Studies in Educational Evaluation 58 (2018) 51–59

Contents lists available at ScienceDirect

Studies in Educational Evaluation

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

Accuracy of primary school children's immediate and delayed judgments of learning about problem-solving tasks

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

[a] Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, The Netherlands
[b] Department of Education, Utrecht University, The Netherlands
[c] Department of Educational Development and Research, Maastricht University, The Netherlands
[d] Early Start Research Institute, University of Wollongong, Australia

ARTICLE INFO

Keywords:
Judgments of learning
Monitoring accuracy
Problem solving
Mental effort
Primary education

ABSTRACT

Accuracy of students' judgments of learning (JOLs) plays an important role in self-regulated learning. Most studies on JOL accuracy have focused on learning word pairs and text but problems-solving tasks are also very important in education. This study investigated whether children in grade 3 could differentiate in their JOLs between problem-solving tasks that varied in complexity. Participants ($N = 76$, 8–10 years old) 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. 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.

1. Introduction

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, quizzes, or short expository texts (Bol, Hacker, O'Shea, & Allen, 2005; Hacker, Bol, & Bahbahani, 2008; 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 Efklides, 2002).

Ackerman and Thompson (2014) described meta-reasoning as the process by which learners monitor and control reasoning, problem solving and decision-making processes. 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 important when learning to solve problems, few studies have investigated these processes in problem-solving tasks. Here, we take a first step towards 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.

1.1. 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, because successful self-regulated learning depends on accurate monitoring and control processes (e.g., Winne & Hadwin, 1998). Research has shown that when metacognitive knowledge, monitoring and control are adequate, learning is enhanced

* Corresponding author at: Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands.
E-mail address: baars@essb.eur.nl (M. Baars).

https://doi.org/10.1016/j.stueduc.2018.05.010
Received 21 February 2018; Received in revised form 9 May 2018; Accepted 30 May 2018
Available online 02 June 2018
0191-491X/

(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 monitoring affects subsequent control (or regulation) of the learning process. Control or regulation of the learning process can involve choices about which items need to be studied next or practiced further and how much time one should spend on them. 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. The difference between these two types of JOLs is usually related to the type of materials that are used in a study. That is, to monitor more complex materials such as expository text, a student should not only monitor memory but also whether he or she understood the text (i.e., comprehension).

In typical studies on monitoring accuracy using JOLs (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 et al., 2003; Thiede, Dunlosky, Griffin, & Wiley, 2005), participants study items such as word pairs or expository 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, much subsequent research has focused on finding ways to improve monitoring accuracy.

1.2. Improving monitoring accuracy

In a well-known study with word pairs, Nelson and Dunlosky (1991) established that when asking about JOLs not immediately after studying a word-pair but after studying all word pairs, relative accuracy was higher. Nelson and Dunlosky called this the delayed-JOL effect, which they explained based on memory systems involved in making JOLs. 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. Schneider, Visé, Lockl, and Nelson (2000) found the delayed-JOL effect 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 the delayed-JOL effect did not apply to materials that were 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 is comparable 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., 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, Van Merriënboer, & Paas, 1998; Sweller, 2010). 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).

Only a few studies investigated the accuracy of monitoring judgments (e.g., JOLs) in classroom settings. Several studies showed that children's monitoring judgments and regulation skills improved during school years (for a review see, Schneider, 2008). Primary school children were shown to be able to monitor whether their answers were correct or incorrect and regulate their learning accordingly, when

learning from an educational movie on sugar production (Krebs & Roebers, 2012; Roebers, Schmid, & Roderer, 2009). Also, Schneider et al. (2000) showed that -just like adults- children of 6, 8 and 10 years old made more accurate JOLs after a delay when learning with picture-word cards. Furthermore, recently a few studies showed that generation strategies can also improve monitoring accuracy for children in primary and secondary education studying more complex materials. A study by de Bruin, Thiede, Camp, and Redford (2011) showed that generating keywords at a delay after reading a text improved JOL accuracy for 9–10 and 12–13 years olds, which is in line with the delayed-generation effect found with adults (Thiede et al., 2003). Also, Redford, Thiede, Wiley, and Griffin (2012) showed that making concept maps of a text was found to improve JOL accuracy for 12–13 year olds.

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. In line with Ackerman and Thompson (2014), we will argue below that 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) but there are also important differences.

1.3. Judgments of learning (JOLs) about problem-solving tasks

To the best of our knowledge, there are only a few studies that have investigated JOLs in problem-solving tasks (Baars, Van Gog, De Bruin, & Paas, 2014; 2017; Baars, Visser, Van Gog, De Bruin, & Paas, 2013; De Bruin, Rikers, & Schmidt, 2005; De Bruin, Rikers, & Schmidt, 2007). However, most of these studies investigated JOL accuracy after studying a combination of worked examples and practice problems (Baars et al., 2014; 2017) or completion problems (Baars et al., 2013). In the classroom children often learn by simply practicing problem-solving tasks. 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. This means it is an open question whether children are able to accurately monitor learning when practicing with problem-solving tasks. 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 their understanding of math problem-solving tasks of different complexity. 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 (research question 1)? Task complexity can partly be determined objectively, by looking at the number of interacting elements a task contains. However, task complexity is also dependent on prior knowledge which can differ between students. 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 & Sweller, 2004; Kalyuga, 2007). Consequently, there will be individual differences in experienced cognitive load within objectively identified levels of task complexity. Indeed, using completion problems and worked examples, Baars et al. (2013) found a negative correlation between experienced cognitive load and JOLs. In line with these findings we expect that 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, Hypothesis 1).

The second goal of this study was to investigate whether the timing of JOLs about problem-solving tasks (i.e., immediate vs. delayed) affects judgment accuracy (research question 2). 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 understanding of a problem-solving procedure which is represented in 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, 2008) 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 could 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. Thus we expected immediate JOLs to be more accurate than delayed JOLs (Hypothesis 2).

2. Method

2.1. Participants and design

Participants were 131 Dutch primary education students in grade three (8–10 years old, 39 boys and 37 girls) from five different classrooms from five schools in a medium sized town in the Southwest of the Netherlands. The SES of inhabitants of this town was comparable to the average SES in the Netherlands (Rijksinstituut voor Volksgezondheid en Milieu, 2010). Only students with scores of B, C, or D on a standardized math test taken shortly before the study were included in the analysis (n = 76). This excludes the very low [E] or very high [A] math ability students because the learning materials used in this study presumably were too complex or too easy, respectively, for these students to find sufficient variation in JOLs and test performance. Four students were

excluded from the analysis because they did not follow instructions during the experiment. 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). Furthermore, a pretest was used to check whether randomization of students with different ability levels over the conditions was successful.

2.2. Materials

All materials were paper-based and children could use their own pencils to write down their answers. Also, for developing the arithmetic problem-solving tasks, we used existing school tasks that were adapted for this study in collaboration with one of the teachers.

2.2.1. Pretest

To measure children's knowledge and ability on the arithmetic problems used in this study, a pretest was administered. The pretest consisted of four arithmetic problems that could be solved using the same procedure as for the problems used in the practice phase (see the next paragraph) but with different numbers (i.e., isomorphic problems, see Table 1).

2.2.2. Practice problems

Four arithmetic problems were used which teachers considered to differ in complexity and solving procedure, one of each of the following types (in order of increasing complexity): addition without carrying (Level 1: e.g., 414 + 135 + 250), subtraction with borrowing tens (Level 2: e.g., 676 − 139), addition with carrying (Level 3: e.g., 119 + 313 + 238), and subtraction with borrowing tens and hundreds (Level 4: e.g., 634 − 497). The problem solving procedure was different for each level and children had to understand each procedure and know when to apply it, to be able to solve the problem-solving task at a certain level. According to the teachers, the children were familiar with the procedures used in addition and subtraction with borrowing tens with three digits numbers (Level 1 and 2) but had not yet practiced with addition with carrying and subtraction with borrowing tens and hundreds (Level 3 and 4). In Table 1, an overview of the different levels of arithmetic problems is provided.

2.2.3. Posttest

The posttest consisted of four arithmetic problems that could be solved using the same procedure as for the problems used in the pretest and practice phase but with different numbers (i.e., isomorphic problems, see Table 1). Posttest scores were used to calculate JOL accuracy (see the section Data Analysis, Relative accuracy; Absolute accuracy).

2.2.4. Mental effort ratings

Directly after each problem, students rated the amount of mental effort they invested in attempting to solve that problem on a five-point rating scale, ranging from (1) very low mental effort, to (5) very high mental effort (cf. Paas, 1992). The original nine-point rating scale developed by Paas (1992) was adjusted 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?*

2.2.5. JOLs

JOLs were provided on a five-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).

2.3. Procedure

This experiment was run in small group sessions ranging from ten 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 the students 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 problem 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.

2.4. Data analysis

2.4.1. Test performance and complexity

Children's performance on the test problems was judged as either incorrect (0) or correct (1). To investigate whether the items on the posttest all measured one component (i.e., performance on arithmetic tasks), a principal component analysis (PCA) was conducted on the 4 items of the posttest with oblique rotation (promax). The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis, KMO = .592, and all KMO values for individual items were > 0.65 which is above the acceptable limit of .5 (Field, 2009). Bartlett's test of sphericity, χ^2 (6) = 22.017, p = .001, indicated that correlations between items were sufficiently large for PCA. An initial analysis was ran to obtain eigenvalues for each component in the data. One component had an eigenvalue over Kaiser's criterion of 1 and explained 41.34% of the variance. Table 2 shows the component matrix. All items cluster on one component which shows how well students performed on the arithmetic tasks.

To analyze whether the dependent variables (JOLs, Mental Effort, and Performance) were affected by the complexity of the problem-solving tasks, a repeated measure ANOVA was used with the four different tasks, which increased in complexity, as levels of the within-subject factor.

Table 1
Overview of arithmetic problem-solving tasks.

Level	Pretest	Practice	Posttest
1: Addition without carrying	414 + 135 + 250 =	111 + 115 + 361 =	220 + 310 + 431 =
2: Addition with carrying	119 + 313 + 238 =	210 + 207 + 564 =	159 + 216 + 300 =
3: Subtraction with borrowing tens	676 – 139 =735 – 209 =653 –		434 =
4: Subtracting with borrowing tens and hundreds	634 – 497 =848 – 179 =624 –		196 =

Table 2
Summary of exploratory factor analysis for the posttest.

Item	1: Arithmetic
Level 1	−.189
Level 2	.301
Level 3	.835
Level 4	.860
Eigenvalues	1.679
% of variance	41.34
α	.62

2.4.2. 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). Relative accuracy expressed as 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, Shield, Wheeler, & Zacchilli, 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 $n = 26$).

2.4.3. Absolute accuracy

Next to relative accuracy, which shows whether students were able to discriminate between different problem types, absolute accuracy shows what the actual deviation between the JOL and test performance on a problem type is. Thus absolute accuracy shows the calibration of students' JOLs with their performance. Because the JOL scale 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. The scoring system is shown in Table 3. The absolute accuracy score varied between 0 and 1 (i.e., 0, 0.25, 0.50, 0.75, and 1), based on each possible combination of JOL (1–5) and test performance (0 or 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 3
Absolute monitoring accuracy scoring system.

Judgments of Learning:	Test performance:	
	Incorrect	Correct
1	1	0
2	0.75	0.25
3	0.50	0.50
4	0.25	0.75
5	0	1

3. Results

To check whether the randomization to the conditions had been successful, the pretest performance data were compared, which -as expected- showed no differences between the Immediate and Delayed JOL Condition, $t(74) = 0.89$, $p = .38$. 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 4.

3.1. Monitoring task complexity

A repeated measures ANOVA with JOLs as dependent variable, Complexity (four levels) as within-subjects factor and Condition (Immediate vs. Delayed JOLs) as between-subjects factor, showed a main effect of Complexity, $F(3, 219) = 3.29$, $p = .021$, ηp2 = η_p^2.04. Contrasts revealed that JOLs were significantly lower for the fourth level of complexity compared to the first level, $F(1, 73) = 6.89$, $p = .011$, ηp2 = $.\eta_p^2$09, and the second level, $F(1, 73) = 5.25$, $p = .025$, ηp2 = η_p^2.07, but not compared to the third level, $F(1, 73) = 2.32$, $p = .132$, ηp2 = η_p^2.03. However, there was no significant main effect of Condition, $F(1, 73) = 2.81$, $p = .098$, ηp2 = η_p^2.04, nor an interaction effect, $F(3, 219) = 1.38$, $p = .250$, ηp2 = η_p^2.02.

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, $F(3, 207) = 4.60$, $p = .004$, ηp2 = η_p^2.06. Contrasts revealed that mental effort ratings were significantly higher for the fourth level of complexity compared to the first level, $F(1, 69) = 8.55$, $p = .005$, ηp2 = η_p^2.11, the second level, $F(1, 69) = 13.76$, $p\,\eta_p^2$.17, and the third level, $F(1, 69) = 4.91$, $p = .030$, ηp2 = η_p^2.07. As expected, there was no main effect of Condition, $F(1, 69) = 0.37$, $p = .546$, ηp2 = η_p^2.01, nor an interaction, $F(3, 207) = 0.26$, $p = .856$, ηp2 < η_p^2.01.

A repeated measures ANOVA with performance in the practice 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, $F(3, 222) = 24.37$, $p\,\eta_p^2$.25. Contrasts revealed that performance was significantly lower for the fourth level of complexity compared to the first level, $F(1, 74) = 64.05$, $p\,\eta_p^2$.46, and the second level, $F(1, 74) = 26.73$, $p\,\eta_p^2$.27, but not compared to the third level, $F(1, 74) = 1.22$, $p = .274$, ηp2 = η_p^2.02. As expected, there was no main effect of Condition, $F(1, 74) = 0.81$, $p = .371$, ηp2 = η_p^2.01, nor an interaction effect, $F(3, 222) = 0.74$, $p = .528$, ηp2 = $_p^2$.01.

Moreover, in line with Hypothesis 1, 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, $r = -.59$, $t(37) = -4.28$, $p < .001$, and the Immediate JOLs condition, $r = -.67$, $t(34) = -5.12$, $p < .001$.

3.2. Immediate vs. delayed JOL accuracy

3.2.1. Relative monitoring accuracy

The mean gamma correlation of the Immediate JOL condition differed significantly from zero, $t(25) = 2.63$, $p = .015$, whereas that of the Delayed JOL condition did not differ significantly from zero, $t(25) = -.16$, $p = .878$. That is, in contrast to delayed JOLs, immediate JOLs were more accurate than chance. A *t*test showed a trend indicating a difference in gamma correlations between the conditions, $t(50) = 1.86$, $p = .068$, Cohen's $d = 0.52$ (medium effect size). As Fig. 1 shows, monitoring accuracy tended to be higher in the Immediate JOLs condition ($M = .38$, SD = *SD*.75) than in the Delayed JOLs condition ($M = -.03$, SD = *SD*.84).

Table 4
Percentages of correct performance, mean subjective mental effort ratings (range: 1–5), and mean Judgments of Learning (JOLs, range: 1–5) during the practice phase for the different problem categories (1 = least complex; 4 = most complex) for both conditions.

	Immediate JOL condition				Delayed JOL condition			
Complexity levels	Pretest	Percentage correct	Mean mental effort (SD)	Mean JOLs (SD)	Pretest	Percentage 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)

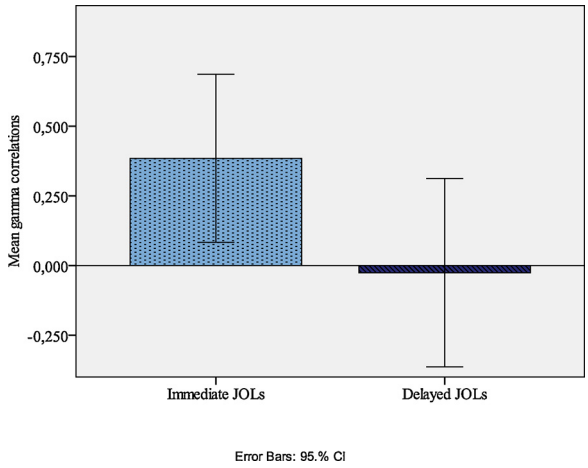

Fig. 1. Mean relative monitoring accuracy presented by group. Error bars represent standard errors of the mean.

3.2.2. Absolute monitoring accuracy

Concerning absolute accuracy, there was a trend in the same direction with the absolute accuracy for immediate JOLs being higher, but a *t*-test showed that the difference between the two conditions was not significant (Immediate: $M = 0.62$, $SD = 0.17$, Delayed: $M = 0.56$, $SD = 0.18$; $t(72) = 1.56$, $p = 0.123$).

4. Discussion

This study explored whether third graders' monitoring judgments about problem-solving tasks would be sensitive to variations in task complexity as reflected in invested mental effort (Research question 1), as well as whether the accuracy of immediate and delayed judgment of learning (JOLs) would differ (Research question 2). The results suggested 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. Yet, effect sizes were small (for JOLs: $\eta_p^2 = .04$ and for mental effort ratings: $\eta_p^2 = .06$). Also, as complexity of the items became larger, performance seemed to decrease much more than JOLs. 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. Although this is a correlation, 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. Indeed, according to Koriat, Nussinson, and Ackerman (2014) study effort can be used as a cue for JOLs. Another explanation could be that the differences in complexity were the common cause for the changes in mental effort and JOLs. Undorf and Erdfelder (2011) found that when participants had no information on task difficulty, encoding fluency was used as a cue to make JOLs. Future research could disentangle the relation between complexity, mental effort and JOLs by using the experimental set-up Undorf and Erdfelder used to control for the influence of item difficulty (see Undorf & Erdfelder, 2011).

Nevertheless, relative accuracy of JOLs, which shows the ability to discriminate between tasks, did not seem to be very high. In line with Hypothesis 2, 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 trend showed a difference in relative monitoring accuracy between the immediate JOLs and delayed JOLs condition, in favor of the immediate JOLs condition. However, in contrast to Hypothesis 2, absolute accuracy did not differ significantly between immediate and delayed JOL conditions. Although, measuring both relative and absolute accuracy makes it more difficult to interpret the findings, it is recommended one use multiple measures of accuracy because this creates the opportunity to analyze different aspects of monitoring accuracy (Schraw, 2009). That is, relative accuracy showed whether students were able to discriminate between the different types of math tasks whereas absolute accuracy showed how big the deviation between the JOL and the actual performance on the math task was.

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 showed 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, 1998a).

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 easily 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 post-dictions of practice problems could be predictive of performance on the test problems as well. This is also in line with the Memory for Past Test performance heuristic (Finn & Metcalfe, 2007, 2008), which states that people use their memories of past test performance as a heuristic to make JOLs on subsequent trials. Ackerman and Thompson (2014) also suggested that monitoring judgments about problem-solving tasks are probably based on a heuristic cue (e.g., fluency). These heuristics cues could have been more salient immediately after solving a problem compared to after a delay. Yet, it is possible that students also used other cues and information to base their JOLs on, for example, prior experience with similar tasks in the classroom. Therefore, future research could investigate the cues and information students use to make JOLs.

When solving arithmetic problems like the ones in the current study, students could use the wrong strategy and fill out a wrong number as their answer. In this case, students probably thought they solved the problem, and based their JOL on this idea. Indeed, Van Loon, de Bruin, van Gog, and van Merriënboer (2013) found that commission errors led to overconfidence in JOLs. To be able to analyze this for problem-solving tasks like arithmetic problems, future research could also investigate worked-out answers to arithmetic problem-solving tasks and take into account what kind of errors children make when solving and how this relates to JOL accuracy.

In sum, if students used 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. Also, students were only given the problem category description when making delayed JOLs, which may have made it harder for them to use their experiences during the task to make a specific JOL. 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 carrying is necessary 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. In the current study, there were no questions or comments from the children about the problem category descriptions being unclear. Furthermore, because the delayed JOLs were prompted sequentially after all the practice problems were solved, the delays between each practice problem and JOL were not consistent. Yet, the length of the delay might be of influence on JOL accuracy. 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 the issue of linking delayed JOLs to the right problem and consistent delays between practice problems and JOLs.

There are several potential limitations to this study. Firstly, gamma correlations could not be computed for a number of participants due to invariance in scores (i.e., ties). 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 (e.g., Anderson & Thiede, 2008). The exclusion of ties can lead to a systematic bias in gamma correlations (Masson & Rotello, 2009). Furthermore, using a low number of items was found to lead to underestimation of calibration between judgments and performance (Weaver, 1990) and a skewed distribution of gamma correlations (Nietfeld, Enders, & Schraw, 2006). Therefore, future research could attempt to replicate these findings regarding relative accuracy using more problem-solving tasks. Second, the absolute accuracy measure combined two different scales by creating five intervals in the performance scale (0, 0.25, 0.50, 0.75, 1) to be able to compare it to the five point JOL scale. Possibly, these intervals did not match the JOL the students gave on the JOL scale. Future research could use the same scales for JOLs and performance. Third, 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 problems as a cue for their JOL. However, a conceptual replication of our study by Meijaard, Baars, De Maeyer and Gijbels (2018) *without* mental effort ratings, revealed that immediate JOLs were more accurate than delayed JOLs for complex problem-solving tasks in vocational education. These results, which were obtained without potential contamination of JOL, suggest that mental effort ratings in our study did not contaminate the JOL accuracy findings. In addition, whereas the time interval between the mental effort ratings and JOLs substantially differed between the immediate and delayed JOL condition, the correlation between these measures did not differ significantly. This can also be considered suggestive evidence against contamination of JOLs by mental effort ratings. In addition, an alternative explanation for the correlation between JOLs and mental effort could be that students first judged their understanding implicitly and use that to make a mental effort rating. Therefore, future research could address the question of whether or not JOLs and mental effort ratings would differ when JOLs are made prior to effort ratings. Also, because mental effort ratings are subjective measures of cognitive load, it would be interesting to include a behavioral measure of invested effort, such as reaction time. Fourth, because we used isomorphic test problems that had exactly the same problem-solving procedure but different numbers, it is unclear to what extent problem format affects JOL accuracy. Future research could provide more clarity on this matter by using identical as well as isomorphic test problems, although for education being able to show high monitoring accuracy regarding the latter is much more interesting since children are hardly ever requested to solve the exact same problem again on a test. In addition, the removal of participants with high and low math ability scores could have restricted the current study and could reduce generalizability. Future research could use a larger range of complexity within the tasks to enable children from all math ability levels to participate.

In conclusion, despite the bulk of research on accuracy of JOLs about language tasks, to the best of our knowledge, this study was one of the first to explore JOLs about the kind of procedural problem-solving tasks typically encountered in important school domains such as math in a real primary school classroom. The findings that third 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, 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 (Metcalfe, 2009) 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 et al., 2012Kostons, 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.

Declaration of interest

The authors have no competing interests to declare.

Acknowledgements

This research was funded by the Netherlands Organization for Scientific Research (project # 411-07-152). The authors would like to thank the schools and teachers involved in this study for their participation.

References

Ackerman, R., & Thompson, V. A. (2014). Meta-reasoning. In A. Feeney, & V. A. Thompson (Eds.). *Reasoning as memory* (pp. 164–182). Psychology Press.

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

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

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*(4), 395–406. http://dx.doi.org/10.1016/j.cedpsych.2013.09.001.

Baars, M., Van Gog, T., Bruin, A., & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology, 28*(3), 382–391. http://dx.doi.org/10.1002/acp.3008.

Baars, M., Van Gog, T., de Bruin, A., & Paas, F. (2017). Effects of problem solving after worked example study on secondary school children's monitoring accuracy. *Educational Psychology, 37*, 810–834. http://dx.doi.org/10.1080/01443410.2016.1150419.

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. http://dx.doi.org/10.3758/BF03193813.

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. http://dx.doi.org/10.1016/j.learninstruc.2009.03.002.

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*, 269–290. http://dx.doi.org/10.3200/JEXE.73.4.269-290.

de Bruin, A. B., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*, 294–310. http://dx.doi.org/10.1016/j.jecp.2011.02.005.

De Bruin, A. B. H., Rikers, R. M. J. P., & Schmidt, H. G. (2005). Monitoring accuracy and self-regulation when learning to play a chess endgame. *Applied Cognitive Psychology, 19*, 167–181. http://dx.doi.org/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. http://dx.doi.org/10.1080/09541440701326204.

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

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

Efklides, 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.

Field, A. (2009). *Discovering statistics using SPSS.* Sage Publications.

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. http://dx.doi.org/10.1037/0278-7393.33.1.238.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*(1), 19–34. http://dx.doi.org/10.1016/j.jml.2007.03.006.

Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American Psychologist, 34*, 906–911. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.3758/MC.37.7.1001.

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

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning, 3*, 101–121. http://dx.doi.org/10.1007/s11409-008-9021-5.

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. http://dx.doi.org/10.1007/s10648-007-9054-3.

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

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

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009a). The easily learned, easily remembered heuristic in children. *Cognitive Development, 24*, 169–182. http://dx.doi.org/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. http://dx.doi.org/10.1016/j.jecp.2008.10.005.

Koriat, A., Nussinson, R., & Ackerman, R. (2014). Judgments of learning depend on how learners interpret study effort. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1624–1637. http://dx.doi.org/10.1037/xlm0000009.

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. http://dx.doi.org/10.1037/0278-7393.32.3.609.

Kostons, D., Van Gog, T., & Paas, F. (2012). Trainingself-assessment and task-selection skills: A cognitive approach to improvingself-regulated learning. *Learning and Instruction, 22*, 121–132. http://dx.doi.org/10.1016/j.learninstruc.2011.08.004.

Krebs, S. S., & Roebers, C. M. (2012). The impact of retrieval processes, age, general achievement level, and test scoring scheme for children's metacognitive monitoring and controlling. *Metacognition and Learning, 7*(2), 75–90. http://dx.doi.org/10.1007/s11409-011-9079-3.

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. http://dx.doi.org/10.3758/BF03201176.

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. http://dx.doi.org/10.1037/0022-0663.97.4.723.

Masson, M. E., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 509–527. http://dx.doi.org/10.1037/a0014876.

Meijaard, C., Baars, M., De Maeyer, S., Gijbels, D. (2018). Do i understand what I just did? Timing of judgments of learning (JOLs) by vocational education students.

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. http://dx.doi.org/10.3758/MC.38.4.441.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*, 159–163. http://dx.doi.org/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. http://dx.doi.org/10.1037/a0012580.

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. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00147.x.

Nelson, T. O. (1984). A comparision of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133. http://dx.doi.org/10.1037/0033-2909.95.1.109.

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

Nietfeld, J. L., Enders, C. K., & Schraw, G. (2006). A Monte Carlo comparison of measures of relative and absolute monitoring accuracy. *Educational and Psychological Measurement, 66*, 258–271. http://dx.doi.org/10.1177/0013164404273945.

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

Paas, F., 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. http://dx.doi.org/10.1207/S15326985EP3801_8.

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Meta

comprehension accuracy improves across reading trials. *Memory & Cognition, 28*, 1004–1010. http://dx.doi.org/10.3758/BF03209348.

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. http://dx.doi.org/10.1016/j.learninginstruc.2011.10.007.

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. http://dx.doi.org/10.1037/a0021705.

Rijksinstituut voor Volksgezondheid en Milieu (2010). *Gemeentelijk gezondheidsprofiel.* Retrieved fromhttp://www.rivm.nl/media/profielen/gemeentelijst.html.

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*(4), 749–767. http://dx.doi.org/10.1348/978185409X429842.

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education, 2*(3), 114–121. http://dx.doi.org/10.1111/j.1751-228X.2008.00041.x.

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. http://dx.doi.org/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. http://dx.doi.org/10.1007/s11409-008-9031-3.

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.

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

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

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. http://dx.doi.org/10.1037/0022-0663.95.1.66.

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. http://dx.doi.org/10.1037/0278-7393.31.6.1267.

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.). *Handbook of metacognition and education* (pp. 85–106). New York: Routledge.

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. http://dx.doi.org/10.1080/01638530902959927.

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1264–1269. http://dx.doi.org/10.1037/a0023719.

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*, 16–26. http://dx.doi.org/10.1080/00461520701756248.

Van Loon, M. H., de Bruin, A. B., van Gog, T., & van Merriënboer, J. J. (2013). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction, 24*, 15–25. http://dx.doi.org/10.1016/j.learninstruc.2012.08.005.

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. http://dx.doi.org/10.1080/10494821003755510.

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. http://dx.doi.org/10.1016/0749-596X(85)90034-8.

Weaver, C. A., III (1990). Constraining factors is calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 214–222. http://dx.doi.org/10.1037/0278-7393.16.2.214.

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

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