLs or test items. The mean absolute accuracy was calculated for each condition (immediate JOLs: $n = 34$; delayed JOLs $n = 40$).

Table 1

Absolute monitoring accuracy scoring system

Test: JOL:	0	1
1	1	0
2	0.75	0.25
3	0.50	0.50
4	0.25	0.75
5	0	1

Results

As a check on randomization, the pretest performance data were compared, which -as expected- showed no differences between the Immediate and Delayed JOL Condition, $t(74) = .89, p = .38$. The pretest scores, percentage of correct responses as well as the mean JOLs, and mean mental effort ratings during the practice phase are presented in Table 2.

Monitoring Task Complexity

A repeated measures ANOVA with JOLs as dependent variable, Complexity (4 levels) as within-subjects factor and Condition (Immediate vs. Delayed JOLs) as between-subjects factor, showed a main effect of Complexity, $F(3, 219) = 3.29, p = .02, \eta_p^2 = .04$. Contrasts revealed that JOLs were significantly lower for the fourth level of complexity compared to the first level, $F(1, 73) = 6.89, p = .01, \eta_p^2 = .09$, and the second level, $F(1, 73) = 5.25, p = .03, \eta_p^2 = .07$, but not compared to the third level, $F(1, 73) = 2.32, p = .13, \eta_p^2 = .03$. However, there was no significant main effect of Condition, $F(1, 73) = 2.81, p = .10, \eta_p^2 = .04$, nor an interaction effect, $F(3, 219) = 1.38, p = .25, \eta_p^2 = .02$.

Table 2

Percentages of correct performance, mean subjective mental effort ratings (*range*: 1-5), and mean JOLs (*range*: 1-5) during the learning phase for the different problem categories, which differed in complexity (1 = lowest; 4 = highest).

Complexity levels	Immediate JOL condition				Delayed JOL condition			
	Pretest (SD)	% correct	Mean mental effort (SD)	Mean JOLs (SD)	Pretest (SD)	% correct	Mean mental effort (SD)	Mean JOLs (SD)
1	0.86 (0.32)	85.2	2.23 (1.19)	4.29 (.83)	0.85 (0.37)	85.4	2.03 (1.14)	3.89 (1.05)
2	0.63 (0.49)	62.9	2.09 (1.11)	4.29 (.94)	0.71 (0.46)	80.5	2.10 (1.17)	3.66 (1.13)
3	0.57 (0.50)	40.0	2.23 (1.35)	3.97 (1.15)	0.39 (0.49)	43.9	2.29 (1.17)	3.78 (1.19)
4	0.37 (0.49)	34.4	2.71 (1.41)	3.74 (1.20)	0.27 (0.45)	34.1	2.44 (1.37)	3.59 (1.38)

A repeated measures ANOVA with mental effort ratings as dependent variable, Complexity (4 levels) as within-subjects factor and Condition (Immediate vs. Delayed JOLs) as between-subjects factor, showed that mental effort ratings increased when problem complexity increased, $F(3, 207) = 4.60, p = .01, \eta_p^2 = .06$. Contrasts revealed that mental effort ratings were significantly higher for the fourth level of complexity compared to the first level, $F(1, 69) = 8.55, p = .01, \eta_p^2 = .11$, the second level, $F(1, 69) = 13.76, p < .001, \eta_p^2 = .17$, and the third level, $F(1, 69) = 4.91, p = .03, \eta_p^2 = .07$. As expected, there was no main effect of Condition, $F(1, 69) < 1$, nor an interaction, $F(3, 207) < 1$.

A repeated measures ANOVA with performance in the learning phase as dependent variable, Complexity (4 levels) as within factor and Condition (Immediate vs. Delayed JOLs) as between factor, showed that performance decreased with increasing complexity, $F(3, 222) = 24.37, p < .001, \eta_p^2 = .25$. Contrasts revealed that performance was significantly lower for the fourth level of complexity compared to the first level, $F(1, 74) =$

64.05, $p < .001$, $\eta_p^2 = .47$, and the second level, $F(1, 74) = 26.73, p < .001, \eta_p^2 = .27$, but not compared to the third level, $F(1, 74) = 1.22, p = .27, \eta_p^2 = .02$. As expected, there was no main effect of Condition, $F(1, 74) < 1$, nor an interaction effect, $F(3, 222) < 1$.

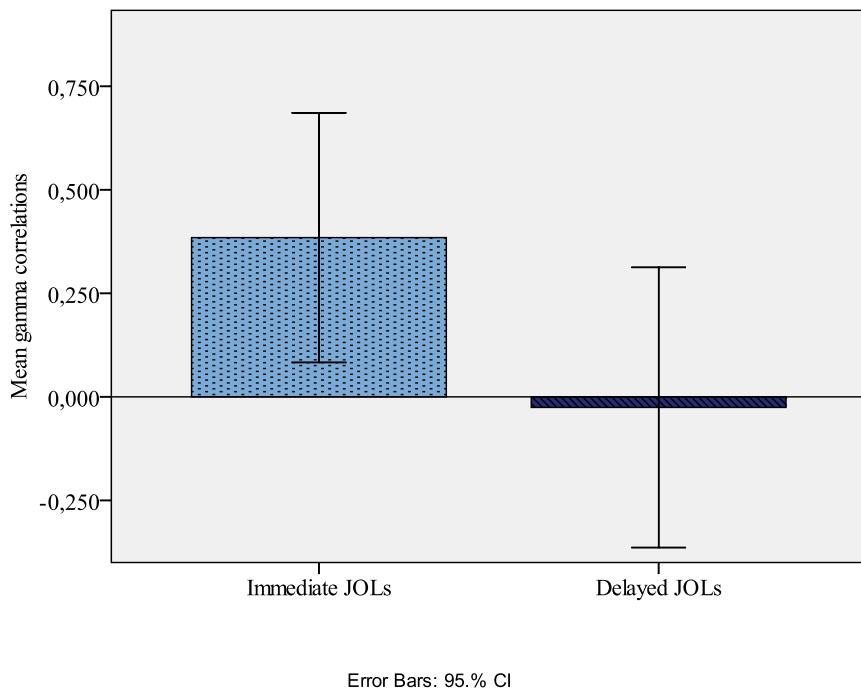
Moreover, in line with our hypothesis, mental effort showed a significant negative correlation with JOLs. That is, when invested mental effort was high, students judged their comprehension to be low. This correlation did not differ significantly between the Delayed JOLs condition, $r = -.59, t(37) = -4.28, p < .01$, and the Immediate JOLs condition, $r = -.67, t(34) = -5.12, p < .01$.

Immediate vs. Delayed JOL Accuracy

Relative monitoring accuracy. The mean gamma correlation of the Immediate JOL condition differed significantly from zero, $t(25) = 2.63, p = .02$, whereas that of the Delayed JOL condition did not differ significantly from zero, $t(25) = -.16, p = .88$. That is, in contrast to delayed JOLs, immediate JOLs were more accurate than chance. A t-test showed a marginally significant difference in gamma correlations between the conditions, $t(50) = 1.86, p = .068$, Cohen's $d = 0.52$ (medium effect size). As Figure 1 shows, monitoring accuracy was higher in the Immediate JOLs condition ($M = .38, SD = .75$) than in the Delayed JOLs condition ($M = -.03, SD = .84$).

Figure 1

Mean monitoring accuracy presented by group. Error bars represent standard errors of the mean.



Absolute monitoring accuracy. On absolute accuracy, there was a trend in the same direction, but a t-test showed that the difference between the two conditions was not significant (Immediate: $M = 0.62$, $SD = 0.17$, Delayed: $M = 0.56$, $SD = 0.18$; $t(72) = 1.56$, $p = .123$).

Discussion

This study explored whether 3rd graders' monitoring judgments about problem-solving tasks would be sensitive to variations in task complexity as reflected in invested mental effort, as well as whether the accuracy of immediate and delayed JOLs would differ. Our results suggest that children indeed seem to be able to monitor the complexity of the

problems solving tasks. That is, with increasing complexity of the problem-solving tasks, performance decreased, and subjective ratings of mental effort increased while JOLs decreased. Furthermore, there was a significant negative correlation between mental effort and JOLs, indicating that the higher the mental effort invested, the lower the JOLs were. So, it seems that students may have used the mental effort they invested in solving a problem as a cue to give a judgment about their comprehension of the problem.

Nevertheless, relative accuracy of JOLs, which shows the ability to discriminate between tasks, did not seem to be very high. The mean gamma correlation in the immediate JOLs condition differed significantly from zero but this was not the case for the delayed JOLs condition. That is, immediate JOLs were moderately accurate whereas delayed JOLs were not accurate. Furthermore, a marginally significant difference in relative monitoring accuracy between the immediate JOLs and delayed JOLs condition was found, in favor of the immediate JOLs condition. The absolute accuracy, however, even though it showed a numerical trend in the same direction with accuracy of immediate JOLs being higher, did not differ significantly between immediate and delayed JOL conditions ($p = .12$).

The direction of these findings is surprisingly different from findings regarding JOLs about language tasks such as word pairs and texts. For word pairs, the delayed-JOL effect shows higher accuracy of delayed JOLs compared to immediate JOLs (Nelson & Dunlosky, 1991), even though Rhodes and Tauber (2011) showed that the effect of delayed JOLs was smaller for children, it was still present. For texts, this delayed-JOL effect was not found when no additional instructions were added, in fact, no differences between the immediate and delayed JOL conditions were found. Maki (1998b) studied immediate and delayed JOLs about texts, and did not give any 'generation instructions', and found gamma correlations of .05 for immediate JOLs and .02 for delayed JOLs. This is much lower than the average gamma correlations reported by Thiede et al. (2005) who found a gamma correlation of .29 for immediate JOLs after keyword generation, Dunlosky and Lipko (2007), who reported an average gamma correlation of .27 across different conditions in laboratory studies, and Thiede et al. (2009) who reported an average gamma correlation of .27 for immediate JOLs in different conditions. However, these averages included conditions that were designed to improve JOL accuracy by 'generation instructions' (e.g., generating keywords), rather than only varying the timing of JOLs. In our study, the gamma correlation in the immediate judgment condition was still moderate ($M = .38$), but much

higher than immediate JOLs about texts without generation instructions (i.e., the .05 reported by Maki).

A possible explanation for the (numerical) difference in accuracy between immediate and delayed JOLs might lie in the cognitive processes associated with problem-solving tasks. When attempting to solve a problem, a problem schema becomes activated – if available. If an immediate JOL has to be made, the learner should be able to judge relatively easy whether or not a problem schema was available from his or her ability to solve the problem; the problem-solving process itself provides direct feedback to a learner (e.g., effort required, experiences of success or failure) on which JOLs can be based (i.e., what Griffin et al., 2009, call the 'postdiction route'), but the saliency of such cues will be diminished after a delay. Moreover, the JOL prompt used in this study did not explicitly ask students to predict their future test performance but asked them about their comprehension of the task (cf. Thiede et al., 2003; 2005 for texts), which makes it even more likely that participants based their immediate JOLs on postdiction about problem solving performance. Because our study did not involve instructions on the problem-solving tasks, these postdictions of practice problems could be predictive of performance on the test problems as well.

In sum, if students use their experiences of ease, failure, or success as a cue for monitoring, it is likely that immediate JOLs would make a better distinction between items that are performed well and items that are not performed well than delayed JOLs, because at a delay these experiences may no longer be very salient anymore. Moreover, students were only given the problem category description when making delayed JOLs, which may have made it hard for them to link their experiences to a specific problem. Future research might therefore investigate whether delayed JOLs would be more accurate when students get to see the initial state of the problem again. This was not possible with the type of problems we used, because providing students with the actual problem again would give them the opportunity to start solving the problem again. In fact, this would even be necessary in order to recognize the problem category (i.e., that the necessity of carrying is not immediately apparent for a learner from the problem statement). However, if learners start solving the problem again, they would no longer be making a delayed JOL, but an immediate one. Perhaps the use of another design in which the delay between problem

solving and JOL consists of a filler task instead of solving other problems might be a way for future research to solve this issue of linking delayed JOLs to the right problem.

There are several potential limitations to this study. First, gamma correlations could not be computed for a number of participants due to invariance in scores. This is presumably due to the low number of tasks (four) used in this study, and this problem has also been described in other studies in which a small number of texts (five to six) was used for instance (e.g., Anderson & Thiede, 2008). Therefore, future research should attempt to replicate these findings regarding relative accuracy using more problem-solving tasks. Second, because effort was rated prior to making a JOL in the immediate condition, students may have been primed to use the mental effort they invested in solving a problem as a cue for their JOL. On the other hand, the correlation between effort and JOLs did not differ significantly between the immediate and delayed JOL condition, and the two ratings were not made in close proximity in the delayed condition. Nevertheless, future research could address the question of whether or not JOLs would differ when they are made prior to effort ratings.

In conclusion, despite the bulk of research on accuracy of JOLs about language tasks, to the best of our knowledge, this study was the first to explore JOLs about the kind of procedural problem-solving tasks typically encountered in important school domains such as math and science. The findings that 3rd graders seem able to monitor their comprehension as a function of task complexity and that immediate JOLs seemed somewhat more accurate than delayed JOLs, are interesting and should be followed-up on in future research. Further insights into how students monitor their problem-solving skills in school domains like math or science, could inspire instructional methods or help teachers to improve self-regulated learning in students. Given that the gamma correlations in the immediate JOLs condition, despite being higher than in the delayed JOL condition, were still moderate, there is room for improvement. Moreover, as mentioned in the introduction, accurate monitoring can inform regulation of study and lead to better learning outcomes (Thiede et al., 2003). Future research might investigate whether additional instructional strategies that would allow learners to better judge their schemas, could improve accuracy of both immediate and delayed JOLs, much like instructions to generate keywords (Thiede et al., 2005), summaries (Anderson & Thiede, 2008; Thiede et al., 2009), or concept maps (Thiede et al., 2010) do for texts. Future research could also investigate whether means of

improving post-dictions on practice problems, for instance, by training students to self-assess their performance (cf. Kostons, Van Gog, & Paas, 2012) could improve their accuracy, and how such post-dictions of comprehension of practice problems would relate to predictions of future test performance.

Chapter 3

Effects of Problem Solving after Worked Example Study on Primary School Children's Monitoring Accuracy²

² This chapter was 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. doi: 10.1002/acp.3008

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

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 worked 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.

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 effects of practice and item difficulty, immediate and delayed JOLs seem to be affected differently (Scheck, Meeter, & Nelson, 2004; Scheck & Nelson, 2005). For instance, in the study by Scheck and Nelson (2005) on the underconfidence with repeated practice effect, students studied easy and difficult English-Swahili word pairs, gave an immediate or delayed JOL and took a self-paced recall test in the first study-test cycle which was repeated in a second study-test cycle. For easy word pairs, both immediate and delayed JOLs showed underconfidence in the second study-test cycle. However, for the difficult word pairs in the second study-test cycle, immediate JOLs did *not* show over- or underconfidence, while delayed JOLs resulted in overconfidence. This shows that delayed JOLs are not always more accurate than immediate JOLs with items such as word pairs.

Other studies have also 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.

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, 1998a). 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, Baars, Van Gog, De Bruin, and Paas (2013a) 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 JOLs, 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 (Nieuwstein, 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 Merriënboer, 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 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).

Although studies on JOL accuracy with word pairs showed that delayed JOLs were more accurate (i.e., delayed-JOL effect, Rhodes & Tauber, 2011) and studies on JOL accuracy with text did not find a difference between immediate and delayed JOLs (Maki, 1998a), there are some indications that accuracy of immediate JOLs tends to be higher for problem-solving tasks (Baars et al., 2013a) and our first hypothesis is therefore that immediate JOLs will be more accurate than delayed JOLs when learning to solve problems by studying worked examples (i.e., JOL accuracy in condition 1 > condition 2), that is, judging comprehension of a *procedure* may be more easily done immediately than at a delay.

Second, 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 1.

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 2. 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.

Table 1

Overview of the 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	<i>Filler task</i>	Problem	Problem	<i>Filler task</i>
<i>Filler task</i>	JOL	JOL	<i>Filler task</i>	Problem
<i>Filler task</i>	<i>Filler task</i>	<i>Filler task</i>	JOL	JOL
Restudy choices				
Test				

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) and 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 one point for each correct step (i.e., maximally six 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 (Dunlosky & Thiede, 1998; Thiede & Dunlosky, 1999), which states that items that are more difficult to learn are more often selected for restudy than items that are easier 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.

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

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.

Table 3

Mean JOL (*range*: 0-6), mean bias (*range*: -6-6), mean absolute deviation (*range*: 0-6), restudy accuracy (*range*: 0-6), number of restudy choices (*range*: 0-6), and mean test performance (*range*: 0-6).

Overview of the data

	Immediate JOL <i>N</i> = 26	Delayed JOL <i>N</i> = 27	Immediate problem and immediate JOL <i>N</i> = 29	Immediate problem and delayed JOL <i>N</i> = 28	Delayed problem and immediate JOL <i>N</i> = 25
Mean practice problem performance	-	-	4.34 (1.56)	4.57 (1.50)	3.77 (1.28)
Mean JOL (<i>range</i>: 0-6)	4.37 (1.03)	3.97 (1.35)	3.83 (1.20)	4.16 (1.00)	3.72 (1.85)
Mean bias (<i>range</i>: -6 -6)	0.69 (1.39)	0.78 (1.64)	0.27 (1.37)	0.21 (1.10)	-0.04 (1.51)
Mean absolute deviation (<i>range</i>: 0- 6)	1.98 (0.71)	2.04 (0.70)	2.09 (0.64)	1.90 (0.72)	1.79 (0.73)
Regulation accuracy (<i>range</i>: 0- 6)	0.56 (0.19)	0.54 (0.26)	0.56 (0.20)	0.49 (0.22)	0.62 (0.24)
Number of restudy choices	1.85 (1.22)	2.41(1.81)	2.10 (1.61)	2.18 (1.56)	1.72 (1.57)
Mean test performance (<i>range</i>: 0-6)	3.71 (1.17)	3.15 (1.31)	3.65 (1.06)	4.00 (0.90)	3.77 (1.28)

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 (Baars et al., 2013a), do not seem to apply to JOLs about worked examples. It should be noted though, that the difference with the prior study was a trend only (i.e., not significant) and that the present study 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) or the content of the problem-solving tasks. However, the lack of

difference between immediate and delayed JOLs in worked examples 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, 1998a) unless a generation strategy was added (Griffin et al., 2008; Thiede et al., 2003; Thiede et al., 2009; Thiede & Anderson, 2003). Because worked examples are also text-based, and do not require any generation of solution steps as problem-solving does (which would give cues about actual understanding of the procedure), we feel that it is likely that it is due to the format, but future research should establish this. 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; after practice: 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.

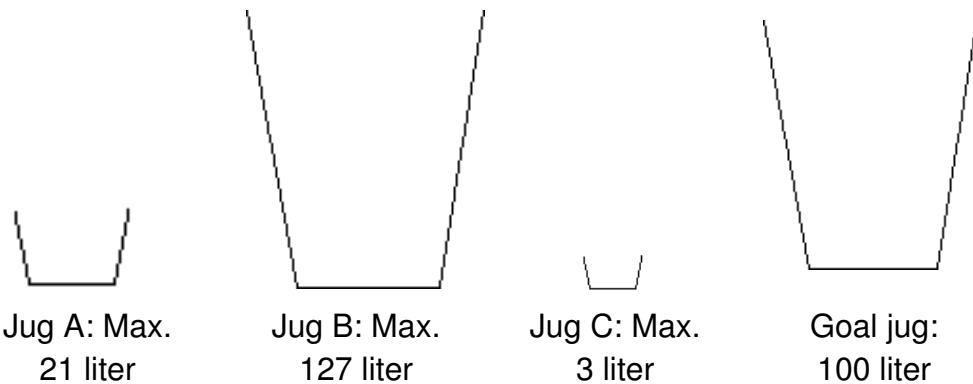
This was the first study to investigate how to improve JOL accuracy when studying worked examples in primary education. However there are some limitations that should be mentioned. First, whereas many studies have used gamma correlations to measure JOL accuracy (e.g., Griffin, Wiley, & Thiede, 2008; 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 calculate 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.

Appendix 1

The goal is to pour 100 liter of water in the goal jug.



Step 1: Jug A can hold 21 liter of water. Jug B can hold 127 liter of water. Jug C can hold 3 liter of water. The aim is to pour 100 liter of water in the goal jug.

Step 2: Fill jug B with 127 liter of water.

$$B = 127$$

Step 3: Pour jug B into Jug A until jug A is full. Jug A contains 21 liter of water. And $127 - 21 = 106$ liter so there is 106 liter left in jug B.

$$B \rightarrow A \text{ so } 127 - 21 = 106$$

Step 4: Pour jug B into jug C until jug C is full. Jug C contains 3 liter of water. And $106 - 3 = 103$ liter so there is 103 liter of water left in jug B.

$$B \rightarrow C \text{ so } 106 - 3 = 103$$

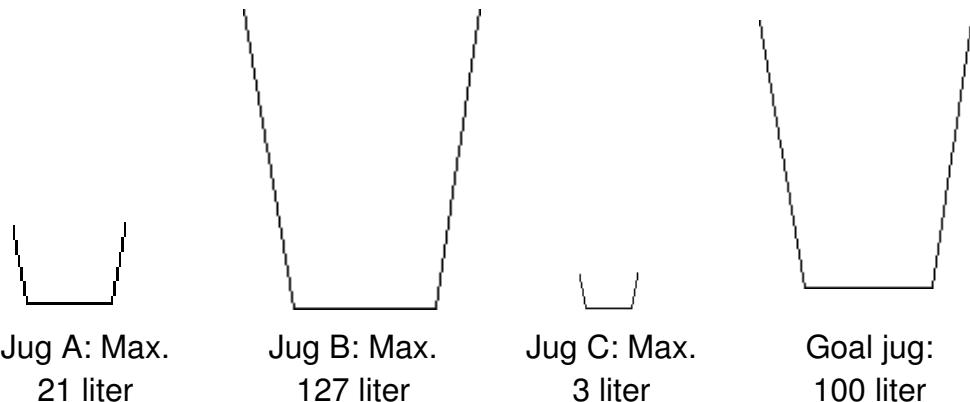
Step 5: Empty jug C and fill it again with 3 liter water from jug B. $103 - 3 = 100$ liter so there is 100 liter of water left in jug B.

$$B \rightarrow C \text{ so } 103 - 3 = 100$$

Step 6: Empty jug B in the goal jug. Now there is 100 liter of water in the goal jug. The goal amount in the goal jug is reached. So you are done with this task.

$$B \rightarrow \text{goal jug}$$

Appendix 2



The goal is to pour 100 liter of water in the goal jug using the other jugs.

Step 1:	
Step 2:	
Step 3:	
Step 4:	
Step 5:	
Step 6:	

Chapter 4

Effects of Problem Solving after Worked Example Study on Monitoring Accuracy in Secondary Education³

³ This chapter is submitted for publication as Baars, M., Van Gog, T., De Bruin, A., & Paas, F. (2013b). *Effects of problem solving after worked example study on secondary school children's monitoring accuracy*.

Effects of Problem Solving after Worked Example Study on Monitoring Accuracy in Secondary Education

Effective self-regulated learning is based on students' ability to monitor their own learning and to allocate study time accordingly. Monitoring accuracy, measured by judgments of learning (JOLs), has been found to be low to moderate, with students often displaying overconfidence (i.e., JOL > actual test performance). When using additional instructional strategies that focus learners' attention on relevant cues for making JOLs, JOL-accuracy is often improved. Primary school children's overconfidence was recently shown to diminish when they practiced problem solving after studying worked examples, but this had no effect on regulation accuracy. The current study aimed to extend this research by investigating whether practicing problem (PP) solving after worked example (WE) 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 PP were provided, and timing of PP provided: 1) WE – JOL, 2) WE – delay – JOL, 3) WE – PP – JOL, 4) WE – PP – delay – JOL, or 5) WE – delay – PP – JOLs. Results showed that practice problems improved absolute accuracy of JOLs as well as regulation accuracy. No differences in final test performance were found.

Students can only learn effectively in a self-regulated way if they have accurate knowledge about their own learning process. Thinking about one's own learning process is called metacognition (Flavell, 1979; Serra & Metcalfe, 2009). Two key metacognitive skills are monitoring and regulating the learning process. Monitoring, that is, keeping track of one's own performance during the learning process, provides the learner with information about the quality of the learning process, which can subsequently be used to regulate further study (Serra & Metcalfe, 2009). That is, 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, that is, lead to better learning outcomes (Thiede, Anderson, & Therriault, 2003; Winne & Hadwin, 1998).

Generally, monitoring is not very accurate, but it can be improved through addition of certain instructional strategies (Dunlosky & Lipko, 2007; Maki, 1998; Thiede, Griffin, Wiley, & Redford, 2009). For instance, research on memory tasks (e.g., word pairs) shows that monitoring accuracy can be improved by delaying monitoring judgments (Rhodes & Tauber, 2011) or, when learning from expository text, by using generation strategies (Thiede et al., 2009) that help students to get an idea of their understanding of the learning material (we will discuss these strategies in more detail later on in this

introduction). 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 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, that is, would improve secondary education students' monitoring and regulation accuracy.

Before describing our approach of the current study in more detail, we will 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

Monitoring accuracy is often measured by asking students to make judgements of learning (JOLs; De Bruin, Thiede, Camp, & Redford, 2011; Lipko et al., 2009; Van Loon, De Bruin, Van Gog, & Van Merriënboer, 2013a; 2013b; Nelson & Dunlosky, 1991; Maki, 1998; Metcalfe & Finn, 2008; Thiede et al., 2003; Thiede, Dunlosky, Griffin, & Wiley, 2005; Thiede et al., 2009). For example, students might be asked to predict their future test performance on the learning materials that were studied (e.g., Nelson & Dunlosky 1991) or might be asked how well they understand the learning material (e.g., Thiede et al., 2003). To determine JOL accuracy, the JOL is compared to the actual performance on a subsequent test. JOL accuracy can be calculated as a relative or as an absolute measure (Mengelkamp & Bannert, 2010; Schraw, 2009). Relative accuracy shows whether students were able to discriminate between the different items that were judged. It is expressed by the Goodman-Kruskal gamma correlation, which correlates the JOLs and test performance pair wise, and ranges between -1, indicating poor accuracy, and +1, indicating perfect accuracy. Absolute accuracy is frequently analyzed by calculating bias or absolute

deviation. Absolute deviation concerns the deviation between JOLs and test performance without a direction showing how well students JOLs were calibrated to their test performance. Bias does take into account the direction of the deviation between JOL and performance and shows whether participants make an under- or overestimation in their JOL.

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 different memory systems involved in making JOLs immediately or at a delay. When making an immediate JOL, the JOL can be based on information on a word pair that is still available from working memory (WM). However, not all information from working memory is also learned, that is, stored in long-term memory (LTM); and on a future test, students have to rely on information available in LTM. When making delayed JOLs, this can only be done based on information available in LTM, which is a more valid source to base JOLs about future test performance on. This delayed-JOL effect has been replicated and 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.

Next to effects of timing of JOLs, 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 judgments were less accurate when item difficulty was higher. Also, difficult items yielded overconfidence whereas easy items yielded underconfidence (Lichtenstein & Fischhoff, 1977; Scheck & Nelson, 2005).

For more complex learning 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 et al., 2005) at a delay after reading a text improved relative JOL accuracy. Generating keywords after reading a text was found to improve relative JOL accuracy for children as well (De Bruin et al., 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, Wiley, & Thiede, 2008) and making concept maps (Thiede, Griffin, Wiley, & Anderson, 2010) directly after reading a text, were also found to improve JOL accuracy.

All of the aforementioned generation strategies are assumed to help students judge their deeper understanding of the text and consequently make better JOLs about the texts, but they probably differ in terms of underlying cognitive mechanisms. Both generating keywords and making summaries were found to improve JOL accuracy most when done at a delay after reading a text. Similar to word pairs, this can be explained by the fact that delayed generation of keywords or summaries rely on retrieving information from LTM, which is more indicative of future test performance than retrieving information from WM. When generating self-explanations or concept maps, however, students can get valid information about their understanding of the text even directly after reading it. Thiede et al. (2009) explained the effect of both immediate and delayed generation strategies by the situation model approach. According to this approach, using information from the situation model helps students to make more accurate JOLs because the deeper level of understanding about the text which students also use during a 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 get access to 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 also generally low, but improvable by generation strategies (De Bruin et al., 2011; Redford, Thiede, Wiley, & Griffin, 2012; Thiede, Redford, Wiley, & Griffin, 2012). De Bruin et al. (2011) found that JOLs were more accurate when 9-10 and 12-13 year olds had to generate keywords at a delay after reading a text, which is in line with the delayed-generation effect found with adults (Thiede et al., 2003). Also, Redford et al. (2012) showed that an immediate generation strategy improved JOL accuracy, that is, making concept maps of a text was found to improve JOL accuracy for 12-13 year olds. So, just like for adults, generation strategies were found to improve JOL accuracy for children. 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 (9 years old; De Bruin et al., 2011). Studies on monitoring and regulation in which children regulated their learning processes by withdrawing answers from a test, also showed that older children (11- 12 years old) were better able to regulate their learning processes than younger children (8- 10 years old; Krebs & Roebers, 2010; Roebers, Schmid, & Roderer, 2009). Because younger children's monitoring was also less optimal compared to older children's, their control processes were possibly less well informed and therefore less effective (Roebers et al., 2009).

What is also clear from this description of studies on improving JOL accuracy in children and adults, is that the majority of these studies have focused on learning items (e.g., word pairs) and learning from expository texts (De Bruin et al., 2011; Rhodes & Tauber, 2011; Thiede, et al., 2009). Only very few studies have investigated JOL accuracy when learning to solve problems and these have mostly focussed on adults (e.g. De Bruin et al., 2005; 2007)⁴. Moreover, 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. Similar to monitoring learning from texts, monitoring how well

⁴ There are some studies that investigated other monitoring judgments than JOLs within the domain of problem solving, for example Feeling of Knowing (FOK) judgments (Metcalfe, 1986; Metcalfe & Wiebe, 1987).

one has learned to solve such well-structured problems concerns more than monitoring memory; students have to monitor whether they *understand* the problem-solving *procedure*.

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, 2013a). 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 (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. However, this improved monitoring accuracy did not affect their regulation; that is, these primary school children did not become better at determining which items they should restudy.

In order to determine whether solving a practice problem after worked example study is an effective generation strategy for other learners as well, it is important to replicate these findings with adolescent secondary education students. 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, we hypothesize in line with the findings by Authors (2013a) 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 hypothesize 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, Van Gog, De Bruin, and Paas (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 hypothesize, 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 et al., 2003; Thiede & Anderson, 2003) led to more accurate JOLs compared to immediate keyword or summary generation. Therefore, it was hypothesized that delayed practice problems would enhance JOL accuracy more than immediate practice problems (i.e.,

Hypothesis 1d: Condition 5 > Conditions 3 and 4). Again, it should be noticed that this is in line with the hypothesis of Baars et al. (2014) but not with their findings; they found no effects of timing of practice problems.

According to models of self-regulation (e.g., Winne & Hadwin, 1998), improved JOL accuracy should result in improved regulation of study, which should result in improved test performance, which was shown for adults (Son & Metcalfe, 2000; Thiede, 1999; Thiede & Dunlosky, 1999; Thiede et al., 2003) as well as for children (De Bruin et al., 2011) when learning items or learning from texts. 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. Task complexity has been found to affect monitoring accuracy of items (e.g., Griffin & Tversky, 1992; Koriat et al., 1980; Lichtenstein & Fischhoff, 1977; Scheck & Nelson, 2005). It may also play a role in monitoring problem solving because monitoring requires working memory (WM) resources (e.g., Griffin et al., 2008; Van Gog, Kester, & Paas, 2011a) and WM resources are assumed to be limited (Baddeley, 1986; Cowan, 2001; Miller, 1956). 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 judgments (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 analyzing whether students' performance improved from criterion to final test.

Method

Participants and Design

Participants were 143 Dutch eleventh-grade students (which would be USA 9th grade) from six different classrooms of two secondary schools ($M_{age} = 14.63$ years, $SD = 0.58$; 79 boys and 64 girls). Participants within each classroom were randomly assigned to one of the five conditions: (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 ($n = 28$). See Table 1 for an overview of the design.

Table 1

Overview of design (WE = Worked Example; JOLs = Judgments of Learning)

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) JOL <i>Filler task</i>	WE (3 min) <i>Filler task</i> JOL <i>Filler task</i>	WE (3 min) Problem (3 min) JOL <i>Filler task</i> 3x	WE (3 min) Problem (3 min) <i>Filler task</i> JOL	WE (3 min) <i>Filler task</i> Problem (3 min) JOL
Restudy choices				
Criterion test (9 min)				
Restudy phase (9 min)				
Final test (9 min)				

Material

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 biology problems in the domain of heredity (laws of Mendel; cf. Kostons, Van Gog, & Paas, 2012). 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.

Pretest. The pretest consisted of 9 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 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 were used which provided a step-by-step demonstration of how to solve biology problems in the domain of heredity (laws of Mendel). The problems were at three different complexity levels, from lowest to highest: (1) 1 generation with an unknown child, (2) 1 generation with an unknown mother, and (3) 2 generations with an unknown child (see also Kostons et al., 2012). 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 biology 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 7-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 three min. delay (condition 2), immediately after the practice problem (condition 3), after a three 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 good 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 to 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 to the ones explained in the worked examples and to the ones practiced.

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 (Conditions 1-2) or after the practice problem (Conditions 3-5), students gave a JOL (Appendix 3). This study-JOL or study-practice-JOL cycle was repeated three times, 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).

Data analysis

Performance scores. Test performance on the criterion and final test was scored 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.

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: $n = 26$.

Table 2

Scoring of absolute monitoring accuracy per step.

	Correct (1)	Incorrect (0)
Criterion test performance per step:		
<u>JOL rating</u>		
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

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 3, 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 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: $n = 27$.

Table 3

Scoring of absolute regulation accuracy per problem.

Restudy choice:	No (0)	Yes (1)
Mean JOL		
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

Results

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

Table 4

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), and final test performance are presented.

	Immediate JOLs	Delayed JOLs	Immediate practice problem and JOLs	Immediate practice problem and delayed JOLs	Delayed practice problem and JOLs
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)
Criterion test performance	3.60 (1.58)	3.75 (1.55)	4.20 (1.56)	4.08 (1.37)	3.83 (1.37)
Absolute accuracy JOLs	0.60 (0.15)	0.58 (0.12)	0.71 (0.16)	0.64 (0.16)	0.61 (0.15)
Relative accuracy JOLs	0.33 (0.48)	0.20 (0.45)	0.39 (0.53)	0.34 (0.50)	0.17 (0.38)
Regulation accuracy	0.50 (0.15)	0.54 (0.13)	0.64 (0.20)	0.60 (0.16)	0.60 (0.19)
Final test performance	4.15 (1.68)	4.03 (1.49)	4.49 (1.69)	4.21 (1.41)	4.50 (1.42)

Monitoring Accuracy

Relative accuracy. Planned comparisons were conducted to test our hypotheses. The first planned comparison (Hypothesis 1a: Condition 3 vs. 4), showed that there was no significant difference in relative accuracy between conditions that gave an immediate vs. delayed JOL after practice problems, $t(118) < 1$, $p = .747$. The second planned comparison (Hypothesis 1b: Condition 1 vs. 2), showed that there was no significant difference in relative accuracy between conditions that gave an immediate vs. delayed JOL after worked

example study, $t(118) = 1.02, p = .308$. The third planned comparison (Hypothesis 1c: Condition 1 & 2 vs. Condition 3, 4 & 5) showed that relative accuracy did not differ between conditions in which students solved problems after worked example study (3, 4, & 5) and conditions in which students did not solve problems (1 & 2), $t(118) < 1, p = .700$. The fourth planned comparison (Hypothesis 1d: Condition 3 & 4 vs. 5), showed that there was a difference in relative accuracy between delayed and immediate problem solving which was marginally 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 marginally 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 (Hypothesis 1a: Condition 3 vs. 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. 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 & 2 vs. Condition 3, 4 & 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, $t(1131) = 2.39, p = .018$, Cohen's $d = -0.42$. Conditions in which students could solve practice problems after studying worked examples, showed higher absolute accuracy. The fourth planned comparison (Hypothesis 1d: Condition 3 & 4 vs. 5) showed that there was a difference between delayed and immediate problem solving which did not reach significance, $t(131) = -1.71, p = .092$. The conditions in which students could solve practice problems directly after worked example study, showed a marginally higher absolute accuracy than the conditions in which students solved practice problems at a delay after worked example study.

Furthermore, a repeated measures ANOVA with Complexity (3 levels) as within-subjects factor and Condition as between-subjects factor showed that absolute accuracy significantly changed over the levels of Complexity, $F(1, 262) = 6.04, p = .003, \eta_p^2 = 0.04$, in all Conditions, $F(1, 39) < 1, p = .840$. Absolute accuracy was higher for the third and

most complex task. Furthermore, there was a significant difference between Conditions, $F(1, 262) = 3.12, p = .017, \eta_p^2 = 0.09$.

Regulation Accuracy

To test our hypotheses about regulation accuracy, we conducted planned comparisons. The first hypothesis (Hypothesis 2a: Condition 3 vs. 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. 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 & 2 vs. Condition 3, 4 & 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 $d = 0.48$. Regulation accuracy was higher for the conditions with practice problems. The fourth planned comparison (Hypothesis 2d: Condition 3 & 4 vs. 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.

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. 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 & 2 vs. Conditions 3, 4 & 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 & 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 & 2 vs. condition 3, 4 & 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, $F(1, 138) = 29.58$, $p < .001$, $\eta_p^2 = 0.18$, but there was no significant difference among Conditions, $F(4, 138) < 1$, $p = .702$ and no interaction between Test Moment and Conditions, $F(4, 138) = 1.84$, $p = .125$.

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 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 seeming 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).

In the present study we also investigated the effects of practice problems as a generation strategy. In line with our expectation (Hypothesis 1c), practice problems helped students to make more accurate JOLs, that is, 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. 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), relative accuracy was higher for students who solved practice problems *immediately* after worked example study. Absolute accuracy also seemed to be higher for the immediate practice problems condition but this difference did not reach statistical significance. Regulation accuracy and final test performance did not differ between conditions with immediate or delayed practice problems.

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). The current study with adolescents in secondary education not only showed higher absolute accuracy in JOLs but also higher accuracy in regulation when using practice problems after worked examples. 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.

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 et al., 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.

Interestingly, timing of practice problems does seem to be important, but in contrast to our expectation, we found that both relative and absolute accuracy tended to be marginally significantly higher for immediate practice problems than delayed practice problems, even though students have to fully rely on information from LTM when solving delayed practice problems. One might argue that on immediate practice problems, students recall the example better, but if anything, one would expect this to affect learning, 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. In other words, spending more time on learning tasks, did not improve learning outcomes. Prior studies comparing worked example study only with example-problem pairs did not find differences in learning outcomes on an immediate test (Van Gog & Kester, 2012; Van Gog et al., 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 less learning tasks. It is therefore quite surprising that the additional opportunity to practice 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.

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 leads to better regulation accuracy and to better test performance if students have the opportunity to control their study time allocation. Although we did find differences in regulation accuracy

that showed JOLs fitted regulation choices better for the conditions in which students were provided with practice problems, final test performance was not significantly higher for students who were provided with practice problems. Note, final test tasks were isomorphic to the ones studied and practiced which could also explain why students who showed higher monitoring and regulation accuracy did not perform better on the final test. Possibly, because JOLs were not perfectly accurate, regulation choices might have been suboptimal. Also, some 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.

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. That is, it is not clear whether complexity or experience caused JOLs to be most accurate at the third and most complex problem-solving task. Future research should try to disentangle these possible causes.

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 by 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 in the relation between JOLs, regulation of study and performance.

Appendix 1
Worked example
1 generation with a homozygote parent and a heterozygote parent

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
<p>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.</p> <p>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).</p> <p>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.</p>	HH
<p>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.</p> <p>We also know that the mother is heterozygote for hair. When a person is heterozygote for a feature then both genes in the genotype are different. In this example it means that the mother has the genotype Hh.</p>	Hh
<p>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.</p>	<p>Answer</p>

<p>Step 4. Make a crosstable to mix the genotypes of the parents and put down the possible genotypes for their children</p> <ol style="list-style-type: none"> 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. Fill out the crosstable by combining the genes of the father and the mother. 	<p>Answer</p> <table border="1" data-bbox="890 193 1167 361"> <tr> <td></td> <td></td> <td colspan="2">Annie Hh</td> </tr> <tr> <td></td> <td></td> <td>H</td> <td>h</td> </tr> <tr> <td>Josh</td> <td>H</td> <td>HH</td> <td>Hh</td> </tr> <tr> <td>HH</td> <td>H</td> <td>HH</td> <td>Hh</td> </tr> </table>			Annie Hh				H	h	Josh	H	HH	Hh	HH	H	HH	Hh
		Annie Hh															
		H	h														
Josh	H	HH	Hh														
HH	H	HH	Hh														
<p>Step 5. Determine the possible genotypes for hair for the children and the chance to get those genotypes</p> <p>GENOTYPE = GENES</p> <table border="1" data-bbox="287 608 563 765"> <tr> <td></td> <td></td> <td colspan="2">Annie Hh</td> </tr> <tr> <td></td> <td></td> <td>H</td> <td>h</td> </tr> <tr> <td>Josh</td> <td>H</td> <td>HH</td> <td>Hh</td> </tr> <tr> <td>HH</td> <td>H</td> <td>HH</td> <td>Hh</td> </tr> </table> <p>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.</p> <p>In this example: two cells have HH = 50% and two cells have Hh = 50%.</p>			Annie Hh				H	h	Josh	H	HH	Hh	HH	H	HH	Hh	<p>Answer</p> <p>50% HH and 50% Hh</p>
		Annie Hh															
		H	h														
Josh	H	HH	Hh														
HH	H	HH	Hh														
<p>Step 6. Determine the possible phenotypes for hair for the children and the chance to get those phenotypes</p> <p>PHENOTYPE = LOOKS</p> <p>Genotype HH means that the dominant feature will show (H = curly hair).</p> <p>Genotype Hh means that the dominant feature will show (H = curly hair).</p> <p>Genotype hh mean that the recessive feature will show (h = straight hair)</p> <p>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.</p>	<p>Answer</p> <p>100% curly hair</p>																

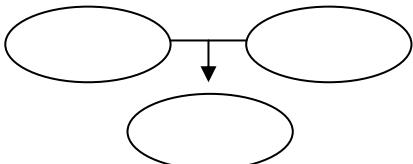
Appendix 2
Practice problem

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 <table border="1"><tr><td></td><td></td><td></td><td>Annie</td></tr><tr><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td></tr><tr><td>Josh</td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td></tr></table>				Annie									Josh							
			Annie																		
Josh																					
Step 5. Determine the possible genotypes for freckles for the children and the chance to get those genotypes GENOTYPE = GENES <table border="1"><tr><td></td><td></td><td></td><td>Annie</td></tr><tr><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td></tr><tr><td>Josh</td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td></tr></table>				Annie									Josh								Answer
			Annie																		
Josh																					
Step 6. Determine the possible phenotypes for freckles for the children and the chance to get those phenotypes	Answer																				

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 1. Determine the genotype for father's hair

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

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

Step 2. Determine the genotype for mother's hair

How well would you be able to solve the step in which the genotype of the mother has to be determined 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 a family tree has to be filled out during a future test?

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

Not at all  Very well

Step 4. Make a cross table 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 genotypes are mixed in a crosstable 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 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 possible genotypes of the child and the chance to get those has to be determined 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 possible phenotypes of the child and the chance to get those has to be determined during a future test?

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

Not at all							Very well
							

Chapter 5

Completion of Partially Worked-out Examples as a Generation Strategy for Improving Monitoring Accuracy⁵

⁵ This chapter was published as 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: <http://dx.doi.org/10.1016/j.cedpsych.2013.09>

Completion of Partially Worked-out Examples as a Generation Strategy for Improving Monitoring Accuracy

Students' Judgments of Learning (JOLs) are often inaccurate: students often overestimate their future test performance. Because of the consequences that JOL inaccuracy can have for regulating study activities, an important question is how JOL accuracy can be improved. When learning texts, JOL accuracy has been shown to improve through 'generation strategies', such as generating keywords, summaries, or concept maps. This study investigated whether JOL accuracy can also be improved by means of a generation strategy (i.e., completing blank steps in the examples) when learning to solve problems through worked example study. Secondary education students of 14-15 years old (cf. USA 9th grade) either studied worked examples or completed partially worked examples and gave JOLs. It was found that completion of worked examples resulted in underestimation of future test performance. It seems that completing partially worked-out examples made students less confident about future performance than studying fully worked examples. However, this did not lead to better regulation of study.

To effectively regulate their own learning process, students must monitor their progress toward learning goals and use this information to regulate further study (e.g., Metcalfe, 2009; Winne & Hadwin, 1998; see also recent special issues by Alexander, 2013; De Bruin & Van Gog, 2012). Monitoring is frequently measured by asking students to provide a *Judgment of Learning* (JOL), that is, a judgment of how well information has been learned in terms of a prediction of future 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). Because students seem to use monitoring judgments to regulate their study activities, JOLs need to be accurate if students are to make an accurate judgment about what information needs to be restudied (Dunlosky, Kubat-Silman, & Hertzog, 2003; Thiede, 1999; Thiede et al. 2003). In other words, JOL accuracy may affect the quality of self-regulated learning (Kornell & Metcalfe, 2006; Metcalfe, 2009; Thiede et al., 2003).

Because JOL accuracy when learning word pairs and learning from expository texts was often found to be low, a lot of studies have investigated how JOL accuracy can be improved by certain instructional interventions (for reviews, see Dunlosky & Lipko, 2007; Rhodes & Tauber, 2011; Thiede, Griffin, Wiley, & Redford, 2009). For instance, Nelson and Dunlosky (1991) found that JOL accuracy when learning word pairs could be improved

by asking a JOL for each pair after a whole set of word pairs was studied, instead of immediately after studying each word pair. This effect has become known as the delayed-JOL effect.

For learning from expository texts, however, the delayed-JOL effect could not be replicated (Maki, 1998; Thiede et al., 2009), unless instructional interventions, such as generating keywords (Thiede et al., 2003), or generating summaries (Thiede & Anderson, 2003), were added prior to making delayed JOLs. This ‘delayed generation-effect’ can be explained from a cue-utilization perspective (see Koriat, 1997) which states that the accuracy of monitoring judgments is dependent on the degree to which the cues used for monitoring correspond with actual performance. Because performance is dependent on the mental representation (i.e., situation model; Kintsch, 1998) of the text that readers build when reading a text, cues based on this mental representation will be the most valid cues for monitoring judgments like JOLs (Thiede et al., 2009). Delayed generation strategies like generating keywords or summaries, require students to access their mental representation of the text from long term memory (LTM); interference from surface-level information about the text from working memory (WM) is no longer present at a delay. This information about the mental representation retrieved from LTM is more indicative of future test performance than information retrieved from WM, and therefore is a cue that helps students to make a more accurate JOL.

However, there are also *immediate* generation strategies such as self-explaining a text (Griffin, Wiley & Thiede, 2008) or making concept maps (Thiede, Griffin, Wiley, & Anderson, 2010) that can enhance immediate JOL accuracy. Thiede et al. (2009) have suggested that the effectiveness of these immediate strategies can also be ascribed to cue-utilization. Immediate strategies such as self-explaining or concept mapping provide students with good cues about their understanding of a text. That is, focusing learners’ attention on the deeper structure of the text rather than on surface features, provides them with cues regarding the quality of their mental representation of the text while they are still studying.

The vast majority of research on how to improve JOL accuracy was conducted with (young) adults, but several studies with children have shown similar findings. That is, children’s JOL accuracy is generally low (De Bruin, Thiede, Camp, & Redford, 2011; Redford, Thiede, Wiley, & Griffin, 2012; Thiede, Redford, Wiley, & Griffin, 2012), but 6-

10 year old children's JOL accuracy for item recall (i.e., unrelated concrete objects presented on picture cards) can also be improved by delaying JOLs (Schneider, Visé, Lockl, & Nelson, 2000) and children's JOL accuracy for more complex materials like text can be improved by generation strategies. For instance, generating keywords at a delay improved 9-13 year old children's JOL accuracy for texts (De Bruin et al., 2011), immediately generating concept maps improved 12- 13 year old children's JOL accuracy for texts (Redford et al., 2012), and generating sentences at a delay improved 10 and 12 year old children's JOL accuracy for idioms (Van Loon, De Bruin, Van Gog, & Van Merriënboer, 2013b).

Moreover, the vast majority of research on improving JOL accuracy, whether it was conducted with adults or with children, has mainly focused on learning word pairs and learning from expository texts (see Rhodes & Tauber, 2011; Thiede, et al., 2009). Problem-solving tasks, however, also play a very important role in education, for instance in math, science, biology, or economics. They come in many forms, varying from insight problems to well-structured transformation problems to ill-structured problems (Jonassen, 2011). In secondary education curricula of math, science, biology or economics, well-structured problems are most common. Such problems have a well-defined initial state, a known goal state, and can be solved using a constrained set of logical operators (Jonassen, 2011). Despite the important role that problem-solving tasks play in education, only a few studies have investigated JOL accuracy, and to the best of our knowledge there are no studies on the *improvement* of JOL accuracy when learning to solve problems. For example, De Bruin, Rikers, and Schmidt (2005, 2007) investigated the accuracy of JOLs when learning to play a chess end-game. They found that JOL accuracy was only low to moderate and was affected by learner expertise. Although these studies focused on JOL accuracy, they did not consider strategies to improve JOL accuracy.

On the one hand, providing JOLs about the kind of well-structured problem-solving tasks frequently encountered in education, shows marked differences with providing JOLs about word pairs or texts: learners need to judge how well they have learned each of the steps that make up the solution procedure, as all those steps are needed to solve that type of problem in the future. On the other hand, because having learned the solution procedure is important and the surface features of a problem are not, it can be expected that (as with texts) strategies that redirect learners' attention away from surface

features and towards cues that give indications of their understanding of the solution procedure, will help them make more accurate JOLs. Therefore, using problem-solving tasks in a school domain (biology), the present study investigated whether 14-15 year old secondary education students' JOL accuracy when learning to solve problems through worked example study can indeed be improved by means of an immediate generation strategy.

JOL Accuracy When Learning to Solve Problems

Studying worked examples (possibly alternated with problem solving) has been found to be a much more effective and efficient way of acquiring problem-solving skills for novices than only engaging in problem-solving practice (for reviews see, Atkinson, Derry, Renkl, & Wortham, 2000; Renkl, 2011; Sweller, Van Merriënboer, & Paas, 1998; Van Gog & Rummel, 2010).

Because it was assumed that students may not always study worked examples very deeply, completion problems (e.g., Paas, 1992; Van Merriënboer, 1990) were introduced, which present students with a partially worked-out solution procedure that they have to complete (Sweller, et al., 1998). This helps learners to process the worked-out solution steps more thoroughly, because otherwise they will not be able to complete the steps that are not worked-out. In the domain of computer programming, Van Merriënboer (1990; see also Van Merriënboer & Krammer, 1990) showed that completion problems led to higher learning outcomes compared to conventional problems (i.e., generating a complete program). A similar result was found by Van Merriënboer, Schuurman, De Croock, and Paas (2002; Experiment 3). In a study by Paas (1992), the effectiveness of completion problems was not only compared to conventional problems, but also to worked examples. This study was conducted in the domain of statistics. Both worked examples and completion problems led to better far transfer test performance than conventional problems. However, no difference was found between worked examples and completion problems on far transfer test performance, which suggests that the additional processing required in the completion condition did not result in better learning outcomes.

In sum, completion problems in which only a partially worked-out solution procedure is given and the missing steps have to be generated by the learner, would seem to require deeper processing of the given steps than when studying fully-worked examples.

More importantly, by attempting to generate the missing steps, learners can be expected to become more aware of their understanding of the problem-solving procedure and the quality of their schema of the solution procedure. That is, generating missing steps will provide them with cues regarding their ability to solve that type of problem on a future test, such as the ease with which they could generate a step, or their feelings of success in generating a step, which are unavailable when merely studying an example. While the additional processing evoked by completion does not necessarily lead to better learning outcomes than when studying fully worked-out examples (Paas, 1992), it might make completion an effective immediate generation strategy to enhance JOL accuracy when acquiring problem-solving skills from worked examples. Despite differences with learning from texts, completing steps in a problem-solving procedure would seem to have effects analogous to the immediate generation strategies (e.g., self-explaining, concept mapping) used in research on learning from expository texts, namely increasing deep processing and focussing learners on their understanding of the underlying structure of the text or problem. Findings by Maki, Foley, Kajer, Thompson, and Willert (1990) are particularly interesting in this respect: they found that JOLs were more accurate when letters in a text were deleted. Because of the deleted letters, these texts supposedly led to increased processing, which provided cues to the learners about their understanding of the text that they could use to make more accurate JOLs. Thus, the following research question was investigated in this study: ‘Does completion of partially worked-out examples (i.e., completion problems) improve JOL accuracy and subsequent regulation accuracy?’.

Because example completion is assumed to lead to deeper processing of the problem solving procedure that has to be learned and to provide cues regarding learners’ understanding of the procedure, our first hypothesis was that the accuracy of JOLs would be enhanced when completing partially worked-out examples compared to studying fully worked-out examples (Hypothesis 1). This hypothesis is based on prior research on the effects of immediate generation strategies on JOL accuracy with texts (Griffin et al., 2008; Redford et al., 2012; Thiede et al., 2010). There is one important difference with these prior studies, however, because they measured relative accuracy (the Goodman-Kruskal gamma correlation), which indicates whether students can discriminate among items, but says nothing about the relationship between their predicted and actual performance per item, which is measured by absolute accuracy (Mengelkamp & Bannert, 2010; Schraw, 2009).

Absolute accuracy is often measured by bias (JOL for an item minus performance on that item; 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). We will focus on bias and absolute deviation here, for a number of reasons. First of all, bias gives information about over- or underconfidence about future test performance for each item, which is of interest given that overconfidence could lead to premature termination of study and may therefore have implications for the success of future learning (Rawson & Dunlosky, 2012). Secondly, because acquiring problem-solving skill is an incremental process, the precision of JOLs, as shown in bias and absolute deviation, is more interesting for problem-solving tasks than the ability to discriminate between tasks (e.g., relative accuracy). Thirdly, gamma can only be reliably computed when many items are used (Nelson, 1984; Schraw, Kuch, & Roberts, 2011), which is not practically feasible with problem-solving tasks.

Our second hypothesis is contingent on the first: if completion would indeed lead to higher JOL accuracy, and if JOLs would be used to make restudy decisions, it can be expected that the accuracy of restudy decisions would be superior in the completion condition (Hypothesis 2; cf. Thiede et al., 2003).

We also address whether completion examples lead to increased processing by analyzing whether completing examples affects invested mental effort. On the one hand, findings by Paas (1992) show no difference between worked examples and completion problems, but on the other hand, it can be expected that processing demands and therefore invested mental effort would be higher for completion problems compared to worked examples (Hypothesis 3). Furthermore, we will analyze whether test performance would be affected by completion examples; findings by Paas (1992) suggested this is not the case (Hypothesis 4).

In this study, tasks at three levels of complexity are used, because there are indications that task complexity affects monitoring. Under the assumption that working memory (WM) resources are limited (Baddeley, 1986; Cowan, 2001; Miller, 1956) and monitoring requires WM resources (e.g., Griffin et al., 2008; Van Gog, Kester, & Paas, 2011a), it can be argued that the more complex a task is, the more resources would be needed to perform it, and the less resources are available for monitoring performance during the task, which might affect the cues available for making JOLs after the task is

completed (Kostons, Van Gog, & Paas, 2009). Therefore, the effect of task complexity on JOL accuracy (Question 1a) and test performance (Question 1b) is explored.

In addition, two types of test problems are used: identical and isomorphic problems. Identical problems are exactly the same as the ones explained in the (partially) worked examples, whereas isomorphic problems have different surface features, but can be solved using the same *solution procedure* that was studied. Because a JOL is made after studying a specific example, monitoring accuracy might be higher when the test task (to which the JOL is compared in order to establish accuracy) consists of an identical problem to be solved than when it concerns an isomorphic problem to be solved, because the learner might make a JOL (at least partially) based on surface features of the examples, which are relevant for identical, but not for isomorphic test tasks. Therefore, we will explore the difference in JOL accuracy (Question 2a) and test performance (Question 2b) between identical and isomorphic test tasks, expecting that monitoring accuracy and test performance will be higher for identical test problems than for isomorphic test problems.

Method

Participants and Design

Participants were 66 Dutch secondary education students ($M_{age} = 14.61$ years, $SD = 0.52$; 24 boys and 42 girls) from three different classrooms of one school. They were in their third year (comparable to ninth grade in the USA) of pre-university education, which has a total duration of six years and is the highest of three levels of secondary education in the Netherlands. They were novices with regard to the learning materials used in this study, as this had not yet been taught in their biology curriculum. Participants within each classroom were randomly assigned to one of the two conditions: worked-out examples ($n = 33$) or completion problems ($n = 33$).

Materials

Pretest. The pretest consisted of nine open-ended questions measuring conceptual knowledge about heredity. A question in this pretest was for example: ‘What is a genotype in reference to a hereditary trait?’.

Worked examples and completion problems. Three worked examples were used that provided a step-by-step demonstration of how to solve biology problems in the domain of heredity (laws of Mendel). The problems could be solved in six steps (see also Kirschner, Paas, & Kirschner, 2009; Kostons et al., 2012): (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 the Punnett square, (6) determining the phenotype of the child. The problems were at three different complexity levels, from lowest to highest: 1) 1 generation with an unknown child, 2) 1 generation with an unknown mother, and 3) 2 generations with an unknown child (see also Kostons et al., 2012).

The completion problems consisted of the same examples, but with steps 4 and 5 left blank for the students to complete. The appendix provides an example of a worked example and completion problem, respectively.

Mental effort rating. Invested mental effort was measured using a 9-point subjective rating scale, which asked participants to rate ‘How much effort did you invest in studying this example?’ with the answer scale ranging from 1 to 9 (Paas, 1992; for information on reliability and sensitivity, see Paas, Van Merriënboer, & Adam, 1994). The scale was presented horizontally and only the first and last answer options were labelled: (1) *very, very low mental effort*, to (9) *very, very high mental effort*.

JOL rating. JOLs were provided on a 7-point rating scale, which asked participants to rate ‘How many steps of a similar problem do you expect to be able to solve correctly on a future test?’ with the answer scale ranging from 0 to 6. The scale was presented horizontally and only the first and last answer options were labelled: (0) *no steps* to (6) *all steps* (for information on consistency of JOLs, see Kelemen, Frost, & Weaver, 2000).

Indication of restudy. Participants were asked to indicate whether they felt they would need to study or complete the problem again, using a yes or no answer format.

Posttest. The posttest consisted of six problem-solving tasks: The three problems encountered in the learning phase in example/completion format (i.e., identical), plus at each of the three complexity levels an isomorphic task was given, which could be solved

using the same procedure, but had different surface characteristics as the problem in the example/completion format at that complexity level.

Procedure

The study was run in three group sessions in students' classrooms, which lasted approximately 50 minutes. First, the experiment leader informed the students about the procedure of the session and students were asked to fill out their demographic data. Then students completed the conceptual knowledge pretest (5 min., which pilot testing had shown to be sufficient). Subsequently, participants either studied the three fully worked-out examples or completed the partially worked-out examples, depending on their assigned condition. To make sure students in both conditions were equally exposed to the learning material, time to study the worked-out examples or to complete the partially worked-out examples was fixed and indicated by the experiment leader. Students had three minutes for studying or completing each worked-out example, which pilot testing had shown to be sufficient. Immediately after studying or completing a worked example, participants had 90 seconds to rate how much effort they invested in studying or completing, made a JOL, and indicated if they felt they would need to study or complete that example again (note that they did not actually get to restudy the examples). Finally, students completed the posttest (max. 18 min.) in which they had approximately three minutes per test item (which a pilot study had shown to be sufficient for solving the problem when the procedure had been learned). It was indicated by the experiment leader when three minutes had passed, but students were allowed to go faster if they had finished before time was up.

Data Analysis

Performance scores. Pretest performance was scored using a standard (based on text book definitions) of the correct answers (cf. Kostons, Van Gog, & Paas, 2012). For each correct answer on question 1-8, one point was assigned (no partial credit was assigned). On question 9, the answer was twofold and therefore 2 points could be obtained (if only one part of the answer was given correctly, one point was assigned), adding up to a maximum score of 10 points for the whole pretest. Answers were judged as being correct if students paraphrased the content of the correct answer. For example, the correct answer to the question: 'What is a genotype in reference to a hereditary trait?' according to the 98

standard would be: ‘A gene pair for a certain trait that is inherited from genes of the mother and father’ but the answer: ‘Two genes from the mother and father’ or: ‘The genes for a trait that are inherited from the mother and father’ were also judged as correct.

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

Monitoring accuracy. Bias was calculated per test problem by subtracting test performance (range: 0 to 6) from the JOL (range: 0 to 6) that was given for that problem type. This results in a positive or negative deviation score, indicating an over- or underestimation of performance, respectively (range: -6 to 6; with 0 meaning perfect accuracy, i.e., no difference between JOL and performance; -6 meaning the largest underestimation possible, i.e., a JOL predicting that 0 steps were performed correctly, while actually performing all steps correctly; and 6 meaning the largest overestimation possible, i.e., a JOL predicting that all steps were performed correctly, while actually performing none correctly). The mean bias over the three identical test tasks and the three isomorphic test tasks was calculated for each student. 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 between JOLs and test performance, but not of the absolute magnitude of the difference. Therefore, we also calculated the absolute deviation, that is, the square root of the squared bias for each item (range: 0 to 6, with 0 meaning perfect accuracy, i.e., no difference between predicted and actual performance, and 6 meaning the largest deviation possible between predicted and actual performance). Note that a reduction in bias in one condition compared to another, would only translate into a reduction in absolute deviation when bias goes from overestimation (positive value) closer to zero, but not when it goes from overestimation to underestimation.

Regulation accuracy. In most studies the accuracy of restudy indications is analyzed using the Goodman-Kruskal Gamma correlation between JOLs and restudy choices (e.g., De Bruin et al., 2011; Thiede et al., 2003). Since we only used three tasks, which also limited the restudy choices to three, we could not compute a reliable gamma correlation. Therefore, we developed an absolute measure of regulation 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 1. As can be seen from the Table, lower JOLs combined

with a choice to restudy result in gradually higher accuracy, whereas lower JOLs combined with a choice not to restudy results in gradually lower accuracy, and vice versa for the higher JOLs. In total three restudy choices could be made, and therefore the total (summed) regulation accuracy score could lie between 0 and 3.

Table 1

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

Missing data. Four students in the completion condition failed to fill out one or more JOLs, and therefore they could not be included in analyses involving JOL accuracy (i.e., monitoring and regulation accuracy). Furthermore, one student in the worked examples condition gave an unclear response to the restudy choices, and this student's data had to be excluded from the regulation accuracy analysis.

Results

As a check on randomization, the pretest performance scores were compared, which - as expected - showed no differences between conditions, $t(63.47) < .001$, $p = 1.000$. The mean JOLs, mean mental effort ratings during the practice phase and mean test performance for identical and isomorphic test tasks are presented in Table 2. In addition, JOLs and test performance over the three different complexity levels are presented in Figure 1.

Table 2

JOLs (*range*: 0-6), invested mental effort (*range*: 1-9) and test performance on identical and isomorphic tasks (*range*: 0-6) for the different problem categories, which differed in complexity (1 = lowest; 3 = highest).

Complexity levels	Mean JOLs		Mean Mental effort		Mean Test performance		
	Worked example (SD)	Completion problem (SD)	Worked example (SD)	Completion problem (SD)	Worked example (SD)	Completion problems (SD)	
1	2.94 (1.52)	1.90 (1.65)	4.85 (2.18)	6.5 (1.76)	Identical	3.52 (1.77)	3.21 (1.76)
					Isomorph	3.40 (1.97)	3.00 (1.94)
2	3.27 (1.75)	1.77 (1.48)	4.58 (2.32)	7.03 (1.72)	Identical	3.18 (1.55)	2.45 (1.46)
					Isomorph	2.94 (1.87)	2.67 (1.65)
3	3.67 (1.88)	2.24 (1.94)	4.30 (2.34)	5.91 (2.23)	Identical	2.90 (1.97)	2.52 (1.97)
					Isomorph	2.67 (2.09)	2.76 (1.92)

Figure 1. JOLs, test performance, and complexity levels.

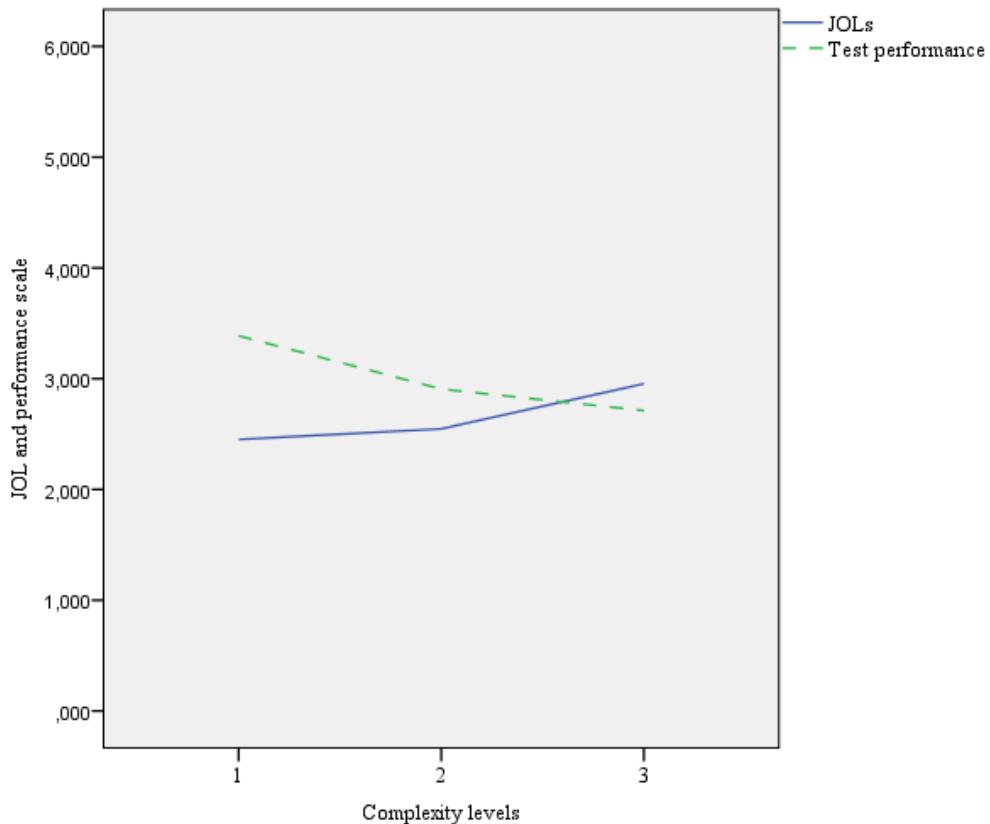


Figure 1: JOLs (*range*: 0-6) and test performance (*range*: 0-6) over the three complexity levels.

Monitoring Accuracy

JOLs. For clarity, we first analyze the JOLs before turning to the bias and absolute accuracy measures. A t-test showed that students in the worked examples condition ($M = 3.29$, $SD = 1.64$) gave higher JOLs than students in the completion problems condition ($M = 2.00$, $SD = 1.50$), $t(60) = 3.22$, $p = .002$, $d = 0.82$.

Bias. To test Hypothesis 1 and explore the effects of complexity (Question 1a) and type of task (Question 2a) on monitoring accuracy in terms of bias, a repeated measures ANOVA with complexity (3 levels) and sort of test problems (identical versus isomorphic

problem-solving tasks, 2 levels) as within-subjects factors and condition as between-subjects factor, showed that there was a main effect of condition, indicating a significant difference in bias between conditions, $F(1, 60) = 5.57, p = .022, \eta_p^2 = .09$. Follow-up *t*-tests comparing bias in each condition to zero, showed that students who completed examples significantly underestimated their future test performance on problem-solving tasks, $t(28) = 2.72, p = .011$, whereas students who studied worked examples did not, $t(32) < 1$.

Furthermore, bias changed significantly as the test problems increased in complexity, $F(2, 120) = 15.53, p < .001, \eta_p^2 = .21$. Post-hoc Bonferroni tests revealed that bias for complexity level one ($M = -0.87$ and $SE = 0.25$) significantly differed from bias for complexity level two ($M = 0.37$ and $SE = 0.23$), $p = .024$, and bias for complexity level three ($M = .23$ and $SE = 0.28$), $p < .001$. The difference between complexity level two and three was also significant ($p = .007$). There was no significant interaction between the complexity levels and condition, $F(2, 120) < 1$. No differences in bias were found between identical and isomorphic test tasks, $F(1, 60) < 1$, and there was no significant interaction between the sort of test problems and condition, $F(1, 60) = 1.98, p = .164$. Also, no interaction effect was found between complexity and sort of test problems, $F(1, 60) < 1$, or between complexity, sort of test problem, and condition, $F(1, 60) = 1.16, p = .318$.

Absolute deviation. To test Hypothesis 1 and explore the effects of complexity (Question 1a) and type of task (Question 2a) on monitoring accuracy in terms of absolute deviation, a repeated measures ANOVA with complexity (3 levels) and sort of test problems (identical versus isomorphic problem-solving tasks, 2 levels) as within-subjects factors and condition as between-subjects factor, showed that there was no significant difference in absolute deviation between conditions, $F(1, 60) < 1$. Furthermore, absolute deviation did not change significantly as the test problems increased in complexity, $F(2, 120) = 1.52, p = .222$. There was no significant interaction between the complexity levels and condition, $F(2, 120) < 1$. The absolute deviation between JOL and test performance was greater for isomorphic test tasks than for identical test tasks, $F(1, 60) = 6.28, p = .015, \eta_p^2 = .10$, which showed that JOLs were more accurate for identical test tasks. There was no significant interaction between the sort of test problems and condition, $F(1, 60) = 1.16, p = .287$. However, there was a significant interaction between complexity and sort of test problems, $F(1, 60) = 3.73, p = .027, \eta_p^2 = .06$. Simple contrasts with the least complex test task as the reference category revealed that the difference between identical and isomorphic

test tasks was different for the least complex problems compared to the more complex problems. Absolute deviations between identical and isomorphic test problems were larger on the second, $F(1, 60) = 10.12, p = .002, \eta_p^2 = .14$, and third, $F(1, 60) = 4.38, p = .041, \eta_p^2 = .07$, complexity levels than on the first complexity level. No significant interaction between complexity, sort of test problem and condition was found, $F(1, 60) < 1$. See Table 3 for mean absolute deviation per complexity level calculated with identical and isomorphic test tasks for both conditions.

Table 3

Mean absolute deviation per complexity level calculated with identical and isomorphic test tasks for both conditions.

Complexity level		Absolute deviation	
		Worked examples	Completion problems
1	Identical	1.73 (1.13)	2.03 (1.32)
	Isomorph	1.61 (1.14)	1.93 (1.44)
2	Identical	1.30 (1.02)	1.48 (1.06)
	Isomorph	1.60 (1.46)	1.86 (1.09)
3	Identical	1.79 (1.60)	1.52 (1.27)
	Isomorph	1.91 (1.68)	2.00 (1.60)

Regulation Accuracy

To test Hypothesis 2, a t-test and a Mann-Whitney U test were used. A t-test showed that regulation accuracy did not differ between the two conditions (worked examples: $M = 1.45, SD = 0.73$, completion problems: $M = 1.45, SD = 0.86; t(59) < 1$). A Mann-Whitney U test showed that the number of tasks selected for restudy did not differ between conditions either (worked examples: $M = 2.22, SD = 1.13$, completion problems: $M = 1.66, SD = 1.34, U = 365.00, p = .119$).

Invested Mental Effort

Hypothesis 3 was tested using a t-test, which showed that students in the worked examples condition ($M = 4.58, SD = 2.10$) invested less mental effort during the learning phase than students in the completion problems condition ($M = 6.37, SD = 1.64$), $t(59.66) = 3.79, p < .001, d = 0.95$. In both conditions, students' mean mental effort rating was 104

significantly negatively correlated with their mean JOL (worked examples: $r = -.72, p < .001$; completion problems: $r = -.68, p < .001$).

Test Performance

To test Hypothesis 4 and explore the effects of complexity (Question 1b) and type of task (Question 2b), a repeated measures ANOVA with complexity (3 levels) and sort of task (identical versus isomorphic test tasks, 2 levels) showed that test performance diminished significantly as the complexity of the problems increased, $F(2, 128) = 7.59, p = .001, \eta_p^2 = .11$. There were no significant differences in performance on identical and isomorphic test tasks, $F(1, 64) < 1$ and there were no significant differences between conditions, $F(1, 64) < 1$.

Discussion

This study investigated whether the accuracy of JOLs when acquiring problem-solving skills through example study, would improve from completing partially worked examples compared to studying fully worked examples. Results show that completing partially worked-out examples led to significant underestimation of future performance compared to studying worked examples. In other words, completion problems led to less confidence in test performance, even though there were no significant differences in test performance between the two conditions.

A possible explanation for the underconfidence in the completion condition could be that participants who completed examples based their JOL on (dis)fluency or feelings of failure in completing those steps without considering what they had learned from the example as a whole (JOLs were generally lower in the completion condition). Indeed, the data-driven view of self-regulation suggests that monitoring can be based on feedback from performance (Koriat & Ackerman, 2010; Koriat, Ma'ayan, & Nussinson, 2006). The finding that there was a significant negative correlation between invested mental effort and JOLs suggests that effort invested in studying was also used as a cue when making a JOL. Example completion resulted in higher mental effort (in contrast to findings by Paas, 1992), which might have made students in this condition underconfident about their future performance. The higher effort investment in the example completion condition also

suggests that -in line with the findings by Maki et al. (1990) on deleted letters in text-completion indeed resulted in increased processing. However, in line with the findings by Paas (1992), this did not lead to improved test performance compared to example study only.

In contrast to bias, the absolute deviation did not differ significantly between conditions. This makes sense because students in the fully worked-out examples condition were on average overconfident, while those in the completion condition were on average underconfident, so while there was a difference in the *direction* of the deviation between predicted and actual performance, there was no difference in the magnitude of the deviation.

The difference in bias between conditions that was found in the current study, did not seem to affect restudy choices, as neither restudy accuracy nor the number of tasks selected for restudy differed between the conditions. Possibly, with problem-solving tasks, the ideas students have about restudy might differ compared to other learning materials. Future research should attempt to shed light on this.

Regarding task complexity, we found an effect of the complexity of the tasks on bias: averaged over both conditions students underestimated their future test performance on the lowest complexity level but not on the second and third complexity level. This might be due to the fact that on average, JOLs tended to increase but performance tended to decrease when the tasks became more complex (see Figure 1). It might be the case that increased familiarity with the task caused JOLs to increase. This would be in line with the cue familiarity hypothesis, which states that the recognition of cues is used as a source to predict future memorability (Metcalfe & Finn, 2008; Schwartz, 1994). However, it should be noted that the absolute deviation between JOLs and test performance did not differ across complexity levels.

In line with our expectation, we found an effect of the sort of problem that was tested. That is, monitoring accuracy in terms of absolute deviation was better for test problems identical to the examples compared to isomorphic test problems. This is interesting because it seems to suggest that the JOL reflects students' judgment of how well they have learned that particular problem in the example, rather than how well they have learned (or can apply) the solution procedure demonstrated in that problem. Note though,

that the latter is what is actually required in educational contexts (students seldom get the exact same practice tasks on an exam).

One limitation of this study is the small number of problem-solving tasks used in the learning phase; ideally, multiple tasks at each complexity level would have been preferable. However, in contrast to word pairs and texts, where each item will have completely different content, and therefore does not affect the JOL or performance on another item, multiple problem-solving tasks within one complexity level would interact with each other, which might influence ratings. Exactly how that would affect ratings, is an interesting question in its own right, and one that future, research might address. Another limitation is that in the classroom context, we were not able to collect process data, such as think aloud protocols that could have shed light on what cues students are using exactly when making JOLs.

Despite these limitations, this study makes several contributions to the literature on monitoring accuracy. First, it extends research on how to improve JOL accuracy from word pairs and texts (Rhodes & Tauber, 2011; Thiede et al., 2009) toward the kind of problem-solving tasks that play an important role in education. Second, it showed that, in line with findings on learning from texts (Griffin et al., 2008; Redford et al., 2012, Thiede et al., 2010), completion as an immediate generation strategy did affect JOL accuracy in terms of bias, but not absolute deviation, as it resulted in underconfidence. Third, the vast majority of research on improving monitoring accuracy was conducted with (young) adults, and this study adds to the relatively small number of studies that investigated the effects of generation strategies on JOL accuracy with children and adolescents (De Bruin et al., 2011; Redford et al., 2012). For educational practice, these findings are interesting because completion is an easy strategy to implement. If future research would replicate these findings and shed more light on whether and how immediate generation strategies such as completion could also affect regulation of study, completion problems can be easily implemented in text books and teaching methods in domains such as math, science, biology, or economics.

Future research should attempt to gain more insight into causes of the underconfidence found with completion problems and why this does not translate into effects on regulation of study, for instance by asking students to think aloud while making JOLs. Given that there is as yet very little research on JOL accuracy when learning to solve

problems, future studies should also investigate JOL accuracy and how to approve it with practice problems and worked examples in other domains. In addition, it should be investigated whether other generation strategies such as having students engage in solving a problem immediately after studying an example (which is a common strategy in example-based learning, known as example-problem pairs; see Van Gog, Kester, & Paas, 2011b), would also result in underconfidence.

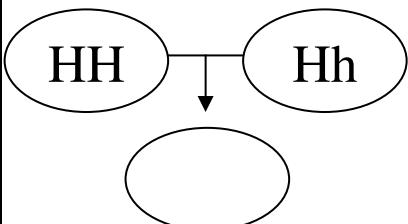
Appendix

Worked example: 1 generation with a homozygote parent and a heterozygote parent

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
<p>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.</p> <p>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).</p> <p>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.</p>	HH
<p>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.</p> <p>We also know that the mother is heterozygote for hair. When a person is heterozygote for a feature then both genes in the genotype are different. In this example it means that the mother has the genotype Hh.</p>	Hh
<p>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.</p>	Answer 

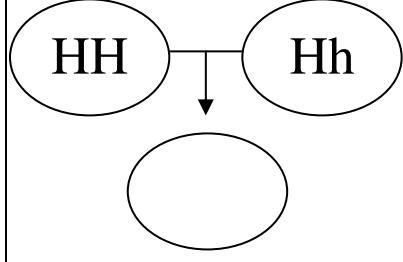
<p>Step 4. Make a table to mix the genotypes of the parents and put down the possible genotypes for their children</p> <p>c. Make a crosstab and divide the genes of the genotypes of the mother in the two cells of the upper row and the genes of the genotypes of the father in the left column.</p> <p>d. Fill out the crosstab by combining the genes of the father and the mother.</p>	<p>Answer</p> <table border="1" data-bbox="901 190 1171 361"> <tr> <td></td><td colspan="2">Annie Hh</td></tr> <tr> <td></td><td>H</td><td>h</td></tr> <tr> <td>Josh</td><td>H</td><td>HH</td><td>Hh</td></tr> <tr> <td>HH</td><td>H</td><td>HH</td><td>Hh</td></tr> </table>		Annie Hh			H	h	Josh	H	HH	Hh	HH	H	HH	Hh		
	Annie Hh																
	H	h															
Josh	H	HH	Hh														
HH	H	HH	Hh														
<p>Step 5. Determine the possible genotypes for hair for the children and the chance to get those genotypes</p>	Answer																
<p>GENOTYPE = GENES</p> <table border="1" data-bbox="298 572 568 736"> <tr> <td></td> <td></td> <td colspan="2">Annie Hh</td> </tr> <tr> <td></td> <td></td> <td>H</td> <td>h</td> </tr> <tr> <td>Josh</td> <td>H</td> <td>HH</td> <td>Hh</td> </tr> <tr> <td>HH</td> <td>H</td> <td>HH</td> <td>Hh</td> </tr> </table> <p>You can get this information from the crosstab you just made. In the four cells of the crosstab 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.</p> <p>In this example: two cells have HH = 50% en 2 cells have Hh = 50%.</p>			Annie Hh				H	h	Josh	H	HH	Hh	HH	H	HH	Hh	50% HH and 50% Hh
		Annie Hh															
		H	h														
Josh	H	HH	Hh														
HH	H	HH	Hh														
<p>Step 6. Determine the possible phenotypes for hair for the children and the chance to get those phenotypes</p> <p>PHENOTYPE = LOOKS</p> <p>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)</p> <p>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.</p>	Answer 100% curly hair																

Completion example: 1 generation with a homozygote parent and a heterozygote parent

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
<p>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.</p> <p>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).</p> <p>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.</p>	HH
<p>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.</p> <p>We also know that the mother is heterozygote for hair. When a person is heterozygote for a feature then both genes in the genotype are different. In this example it means that the mother has the genotype Hh.</p>	Hh
<p>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.</p>	Answer 

<p>Step 4. Make a crosstable to mix the genotypes of the parents and put down the possible genotypes for their children</p>	<p>Answer</p> <table border="1" data-bbox="917 226 1256 518"> <tr> <td></td><td></td><td colspan="2">Annie</td></tr> <tr> <td></td><td></td><td></td><td></td></tr> <tr> <td colspan="2">Josh</td><td></td><td></td></tr> <tr> <td></td><td></td><td></td><td></td></tr> </table>			Annie						Josh							
		Annie															
Josh																	
<p>Step 5. Determine the possible genotypes for hair for the children and the chance to get those genotypes</p> <p>GENOTYPE = GENES</p> <table border="1" data-bbox="298 727 679 1019"> <tr> <td></td><td></td><td colspan="2">Annie</td> </tr> <tr> <td></td><td></td><td></td><td></td> </tr> <tr> <td rowspan="2">Josh</td><td></td><td></td><td></td> </tr> <tr> <td></td><td></td><td></td> </tr> </table>			Annie						Josh							<p>Answer</p>	
		Annie															
Josh																	
<p>Step 6. Determine the possible phenotypes for hair for the children and the chance to get those phenotypes</p> <p>PHENOTYPE = LOOKS</p> <p>Genotype HH means that the dominant feature will show (H = curly hair).</p> <p>Genotype Hh means that the dominant feature will show (H = curly hair).</p> <p>Genotype hh mean that the recessive feature will show (h = straight hair)</p> <p>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.</p>	<p>Answer</p> <p>100% curly hair</p>																

Chapter 6

Effects of Training Self-assessment and Using Assessment Standards on Retrospective and Prospective Monitoring of Problem Solving⁶

⁶ This chapter is submitted for publication as Baars, M., Vink, S., Van Gog, T., De Bruin, A., & Paas, F. (2013). *Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving*.

Effects of Training Self-assessment and Using Assessment Standards on Retrospective and Prospective Monitoring of Problem Solving

Both retrospective and prospective monitoring are considered important for self-regulated learning of problem-solving skills. Retrospective monitoring (or self-assessment; SA), refers to students' assessments of how well they performed on a problem just completed. Prospective monitoring (or Judgments of Learning; JOLs), refers to students' judgments about how well they will perform on a (similar) problem on a future test. We investigated whether secondary education students' SA accuracy could be improved by training (Experiment 1 and 2), or by providing assessment standards (Experiment 2), and whether this would also affect the accuracy of JOLs. Accurate assessment of past performance might provide a good cue for judging future performance. Both Experiment 1 and 2 showed no effect of training on SA or JOL accuracy, but SA and JOLs were positively correlated with each other and negatively with effort. Providing standards did improve SA and JOL accuracy on identical, and performance on all problems.

Self-regulated learning can only be optimally effective for learning outcomes when students are able to accurately monitor their own performance and use this information to choose what to study again or what to study next (e.g., Metcalfe, 2009; Winne & Hadwin, 1998; see also recent special issues by Alexander, 2013; De Bruin & Van Gog, 2012). As such, accurate monitoring seems to be a pivotal aspect of self-regulated learning. Monitoring can be measured both *retrospectively*, by asking students to judge their performance on a task just completed, which is also known as self-assessment (Kostons, Van Gog, & Paas, 2012) or a self-score judgment in verbal tasks (Lipko et al., 2009; Rawson & Dunlosky, 2007) and *prospectively*, by asking students to predict their performance on that task on a future test, which is also known as a *Judgment of Learning* (JOL; e.g., Koriat, Ackerman, Lockl, & Schneider, 2009a, 2009b; Metcalfe & Finn, 2008; Nelson & Dunlosky, 1991). Monitoring accuracy can then be determined by comparing students' self-assessed or predicted performance with their actual performance on a task. The more accurate monitoring is, the better participants are assumed to be able to keep track of their learning process, and the better they might be able to regulate it. Research has shown, however, that accurately monitoring their own performance is hard for students, and accuracy of both self-assessments (Bjork, 1999; Dunlosky, Rawson, & McDonald, 2002) and JOLs (Dunlosky & Lipko, 2007; Maki, 1998; Serra & Metcalfe, 2009; Thiede, Griffin, Wiley, & Redford, 2009) is often low, but there are instructional techniques that seem to improve accuracy. In the next sections, we will discuss research showing how accuracy can be improved, in light

of the aims of the present study, which were to investigate whether the accuracy of secondary education students' self-assessment of problem-solving tasks can be enhanced by training, providing standards, or both, and whether increased self-assessment accuracy (retrospective) would also enhance JOL accuracy (prospective).

Improving Self-assessment Accuracy

Two techniques that have been shown to improve self-assessment (SA) accuracy are training and using standards to assess performance. Kostons et al. (2012) showed that SA skills can be trained by means of modeling examples that show students how to assess performance on biology problem-solving tasks. They trained secondary school students to assess their own performance and, based on that assessment, to select an appropriate task for further studying. In this training students were shown four computer screen recordings of human models who verbally explained how they: 1) solved a heredity problem, 2) assessed their own performance on that problem by assigning 1 point for each step in the procedure they felt they had performed correctly (i.e., no standard was used), and 3) selected a new task to study next, at an appropriate level of difficulty given their assessed performance in combination with the amount of mental effort they had to invest to reach that performance. They also investigated the effects of training only SA skills, with modeling examples showing only step 1 and 2, or training only task selection (step 1 and 3). Kostons et al. found that SA and task-selection skills improved when these had been trained, and that it was necessary to train both; they found no effects of training SA skills on task selection accuracy or vice versa (Experiment 1). Moreover, when students engaged in self-regulated learning after a SA and task selection training, their learning gains were higher than for students who had not been trained (Experiment 2).

Another technique that was found to improve SA accuracy, at least when learning from text, is using standards. Rawson and Dunlosky (2007) showed that students were better able to assess the correctness of their own test performance when they were provided with an assessment standard, that is, a description of the correct answer to compare their own answer to. They asked college students to self-assess the quality of their recall of key concepts from textbooks, by assigning themselves no credit, full credit, or partial credit. Students overestimated their own recall performance, but when they were provided with a standard, consisting of the correct definition of the key concepts, their overestimation was

smaller. Lipko et al. (2009) replicated these findings with middle school children. Thus, by having correct definitions available as standard for evaluation, students are better able to recognize incorrect responses, which reduces their overconfidence and leads to better calibration of their assessment and their actual performance. In a cyclical learning process, in which learners continue studying after the self-assessment, standards might also improve learning outcomes, because they also provide learners with feedback regarding their own performance and correct responses (Butler & Winne, 1995). Indeed, Rawson and Dunlosky (2007) found that performance on a criterion test improved when students had used standards to assess their performance on a practice test of definitions of the key concepts. To the best of our knowledge, the effects of standards on calibration of SA, JOLs, and learning outcomes, have not yet been tested with problem-solving tasks.

In sum, SA accuracy can be improved by training (Kostons et al., 2012) or by using standards (Lipko et al., 2009; Rawson & Dunlosky, 2007). It would be interesting to investigate whether the findings by Kostons et al. can be replicated using written worked examples to train SA (instead of video-based modeling examples) as these might be easier to create and implement, and whether *combining* training prior to making SAs of problem-solving tasks with standards provided while making those assessments, would be more effective than either method alone.

Improving Accuracy of Judgments of Learning

Many studies have investigated the accuracy of JOLs when learning word pairs (for a review see Rhodes & Tauber, 2011) or when learning from expository texts (for a review see Thiede et al., 2009). They have shown that relative accuracy of JOLs when learning from more complex materials like expository text or problem-solving tasks is very low, but that adding so-called ‘generation strategies’, helps students to make more accurate JOLs. Such strategies make students actively generate (part of) the learning materials after studying them, focusing their attention on the gist of the material or the underlying structure of the material. For example, it was found that generating keywords (De Bruin, Thiede, Camp, & Redford, 2011; Thiede, Anderson, & Therriault, 2003), making summaries (Thiede & Anderson, 2003), making concept maps (Thiede, Griffin, Wiley, & Anderson, 2010), and self-explaining (Griffin, Wiley, & Thiede, 2008) improved the accuracy of JOLs when learning from text. In addition, generating keywords did not only improve JOL

accuracy but also affected regulation and led to greater test performance (Thiede et al., 2003).

The cue utilization framework (Koriat, 1997) can explain the effect of generation strategies on JOL accuracy. According to this framework, JOL accuracy is the result of the cues that are used to make a JOL and the extent to which these cues are diagnostic for future test performance. Generating keywords or summaries, or self-explaining a text, are all activities that provide participants with insight into the quality of their representation of the text, and they can use this information when making JOLs. The cues provided by such generation strategies are more indicative for future test performance than cues learners would spontaneously use, and therefore lead to more accurate JOLs (Thiede, Dunlosky, Griffin, & Wiley, 2005; Thiede et al., 2009). It should be noted that accuracy in those studies was mostly defined as *relative* accuracy, measured by calculating a Goodman-Kruskal gamma correlation (Thiede et al., 2009). This shows whether participants are able to discriminate between different items, it does not give any information on how accurate they were in predicting their performance per item. However, *absolute* measures of accuracy that do show the precision of JOLs, like bias (i.e., the difference between self-assessed and actual performance, with positive values indicating overestimation and negative values underestimation) or absolute deviation scores (i.e., without direction), have also been used to analyze JOL accuracy (Authors, 2013; Maki, Shields, Wheeler, & Zaccchili, 2005; Mengelkamp & Bannert, 2010; Schraw, 2009).

With regard to problem-solving tasks, only few studies have investigated JOL accuracy and how to improve it (e.g., De Bruin, Rikers, & Schmidt, 2005, 2007). Recent research showed that when acquiring problem-solving skills from worked examples, which is an effective and efficient way of learning to solve problems compared to engaging in problem-solving practice (Atkinson, Derry, Renkl, & Wortham, 2000; Renkl, 2013; Sweller, Van Merriënboer, & Paas, 1998; Van Gog & Rummel, 2010), the use of a generation strategy was effective for improving JOL accuracy in terms of bias (Baars, Van Gog, De Bruin, & Paas, 2014). That is, when students tested their knowledge after studying a worked example, by means of solving a problem on their own, bias was reduced. Presumably, engaging in problem solving after worked example study provides a learner with relevant cues on which to base their JOLs.

This would mean, basically, that students use cues from past performance (i.e., the problem they just solved) to judge future performance, which is what Griffin, Jee, and Wiley (2009) refer to as the postdiction route to JOLs. The postdiction route is based on the Memory for Past Test (MPT) heuristic (Finn & Metcalfe, 2007). According to the MPT heuristic, when making a JOL about word pairs following a practice test, learners will use their feeling of success or failure in getting the item right on that test as a cue for the JOL. Such cues could help students to make more accurate JOLs, but only if they are able to make accurate (retrospective) judgments about their performance on a practice test. It would be interesting to investigate whether this would mean that when the accuracy of SA (i.e., retrospective monitoring) of performance on a problem can be improved, for instance through training or using standards, JOL accuracy (i.e., prospective monitoring) would improve as well.

The Present Study

The present study consists of two experiments, aimed at investigating whether we could improve SA accuracy and whether this would also improve JOL accuracy and subsequent regulation accuracy and test performance, because a more accurate SA might also provide students with more valid cues to predict their future performance and regulate it accordingly. Experiment 1 investigated whether SA training could improve the accuracy of SAs and JOLs, and explored whether regulation choices (i.e., decisions about the need to restudy) made after practice problems and test performance would also improve. SA and JOL accuracy were operationalized in terms of bias (i.e., the difference between self-assessed and actual performance, with positive values indicating overestimation and negative values underestimation) and in terms of the absolute deviation (i.e., without direction) between self-assessed and actual performance scores (Maki et al., 2005; Mengelkamp & Bannert, 2010; Schraw, 2009). Experiment 2 used a slightly older (ninth grade) student population to investigate the effects of SA training and the use of standards on the accuracy of SA and JOLs, and again explored effects on regulation and test performance.

Experiment 1

Experiment 1 aimed to investigate whether the findings by Kostons et al. (2012) regarding the effect of self-assessment (SA) training on SA accuracy, could be replicated with training consisting of written worked examples of another person's performance self-assessments instead of video-based modeling examples. The same type of learning task was used as in the Kostons et al. study: biology problems on heredity. Whereas Kostons et al. (2012) used four video-based modeling examples for the SA training, we used a short, paper-based SA training with two written worked examples containing the same information as in the SA training used by Kostons et al. (2012). Using two written worked examples instead of four video-based modeling examples could be advantageous because it provides a shorter training which gives students the opportunity to review the example multiple times because information is not transient like in videos. Furthermore, paper-based examples are more easily implemented in educational material.

It was hypothesized that SA training would improve SA accuracy (Hypothesis 1a) and thereby provide students with cues to make JOLs which should lead to improved JOL accuracy (Hypothesis 1b). Because the hypothesis that enhanced SA accuracy would also lead to enhanced JOL accuracy depends on a relationship between the two, we explored this relationship by calculating the correlation between SAs and JOLs (Question 2).

We explored effects of SA training on regulation accuracy. Kostons et al. (2012) only found effects on regulation (i.e., task selection) accuracy and learning outcomes when task selection was trained; SA training alone did not improve regulation. Thiede et al. (2003) did find that improved JOL accuracy led to more effective regulation and greater overall test performance, but they used very different learning materials. It is therefore an open question whether we will find beneficial effects of SA training on regulation accuracy (Question 3) and posttest performance (Question 4).

Another explorative question was whether the nature of the test problems would play a role in JOL accuracy, by using test problems that were identical and isomorphic to the ones seen in the learning phase (Question 5). Being able to accurately predict performance on isomorphic problems, which have different surface features but can be solved using the same *solution procedure* that was studied, is more relevant for educational practice. However, when SA accuracy would affect JOL accuracy, this might only be the case for identical problem-solving test tasks that are exactly the same as the one on which

performance was assessed. Even though learners could use the same procedure for solving isomorphic problems, performance on isomorphic test problems is less likely to profit from performance on the practice problems than performance on identical test problems.

Finally, because earlier research showed that the perceived amount of invested mental effort might be used as a cue for making JOLs, with effort and JOLs being negatively related (Baars, Visser, Van Gog, De Bruin, & Paas, 2013), we also investigated the relation between invested mental effort and JOLs and between invested mental effort and SA. In line with earlier research, it was expected that invested mental effort would show a negative correlation with SA's (Hypothesis 6a) and JOL's (Hypothesis 6b).

Method

Participants and Design

Participants were 44 Dutch eighth grade students ($M_{age} = 12.95$ years, $SD = 0.46$; 21 boys and 23 girls) from two different classrooms of one school. Participants within each classroom were randomly assigned to one of the two conditions: SA training ($n = 23$) or no SA training ($n = 21$).

Material

All materials were paper-based and each worked example, problem-solving task, and rating scale was presented on a new page. In this experiment students had to learn to solve biology problems in the domain of heredity (laws of Mendel; cf. Kostons et al., 2012). 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 the Punnett square, (6) determining the phenotype of the child. We used an additional step compared to the procedure used by Kostons et al.; that is, step 3 (i.e., making a genealogical tree) was added based on the procedure to solve heredity problems that was taught in a study by Kirschner, Paas, and Kirschner (2009). Whereas Kostons et al. used tasks at five

levels of complexity, with SA training involving demonstrations of problems at level 1 and 2, the examples and problems used in this experiment were at three levels of complexity (i.e., the second, third, and fourth level of complexity of Kostons et al.), and training involved demonstrations of problems at level 2.

Pretest. The pretest consisted of 9 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?’

SA training phase. Students in the SA training condition were given instructions on paper that one point should be assigned to each step of the practice problem that was correctly completed, and to sum these to determine the final SA score (ranging from 0, no steps correct to 6, all steps correct). After reading the written instructions, students in the SA training condition were shown worked-out solutions to two problems (at complexity level 2) made by a fictive student who made mistakes. The worked-out solutions were followed by SAs by the fictive student in which the performance on the problems was accurately self-assessed by indicating which steps were performed correctly (yes or no format) and giving a total score of correctly performed steps. For an example see Appendix 2.

Students in the condition without SA training were shown the same worked-out solution with mistakes and they were instructed to read the examples and find the mistakes.

Learning phase. The learning phase consisted of three example-problem pairs at three complexity levels, ordered 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). In each pair, a worked example first provided a step-by-step demonstration of how to solve that kind of problem; students then received an isomorphic practice problem, which they had to solve themselves. An example of an example-problem pair (at complexity level 1) can be found in Appendix 1.

Mental effort rating. Invested mental effort was measured using a 9-point subjective rating scale, which asked participants to rate how much effort they invested in practicing the problem, ranging from (1) *very, very low mental effort*, to (9) *very, very high mental effort* (Paas, 1992).

SA rating. SA ratings were provided on a 7-point rating scale, which asked participants to rate how many steps they performed correctly, ranging from (0) *none* to (6) *all steps correct*.

JOL rating. JOLs were provided on a 7-point rating scale, which asked students to predict how many steps of a similar problem they expected they could solve correctly on a future test, ranging from (0) *none* to (6) *all steps*.

Indication of restudy. Participants were asked to indicate whether they wanted to study the worked example again to be able to solve it during a test, after they made a SA and JOL about a problem they studied and practiced, using a yes or no answer format.

Posttest. The posttest consisted of six problem-solving tasks: The three problems that were encountered in the learning phase as practice problems (i.e., identical problems), plus an isomorphic problem at each of the three complexity levels, which could be solved using the same procedure as demonstrated in the worked example and practiced on the practice problem of that complexity level, but had different surface characteristics than the example and practice problems. The posttest problems were ordered by complexity starting with the test problem identical to the practice problem at the lowest complexity level, followed by the isomorphic problem at that level, and so on for the other two complexity levels.

Procedure

The study was run in two group sessions in students' classrooms, which lasted approximately 70 minutes, and both conditions, to which students were randomly assigned, were present in each session. First, students made a five min pretest with nine items measuring their conceptual knowledge about heredity. Next, half of the participants got a five min training about how to self-assess their test performance, while the other half just read two worked-out solutions of the two problem-solving tasks that were also used in the training and were asked to find the errors in those examples. Then in the study phase (21 min) all participants worked on three example-problem pairs, which entailed: studying the worked example for three min, solving a practice problem for three min, indicating how much effort they invested in solving that problem, making a SA, giving a JOL, and indicating if they needed to study a worked example again. Finally, the posttest (max. 18 min) was completed. During the posttest the experiment leader indicated every three min

that it was time to move on to the next problem-solving task, but students were free to move on if they were done before the three min were up. The allotted time was based on pilot tests and a prior study in which the same tasks were used (Baars et al., 2013).

Data analysis

Performance scores. Test performance was scored by assigning 1 point for each correct step (i.e., range per problem: 0-6 points).

Monitoring accuracy. The accuracy of SAs and JOLs was analysed by calculating bias and absolute deviation scores. For SAs, bias was calculated per *practice* problem by subtracting performance on a practice problem from the SA that was given for that problem (i.e., retrospective judgment). For JOLs, bias was calculated per *test* problem by subtracting test performance from the JOL that was given for that problem type (i.e., prospective judgment). This results in a positive or negative deviation score, indicating an over- or underestimation of performance, respectively. The mean bias in JOLs over the three identical test tasks and the three isomorphic 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 SA or JOLs and performance. Therefore, we also calculated this absolute deviation for SAs and JOLs, that is, we calculated the square root of the squared bias for each item that was self-assessed or given a JOL. Again, the absolute deviation in JOLs was averaged for the three identical and the three isomorphic items separately (min. = 0; max. = 6).

Regulation accuracy. In most studies the accuracy of restudy indications is analysed using the Goodman-Kruskal Gamma correlation between JOLs and restudy choices (e.g., De Bruin et al., 2011; Thiede et al., 2003). Since we only used three tasks, which also limited the restudy choices to three, we could not compute a reliable gamma correlation. Therefore, we developed an absolute measure of regulation 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 1. As can be seen from the Table, lower JOLs combined with a choice to restudy result in higher accuracy, whereas lower JOLs combined with a choice not to restudy result in lower accuracy, and vice versa for the higher JOLs.

Table 1

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

Missing data. Three students in the SA training condition failed to fill out SA, JOL, and restudy indication on one of the tasks, and therefore they were not included in analyses involving those measures. Furthermore, three students in the no SA training condition and two students in the SA training condition did not fill out one or more mental effort ratings and were not included in the analysis on mental effort.

Results

Table 2 presents the data per condition, both the original and computed measures. As a check on randomization, the pretest performance scores were compared, which –as expected- showed no significant differences between conditions, $t(28.11) = 1.32$, $p = .197$.

Table 2

The mean SA (*range*: 0-6), practice problem performance (*range*: 0-6), SA bias (*range*: -6-6), SA absolute deviation (*range*: 0-6), JOLs (*range*: 0-6), posttest performance on Identical and Isomorphic Test Tasks (*range*: 0-6), JOL bias on Identical and Isomorphic Test Tasks (*range*: -6-6), JOL absolute deviation on Identical and Isomorphic Test Tasks (*range*: 0-6), invested mental effort during the learning phase (*range*: 0-9) and regulation accuracy (*range*: 0-1) from Experiment 1 are presented.

	No SA training	SA training
SA	3.46 (1.45)	3.80 (1.23)
Practice problem performance	2.25 (1.48)	2.51 (1.45)
SA bias	1.15 (1.30)	1.21 (0.84)
SA absolute deviation	1.82 (0.86)	1.68 (0.90)
JOL	3.54 (1.40)	3.67 (1.40)
Posttest performance Identical Tasks	2.59 (1.50)	2.46 (1.62)
Bias Identical Test Tasks	0.95 (1.34)	1.15 (1.77)
Absolute deviation Identical Test Tasks	1.97 (0.91)	2.28 (0.75)
Posttest performance Isomorphic Tasks	2.56 (1.33)	2.59 (1.63)
Bias Isomorphic Test Tasks	0.98 (1.66)	0.93 (1.52)
Absolute deviation Isomorphic Test Tasks	2.22 (0.85)	1.97 (0.94)
Invested mental effort (after practice problem)	6.05 (1.40)	6.09 (2.16)
Regulation accuracy	0.38 (0.22)	0.39 (0.29)

Monitoring Accuracy

SA accuracy. Hypothesis 1a that SA training would improve SA accuracy was tested with t-tests which showed that there were no significant differences between conditions in SA accuracy either in terms of bias, $t(39) < 1$, $p = .869$, or absolute deviation, $t(39) < 1$, $p = .630$.

JOL accuracy. To test Hypothesis 1b that SA training would improve JOL accuracy and explore Question 5 about whether the nature of the test problems would play a role in JOL accuracy, repeated measures ANOVA with Type of Test Task (Identical vs. Isomorphic) as within-subjects factor and Condition (SA Training vs. no Training) as between-subjects factor was conducted. It showed no main effect of Type of Test Task on bias $F(1, 39) < 1$, $p = .507$, or absolute deviation, $F(1, 39) < 1$, $p = .809$, and no main effect of Condition on bias, $F(1, 39) < 1$, $p = .878$, or absolute deviation, $F(1, 39) < 1$, $p = .902$. There was no significant interaction on bias, $F(1, 39) < 1$, $p = .376$, but there was a significant interaction between Type of Test Task and Condition on absolute deviation

scores, $F(1, 39) = 4.99, p = .031, \eta_p^2 = 0.11$. Whereas absolute deviation in the No Training condition was bigger for isomorphic than for identical test tasks, the reverse was found for the Training condition. However, post-hoc t-tests showed no significant difference in absolute deviation scores on Identical and Isomorphic Test Tasks within the No Training, $t(20) = 1.53, p = .141$, or the SA Training condition, $t(19) = 1.62, p = .122$, and no significant differences between the two conditions in absolute deviation scores on Identical Test Tasks, $t(39) = 1.20, p = .236$, and Isomorphic Test Tasks, $t(39) < 1, p = .367$.

SA and JOLs. To explore Question 2 about the relation between SAs and JOLs, the correlation between SA ratings and JOL ratings had to be calculated. Because these data were nested (i.e., multiple measures within persons), we first calculated the intraclass correlation coefficient (ICC) to examine the amount of variance at both levels. The ICC reflects the amount of between-person variability compared to the total amount of variability (both between- and within-person variability). The ICC's indicated that the variance explained by the person level ranged from 39.6% for SA ratings to 43.5% for JOL ratings. Therefore, correlations between SA ratings and JOL ratings in both conditions were calculated using multi-level analysis in MPlus (Muthén & Muthén, 1998-2010).

In both conditions SA ratings were significantly correlated to JOLs (SA Training: $r = .99, p < .001$ and No Training: $r = .83, p < .001$), Fisher's $z = 4.49, p < .001$.

Regulation Accuracy

Question 3 about beneficial effects of SA training on regulation accuracy was explored by performing a t-test which showed that there was no significant difference in regulation accuracy between both conditions, $t(39) < 1, p = .842$.

Invested Mental Effort

To test Hypotheses 6a and b, that mental effort ratings would show a negative correlation with SA's (a) and JOLs (b), mental effort ratings given after the practice problems were correlated to SA and JOL ratings. Again, we used multi-level analysis in Mplus to calculate correlations between SAs, JOLs and mental effort ratings because the data was nested. The ICCs indicated that the variance explained by the person level ranged from 39.6% for SA ratings, 43.5% for JOL ratings, to 41.8% for mental effort ratings. In the No Training condition SA was significantly correlated to mental effort ratings ($r = -.98, 126$

$p < .001$) but this was not the case for the SA Training condition ($r = -.05$, $p = .931$), Fisher's $z = -6.92$, $p < .001$. Also, mental effort ratings were significantly negatively related to JOLs in the No Training condition ($r = -.76$, $p < .001$) but not in the SA Training condition ($r = -.27$, $p = .688$), Fisher's $z = -2.21$, $p = .03$.

Posttest Performance

To explore Question 4 about beneficial effects of SA training and Question 5 about the role of the nature of the test problems on test performance, a repeated measures ANOVA on posttest performance with Type of Test Task (Identical vs. Isomorphic) as within-subjects factor and Condition (SA Training vs. No Training) as between-subjects factor showed no main effect of Type of Task, $F(1, 42) < 1$, $p = .714$, no main effect of Condition, $F(1, 42) < 1$, $p = .924$, and no interaction, $F(1, 42) < 1$, $p = .548$.

Discussion

Results of Experiment 1 show that, in contrast to our hypothesis (Hypothesis 1a), training did not improve SA accuracy. Consequently, it is not surprising that we did not find any evidence for improved JOL accuracy (Hypothesis 1b), regulation accuracy (Question 3), or performance (Question 4) either. Although the ANOVA showed an interaction effect of Condition and Type of Test Task on absolute deviation scores (Question 5), which, according to Table 2, indicated that in the no training condition the absolute deviation score was higher on isomorphic tasks than on identical tasks, while in the SA training condition this was the other way around, post hoc analyses showed that these differences were not significant.

There was a significant correlation between SA ratings and JOLs, though, suggesting that students do seem to use their SA to make a JOL (Question 2). Interestingly, while, in line with our hypothesis (Hypotheses 6a and 6b), invested mental effort was negatively correlated with SAs and JOLs in the no training condition, this was not the case for the condition in which students received a SA-training. This might suggest that the SA training, in which students were shown that they should first evaluate their performance per step and then use the number of steps they felt they performed correctly as a SA rating, reduces the reliance on invested mental effort as a cue for SAs and JOLs.

The question is why SA training, which was very effective in the study by Kostons et al. (2012), did not enhance SA accuracy in our study. Possibly, this is due to the differences in the way in which the training was implemented, using two instead of four examples and using worked examples instead of video-based modeling examples. However, another potential explanation might be that the eighth grade students (as compared to the 10th grade students in the Kostons et al. study) found these tasks very challenging; posttest scores showed they learned relatively little from the worked examples and practice problems (mean score < 50%). As a consequence, they may have learned from the training that they should assess performance on each of the steps in the procedure, but they may not have been able to accurately assess their performance on each of the steps. Because no standards were given, students had only their prior knowledge to rely on when making self-assessments, which may have been insufficiently developed to accurately judge their performance (for a discussion of the relationship between competence and self-assessment, see e.g., Dunning, Johnson, Erlinger, & Kruger, 2003). Therefore, Experiment 2 used a slightly older (ninth grade) student population and also investigated the effects of the use of standards next to SA training on the accuracy of SAs and JOLs.

Experiment 2

We expected that standards would help students to judge whether their performance on a step was correct or not, thereby improving SA accuracy (Hypothesis 1b), JOL accuracy (Hypothesis 2b), regulation accuracy (Question 3b), and posttest performance (Question 4b). If the SA training did not have a significant effect on SA accuracy in Experiment 1 because the problem-solving tasks were too difficult for the eighth grade students, then in this second experiment with ninth grade students we expected a main effect of SA training on SA accuracy (Hypothesis 1a), JOL accuracy (Hypothesis 2a), regulation accuracy (Question 3a), and posttest performance (Question 4a). Possible interactions between standards and training with regard to SA accuracy (Hypothesis 1c), JOL accuracy (Hypothesis 2c), regulation accuracy (Question 3c), and performance (Question 4c) were explored.

As in Experiment 1, we expected a positive correlation between SA and JOL ratings (Hypothesis 5), and a negative correlation between SA and effort ratings and

between JOL and effort ratings in the control condition (Hypothesis 6), which could possibly be qualified by training or standards in the other conditions (Question 7). Finally, like in Experiment 1, we explored whether identical and isomorphic problems affected JOL accuracy differently (Question 8).

Method

Participants and Design

Participants were 133 Dutch ninth grade students ($M_{age} = 14.17$ years, $SD = 0.49$; 61 boys and 72 girls) from six different classrooms of two schools. Participants within each classroom were randomly assigned to one of the four conditions resulting from a 2×2 design with between-subjects factors Training and Standards: (1) control ($n = 35$), (2) SA training only ($n = 32$), (3) standards only ($n = 33$), and (4) SA training with standards ($n = 33$).

Materials

The materials and procedure in conditions 1 and 2 were the same as in Experiment 1. In conditions 3 and 4 the materials were the same except for the standards that were added to the training and learning phase. This meant that in the training phase, the students in the standards only condition had to find the mistakes in the worked-out solutions with the aid of a standard showing the correct solution to the problem. Students in the SA training with standard condition saw a self-assessment by the fictive student in which that student was aided by a standard (Appendix 3). Students in the standard only condition and in the SA training with standard condition could use a standard to make their SA in the learning phase.

Procedure

The study was run in five group sessions in students' classrooms at two schools, which lasted approximately 70 minutes, and all conditions, to which students were randomly assigned, were present in each session. The procedure was the same as in Experiment 1.

Data analysis

The procedure and the data analysis were the same as in Experiment 1.

Missing data. One student in the control condition, four students in the SA training only condition, and three students in the standards only condition, failed to fill out one or more SAs and therefore they were not included in analyses involving SA accuracy.

One student in the control condition, three students in the SA training only condition, three students in the standards only condition, and one student in the SA training with standards condition failed to fill out one or more JOLs, and therefore they were excluded in analyses involving JOL and regulation accuracy.

Two students in the control condition, two students in the standards only condition, and two students in the SA training with standards condition did not fill out one or more mental effort ratings and were therefore excluded from the analyses of invested mental effort.

Results

Table 3 presents the data per condition, both the original and computed measures.

As a check on randomization, we tested whether pretest performance differed between conditions. The assumption of homogeneity of variance was violated, therefore, a Kruskall-Wallis test was conducted which showed no significant differences among conditions, $H(3) = 4.51, p = .212$.

Table 3

The mean SA (*range*: 0-6), practice problem performance (*range*: 0-6), SA bias (*range*: -6-6), SA absolute deviation (*range*: 0-6), JOLs (*range*: 0-6), posttest performance on Identical and Isomorphic Test Tasks (*range*: 0-6), JOL bias on Identical and Isomorphic Test Tasks (*range*: -6-6), JOL absolute deviation on Identical and Isomorphic Test Tasks (*range*: 0-6), invested mental effort during the learning phase (*range*: 0-9) and regulation accuracy (*range*: 0-1) from Experiment 2 are presented.

	Control (1)	SA Training only (2)	Standards only (3)	SA training with standards (4)
SA	4.31 (1.44)	4.52 (1.04)	4.26 (1.06)	3.97 (1.32)
Practice problem performance	3.58 (1.50)	3.54 (1.67)	3.93 (1.07)	3.79 (1.35)
SA bias	0.76 (1.44)	0.81 (1.70)	0.30 (0.48)	0.18 (0.55)
SA absolute deviation	1.47 (1.02)	1.52 (1.16)	0.41 (0.46)	0.40 (0.57)
JOL	3.87 (1.67)	4.34 (1.15)	4.74 (0.95)	4.89 (0.99)
Posttest performance Identical Tasks	3.90 (1.63)	3.79 (1.74)	5.09 (1.10)	4.86 (1.30)
JOL bias Identical Test Tasks	0.04 (1.43)	0.55 (1.78)	-0.42 (1.30)	0.04 (1.17)
JOL absolute deviation Identical Test Tasks	1.39 (0.97)	1.70 (1.01)	1.18 (0.86)	1.04 (0.80)
Posttest performance Isomorphic Tasks	4.14 (1.44)	4.04 (1.55)	4.86 (1.29)	4.96 (1.25)
JOL bias Isomorphic Test Tasks	0.33 (1.54)	0.75 (1.78)	0.57 (1.20)	0.76 (1.15)
JOL absolute deviation Isomorphic Test Tasks	1.51 (1.00)	1.74 (1.01)	1.41 (0.82)	1.55 (1.01)
Invested mental effort (after practice problem)	5.06 (2.00)	4.63 (1.84)	4.27 (1.63)	4.46 (1.84)
Regulation accuracy	0.32 (0.26)	0.33 (0.29)	0.43 (0.33)	0.38 (0.34)

Monitoring Accuracy

SA accuracy. A 2 x 2 ANOVA with between-subjects factors Training and Standards showed no main effect of Training on bias, $F(1, 121) < 1, p = .861$, and absolute deviation, $F(1, 121) < 1, p = .880$ (Hypothesis 1a). There was a main effect of Standards on both bias (No Standards: $M = 0.78, SD = 1.55$ and Standards: $M = 0.24, SD = 0.52$), $F(1,$

121) = 6.87, $p = .010$, $\eta_p^2 = 0.05$ ⁷, and absolute deviation (No Standards: $M = 1.52$, $SD = 1.16$ and Standards: $M = 0.40$, $SD = 0.57$), $F(1, 121) = 51.10$, $p < .001$, $\eta_p^2 = 0.30$ ⁸, showing that bias and absolute deviation were lower (i.e., accuracy was higher) when students could use a standard (Hypothesis 1b). There was no interaction effect on bias, $F(1, 121) < 1$, $p = .696$, or absolute deviation, $F(1, 121) < 1$, $p = .844$ (Hypothesis 1c).

JOL accuracy. A repeated measures ANOVA with Type of Test Task (Identical vs. Isomorphic) as within-subjects factor and Training and Standard as between-subjects factors showed a main effect of Type of Test Task, indicating that both bias (Identical: $M = 0.05$, $SD = 1.45$ and Isomorphic: $M = 0.59$, $SD = 1.43$), $F(1, 121) = 43.82$, $p < .001$, $\eta_p^2 = 0.27$, and absolute deviation (Identical: $M = 1.32$, $SD = 0.94$ and Isomorphic: $M = 1.55$, $SD = 0.96$), $F(1, 121) = 9.65$, $p = .002$, $\eta_p^2 = 0.07$, were smaller (i.e., accuracy was higher) for Identical than for Isomorphic test tasks (Question 8). There was no significant main effect of Training on bias, $F(1, 121) = 2.66$, $p = .106$, or absolute deviation, $F(1, 121) < 1$, $p = .378$ (Hypothesis 2a). There was no significant main effect of Standards on bias, $F(1, 121) < 1$, $p = .457$ or absolute deviation, $F(1, 121) = 3.60$, $p = .060$, $\eta_p^2 = 0.03$ (Hypothesis 2b), nor was there a significant interaction between Training and Standards on bias, $F(1, 121) < 1$, $p = .783$, and absolute deviation, $F(1, 121) < 1$, $p = .386$ (Hypothesis 2c).

Whereas there was no interaction between Type of Test Task and Training on bias, $F(1, 121) = 1.24$, $p = .269$, $\eta_p^2 = 0.01$, or absolute deviation, $F(1, 121) < 1$, $p = .503$, the analysis did reveal a significant interaction between Type of Test Task and Standards on both bias, $F(1, 121) = 13.47$, $p < .001$, $\eta_p^2 = 0.10$, and absolute deviation, $F(1, 121) = 4.21$, $p = .042$, $\eta_p^2 = 0.03$ (Question 8). These interactions were tested further using (paired) t-tests. There was no difference in bias scores between students who did and did not get standards on Identical, $t(123) = 1.78$, $p = .078$, and Isomorphic Test Tasks, $t(123) < 1$, $p = .578$. Within group paired t-tests showed that the difference in bias between Identical and Isomorphic Test Tasks did not reach significance in the No Standards group (after Bonferroni correction: $.05/4 = .0125$), $t(62) = -2.14$, $p = .036$, but did reach significance in

⁷ The assumption of homogeneity of variance was violated, however, because a Mann-Whitney U test, $U = 1442.00$, $p = .010$, led to the same conclusions, we decided to report the results from the ANOVA.

⁸ The assumption of homogeneity of variance was violated, however, because a Mann-Whitney U test, $U = 624.50$, $p < .001$, led to the same conclusions, we decided to report the results from the ANOVA.

the Standards group (Identical: $M = -0.18$, $SD = 1.25$ and Isomorphic: $M = 0.67$, $SD = 1.17$), $t(61) = -7.22$, $p < .001$, indicating that bias scores were lower (i.e., accuracy was higher) on Identical Test Tasks than on Isomorphic Test Tasks.

For absolute deviation scores, there was a difference between No Standards and Standards groups for Identical Test Tasks (No standard: $M = 1.53$, $SD = 0.99$ and Standard: $M = 1.11$, $SD = 0.83$), $t(123) = 2.61$, $p = .010$, but not for Isomorphic Test Tasks, $t(123) < 1$, $p = .453$. Students in the Standard group showed a lower absolute deviation (i.e., accuracy was higher) on Identical Test Task compared to students in the No standard group. Within group paired t-tests showed that the difference between Identical and Isomorphic Test Tasks was not significant for the No Standards group, $t(62) < 1$, $p = .407$, but for students who could use Standards, absolute differences were lower (i.e., accuracy was higher) on Identical ($M = 1.11$, $SD = 0.83$) than on Isomorphic Test Tasks ($M = 1.48$, $SD = 0.92$), $t(61) = -3.48$, $p = .001$.

No interaction effect of Type of Test Tasks, Training and Standards was found for bias, $F(1, 121) < 1$, $p = .08$, and absolute deviation, $F(1, 121) = 1.56$, $p = .21$.

SA and JOLs. To test Hypothesis 5 that SA's and JOLs would be positively correlated, multi-level analysis in Mplus was used to calculate correlations between SAs and JOLs ratings. The data was nested and the ICC's indicated that the variance explained by the person level ranged from 22.5% for SA ratings to 60.1% for JOL ratings. We calculated these correlations for each condition separately because the data per factor (Training and Standards) would be cross classified, that is, half of the participants who received Training were also provided a Standard and vice versa. In three conditions SAs were significantly correlated to JOLs; control (1): $r = .87$, $p < .000$, SA training only (2): $r = .80$, $p < .000$, SA training with standard (4): $r = .75$, $p < .000$. In the standard only condition there was no significant correlation (3): $r = .63$, $p = .38$. Correlations in condition 1 vs. 2, Fisher's $z = 0.91$, $p = .363$, 1 vs. 3, Fisher's $z = 2.33$, $p = .020$, 1 vs. 4, Fisher's $z = 1.42$, $p = .156$, 2 vs. 3, Fisher's $z = 1.37$, $p = .171$, 2 vs. 4, Fisher's $z = 0.48$, $p = .631$, and 3 vs. 4, Fisher's $z = -0.90$, $p = .368$, did differ significantly (with a Bonferroni corrected alpha set at $p = .008$).

Regulation Accuracy

A 2 x 2 ANOVA showed no main effect of Training, $F(1, 120) < 1, p = .731$ (Question 3a), no main effect of Standards, $F(1, 120) = 1.85, p = .176$ (Question 3b), and no interaction between Training and Standards on regulation accuracy, $F(1, 120) < 1, p = .583$ (Question 3c).

Invested Mental Effort

To test Hypothesis 6 that SAs and JOLs would show a negative correlation with effort ratings in the control condition and explore Question 7 about whether training and standards would affect this correlation, correlations between SA, JOL and mental effort ratings were calculated. Again, we used multi-level analysis in Mplus to calculate correlations because the data was nested and the ICC's indicated that the variance explained by the person level ranged from 22.5% for SA ratings, 60.1% in JOL ratings, to 44.9% for mental effort ratings. We calculated these correlations for each condition separately because the data per factor (Training and Standards) would be cross classified. Mental effort ratings given after the practice problems were significantly negatively correlated to SAs in the control condition (1): $r = -.78, p < .001$, in the SA training only condition (2): $r = -.80, p < .000$, and in the SA training with standards condition (4): $r = -.97, p < .000$, but not in the standards only condition (3): $r = -.35, p = .72$. Correlations in condition 1 vs. 2 did not differ significantly, Fisher's $z = 0.21, p = .834$. Correlations in condition 1 vs. 3, Fisher's $z = 4.12, p < .001$, 1 vs. 4, Fisher's $z = -2.68, p = .007$, 2 vs. 3, Fisher's $z = 3.82, p < .001$, 2 vs. 4, Fisher's $z = -2.82, p = .005$, and 3 vs. 4, Fisher's $z = -6.69, p < .001$, did differ significantly (with a Bonferroni corrected alpha set at $p = .008$). Mental effort ratings were significantly negatively correlated to JOLs in the control condition (1): $r = -.71, p < .000$, in the SA training only condition (2): $r = -.74, p < .000$, in the standard only condition (3): $r = -.55, p = .001$, and in the SA training with standard condition (4): $r = -.71, p < .000$. Correlations in condition 1 vs. 2, Fisher's $z = 0.25, p = .803$, 1 vs. 3, Fisher's $z = -1.06, p = .289$, 1 vs. 4, Fisher's $z = 0.00, p = 1$, 2 vs. 3, Fisher's $z = -1.28, p = .201$, 2 vs. 4, Fisher's $z = -0.02, p = .810$, and 3 vs. 4, Fisher's $z = 1.04, p = .298$, did differ significantly (with a Bonferroni corrected alpha set at $p = .008$).

Posttest Performance

A repeated measures ANOVA with type of Test Task (Identical vs. Isomorphic) as within-subjects factor and Training and Standard as between-subjects factors, showed no main effect of Type of Test Task, $F(1, 129) = 1.81, p = .181$, and no main effect of Training $F(1, 129) < 1, p = .725$ (Questions 4a). There was a main effect of Standards, $F(1, 129) = 16.70, p < .001, \eta_p^2 = 0.12$, indicating that students who used Standards ($M = 4.94, SD = 1.17$) scored higher on the posttest than students who did not ($M = 3.97, SD = 1.53$) (Question 4b). There was no interaction between Training and Standards, $F(1, 129) < 1, p = .939$ (Question 4c).

There was no interaction between Type of Test Task and Training, $F(1, 129) = 1.52, p = .220$, but there was an interaction between type of Test Task and Standards, $F(1, 129) = 5.34, p = .022, \eta_p^2 = 0.04$. Follow-up t-tests showed that students' posttest performance was significantly higher in the Standards group compared to the No Standards group on both Identical (No standard: $M = 3.85, SD = 1.68$ and Standard: $M = 4.97, SD = 1.20$), $t(119.63) = -4.48, p < .001$ ⁹, and Isomorphic Test Tasks (No standard: $M = 4.09, SD = 1.48$ and Standard: $M = 4.90, SD = 1.27$), $t(131) = -3.41, p = .001$. Within the No Standards group, a paired t-test showed that the posttest performance was higher on Isomorphic ($M = 4.09, SD = 1.15$) than Identical Test Tasks ($M = 3.85, SD = 1.68$), $t(66) = -2.63, p = .011$ ¹⁰, but this was not the case for the Standard group, $t(61) < 1, p = .507$. No interaction effect of Type of Test Tasks, Training and Standards was found, $F(1, 129) = 1.48, p = .226$.

General Discussion

Interestingly, providing students with assessment standards was found to improve SA accuracy in Experiment 2 (Hypothesis 1b): both bias and absolute deviation of self-

⁹ The assumption of homogeneity of variance was violated, however, because a Mann-Whitney U test, $U = 3147.00, p < .001$, led to the same conclusions, we decided to report the results from the ANOVA.

¹⁰ The assumption of homogeneity of variance was violated, however, because a Mann-Whitney U test, $U = 634.50, p = .021$, led to the same conclusions, we decided to report the results from the ANOVA.

assessed performance compared to actual performance were smaller when students could use a standard when assessing their performance on the practice problem. Also, standards improved JOL accuracy (Hypothesis 2b) but only on test problems that were identical to the practice problems in the learning phase, not on isomorphic test problems (Question 8). While the use of standards did not affect regulation accuracy (Question 3b), it did improve posttest performance (Question 4b). There were no interactions of standards with training (Hypotheses 1c, 2c, 3c and 4c).

The finding that standards enhance retrospective monitoring accuracy also for problem-solving tasks is in line with and adds to prior research on the use of standards when learning key concepts (Lipko et al., 2009; Rawson & Dunlosky, 2007). Like Rawson and Dunlosky (2007), we found that using standards improved self-assessment and test performance. Another novel and interesting contribution of our study is the finding that standards used in retrospective monitoring can also improve prospective monitoring. However, our results also show that this effect was limited to identical tasks, meaning that a standard used when assessing performance on a task that was just practiced, leads to a more accurate estimation of future performance on that exact same task, but not to more accurate estimation of future performance on similar tasks, which is arguably more important in education. This finding might indicate that the prediction of future test performance might be based on surface characteristics of the practice problem, instead of structural features (i.e., understanding of the solution procedure). Future research might investigate what cues students use when making JOLs (after SAs) and which strategies could focus students' attention more on their deeper understanding of the problem-solving task in order to improve JOL accuracy for isomorphic tasks.

Interestingly, using the standards (which inherently provide feedback about the correct answer), also enhanced students test performance, and not only on identical but also on isomorphic tasks. Note that the standards provided only the correct outcome of each step, they did not provide fully worked-out steps, so this effect on learning does not seem to be an artifact of having studied more examples. Thus, it seems standards are an instructional tool that is useful for fostering learning outcomes as well as monitoring accuracy. Unfortunately, it seems of limited use for fostering self-regulated learning, as our study showed no effects of standards on regulation of study. This suggests that although accurate monitoring can be considered prerequisite for accurate regulation, it does not seem

to be sufficient when it comes to problem-solving tasks (cf. the findings by Kostons et al., 2012, showing that both SA and task selection skills had to be trained).

As for the effects of training, the findings from Experiment 2 are in line with those from Experiment 1. Despite the slightly older students who indeed learned more from these tasks than students in Experiment 1, results from Experiment 2 showed that SA training did not improve SA accuracy (Hypothesis 1a), JOL accuracy (Hypothesis 2a), regulation accuracy (Question 3a), or posttest performance (Question 4a). Thus, across two experiments we failed to replicate the findings by Kostons et al. (2012) regarding training of self-assessment skills, but despite highly similar content, there were two main differences between our training and theirs: we used paper-based examples instead of video-modeling examples, and we used only two instead of four examples in the training phase. Whereas their video-modeling examples progressed step-by-step, making it easy for the learners to follow along without other distracting information, learners might not have processed all of the steps in our paper-based examples equally well or in a coherent order. On the other hand, research comparing learning from written text with learning from videos has shown no differences in terms of learning outcomes or even superiority of texts when learners have no control over the videos (see Merkt, Weigand, Heier, & Schwan, 2011), which suggests that the reduction in the number of training examples might have been the more crucial factor. Given the importance of SA accuracy in self-regulated learning, future research should address which of these factors is crucial for establishing beneficial effects of SA training on SA accuracy.

Furthermore, the explorative analyses yielded some interesting results. First, the finding that students' SAs were highly positively correlated to their JOLs in Experiments 1 and 2 (with the exception of the standards only condition), suggests that students do indeed use their assessment of performance on a practice problem as a cue for predicting future performance (cf. the postdiction route to JOLs, Griffin et al., 2009). However, the finding that SA and JOLs were highly correlated suggests that SA ratings and JOL ratings were filled out the same, possibly because of the design of both experiments in which SA and JOL ratings were made in near proximity or because students did not differentiate between both ratings. Second, we replicated previous findings (Baars et al., 2013) regarding the negative relation between JOLs and invested mental effort (i.e., the higher the effort students invested, the lower their JOLs), for both SAs and JOLs, which might suggest that

students use invested effort as a cue for making both retrospective and prospective monitoring judgments. However, it should be noted that the interesting reduction of this negative relation in the training condition in Experiment 1, suggesting that training directed students attention to other cues than effort upon which to base their judgments, was not replicated in Experiment 2; and in Experiment 2, the standards only condition did not show this negative relation for SAs, only for JOLs. It is not entirely clear what causes these differences, and future research should therefore further investigate whether and how students use invested mental effort in making monitoring judgments.

A limitation to the current classroom study is that because of time limits, only three problem-solving tasks were used in the learning phase. Future research should investigate JOL accuracy with more problems solving tasks, especially because more problem-solving task at each complexity level might influence JOL ratings due to practice.

In sum, this study showed that providing students with standards that show the correct answer to each problem-solving step, which they can use to self-assess their performance on a practice problem not only has positive effects on self-assessment accuracy, but also on predictions of future performance of that same problem, and on posttest performance of the same and similar problems. This finding would need to be replicated in future research, but because it is easy to implement in education, standards might be a promising tool for improving students learning and monitoring accuracy.

Appendix 1

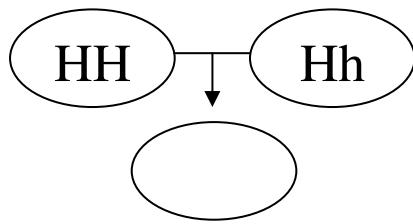
Worked example – practice problem pair

Worked example: 1 generation with a homozygote parent and a heterozygote parent

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
<p>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.</p> <p>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).</p> <p>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.</p>	HH
<p>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.</p> <p>We also know that the mother is heterozygote for hair. When a person is heterozygote for a feature then both genes in the genotype are different. In this example it means that the mother has the genotype Hh.</p>	Hh
<p>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.</p>	Answer 

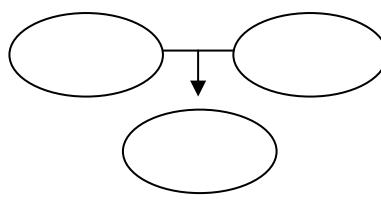
<p>Step 4. Make a crosstable to mix the genotypes of the parents and put down the possible genotypes for their children</p> <p>e. Make a crosstable and divide the genes of the genotypes of the mother in the two cells of the upper row and the genes of the genotypes of the father in the left column.</p> <p>f. Fill out the crosstab by combining the genes of the father and the mother.</p> <p>g.</p>	<p>Answer</p> <table border="1" data-bbox="917 190 1171 359"> <tr> <td></td> <td></td> <td colspan="2">Annie Hh</td> </tr> <tr> <td></td> <td></td> <td>H</td> <td>h</td> </tr> <tr> <td>Josh</td> <td>H</td> <td>HH</td> <td>Hh</td> </tr> <tr> <td>HH</td> <td>H</td> <td>HH</td> <td>Hh</td> </tr> </table>			Annie Hh				H	h	Josh	H	HH	Hh	HH	H	HH	Hh
		Annie Hh															
		H	h														
Josh	H	HH	Hh														
HH	H	HH	Hh														
<p>Step 5. Determine the possible genotypes for hair for the children and the chance to get those genotypes</p>	<p>Answer</p>																
<p>GENOTYPE = GENES</p> <table border="1" data-bbox="287 572 563 729"> <tr> <td></td> <td></td> <td colspan="2">Annie Hh</td> </tr> <tr> <td></td> <td></td> <td>H</td> <td>h</td> </tr> <tr> <td>Josh</td> <td>H</td> <td>HH</td> <td>Hh</td> </tr> <tr> <td>HH</td> <td>H</td> <td>HH</td> <td>Hh</td> </tr> </table> <p>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.</p>			Annie Hh				H	h	Josh	H	HH	Hh	HH	H	HH	Hh	<p>50% HH and 50% Hh</p>
		Annie Hh															
		H	h														
Josh	H	HH	Hh														
HH	H	HH	Hh														
<p>In this example: two cells have HH = 50% en two cells have Hh = 50%.</p> <p>Step 6. Determine the possible phenotypes for hair for the children and the chance to get those phenotypes</p> <p>PHENOTYPE = LOOKS</p> <p>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)</p> <p>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.</p>	<p>Answer</p> <p>100% curly hair</p>																

Practice problem: 1 generation with a homozygote parent and a heterozygote parent

Given:

1. The gene for thick hair (T) dominates the gene for thin hair (t).
2. The father Peter has thick hair.
3. The mother Maria has thin hair too.
4. Peter has a homozygote genotype and Maria has a heterozygote genotype.

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

Step	Answer
Step 1. Determine the genotype for father's hair	
Step 2. Determine the genotype for mother's hair	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 <table border="1"><tr><td></td><td></td><td>Maria</td></tr><tr><td></td><td></td><td></td></tr><tr><td>Peter</td><td></td><td></td></tr><tr><td></td><td></td><td></td></tr></table>			Maria				Peter					
		Maria											
Peter													
Step 5. Determine the possible genotypes for hair for the children and the chance to get those genotypes	Answer												
Step 6. Determine the possible phenotypes for hair for the children and the chance to get those phenotypes	Answer												

Appendix 2
SA Training
Self-assessment

Later on you will study several worked-out examples and can learn how to solve similar problems from these examples. You will be asked to assess your own performance on these problems. Below you will read how you can assess your own performance and you are shown two examples of how two students have correctly judged their own performance.

Each problem consists of 6 steps. Each step is important, so for each step performed correctly 1 point should be assigned. The minimum score is 0 points (no step was performed correctly) and the maximum score is 6 points (all the steps were performed correctly). On the next pages you will see two examples of other students who solved a problem and assessed their own performance. Study these examples carefully, you should be able to assess your own performance later on.

Example 1

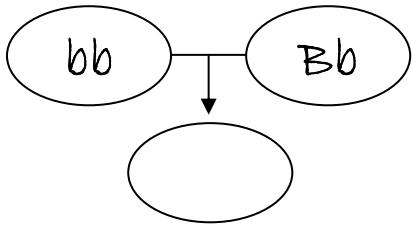
This problem was solved by Pete.

Problem 3

Given:

1. The genotype for brown hair (B) dominates the genotype for blond hair (b).
2. Dad Otto and mum Helen both have blond hair and are homozygous for this trait.
3. The genotype for hair color of daughter Nienke is unknown.
4. Daughter Nienke gets a baby (Paula) together with her husband Roel.
5. Roel has brown hair and is heterozygous for this trait.

Question: What could the genotype and phenotype for hair color of baby Paula be?

Step 1. Determine the genotype for hair color of Paula's mother	Answer bb															
Step 2. Determine the genotype for hair color of Paula's father	Answer Bb															
Step 3. Make a family tree	Answer 															
Step 4. Make a crosstable to combine genotypes of the mother and father and fill out the possible genotypes of their children.	Answer <table border="1" data-bbox="801 1156 1278 1313"> <thead> <tr> <th></th> <th></th> <th colspan="2">Mother</th> </tr> <tr> <th></th> <th></th> <th>B</th> <th>B</th> </tr> </thead> <tbody> <tr> <th rowspan="2">Father</th> <th>B</th> <td>Bb</td> <td>Bb</td> </tr> <tr> <th>b</th> <td>bb</td> <td>bb</td> </tr> </tbody> </table>			Mother				B	B	Father	B	Bb	Bb	b	bb	bb
		Mother														
		B	B													
Father	B	Bb	Bb													
	b	bb	bb													
Step 5. Determine the possible <u>Genotypes</u> for hair color and the chance to get the possible genotypes	Answer 50% Bb and 50% bb															
Step 6. Determine the possible <u>Phenotypes</u> for hair color and the chance to get the possible phenotypes	Answer 100% brown hair															

Below you can see how he self-assessed his performance. He circled per step whether or not he performed it correctly. Look at how he self-assessed his performance in relation to his performance.

Step 1.	Was your answer correct? <input checked="" type="radio"/> Yes <input type="radio"/> No
Step 2.	Was your answer correct? <input checked="" type="radio"/> Yes <input type="radio"/> No
Step 3.	Was your answer correct? <input checked="" type="radio"/> Yes <input type="radio"/> No
Step 4.	Was your answer correct? <input checked="" type="radio"/> Yes <input type="radio"/> No
Step 5.	Was your answer correct? <input checked="" type="radio"/> Yes <input type="radio"/> No
Step 6.	Was your answer correct? <input type="radio"/> Yes <input checked="" type="radio"/> No

How many steps did you perform correctly in total?						
0	1	2	3	4	5	<input checked="" type="radio"/> 6

Appendix 3
SA training with standards

Self-assessment

Later on you will study several worked-out examples and can learn how to solve similar problems from these examples. You will be asked to assess your own performance on these problems. Below you will read how you can assess your own performance and you are shown two examples of how two students have correctly judged their own performance.

Each problem consists of 6 steps. Each step is important, so for each step performed correctly 1 point should be assigned. The minimum score is 0 points (no step was performed correctly) and the maximum score is 6 points (all the steps were performed correctly). On the next pages you will see two examples of other students who solved a problem and assessed their own performance. Study these examples carefully, you should be able to assess your own performance later on.

Example 1

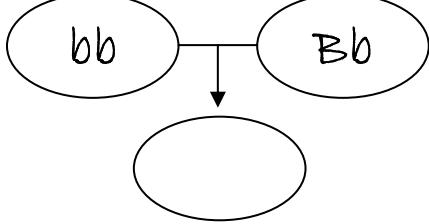
This problem was solved by Pete.

Problem 3

Given:

1. The genotype for brown hair (B) dominates the genotype for blond hair (b).
2. Dad Otto and mum Helen both have blond hair and are homozygous for this trait.
3. The genotype for hair color of daughter Nienke is unknown.
4. Daughter Nienke gets a baby (Paula) together with her husband Roel.
5. Roel has brown hair and is heterozygous for this trait.

Question: What could the genotype and phenotype for hair color of baby Paula be?

Step 1. Determine the genotype for hair color of Paula's mother	Answer bb															
Step 2. Determine the genotype for hair color of Paula's father	Answer Bb															
Step 3. Make a family tree	Answer 															
Step 4. Make a crosstable to combine genotypes of the mother and father and fill out the possible genotypes of their children	Answer <table border="1" data-bbox="801 1179 1278 1336"> <thead> <tr> <th></th> <th></th> <th colspan="2">Mother</th> </tr> <tr> <th></th> <th></th> <th>B</th> <th>B</th> </tr> <tr> <th rowspan="2">Father</th> <th>B</th> <td>Bb</td> <td>Bb</td> </tr> </thead> <tbody> <tr> <th>b</th> <td>bb</td> <td>bb</td> </tr> </tbody> </table>			Mother				B	B	Father	B	Bb	Bb	b	bb	bb
		Mother														
		B	B													
Father	B	Bb	Bb													
	b	bb	bb													
Step 5. Determine the possible <u>Genotypes</u> for hair color and the chance to get the possible genotypes	Answer 50% Bb and 50% bb															
Step 6. Determine the possible <u>Phenotypes</u> for hair color and the chance to get the possible phenotypes	Answer 100% brown hair															

Below you can see how he self-assessed his performance. He circled per step whether or not he performed it correctly. Look at how he self-assessed his performance with the correct answers available.

Step 1.	Answer bb	Was your answer correct? <input checked="" type="radio"/> Yes <input type="radio"/> No																
Step 2.	Answer Bb	Was your answer correct? <input checked="" type="radio"/> Yes <input type="radio"/> No																
Step 3.	Answer 	Was your answer correct? <input checked="" type="radio"/> Yes <input type="radio"/> No																
Step 4.	Answer <table border="1"><tr><td></td><td></td><td colspan="2">Mother</td></tr><tr><td></td><td></td><td>b</td><td>b</td></tr><tr><td>Father</td><td>B</td><td>Bb</td><td>Bb</td></tr><tr><td></td><td>b</td><td>bb</td><td>bb</td></tr></table>			Mother				b	b	Father	B	Bb	Bb		b	bb	bb	Was your answer correct? <input checked="" type="radio"/> Yes <input type="radio"/> No
		Mother																
		b	b															
Father	B	Bb	Bb															
	b	bb	bb															
Step 5.	Answer 50% Bb and 50% bb	Was your answer correct? <input checked="" type="radio"/> Yes <input type="radio"/> No																
Step 6.	Answer 50% brown hair en 50% blond hair	Was your answer correct? <input checked="" type="radio"/> Yes <input type="radio"/> No																

How many steps did you perform correctly in total?

0 1 2 3 4 5 6

Chapter 7

Summary and General Discussion

Summary and General Discussion

In this final chapter the results from the studies presented in this dissertation are summarized and the main findings are discussed, along with ideas for future research.

Monitoring and regulation skills are central to the process of self-regulated learning, that is, to be able to learn in a self-regulated way one must be able to monitor one's own learning process and regulate further learning accordingly (Winne & Hadwin, 1998). Even though there has been quite some research on improving monitoring and regulation skills when learning word pairs (see Rhodes & Tauber, 2011) and learning from text (Thiede, Griffin, Wiley & Redford, 2009), very few studies have investigated monitoring judgments when learning to solve problems. Therefore, the main research questions in this dissertation were: a) whether students are able to make accurate prospective judgments, like Judgments of Learning (JOLs), and use these to control their learning process when solving or learning to solve well-structured problems, b) whether delaying JOLs about problem-solving tasks leads to higher accuracy, c) whether generation strategies are effective for improving JOL accuracy when learning to solve problems by means of worked example study, and d) whether retrospective monitoring (self-assessment) of problem-solving can also be improved by self-assessment training and providing students with standards of the correct solution, and whether improvement in retrospective monitoring will lead to improvement in prospective monitoring accuracy. Furthermore, effects of task complexity on monitoring accuracy, relations between JOL and invested mental effort, and effect of type of test task on monitoring accuracy were explored.

Summary of the Main Findings

Chapter 2 presented an experiment that investigated primary school children's monitoring and regulation accuracy when solving arithmetic problems. Results indicate that the children were able to monitor the complexity of the problem-solving tasks. That is, with increasing complexity of the problem-solving tasks, performance decreased, and subjective ratings of mental effort increased while Judgments of Learning (JOLs) ratings decreased. However, relative accuracy of JOLs, which shows the ability to discriminate between tasks, was moderate to low. The mean gamma correlation in the immediate JOLs condition differed significantly from zero but this was not the case for the delayed JOLs condition.

That is, immediate JOLs were moderately accurate whereas delayed JOLs were not accurate. Furthermore, considering relative accuracy, immediate JOLs seemed to be more accurate than delayed JOLs, but this difference was not statistically significant ($p = .068$). The absolute accuracy data also showed a numerical trend in the same direction with accuracy of immediate JOLs being higher, but again, this difference was not statistically significant ($p = .123$). There was a significant negative correlation between mental effort and JOLs, indicating that the higher the mental effort invested, the lower the JOLs were. This suggests that effort invested in studying could have been used as a cue when making a JOL. Test performance did not differ between the condition in which students gave immediate JOLs and the condition in which students gave delayed JOLs. In sum, results from the study presented in Chapter 2 showed that in contrast to earlier research on learning word pairs and learning from expository texts, immediate JOLs were moderately accurate and tended to be somewhat (though not significantly) more accurate than delayed JOLs when solving problems in primary education.

In Chapters 3 and 4 two experiments were reported, in primary and secondary education, respectively, which investigated the effect of practice problems after worked example study. The main question was whether this would be a useful generation strategy to improve monitoring and regulation accuracy.

The experiment reported in Chapter 3 showed that absolute accuracy of immediate and delayed JOLs after worked example study did not show the same trend found for immediate and delayed JOLs after problem solving as described in Chapter 2. In line with our hypothesis, the opportunity to solve a practice problem after worked example study was found to decrease bias, but this was the case regardless of the timing of the practice problem. That is, on average overconfidence was smaller for students in the conditions in which they were provided with practice problems after worked example study. However, no difference between conditions with and without practice problems in absolute deviation between JOLs and actual test performance was found. Even though providing primary school children with practice problems improved their monitoring accuracy, it did not affect their regulation accuracy. Furthermore, children showed more accurate JOLs (both bias and absolute deviation) on the less complex tasks. This is probably caused by higher JOLs on the more complex tasks while performance was lower there. Although students in the condition with practice problems had an extra learning opportunity (i.e., to apply what they

had learned from the worked example), there were no differences among conditions with and without practice problems in test performance.

In the experiment reported in Chapter 4, JOLs were measured per step of the worked example or practice problem; therefore, relative accuracy and a new measure of absolute accuracy were used to analyze JOL accuracy. Results showed, in line with findings from primary education presented in Chapter 3, no significant difference between immediate and delayed JOL accuracy after practice problems or worked examples. However, numerically absolute accuracy was higher for immediate JOLs compared to delayed JOLs after practice problems. Again, as hypothesized, practice problems functioned as a generation strategy that helped secondary education students to make more accurate JOLs, as indicated by a higher absolute accuracy for students who worked on practice problems after worked examples study. Furthermore, in contrast to our expectation that delayed practice problems would lead to the highest JOL accuracy, both relative accuracy ($p = .090$) and absolute accuracy ($p = .092$) were marginally significantly higher for students who solved practice problems immediately after worked example study. Moreover, and in contrast to the findings reported in Chapter 3, regulation accuracy was higher for students who were provided with practice problems after worked example study. However, results did not show the expected positive effect of enhanced regulation accuracy on final test performance. Possibly, because the final test consisted of problems which were isomorphic to the problems studied or practiced, no effect of improved monitoring and regulation accuracy on final test performance was found.

In sum, results from Chapters 3 and 4 show that using practice problems helped primary and secondary education students monitor their learning. In addition, secondary education students in conditions with practice problems showed higher regulation accuracy than students in conditions without practice problems.

The study in Chapter 5 focused on the effects of an immediate generation strategy, that is, completion of partially worked-out examples, on secondary education students' monitoring and regulation accuracy. Results showed that completing partially worked-out examples led to significant underestimation of future performance compared to studying worked examples. However, there was no significant difference between conditions in absolute deviation between JOL and test performance. Also, regulation accuracy and learning outcomes did not differ between the two conditions. Regarding task complexity, an

effect of the complexity of the tasks on bias was found: averaged over both conditions students underestimated their future test performance on the lowest complexity level but not on the two higher complexity levels. This might be due to the fact that on average, JOLs tended to increase but performance tended to decrease when the tasks became more complex. This study also showed an effect of the type of test task on monitoring accuracy. That is, monitoring accuracy in terms of absolute deviation was better for test problems that were identical to the examples in the learning phase compared to test problems that were isomorphic to the examples in the learning phase. This study also again showed a significant negative correlation between invested mental effort and JOLs (see also Chapter 2), which suggests that effort invested in studying could have been used as a cue when making a JOL. In sum, results from the experiment reported in Chapter 5 showed that completion of partially worked-out examples affects monitoring accuracy in the sense that it caused students to underestimate their future test performance, while regulation accuracy was not affected.

In Chapter 6 a study was presented on the use of self-assessment (SA) training and standards to improve secondary education students' monitoring and regulation accuracy. This study consisted of two experiments. Experiment 2 showed that providing students with standards of the correct answers to each problem-solving step, not only had positive effects on SA accuracy, but also on JOL accuracy for *identical* future test problems (not for isomorphic problems; see also Chapter 5), and on test performance. Experiments 1 and 2 showed that SA training did not improve SA accuracy, JOL accuracy, regulation accuracy, or test performance. SA ratings were highly positively correlated to JOLs in both Experiments 1 and 2 (with the exception of the standards-only condition in Experiment 2), so it seems that students used SA of performance on a practice problem as a cue for predicting their future performance on a similar task. Moreover, in line with results from Chapters 2 and 5, invested mental effort was found to correlate negatively with both SAs and JOLs. That is, the higher the mental effort students invested, the lower their SAs and JOLs were, which suggests that students used invested mental effort as a cue for making both retrospective (SA) and prospective (JOL) monitoring judgments. In sum, whereas SA training did not affect retrospective and prospective monitoring accuracy or regulation accuracy, using standards of the correct solution to assess practice problem performance did

affect both retrospective and prospective monitoring accuracy and test performance, but not regulation accuracy.

Discussion of the Main Findings

In line with the findings from studies on JOL accuracy when learning from text, no delayed-JOL effect (Nelson & Dunlosky, 1991) was found with problem-solving tasks (Chapter 2), with worked example study (Chapters 3 and 4), or with practice problems after worked example study (Chapter 4). Just like the generation strategies that were found to improve JOL accuracy when learning from text (e.g., generating keywords: De Bruin, Thiede, Camp, & Redford, 2011; Thiede, Anderson, & Therriault, 2003; summarizing: Thiede & Anderson, 2003; self-explaining: Griffin, Wiley, & Thiede, 2008; concept mapping: Redford, Thiede, Wiley, & Griffin, 2012), solving practice problems after worked example study (Chapters 3 and 4) and completion of partially worked examples (Chapter 5) were found to influence JOL accuracy when learning to solve problems. Also, in line with findings from text research (Lipko et al., 2009; Rawson & Dunlosky, 2007), providing standards of the correct solution to a problem helped students to make more accurate retrospective (i.e., self-assessments) and prospective judgments (i.e., JOLs) about identical test tasks, and improved test performance (Chapter 6). So, the findings in this dissertation extend the findings from the domain of language learning to the domain of learning to solve problems in primary and secondary education. In addition, some issues and questions have arisen from the results in this dissertation which will be discussed in this section.

Timing of JOLs. One of the main questions investigated in Chapters 2, 3, and 4 concerned the timing of JOLs in accordance to the learning materials. Whereas research on less complex materials that require the learner to remember words or pictures has shown delayed JOLs to be more accurate than immediate JOLs (for a review see Rhodes & Tauber, 2011), it seems to be different for more complex materials like comprehending a text or solving a problem-solving task. Studies on learning from expository texts found that relative accuracy of both immediate and delayed JOLs was very low and not significantly different (Maki, 1998; Thiede, Griffin, Wiley, & Redford, 2009). The problem-solving tasks studied in this dissertation are also complex materials (i.e., require processing of many interacting information elements; Sweller, Van Merriënboer, & Paas, 1998), but in contrast to learning from expository texts, learning to solve problems requires learners to

remember the *procedure* for solving the problem, to *understand* that procedure (because the aim is to be able to solve not just that one problem, but also isomorphic problems) and to *apply* that procedure on similar problems in the future. Monitoring therefore involves judging whether a solution procedure has been understood and how well it will be remembered/can be applied in the future. Interestingly, as with expository texts, no delayed-JOL effect was found with problem-solving tasks. The study described in Chapter 2 even suggested that immediate JOLs after problem-solving tasks were somewhat more accurate. On the one hand, this seems logical because students get feedback from problem solving; that is, when making an immediate JOL they still have information on how fast or how easily they were able to solve a problem and can use that as a cue to make a JOL. On the other hand, it should be noted that studies described in Chapter 3 and 4 did not find immediate JOLs to be more accurate than delayed JOLs after studying worked examples (Chapters 3 and 4) or after solving practice problems following a worked example (Chapter 4). In Chapter 3 it was pointed out that the immediate JOLs might not be more accurate because worked examples are different from problem-solving tasks, however, the results in Chapter 4 show that immediate JOLs were not more accurate after practice problems either. However, in Chapter 4 JOLs were asked per step of the problem-solving task. The steps in the problem-solving task were cumulative and if the sequence of steps was performed correctly, the whole problem was solved. When giving a JOL per step, one focuses on whether one is able to solve that specific step in contrast to a JOL about the whole problem which focuses one on whether he or she is able to solve the whole problem. Whereas it might be easier to judge whether one will be able to solve the whole problem directly after actually solving it, judging whether one can solve a specific step in a problem solving procedure might also be relatively easily done after a delay because it is a more specific question. Nevertheless, it seems that overall we can conclude, in line with findings from monitoring learning from text (Maki, 1998, Thiede et al., 2009), that immediate and delayed JOL accuracy when solving problems and learning to solve problems is low to moderate and does not differ significantly.

Generation strategies. It should be noted though, that monitoring accuracy when learning from texts improved substantially when so-called generation strategies were used, which are instructional activities that provided students with cues about their understanding and therefore helped them monitor their learning (Thiede, Dunlosky, Griffin, & Wiley,

2005; Thiede, Griffin, Wiley, & Redford, 2009). Delayed keyword generation (De Bruin, et al., 2011; Thiede et al., 2003), delayed summary writing (Thiede & Anderson, 2003), immediate self-explaining (Griffin et al., 2008) and making concept maps immediately (Redford et al., 2012) all improved monitoring accuracy when learning from text. Because these strategies gave the students the opportunity to immediate (self-explaining and concept maps) or delayed (delayed keyword generation and summarizing) access to their situation model in which their understanding of the text resides, JOL accuracy was improved (Thiede et al., 2009). Indeed, the studies presented in this dissertation showed that adding a generation strategy also helps students monitor their learning when learning to solve problems by means of worked example study, both in primary and secondary education (Chapters 3, 4, and 5). Solving practice problems *after* worked example study (Chapter 3: Cohen's $d = .036$ and Chapter 4: Cohen's $d = 0.42$) or completing steps in partially worked-out examples *during* example study (Chapter 5: $\eta_p^2 = .09$) had a small to medium effect on monitoring accuracy when learning to solve problems.

In contrast to the findings on delayed keyword generation or summarizing when learning from text, delaying practice problem solving did not lead to more accurate monitoring than practice problem solving immediately after example study. It seems that being able to practice the problem and base their JOL on that experience is what provides students with relevant cues for monitoring their learning of the procedure, regardless of timing. Possibly, this is due to the fact that even when solving an immediate practice problem, students can no longer go back to the worked example, and might already have to rely on information from long term memory (LTM) in order to solve the problem, just as with delayed practice problems. In other words, perhaps the cues obtained from solving immediate practice problems are not different from those obtained when solving delayed practice problems.

Interestingly, the immediate generation strategy, that is, completion of partially worked-out examples *during* example study, was found to lead to *underconfidence* in future test performance. Whereas this can be seen as a positive effect, in the sense that overconfidence, which is known to be detrimental for future study activities (Dunsloky & Rawson, 2012), was reduced, it should be noted that it remains unclear whether this actually was a positive outcome in terms of the self-regulated learning process, because no effects on regulation accuracy were found. If underconfidence would affect regulation,

causing more accurate restudy choices, and if there would be an opportunity to study chosen items again, then test performance could theoretically benefit from underconfidence. Yet, possible effects of underconfidence on motivation also have to be taken into account, that is, if underconfidence discourages students to learn, then effective restudying will probably not take place. Therefore, future research should investigate the effects of bias in JOLs on both motivation and performance.

What future studies should also address, is whether other immediate generation strategies have the same effect. A possible reason why generation *during* example study led to underconfidence, but not generation *after* example study, is that students experienced difficulties when completing steps, used this as cues for making their JOLs (which led to lower JOLs), but may have been so focused on those cues that they are not aware that they are actually learning something by studying and completing the example. Generation *after* example study does allow learners to experience that they did learn at least something from the example (even though it may not be the entire procedure).

Measuring JOL and regulation accuracy. An issue concerning the measurement of JOL accuracy in general is that the measure that one chooses to express the accuracy affects the outcomes. While the use of multiple measures has been advocated because it allows for analyzing different aspects of monitoring accuracy (Schraw, 2009), it does make it more challenging to interpret findings. Relative and absolute measures reflect two different aspects of monitoring ability. That is, relative accuracy shows whether a student is able to discriminate between different items while absolute accuracy shows whether a student's judgment about a task or item is close to actual future test performance on that task or item. Considering the different focus of relative versus absolute accuracy, it is not too surprising that both measures show different results. In Chapter 2 relative accuracy showed a marginally significant difference between immediate and delayed JOLs whereas absolute accuracy did not. It should be noted though, that the measure of relative accuracy in Chapter 2 was based on a very low number of tasks, making it potentially unreliable (see Schraw, 2009), which is why it was not used in later studies, with the exception of the one reported in Chapter 4, in which JOLs were obtained for each *step* in the problem-solving procedure. In Chapter 4, absolute accuracy was higher when students received practice problems after worked example study compared to worked examples only, but relative accuracy did not show the same effect. This suggests that the deviation between JOLs and

test performance became smaller when students were provided practice problems but it did not affect the ability to discriminate between problem solving steps in the problems.

Within absolute measures of accuracy, a further distinction can be made in measures that express the difference between predicted and actual performance in terms of bias (i.e., over- or underestimation) and absolute deviation (i.e., regardless of the direction). Both bias and absolute deviation measures were used in the studies reported in Chapters 3, 5 and 6. 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), results on both measures can differ. For example, 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. In addition, to be able to express absolute accuracy for tasks on which performance scores were not similar to the JOL scale, a newly developed measure of absolute accuracy was used in the studies presented in Chapters 2 and 4. Because the raw data in those two studies did not allow for the calculation of absolute deviation or bias (i.e., JOLs and performance did not have the same scale), a gradual measure of absolute accuracy was used that varies between 0 and 1 based on each possible combination of the JOL and the test performance score. With this measure of absolute accuracy, interpretation is different from absolute deviation or bias: whereas bias and absolute deviation express best accuracy at 0, the newly developed measure shows best accuracy at 1. Future research should investigate the reliability and validity of this new measure of absolute accuracy in comparison to other measures of accuracy.

Next to monitoring, some issues about regulation need further attention as well. In the studies presented in Chapters 2, 3, 5 and 6 regulation accuracy was not affected. When monitoring accuracy did not differ between conditions, no differences in regulation accuracy were to be expected according to models of self-regulation (e.g., Winne & Hadwin, 1998). Yet, monitoring accuracy was significantly affected by using practice problems in Chapter 3, by completion of partially worked out examples in Chapter 5 and by using standards to self-assess performance in Chapter 6. So why did this not lead to improved regulation accuracy in these studies? There are two possible explanations. The first lies in the development of monitoring and regulation skills. Since both monitoring and regulation skills develop during childhood and adolescence (Krebs & Roebers, 2010;

Schneider, 2008), students who participated in the current studies, especially children in primary education, might not have been able to use their JOLs to regulate their learning process. Because the studies reported in this dissertation use problem-solving tasks, which are different from learning word pairs or learning from expository text, it is unclear at what age monitoring can be expected to start to inform regulation. The second explanation lies in how regulation accuracy was measured. In studies on learning from texts, regulation accuracy is usually measured by gamma correlations, which expresses whether JOLs are correlated to restudy choices (e.g., De Bruin et al., 2011; Thiede et al., 2003). Since gamma correlations could not be used with the low number of tasks in the studies reported here, another measure had to be developed for these studies. Even though it is a sensible measure seen from the discrepancy reduction perspective on regulation of study, other hypotheses about regulation have been proposed (e.g., region of proximal learning, Metcalfe, 2002; Metcalfe & Kornell, 2005; agenda-based learning, Ariel, Dunlosky, & Bailey, 2009) that fit this measure less well. Future research could further investigate ways of measuring and analyzing regulation accuracy that are compatible with different hypotheses on regulation. In addition, it could be interesting to conduct developmental studies investigating the effects of monitoring on regulation.

Task complexity and mental effort. Next to the main findings, several issues related to task characteristics were explored. First, in most of the studies described in this dissertation (Chapters 2, 3, 4, and 5), the task complexity was manipulated and its effect on monitoring accuracy was explored. Because complexity of the learning material is dependent on prior knowledge and experience, it is hard to compare the studies presented in Chapters 2, 3, 4, and 5, in which students of different ages participated and worked on different problem-solving tasks. Moreover, findings presented in Chapters 3 and 5 also show that the design of the studies possibly influenced the results on complexity. That is, students in primary and secondary education were found to change from underconfident to overconfident during the learning phase. This change does not necessarily depend on increasing complexity during the learning phase but could also be the result of experience in monitoring their performance on the problem-solving tasks students gained during the learning phase. Therefore, it could be interesting for future research to investigate the role of prior knowledge and experience when monitoring complexity of the learning material.

Second, the mental effort students invested to solve the problems, complete the partially worked-out examples or study the worked examples was found to be negatively related to JOL ratings (Chapters 2 and 5). This seems to suggest that the invested mental effort was used as a cue to make a JOL about the learning materials. For example in the study described in Chapter 5, completion of partially worked-out examples led to higher invested mental effort, which was negatively related to JOLs, and bias showed underconfidence in JOLs for students who completed examples. This raises the questions on what cues do students base their JOLs and which cues will also be a valid source to make accurate JOLs? However, because of the proximity of the mental effort rating and the JOL ratings in the design of the studies described in this dissertation, it is not warranted to conclude that invested mental effort was indeed used as a cue to make a JOL. It would be interesting for future research to investigate cues on which JOLs are based and, more specifically, investigate invested mental effort as a possible source to base a JOL on by manipulating the mental effort that is needed to perform a task.

Third, in the studies in secondary education reported in Chapter 5 and 6 two different types of tasks were used in the tests: identical and isomorphic test tasks. Identical tasks are exactly the same as the ones explained in the worked examples, whereas isomorphic tasks have different surface features, but can be solved using the same *solution procedure* that was studied. In the study reported in Chapter 5 it was found that absolute accuracy was better for the identical than for isomorphic test tasks, and in the study reported in Chapter 6 it was found that using standards of the correct solution improved JOL accuracy for identical but not for isomorphic test tasks. This seems to suggest that the JOL reflects students' judgment of how well they have learned the particular problem demonstrated in a worked example, rather than how well they have learned (or can apply) the solution procedure demonstrated in that example. Perhaps students used superficial cues about the specific problem to base their JOLs on, which would only match identical test tasks. It is also possible that students overestimated their future performance but performed better on identical test tasks compared to isomorphic test tasks. Yet, being able to apply a problem-solving procedure to new and slightly different problems is what is actually required in educational contexts. Therefore, an important question for future research is how to improve monitoring accuracy concerning isomorphic tasks.

Implications for Practice

The studies presented in this dissertation were exploratory in nature, given the lack of prior research on JOLs in problem-solving tasks in educational domains. Nevertheless, because problem-solving tasks play such an important role in both primary and secondary education curricula, and because students are also increasingly expected to engage in self-regulated learning in domains in which problem-solving tasks play an important role, the results from these studies are of interest for educational practice. The studies presented in this dissertation showed that the use of generation strategies (i.e., solving a practice problem or completing steps in an example) and feedback (i.e., standards of correct answers) seems to help both primary and secondary education students to monitor their own learning process when learning to solve problems by means of worked example study. Although these findings regarding monitoring accuracy should be replicated and the relation with regulation accuracy and learning outcomes should be studied further, these findings are very promising because these instructional strategies are relatively easy to implement in educational practice.

Next to exploring several of the issues discussed above in more detail, future research might also attempt to find other delayed or immediate generation strategies that could foster monitoring and regulation accuracy for problem-solving tasks, which would be both theoretically and practically relevant. For example, self-explaining when learning from worked examples could potentially improve monitoring accuracy similar to the results found with learning from text (Griffin, Wiley, & Thiede, 2008). Moreover, self-explaining could not only provide learners with a better idea of their learning process but also provide more insight in the processes of monitoring and regulation for researcher by investigating the verbal protocol. The outcomes of such studies are expected to have a substantial positive impact on research on monitoring and regulation when learning to solve problems and could lead to guidelines for educational practice.

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

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction*, 24, 1-3. doi: 10.1016/j.learninstruc.2012.10.003

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

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology*, 96, 523-535. doi: 10.1037/0022-0663.96.3.523

Baars, M., Van Gog, T., De Bruin, A., & Paas, F. (2013a). *Accuracy of primary school children's immediate and delayed judgments of learning about problem solving tasks*. Manuscript submitted for publication.

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., 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: <http://dx.doi.org/10.1016/j.cedpsych.2013.09>.

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

Baker, J. M. C., & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study-judgments lags on

accessibility effects. *Psychonomic Bulletin & Review*, 13, 60-65. doi: 10.3758/BF03193813

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher, & A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435e459). Cambridge, MA: MITPress.

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

Butler, D. L. & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65, 245- 281. doi: 10.3102/00346543065003245

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., 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

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., & Van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction*, 22, 254-252.
doi:10.1016/j.learninstruc.2012.01.003

Dunlosky, J., Kubat-Silman, A. K., & Hertzog, C. (2003). Training monitoring skills improves older adults' self-paced associative learning. *Psychology and Aging*, 18, 340-345. doi:10.1037/0882-7974.18.2.340

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. & 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

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, England: 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

Dunlosky, 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

Dunning, D., Johnson, K., Erlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, 12, 83-87. doi: 10.1111/1467-8721.01235

Efkides, A. (2002). The systemic nature of metacognitive experiences: Feelings, judgments, and their interrelations. In M. Izaute, P. Chambres, & P.-J. Marescaux (Eds.), *Metacognition: Process, function, and use* (pp. 19-34). Dordrecht, The Netherlands: Kluwer.

Efkides, 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

Finn, B. & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology, 33*, 238-244. doi:10.1037/0278-7393.33.1.238

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist, 34*(10), 906. doi: 10.1037/0003-066X.34.10.906

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116*, 119-136. doi: 10.1037/0096-3445.116.2.119

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, 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 & 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. (2007). Expertise reversal effect and its implications for learner-tailored instructions. *Educational Psychology Review, 19*, 509-519, doi: 10.1007/s10648-007-9054-3.

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

Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology, 96*, 558-568, doi: 10.1037/0022-0663.96.3.558

Kelemen, W. L., Frost, P. J., & Weaver III, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition, 28*, 92-107. doi: 10.3758/BF03211579

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kirschner, F., Paas, F., & Kirschner, P. A. (2009). Individual and group-based learning from complex cognitive tasks: Effects on retention and transfer efficiency. *Computers in Human Behavior, 25*, 306-314. doi: 10.1016/j.chb.2008.12.008

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., (2010). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science, 13*, 441-453. doi: 10.1111/j.1467-7687.2009.00907.x

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

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009b). The memorizing effort heuristic in judgments of learning: A developmental perspective. *Journal of Experimental Child Psychology, 102*, 265-279. doi:10.1016/j.jecp.2008.10.005

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

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect

relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, 135, 36–69. doi: 10.1037/0096-3445.135.1.36

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

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

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle-school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied*, 15, 307–318. doi:10.1037/a0017599

Maki, R. H. (1998a). 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.

Maki, R. H. (1998b). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition*, 26, 959-964. doi: 10.3758/BF03201176

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning Memory, and Cognition*, 16, 609–616. doi: 10.1037/0278-7393.16.4.609

Maki, R. H., Shield, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*, 723- 731. doi: 10.1037/0022-0663.97.4.723

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

Merkt, M., Weigand, S., Heier, A., & Schwan, S. (2011). Learning with videos vs. learning with print: The role of interactive features. *Learning and Instruction, 21*, 687-704. doi: 10.1016/j.learninstruc.2011.03.004

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. *Journal 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 & 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

Muthén, L.K. & Muthén, B.O. (1998-2010). *Mplus User's Guide* (Sixth Edition).

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: 10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., Gerler, D., & Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General, 113*, 282-300. doi: 10.1037/0096-3445.113.2.282

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.2012.12.004

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

Paas, F., Tuovinen, J., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*, 63-71. doi:10.1207/S15326985EP3801_8

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

Paas, F., Van Merriënboer, J. J. G., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79*, 419–430. doi: 10.2466/pms.1994.79.1.419

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.

Rawson, K., & Dunlosky, J. (2012). Overconfidence produces underachievement: Inaccurate self evaluation undermine students' learning and retention. *Learning and Instruction*, 22, 271- 280. doi:10.1016/j.learninstruc.2011.08.003

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition*, 28, 1004-1010. doi: 10.3758/BF03209348

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

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.learninstruc.2011.10.007

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.

Renkl, A. (2013). Towards an instructionally-oriented theory of example-based learning, in press in *Cognitive Science*. 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: 10.1348/096407409X429842

Roediger, H. L., & Karpicke, J. D. (2006). Test enhanced learning: Taking memory tests improves long term retention. *Psychological Review, 113*, 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

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.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychological Bulletin & Review, 1*, 357-375. doi: 10.3758/BF03213977

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: 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

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

Sweller, J. (2010). Element interactivity and intrinsic, extraneous and germane cognitive load. *Educational Psychology Review*, 22, 123-138. doi: 10.1007/s10648-010-9128-5

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 monitoring and self-regulation during multitrial learning. *Psychological 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*. 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 7th and 8th graders. *Journal of Educational Psychology*, 10, 554- 564. doi: 10.1037/a002860

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. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist*, 43, 16-26. doi: 10.1080/00461520701756248

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

Van Loon, M. H., de Bruin, A. B., van Gog, T., & van Merriënboer, J. J. (2013a). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction*, 24, 15-25. doi: 10.1007/s11409-013-9100-0

Van Loon, M. H., de Bruin, A. B., van Gog, T., & van Merriënboer, J. J. (2013b). The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning*, 8, 173-191. doi 10.1007/s11409-013-9100-0

Van Loon-Hillen, N. H., Van Gog, T., & Brand-Gruwel, S. (2012). Effects of worked examples in a primary school mathematics curriculum. *Interactive Learning Environments*, 20, 89-99. doi: 10.1080/10494821003755510

Van Merriënboer, J. J. G. (1990). Strategies for programming instruction in high school: Program completion vs. program generation. *Journal of Educational Computing Research*, 6, 265-287. doi: 10.2190/4NK5-17L7-TWQV-1EHL

Van Merriënboer, J. J. G. & Krammer, H. P. M. (1990). The "completion strategy" in programming instruction: Theoretical and empirical support. In Dijkstra, S., Van Hout-Wolters, B. H. M., & Van der Sijde, P. C. (eds.), *Research on Instruction*, (pp.45-61). Englewood Cliffs, NY: Educational Technology Publications.

Van Merriënboer, J. J. G., Schuurman, J. G., De Croock, M. B. M., & Paas, F. (2002). Redirecting learners' attention during training: Effects on cognitive load, transfer test performance and training efficiency. *Learning and Instruction*, 12, 11-37. doi: 10.1016/S0959-4752(01)00020-2

Vesonder, G. T., & Voss, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language*, 24, 363- 376. doi: 10.1016/0749-596X(85)90034-8

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *The Journal of General Psychology*, 132, 408-428. doi: 10.3200/GENP.132.4.408-428

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.

Nederlandse samenvatting (Dutch summary)

Om effectief zelf-gereguleerd te leren, moeten leerlingen in staat zijn hun eigen leerproces te *monitoren* en op basis daarvan het vervolg van het leerproces te *reguleren* (Winne & Hadwin, 1998). Uit onderzoek blijkt echter dat leerlingen zowel met monitoring als met regulatie veel moeite hebben (Dunlosky & Lipko, 2007). Er is dan ook al veel onderzoek gedaan naar instructiestrategieën om deze kernprocessen in zelf-gereguleerd leren te ondersteunen en verbeteren, maar dit onderzoek beperkte zich tot voor kort vooral tot monitoring- en regulatievaardigheden bij het leren van woordparen (zie Rhodes & Tauber, 2011) en teksten (zie Thiede, Griffin, Wiley, & Redford, 2009). Ondanks het feit dat monitoring en regulatie ook bij het leren probleem-oplossen van belang is –en dus om goed zelf-gereguleerd te kunnen leren in bijvoorbeeld wiskunde, scheikunde, natuurkunde en economie- waren er slechts enkele studies die de accuratesse van monitoring bij het leren probleem-oplossen onderzochten (De Bruin, Rikers, & Schmidt, 2005; Metcalfe, 1986; Metcalfe & Wiebe, 1987).

In dit proefschrift werden daarom de volgende onderzoeksvragen onderzocht tijdens het oplossen of leren oplossen van gestructureerde probleem-oplostaken: a) zijn leerlingen in staat om hun eigen leren (begrip en toekomstige testprestatie) accuraat te beoordelen, dat wil zeggen, zijn hun Judgments of Learning (JOLs) accuraat, en zijn zij in staat deze te gebruiken om hun leerproces te reguleren? b) zorgt het vertragen van JOLs voor hogere accuratesse zoals bij woordparen het geval bleek? c) zijn instructiestrategieën waarbij studenten tijdens of na het bestuderen van uitgewerkte voorbeelden zelf (een deel van) de oplossing dienen te genereren alvorens een JOL te geven (“toepassingstrategieën”), effectief om JOL accuratesse te verbeteren zoals bij het leren van teksten het geval bleek? en d) kunnen *retrospectieve* zelfbeoordelingen (self-assessments) worden verbeterd doormiddel van self-assessment training en het gebruik van antwoordmodellen van de juiste oplossing en leidt deze verbetering ook tot een verbetering in prospectieve monitoring oordelen (JOLs)? Daarnaast werden in de meeste studies de effecten van taakcomplexiteit op monitoring accuratesse, de relatie tussen JOL en geïnvesteerde moeite en het effect van taaktype op monitoring accuratesse geëxploereerd.

Hoofdstuk 2 in dit proefschrift betrof een studie naar monitoring en regulatie accuratesse tijdens het oplossen van probleem-oplostaken. Resultaten lieten zien dat

basisschool leerlingen tijdens het probleemoplossen in staat waren de complexiteit van de taken te monitoren. Daar waar de complexiteit groter werd, werden prestaties lager, geïnvesteerde moeite hoger en de judgments of learning (JOLs) lager. Echter, de relatieve accuratesse van JOLs, die laat zien in welke mate leerlingen kunnen discrimineren tussen taken, was laag tot matig. De gemiddelde gamma correlatie in de conditie met directe JOLs verschilde significant van nul, wat niet het geval was voor de conditie waarin vertraagde JOLs gegeven werden. Met andere woorden, directe JOLs waren matig accuraat terwijl vertraagde JOLs niet accuraat waren. Verder was er een significante negatieve correlatie tussen geïnvesteerde moeite en JOLs die liet zien dat hoe hoger de geïnvesteerde moeite was, hoe lager de JOLs waren. Dit lijkt erop te wijzen dat de geïnvesteerde moeite door de leerlingen gebruikt is als een cue om hun JOLs op te baseren. Testprestaties verschilden niet tussen de conditie waarin leerlingen directe JOLs gaven en de conditie waarin ze vertraagde JOLs gaven. Kortom, de resultaten uit Hoofdstuk 2 laten zien dat in tegenstelling tot resultaten uit eerder onderzoek met woordparen en teksten, directe JOLs matig accuraat waren en daarmee (niet significant) meer accuraat lijken te zijn dan vertraagde JOLs bij het probleemoplossen in het basisonderwijs.

In Hoofdstuk 3 en 4 zijn twee studies gerapporteerd, in het basis- en voortgezet onderwijs, waarin het effect van oefenproblemen na het bestuderen van uitgewerkte voorbeelden werd onderzocht. De hoofdvraag in deze studies was of dit een effectieve toepassingsstrategie zou zijn om monitoring en regulatie accuratesse te verbeteren.

De studie in Hoofdstuk 3 laat zien dat het oplossen van oefenproblemen na het bestuderen van uitgewerkte voorbeelden de bias (de *richting* van het verschil tussen JOL en testprestatie) in JOLs verminderde. Dit was het geval ongeacht het moment waarop de oefenproblemen werden aangeboden. Met andere woorden, de gemiddelde overschatting die leerlingen lieten zien in hun JOLs werd significant verminderd in alle condities waarin leerlingen oefenproblemen kregen. Echter, de absolute accuratesse van JOLs (de *grootte* van het verschil tussen JOL en testprestatie) verschilde niet tussen de condities. Ondanks het effect van het oplossen van oefenproblemen op JOL accuratesse, werd regulatie accuratesse niet significant beïnvloed door het maken van oefenproblemen. Verder lieten leerlingen meer accurate JOLs (zowel bias als absolute accuratesse) zien op de minder complexe taken. Dit werd waarschijnlijk veroorzaakt door hogere JOLs op de meer complexe taken waar de testprestatie lager was. Ondanks de extra oefening in de condities

met oefenproblemen (i.e., het toepassen wat geleerd was van het uitgewerkte voorbeeld), verschilden de testprestaties niet tussen de condities met of zonder oefenproblemen.

In het experiment dat gerapporteerd werd in Hoofdstuk 4, werden JOLs per stap in het uitgewerkte voorbeeld of de probleemopstak gemeten, waardoor relatieve accuratesse en een nieuwe maat van absolute accuratesse gebruikt konden worden om JOL accuratesse te analyseren. In overeenstemming met de resultaten in Hoofdstuk 3, werden er geen verschillen tussen directe en vertraagde JOLs na uitgewerkte voorbeelden of oefenproblemen gevonden. Echter, numeriek was absolute accuratesse hoger voor directe JOLs dan voor vertraagde JOLs na het maken van oefenproblemen (vgl. Hoofdstuk 2). Zoals verwacht, werkten oefenproblemen als een toepassingstrategie waarmee leerlingen betere JOLs konden maken, aangezien absolute accuratesse hoger bleek voor leerlingen die oefenproblemen kregen na het bestuderen van uitgewerkte voorbeelden. In tegenstelling tot de verwachting dat vertraagde oefenproblemen tot de meest accurate JOLs zouden leiden, waren zowel relatieve accuratesse ($p = .090$) als absolute accuratesse ($p = .092$) marginaal significant beter voor leerlingen die direct na het bestuderen van uitgewerkte voorbeelden de oefenproblemen kregen. Bovendien was regulatie accuratesse in deze studie hoger voor leerlingen die oefenproblemen kregen na het bestuderen van uitgewerkte voorbeelden. Echter, het verwachtte effect van betere regulatie op de uiteindelijke testprestatie werd niet gevonden. Dit komt wellicht doordat de uiteindelijke test uit isomorfe (i.e.,zelfde oplossingsprocedure maar andere waarden) testproblemen ten opzichte van de bestudeerde en geoefende problemen bestond.

Kortom, de resultaten uit Hoofdstuk 3 en 4 laten zien dat het oplossen van oefenproblemen na het bestuderen van uitgewerkte voorbeelden leerlingen in het basis en voortgezet onderwijs kan helpen om hun eigen leerproces te monitoren. Daarnaast lieten leerlingen in het middelbaar onderwijs ook betere regulatie zien wanneer zij oefenproblemen oplost na het bestuderen van uitgewerkte voorbeelden.

De studie in Hoofdstuk 5 richtte zich op het effect van een directe toepassingsstrategie, namelijk completeren van gedeeltelijk uitgewerkte voorbeelden, op monitoring en regulatie accuratesse in het voortgezet onderwijs. De resultaten lieten een verschil in bias zien; het completeren van gedeeltelijk uitgewerkte voorbeelden leidde tot een significante onderschatting van toekomstige testprestaties in vergelijking met het bestuderen van volledig uitgewerkte voorbeelden. Echter, er was geen verschil tussen beide

condities in de absolute accuratesse (grootte van het verschil tussen JOLs en testprestaties). Daarnaast verschilden regulatie accuratesse en testprestaties niet tussen beide condities. De complexiteit van de taak had een effect op bias in JOLs, namelijk, gemiddeld over beide condities overschatte leerlingen hun toekomstige testprestaties op de minst complexe taak, maar dit was niet het geval voor de twee meer complexe taken. Dit kan veroorzaakt zijn doordat gemiddelde JOL ratings omhoog gingen terwijl prestatie daalde naarmate de taken meer complex werden. Verder werd in deze studie een effect van het type taak gevonden: monitoring accuratesse gemeten met absolute deviatie was beter voor de testtaken die identiek (i.e.,zelfde oplossingsprocedure en waarden) waren aan voorbeelden in de leerfase in vergelijking met de testtaken die isomorf (i.e.,zelfde oplossingsprocedure maar andere waarden) waren aan de voorbeelden in de leerfase. In deze studie werd ook een significante negatieve correlatie tussen geïnvesteerde moeite en JOLs gevonden (zie ook Hoofdstuk 2) wat suggereert dat geïnvesteerde moeite gebruikt is als een cue om een JOL te maken. Kortom, de resultaten van het experiment dat in Hoofdstuk 5 gerapporteerd wordt, laten zien dat completeren van gedeeltelijk uitgewerkte voorbeelden monitoring accuratesse beïnvloedt, namelijk na completeren onderschatte leerlingen hun toekomstige testprestaties terwijl hun regulatie accuratesse niet beïnvloed werd.

In Hoofdstuk 6 werd een studie gepresenteerd waarin self-assessment (SA) training en antwoordmodellen werden gebruikt om monitoring en regulatie accuratesse te verbeteren in het voortgezet onderwijs. Deze studie bestond uit twee experimenten. Experiment 2 liet zien dat antwoordmodellen met het correcte antwoord voor elke probleemoplossing, een positief effect hadden op SA accuratesse, monitoring accuratesse voor identieke testtaken en op de testprestaties. Experiment 1 en 2 lieten zien dat SA training geen verbetering in SA accuratesse, JOL accuratesse, regulatie accuratesse of testprestatie teweeg bracht. SA ratings waren hoog gecorreleerd met JOL ratings in zowel Experiment 1 als 2 (met uitzondering van de conditie met alleen een antwoordmodel in Experiment 2). Dus waarschijnlijk gebruikten leerlingen hun SA van een oefenprobleem als een cue om hun toekomstige testprestatie te voorspellen op een soortgelijke testtaak. Bovendien, in overeenstemming met de resultaten uit Hoofdstuk 2 en 5, was geïnvesteerde moeite significant negatief gecorreleerd met zowel SA en JOL ratings. Namelijk, hoe hoger de geïnvesteerde moeite was, hoe lager de SA en JOL ratings waren. Dit suggereert dat leerlingen geïnvesteerde moeite als een cue gebruikten voor zowel retrospectieve (SA) als

prospectieve (JOL) monitoring oordelen. Kortom, waar SA training geen effect had op retrospectieve en prospectieve monitoring oordelen, had het gebruik van antwoordmodellen om een SA van de prestatie op oefenproblemen te maken een effect op zowel retrospectieve als prospective monitoring oordelen en testprestaties maar niet op regulatie accuratesse.

De studies die gepresenteerd zijn in dit proefschrift waren exploratief van aard omdat er tot op heden weinig onderzoek naar JOLs over probleem-oplostaken gedaan is. In overeenstemming met de bevindingen uit studies naar monitoring bij het leren van tekst (Thiede, Griffin, Wiley, & Redford, 2009), laten de studies in dit proefschrift zien dat het gebruik van toepassingsstrategieën (i.e., oefenproblemen of completeren van stappen in uitgewerkt voorbeeld) en feedback (i.e., antwoordmodel) leerlingen in het basis- en voortgezet onderwijs helpt om beter hun eigen leerproces te beoordelen tijdens het leren probleemoplossen. Hoewel de bevindingen op het gebied van monitoring gereplieerd moeten worden en de relatie tussen monitoring en regulatie verder onderzocht moet worden, zijn de bevindingen veelbelovend omdat de toepassingsstrategieën relatief gemakkelijk naar de onderwijspraktijk te vertalen zijn.

Publications

2014

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

2013

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: <http://dx.doi.org/10.1016/j.cedpsych.2013.09.001>.

Submitted manuscripts

Baars, M., Van Gog, T., De Bruin, A., & Paas, F. (2013a). *Accuracy of primary school children's immediate and delayed judgments of learning about problem solving tasks*.

Baars, M., Van Gog, T., De Bruin, A., & Paas, F. (2013b). *Effects of problem solving after worked example study on secondary school children's monitoring accuracy*.

Baars, M., Vink, S., Van Gog, T., De Bruin, A., & Paas, F. (2013). *Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving*.

Conference proceedings

Baars, M.A., Van Gog, T. & Paas, F. (2010). Accuracy of immediate and delayed Judgments of Learning during problem solving. In *Instructional design for motivated and competent learning in a digital world: Proceedings of the joint conference of EARLI Special Interest Groups Instructional Design and Learning and Instruction with Computers* (pp. 38-40). Ulm University.

Presentations

Baars, M.A., Visser, S., Van Gog, T., De Bruin, A. & Paas, F. (2013, Augustus). *Completion of partially worked-out examples as a generation strategy to improve monitoring accuracy*. Paper presentation at EARLI conference, München, Germany.

Baars, M.A., Visser, S., Van Gog, T., De Bruin, A. & Paas, F. (2013, June). *Completion of partially worked-out examples leads to underestimation of future test performance*. Paper presentation at Cognitive Load Conference, Toulouse, France.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2013, June). *Completion of partially worked-out examples leads to underestimation of future test performance*. Paper presentation at Onderwijs Research Dagen, Brussels, Belgium.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2012). *Fostering self-monitoring and self-regulation when learning to solve problems*. Presentation at the symposium 'Improving metacognitive skills', Maastricht, Netherlands.

Baars, M.A., Gog, T. van, De Bruin, A. & Paas, F. (2012). *Using self-testing to improve monitoring accuracy when studying worked examples in primary education*. Paper presentation at the 5th Biennial Meeting of the EARLI Special Interest Group 16 Metacognition, Milaan, Italy.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2012, June). *Using self-test to improve monitoring accuracy when studying worked examples in primary education*. Paper presentation at Onderwijs Research Dagen, Wageningen, Netherlands.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2012, April). *Accuracy of immediate and delayed Judgments of Learning in worked examples and problem solving tasks*. Paper presentation at Cognitive Load Conference, Tallahassee, USA.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2011, August). *Accuracy of immediate and delayed Judgments of Learning in worked examples and problem solving tasks*. Paper presentation at EARLI conference, Exeter, United Kingdom.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2011, August). *Accuracy of immediate and delayed Judgments of Learning when studying worked examples*. Paper presentation at JURE conference, Exeter.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2011, June). *Accuracy of immediate and delayed Judgments of Learning about problem solving tasks*. Paper presentation at Onderwijs Research Dagen, Maastricht, Netherlands.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2010). Accuracy of immediate and delayed Judgments of Learning during problem solving. In *Instructional design for motivated and competent learning in a digital world: Proceedings of the joint conference of EARLI Special Interest Groups Instructional Design and Learning and Instruction with Computers* (pp. 38-40). Ulm University.

Poster presentations

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2013, Augustus). *Effects of immediate and delayed problem solving after worked example study on primary school children's monitoring accuracy*. Poster presentation at EARLI conference, München, Germany.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2012, October). *Completion of partially worked-out examples leads to underestimation of future test performance*. Poster presentation at Graduate Research Day of the institute of Psychology at Erasmus University, Rotterdam, Netherlands.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2012, April). *Accuracy of immediate and delayed comprehension judgments about problem solving tasks*. Poster presentation at AERA conference, Vancouver, Canada.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2012, April). *Accuracy of Judgments of Learning and restudy choices when studying worked examples*. Poster presentation at AERA conference, Vancouver, Canada.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2011, November). *Using self-test to improve monitoring accuracy when studying worked examples in primary education*. Poster presentation at ICO ToogDagen, Eindhoven, The Netherlands.

Baars, M.A., Van Gog, T., De Bruin, A. & Paas, F. (2010, September). *Accuracy of immediate and delayed Judgments of Learning about problem solving tasks*. Poster presentation at Graduate Research Day of the institute of Psychology at Erasmus University, Rotterdam, Netherlands.

Dankwoord

Het afmaken van je proefschrift betekent ook terugkijken en je realiseren hoeveel steun en hulp je hebt mogen ontvangen.

Allereerst wil ik graag ‘mijn team’, Tamara, Anique en Fred, bedanken. Tamara, als dagelijks begeleidster en promotor heb jij me op zoveel vlakken dingen geleerd, me laten relativeren, gestimuleerd, me geïntroduceerd in de wondere wereld van de wetenschap en laten lachen. Jouw enthousiasme en doorzettingsvermogen zijn voor mij een voorbeeld van onschatbare waarde. Anique, onze ‘echte’ of Skype-afspraken waarin jij mij dat hart-onder-de-riem of het schouderklopje gaf, gaven mij vaak het zetje wat ik nog nodig had. Bedankt voor de ruimte die je me gaf bij het organiseren van workshops en het symposium met onze PROO groep. Fred, jouw kritische blik en humor riepen me vaak terug en dwongen me om mijn stappen nog eens na te gaan, wat heel leerzaam en waardevol was. Bedankt dat ik altijd je kamer binnen kon lopen met onsaamhangende updates.

Ten tweede wil ik alle basisscholen en middelbare scholen bedanken waar ik in de afgelopen jaren zo gastvrij ontvangen ben en mijn onderzoek kon uitvoeren. Dankzij jullie medewerking en organisatie kon ik mijn onderzoeksplannen daadwerkelijk uitvoeren. Heel veel dank hiervoor!

Daarnaast wil ik graag de projectgroep waarin ik mijn onderzoek heb uitgevoerd bedanken. Jeroen, Peter, Luciana, Mariëtte, Anique, Fred en Tamara bedankt voor jullie feedback en ideeën over de studies die we hebben uitgevoerd. Luciana, thank you for your support, humor and talent for findings us great places to have food and drinks. Mariëtte, congresbezoeken waren met jou een avontuur en een beetje vakantie. Samen sessies bezoeken en daarna die ene prof met vragen overladen maar ook eindeloos kletsen, tutten en borrelen.

Eén groep is geen groep. Ik wil iedereen in de O&O-groep en de pub-groep bedanken voor de constructieve sfeer waarin er altijd ruimte was om je onderzoek of je manuscript te bespreken en vol goede moed weer verder te gaan. De ruimte om samen de andere betekenis van *pub* te verkennen gaf ook altijd veel gezelligheid, motivatie en inspiratie. Ook de groep psychologie AIO’s was een onmisbare groep tijdens mijn promotietijd. Ik wist niet dat in hetzelfde schuitje zitten zo gezellig kan zijn. In het speciaal wil ik Charly, Daniëlle, Gabriela, Gerdien, Jacqueline, Jan, Kim, Kimberley, Lisette, Nicole, Sabine, Vincent, en Wim, bedanken voor alle koffie-momentjes, drankjes, etentjes

en gezelligheid. Ook iedereen van de ICO onderwijscommissie wil ik bedanken voor de leuke en leerzame meetings. Ook de collega's van het Wetenschapsknooppunt van de EUR wil ik bedanken. Het was een uitdagende en welkome afwisseling om samen met jullie, leerlingen en leerkrachten kennis te laten maken met de wetenschap en de psychologie. Nog een groep: aan het einde van mijn promotietraject kon ik 2 maanden aan de Ruhr Universität in Bochum werken, ik kreeg een groep nieuwe collega's erbij. Joachim, Tina, Silke, Ferdi, David, Jessica, Jessica, Nimisha, Alex, Claudia und Katharina: Vielen Dank für alles! Verder wil ik iedereen van het secretariaat van de afdeling Psychologie en van het EBL bedanken voor alle hulp en ondersteuning.

Hoe zouden de afgelopen 4 jaar zijn geweest zonder mijn kamergenoten! Eerst op T11.11 en later op T12.46. We hebben met elkaar gelachen, gehuild, striptekeningen gemaakt, honkbal gespeeld, frietjes gehaald, koffie gedronken (en kortsluiting veroorzaakt) en overal gekke foto's van gemaakt. Jan, jouw luisterend oor, hulp en antwoorden op wat ik je ook voorlegde, zijn heel waardevol geweest voor me. Kimberley, waar wij van kamergenoten naar vriendinnen zijn gegaan, weet ik niet meer. Was het je oog voor mijn kleding en of het wel 'matchte', of het feit dat we allebei vanalles eruit flappen, je geduld met mijn ongeduld, salsa dansen of hardlopen? Dat doet er niet meer toe. Ik weet dat wij nog heel lang vriendinnen blijven ookal moeten we echt vaak afspreken om een beetje aan onze uren te komen tegenwoordig.

Naast mijn werk was er gelukkig ook tijd voor vrienden. Soms bewust en ook heel vaak onbewust zijn jullie voor mij een rots in de branding geweest en gaven jullie me nieuwe energie om door te gaan. In het speciaal, Annemiek, Anniek, Eline, Freek, Han, Hanneke, Karlijn, Marrit, Quirijn, Rens, en Sanne bedankt voor alle leuke feestjes, etentjes en uitjes die de broodnodige ontspanning brachten.

Tenslotte, mijn familie, ik kan jullie niet genoeg bedanken voor onvoorwaardelijk vertrouwen en liefde. Pap, Mam, Annelies, Martijn en Nico, ik vind het geweldig dat ik met jullie een gezin vorm!

Curriculum Vitae

Martine Baars was born in Woerden, the Netherlands, on April 25th 1985. After completing secondary education (VWO) at Hendrik Pierson College in Zetten, she started studying Cultural Anthropology and Developmental Studies at the Radboud University Nijmegen. In 2006 she obtained her Bachelor's degree in Cultural Anthropology. In September 2006 she started studying Pedagogical and Educational Science at the Radboud University Nijmegen. In 2009 she obtained her Master's degree in Educational Science. In December 2009 she started working on her NWO-PROO funded PhD project at the Institute of Psychology at the Erasmus University Rotterdam that resulted in this dissertation on instructional strategies to improve self-monitoring performance when learning to solve problems. During her PhD project, she presented her work at national and international conferences and was involved in teaching a number of courses at the Institute of Psychology at the Erasmus University Rotterdam. In addition, she was a member of the Educational Committee of the Dutch research school ICO (the Interuniversity Center for Educational Research) from 2010 until 2013. Furthermore, she participated in the Wetenschapsknooppunt at Erasmus University Rotterdam to develop and carry out educational activities that aimed to introduce scientific research on various topics to primary and secondary school children and their teachers.

ICO Dissertation Series

In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the ICO Partner Universities: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, University of Amsterdam, University of Twente, Utrecht University, VU University Amsterdam, and Wageningen University, and formerly University of Groningen (until 2006), Radboud University Nijmegen (until 2004), and Tilburg University (until 2002). The University of Groningen, University of Antwerp, University of Ghent, and the Erasmus University Rotterdam have been 'ICO 'Network partner' in 2010 and 2011. From 2012 onwards, these ICO Network partners are full ICO partners, and from that period their dissertations will be added to this dissertation series. Over two hundred dissertations have been published in this series. The most recent ones are listed below.

207. Zitter, I.I. (04-02-2010). *Designing for learning: Studying learning environments in higher professional education from a design perspective*. Utrecht: Utrecht University.
208. Koopman, M. (11-02-2010). *Students' goal orientations, information processing strategies and knowledge development in competence-based pre-vocational secondary education*. Eindhoven: Eindhoven University of Technology.
209. Platteel, T. (11-02-2010). *Knowledge development of secondary school L1 teachers on concept-context rich education in an action-research setting*. Leiden: Leiden University.
210. Mittendorff, K. M. (12-03-2010). *Career conversations in senior secondary vocational education*. Eindhoven: Eindhoven University of Technology.
211. Moolenaar, N.M. (01-06-2010). *Ties with potential: Nature, antecedents, and consequences of social networks in school teams*. Amsterdam: University of Amsterdam.
212. Duijnhouwer, H. (04-06-2010). *Feedback effects on students' writing motivation, process, and performance*. Utrecht: Utrecht University.
213. Kessels, C.C. (30-06-2010). *The influence of induction programs on beginning teachers' well-being and professional development*. Leiden: Leiden University.
214. Endedijk, M.D. (02-07-2010). *Student teachers' self-regulated learning*. Utrecht: Utrecht University.
215. De Bakker, G.M. (08-09-2010). *Allocated online reciprocal peer support as a candidate for decreasing the tutoring load of teachers*. Eindhoven: Eindhoven University of Technology.
216. Groenier, M. (10-09-2010). *The decisive moment: Making diagnostic decisions and designing treatments*. Enschede: University of Twente.
217. Bonestroo, W.J. (24-09-2010). *Planning with graphical overview: Effects of support tools on self-regulated learning*. Enschede: University of Twente.
218. Vos, M.A.J. (30-09-2010). *Interaction between teachers and teaching materials: On the implementation of context-based chemistry education*. Eindhoven: Eindhoven University of Technology.
219. Kostons, D.D.N.M. (05-11-2010). *On the role of self-assessment and task-selection skills in self-regulated learning*. Heerlen: Open University of the Netherlands.

220. Bruin-Muurling, G. (21-12-2010). *The development of proficiency in the fraction domain: Affordances and constraints in the curriculum*. Eindhoven: Eindhoven University of Technology.

221. Slof, B. (28-01-2011). *Representational scripting for carrying out complex learning tasks*. Utrecht: Utrecht University.

222. Fastré, G. (11-03-2011). *Improving sustainable assessment skills in vocational education*. Heerlen: Open University of the Netherlands.

223. Min-Leliveld, M.J. (18-05-2011). *Supporting medical teachers' learning: Characteristics of effective instructional development*. Leiden: Leiden University.

224. Van Blankenstein, F.M. (18-05-2011). *Elaboration during problem-based small group discussion: A new approach to study collaborative learning*. Maastricht: Maastricht University.

225. Dobber, M. (21-06-2011). *Collaboration in groups during teacher education*. Leiden: Leiden University.

226. Jossberger, H. (24-06-2011). *Towards self-regulated learning in vocational education: Difficulties and opportunities*. Heerlen: Open University of the Netherlands.

227. Schaap, H. (24-06-2011). *Students' personal professional theories in vocational education: Developing a knowledge base*. Utrecht: Utrecht University.

228. Kolovou, A. (04-07-2011). *Mathematical problem solving in primary school*. Utrecht: Utrecht University.

229. Beausaert, A.J. (19-10-2011). *The use of personal developments plans in the workplace. Effects, purposes and supporting conditions*. Maastricht: Maastricht University

230. Favier, T.T. (31-10-2011). *Geographic information systems in inquiry-based secondary geography education: Theory and practice*. Amsterdam: VU University Amsterdam.

231. Brouwer, P. (15-11-2011). *Collaboration in teacher teams*. Utrecht: Utrecht University.

232. Molenaar, I. (24-11-2011). *It's all about metacognitive activities; Computerized scaffolding of self-regulated learning*. Amsterdam: University of Amsterdam.

233. Cornelissen, L.J.F. (29-11-2011). *Knowledge processes in school-university research networks*. Eindhoven: Eindhoven University of Technology.

234. Elffers, L. (14-12-2011). *The transition to post-secondary vocational education: Students' entrance, experiences, and attainment*. Amsterdam: University of Amsterdam.

235. Van Stiphout, I.M. (14-12-2011). *The development of algebraic proficiency*. Eindhoven: Eindhoven University of Technology.

236. Gervedink Nijhuis, C.J. (03-2-2012) *Culturally Sensitive Curriculum Development in International Cooperation*. Enschede: University of Twente

237. Thoonen, E.E.J. (14-02-2012) *Improving Classroom Practices: The impact of Leadership School Organizational Conditions, and Teacher Factors*. Amsterdam: University of Amsterdam

238. Truijen, K.J.P (21-03-2012) *Teaming Teachers. Exploring factors that influence effective team functioning in a vocational education context*. Enschede: University of Twente

239. Maulana, R.M. (26-03-2012) *Teacher-student relationships during the first year of secondary education. Exploring of change and link with motivation outcomes in The Netherlands and Indonesia*. Groningen: University of Groningen

240. Lomos, C. (29-03-2012) *Professional community and student achievement*. Groningen: University of Groningen

241. Mulder, Y.G. (19-04-2012) *Learning science by creating models* Enschede: University of Twente

242. Van Zundert, M.J. (04-05-2012) *Optimising the effectiveness and reliability of reciprocal peer assessment in secondary education* Maastricht: Maastricht University

243. Ketelaar, E. (24-05-2012) *Teachers and innovations: on the role of ownership, sense-making, and agency*. Eindhoven: Eindhoven University of Technology

244. Logtenberg, A. (30-5-2012) *Questioning the past. Student questioning and historical reasoning* Amsterdam: University of Amsterdam

245. Jacobse, A.E. (11-06-2012) *Can we improve children's thinking?* Groningen: University of Groningen

246. Leppink, J. (20-06-2012) *Propositional manipulation for conceptual understanding of statistics* Maastricht: Maastricht University

247. Van Andel, J (22-06-2012) *Demand-driven Education. An Educational-sociological Investigation*. Amsterdam: VU University Amsterdam

248. Spanjers, I.A.E. (05-07-2012) *Segmentation of Animations: Explaining the Effects on the Learning Process and Learning Outcomes*. Maastricht: Maastricht University

249. Vrijnsen-de Corte, M.C.W. *Researching the Teacher-Researcher. Practice-based research in Dutch Professional Development Schools* Eindhoven: Eindhoven University of Technology

250. Van de Pol, J.E. (28-09-2012) *Scaffolding in teacher-student interaction. Exploring, measuring promoting and evaluating scaffolding* Amsterdam: University of Amsterdam

251. Phielix, C. (28-09-2012) *Enhancing Collaboration through Assessment & Reflection* [Samenwerking Verbeteren door middel van Beoordeling en Reflectie] Utrecht: Utrecht University

252. Peltenburg, M.C. (24-10-2012) *Mathematical potential of special education students* Utrecht: Utrecht University

253. Doppenberg, J.J. (24-10-2012) *Collaborative teacher learning: settings, foci and powerful moments* Eindhoven: Eindhoven University of Technology

254. Kenbeek, W.K. (31-10-2012) *Back to the drawing board. Creating drawing or text summaries in support of System Dynamics modeling* Enschede: University of Twente

255. De Feijter, J.M. (09-11-2012) *Learning from error to improve patient safety* Maastricht: Maastricht University

256. Timmermans, A.C. (27-11-2012) *Value added in educational accountability: Possible, fair and useful?* Groningen: University of Groningen

257. Van der Linden, P.W.J. (20-12-2012) *A design-based approach to introducing student teachers in conducting and using research*. Eindhoven: Eindhoven University of Technology

258. Noroozi, O. (11-01-2013) *Fostering Argumentation-Based Computer-Supported Collaborative Learning in Higher Education* Wageningen: Wageningen University

259. Bijker, M.M. (22-03-2013) *Understanding the gap between business school and the workplace: Overconfidence, maximizing benefits, and the curriculum* Heerlen: Open University of the Netherlands

260. Belo, N.A.H. (27-03-2013) *Engaging students in the study of physics* Leiden: Leiden University

261. Jong, R.J. de (11-04-2013) *Student teachers' practical knowledge, discipline strategies, and the teacher-class relationship* Leiden: Leiden University

262. Verberg, C.P.M. (18-04-2013) *The characteristics of a negotiated assessment procedure to promote teacher learning* Leiden: Leiden University

263. Dekker-Groen, A. (19-04-2013) *Teacher competences for supporting students' reflection. Standards, training, and practice* Utrecht: Utrecht University

264. M.H. Knol (19-04-2013). *Improving university lectures with feedback and consultation*. Amsterdam: University of Amsterdam

265. Diggelen, M.R. van (21-05-2013) *Effects of a self-assessment procedure on VET teachers' competencies in coaching students' reflection skills* Eindhoven: Eindhoven University of Technology

266. Azkiyah, S.N. (23-5-2013) *The effects of Two Interventions - on Teaching Quality and Student Outcome* Groningen: University of Groningen

267. Taminiau, E.M.C. (24-05-2013) *Advisory Models for On-Demand Learning* Heerlen: Open University of the Netherlands

268. Milliano, I.I.C.M. de (24-05-2013) *Literacy development of low-achieving adolescents. The role of engagement in academic reading and writing* Amsterdam: University of Amsterdam

269. Vandyck, I.J.J. (17-06-2013), *Fostering Community Development in School-University Partnerships*. Amsterdam: VU Universitiy Amsterdam

270. Hornstra, T.E. (17-06-2013) *Motivational developments in primary school. Group-specific differences in varying learning contexts* Amsterdam: University of Amsterdam

271. Keuvelaar-Van den Bergh, L. (26-06-2013) *Teacher Feedback during Active Learning: The Development and Evaluation of a Professional Development Programme*. Eindhoven: Eindhoven University of Technology.

272. Meeuwen, L.W. van (06-09-13) *Visual Problem Solving and Self-regulation in Training Air Traffic Control* Heerlen: Open University of the Netherlands

273. Pillen, M.T. (12-09-2013) *Professional identity tensions of beginning teachers* Eindhoven: Eindhoven University of Technology

274. Kleijn, R.A.M. de, (27-09-2013) *Master's thesis supervision. Feedback, interpersonal relationships and adaptivity* Utrecht: Utrecht University

275. Bezdan, E. (04-10-2013) *Graphical Overviews in Hypertext Learning Environments: When One Size Does Not Fit All* Heerlen: Open University of the Netherlands

276. Bronkhorst, L.H. (4-10-2013) *Research-based teacher education: Interactions between research and teaching* Utrecht: Utrecht University

277. Popov, V. (8-10-2013) *Scripting Intercultural Computer-Supported Collaborative Learning in Higher Education* Wageningen: Wageningen University

278. Dolfing, R. (23-10-2013) *Teachers' Professional Development in Context-based Chemistry Education. Strategies to Support Teachers in Developing Domain-specific Expertise*. Utrecht: Utrecht University

279. Lucero, M.L. (21-11-2013) Considering teacher cognitions in teacher professional development: Studies involving Ecuadorian primary school teachers Ghent: Ghent University
280. Kamp, R.J.A. (28-11-2013) *Peer feedback to enhance learning in problem-based tutorial groups* Maastricht: Maastricht University
281. Cviko, A. (19-12-2013) *Teacher Roles and Pupil Outcomes. In technology-rich early literacy learning* Enschede: University of Twente