Metacognition and Learning (2022) 17:213–239
https://doi.org/10.1007/s11409-021-09281-9

Effects of self-scoring their math problem solutions on primary school students' monitoring and regulation

Sophie Oudman[1] · **Janneke van de Pol**[1] · **Tamara van Gog**[1]

Received: 16 April 2020 / Accepted: 29 September 2021 / Published online: 18 October 2021

Abstract
Preparing students to become self-regulated learners has become an important goal of primary education. Therefore, it is important to investigate how we can improve self-monitoring and self-regulation accuracy in primary school students. Focusing on mathematics problems, we investigated whether and how (1) high- and low-performing students differed in their monitoring accuracy (i.e., extent to which students' monitoring judgments match their actual performance) and regulation accuracy (i.e., extent to which students' regulation judgments regarding the need for further instruction/practice match their actual need), (2) self-scoring improved students' monitoring and regulation accuracy, (3) high- and low-performing students differed in their monitoring and regulation accuracy after self-scoring, and (4) students' monitoring and regulation judgments are related. On two days, students of 9 – 10 years old from 34 classes solved multiplication and division problems and made monitoring and regulation judgments after each problem type. Next, they self-scored their answers and again made monitoring and regulation judgments. On the multiplication problems, high-performing students made more accurate monitoring and regulation judgments before and after self-scoring than low-performing students. On the division problems, high-performing students made more accurate monitoring judgments before self-scoring than low-performing students, but after self-scoring this difference was no longer present. Self-scoring improved students' monitoring and regulation accuracy, except for low- and high-performing students' regulation accuracy on division problems. Students' monitoring and regulation judgments were related. Our findings suggest that self-scoring may be a suitable tool to foster primary school students' monitoring accuracy and that this translates to some extent into more accurate regulation decisions.

Keywords Monitoring accuracy · Regulation accuracy · Unskilled and unaware · Self-scoring · Primary education · Problem solving

Preparing students to become self-regulated learners has become an important goal of primary education. Not only because students are increasingly required to self-regulate their own learning throughout their entire lifetime, for which primary education lays the

✉ Sophie Oudman
v.s.oudman@uu.nl

[1] Department of Education, Utrecht University, P.O. Box 80.140, 3508 TC Utrecht, The Netherlands

foundation, but also because of the beneficial effects of self-regulated learning skills on academic achievement (McClelland & Cameron, 2011). Several models exist that describe the phases and cognitive processes that are involved in self-regulated learning somewhat differently (e.g., Pintrich, 2000; Winne & Hadwin, 1998; Zimmerman, 2000). However, these models share general features including that three phases can be distinguished in self-regulated learning—a forethought, performance and reflection phase—between which learners switch whenever necessary (Panadero, 2017). Two central processes in most models of self-regulated learning, and in switching between the processes, are self-monitoring (evaluating one's own performance) and self-regulation (controlling one's own study activities; De Bruin & Van Gog, 2012; Panadero, 2017; Griffin et al., 2013). Unfortunately, primary school students' self-monitoring and self-regulation are often inaccurate, and researchers are looking for ways to help them improve these processes (e.g., Baars et al., 2014; García et al., 2016; Van Loon & Roebers, 2017). However, relatively little attention has been paid to self-monitoring, and especially self-regulation, when practicing with problem-solving tasks (van Gog et al., 2020), even though problem solving plays an important role in many primary and secondary school subjects such as mathematics and science (for exceptions, see Baars et al., 2014, 2018; Boekaerts & Rozendaal, 2010; García et al., 2016; Rutherford, 2017). To be able to optimally support primary school students' self-monitoring and self-regulation when acquiring problem-solving skills, we need to gain a better understanding of their self-monitoring and self-regulation accuracy and what interventions are successful for improving accuracy. Moreover, the differences in accuracy between different student groups (e.g., low- vs. high-performing students) are of interest, as these may call for a differential focus in interventions. Therefore, the present study aims to make a novel contribution to the literature by investigating (1) how students' monitoring accuracy *and* regulation accuracy when practicing problem solving, differs between low- and high-performing students, (2) whether self-scoring, an intervention that has been shown to be effective on other types of learning tasks (Van Loon & Roebers, 2017), has beneficial effects on students' monitoring and regulation accuracy on problem-solving tasks, (3) whether there is a differential effect of self-scoring for low- and high-performing students, and (4) whether monitoring and regulation are related. Before specifying our research questions in more detail, we first elaborate on prior literature with regard to monitoring and regulation judgments, the differences between low- and high-performing students, and the effects of self-scoring.

Monitoring and regulation judgments

Accurate monitoring and regulation are necessary for effective self-regulated learning (Dunlosky & Rawson, 2012). In the present study we defined monitoring accuracy as the degree to which students know how well they performed on a task, expressed by the absolute difference between students' judgments of how many problems they answered correctly and the number of problems they actually answered correctly (cf. Baars et al., 2014; Dunlosky & Rawson, 2012). Across studies, the average accuracy of primary school students' monitoring judgments varies enormously depending on the type of task (Boekaerts & Rozendaal, 2010; Rutherford, 2017) and students' age (accuracy is higher for older children; Destan & Roebers, 2015; Roebers et al., 2014). The present study focuses on self-regulated learning of problem-solving tasks in upper primary school, more specifically on fourth-grade students practicing computational tasks (US fourth grade is comparable to

Dutch sixth grade; students of approximately 9-10 years old). To the best of our knowledge, only four studies have focused on primary school students' monitoring judgments of problem-solving tasks. Baars et al. (2014) studied the effect of self-testing after studying worked examples of water jug problems (subtracting and adding volumes) on fifth grade students' monitoring accuracy. Rutherford (2017) studied how second to fifth grade students' monitoring accuracy affected their performance on math problems. García et al. (2016) studied fifth and sixth grade students' monitoring judgments on math problems in relation to online measures of students' metacognitive processes during problem solving. Boekaerts and Rozendaal (2010) studied the effects of math problem type, instruction method, judgment timing, and gender on fifth graders' monitoring accuracy. In all four studies, students mostly overestimated their performance, which is a widespread phenomenon in general (e.g., De Bruin et al., 2017; Kruger & Dunning, 1999).

Self-regulated learning theories generally assume that students' monitoring judgments influence their regulation judgments and that accurate monitoring is a necessary (though not sufficient) precondition for accurate regulation (Metcalfe & Finn, 2008; Pintrich, 2000; Dunlosky & Rawson, 2012; Winne & Hadwin, 1998; Zimmerman, 2000). Regulation judgments are decisions on what subsequent learning actions should be taken to reach a learning goal, such as help-seeking or restudying the learning material (Zimmerman, 2000). As regulation judgments directly influence whether and how students continue learning and indirectly influence whether they will master the learning goals or not, making accurate regulation judgments is important. We use the concept "regulation accuracy" to indicate the extent to which the regulation *judgments* are in line with students' *actual need* for regulation, as indicated by experts (thus actual regulation actions were not measured, which is in line with prior studies; e.g., Baars et al., 2014; Van Loon & Roebers, 2017). Making accurate regulation judgments appears to be a skill that is strongly under development during the upper years of primary school: Studies about primary school students practicing recall (of word-pairs or information from a video) showed that regulation judgments of fifth graders are influenced more strongly by their monitoring judgments than those of third graders and were also far more accurate (i.e., unknown words or definitions were more often selected for restudy). The regulation judgments of the third graders seemed to be rather random (Dufresne & Kobasigawa, 1989; Metcalfe & Finn, 2013; Roebers et al., 2014). It remains an open question to what extent primary school students are capable of making accurate regulation judgments in the context of problem solving and whether their regulation judgments are based on their monitoring judgments. If students' regulation decisions are based on their monitoring judgments, then their regulation judgments are presumably too optimistic (given that students' mostly overestimate their performance; Baars et al., 2014; Boekaerts & Rozendaal, 2010; García et al., 2016; Rutherford, 2017). A potential consequence of regulation that is too optimistic is that students might not seek additional instruction or quit practicing too early and therefore learn less than students who make more accurate judgments.

Prior studies on students' regulation accuracy are mostly in the field of word-pair learning, concept learning, and text comprehension; relatively few have focused on problem solving (e.g., in primary education: Baars et al., 2014; in secondary education: Baars et al., 2013, 2017; Kostons et al., 2012; for a review of these studies see Van Gog et al., 2020). In these studies, regulation judgments involved students being asked to select word pairs, definitions, texts or worked examples for *restudy* (e.g., Baars et al., 2013; Baars et al., 2014, 2017; Dunlosky & Rawson, 2012; Metcalfe & Finn, 2008; Van de Pol et al., 2019; Van Loon & Roebers, 2017), *allocate study time* to word pairs (e.g., Dufresne & Kobasigawa, 1989), or *select the complexity* of the subsequent problem-solving task (e.g., Kostons

et al., 2012). However, common regulatory actions for problem solving in (Dutch) primary school are somewhat different (cf. three most used mathematics lesson books in the Netherlands [Baak et al., 2018; Borghouts et al., 2019a, 2019b] and EDI, a widely applied teaching model: Hollingsworth & Ybarra, 2018). When students have not yet mastered specific problem-solving skills two regulatory actions are most common: (1) Students receive or ask for additional instruction (by the teacher or another student) when they do not understand how to solve the problems, or (2) Students receive or decide to complete additional (comparable) practice problems when they understand how to solve the problems, but still need a relatively long time to solve the problems. When students master a certain type of problem, they can continue working on another/subsequent learning goal. In line with this practice, we defined self-regulation judgments in the present study as students' indications of what they would need: additional instruction, additional practice, both, or nothing.

Unskilled and unaware

When investigating monitoring and regulation accuracy, it is important to also investigate whether there are differences in accuracy between low- vs. high-performing students, as this might call for a differential focus in interventions. The *unskilled-and-unaware effect* refers to the well-known phenomenon that low-performing students seem to overestimate their performances more (i.e., make less accurate monitoring judgments) than high-performing students. That is, low-performing students often think they have mastered the learning material whereas they actually have not. In contrast, high-performing students, seem to overestimate to a lesser extent, if at all (Kruger & Dunning, 1999). Overestimating one's learning is problematic because one could terminate practicing and move on to another task before the initial skill has been mastered and therefore additional practice or instruction would be needed to fully master the initial skill. Thus, this finding that low performing students overestimate their performances more than high-performing students is even more problematic because making appropriate choices about subsequent learning activities is arguably even more important for low-performing students as they are furthest away from mastering the learning goals. However, as we will explain below, research on the unskilled-and-unaware effect has hardly addressed potential consequences for regulatory actions.

There are several possible (non-mutually exclusive) explanations for the difference in judgment accuracy between low- and high-performing students. First, high-performing students' knowledge of the task seems to provide them with more information to recognize their competence and potential knowledge gaps (De Bruin et al., 2017; Kruger & Dunning, 1999). Second, because intrinsic cognitive load is lower when students have more prior knowledge of the tasks, the learning tasks are more cognitively demanding for low performing students. High-performing students may have more cognitive capacity available for solving the math problems and simultaneously monitoring (keeping track of) their performance. This provides high-performing students with more information afterwards on which to base their monitoring judgments, and subsequently, their regulation judgments (Van Gog et al., 2011). Third, wishful thinking amongst low-performing students might influence their monitoring accuracy. This is supported by the study of Serra and DeMarree (2016) who showed that students' desired grades impacted their monitoring judgments and that the discrepancy between the desired and actual performance was larger amongst low-performing than amongst high-performing students.

So far, most research on the unskilled-and-unaware effect has focussed on university students and on learning of word pairs or text comprehension. However, there is one study that suggests that this effect also applies to primary school students who are engaged in problem-solving tasks (García et al., 2016); a finding we aimed to replicate in the present study. Moreover, García et al. (2016) suggest that high-performing students may also make more accurate *regulation* judgments (as opposed to monitoring judgments) than low-performing students, but they did not provide empirical evidence for this suggestion. Because regulation judgments of upper primary school students are not necessarily based on their monitoring judgments (see section Monitoring and Regulation Judgments), the unskilled-and-unaware effect as it occurs in students' monitoring judgments, may not necessarily be translated into their regulation judgments. For designing interventions aimed at improving students' regulation judgments, it is relevant to know whether low-performing primary school students indeed differ from their high-performing peers in their regulation judgment accuracy and hence need a different kind of intervention or teacher support.

Improving judgment accuracy: Effects of self-scoring

Self-scoring one's own test responses can lead to improved monitoring judgments. That is, when students compare their test responses to objectively correct information, they have access to information about the correctness of their answers (Rawson & Dunlosky, 2007). Prior studies on self-scoring were mainly conducted in the field of concept learning and showed that self-scoring improved the monitoring accuracy of primary school students (Van Loon & Roebers, 2017), adolescents (Lipko et al., 2009), and adults (Rawson & Dunlosky, 2007). However, students still overestimated their performance after self-scoring. Overestimation after self-scoring can be caused by students' limited ability or motivation to recognise differences between their answers and the objectively correct information (Dunlosky et al., 2005; Rawson & Dunlosky, 2007). Yet, comparing one's answers on problem-solving tasks and specifically on computational tasks (e.g., 6*274) to the correct answers is probably less challenging than assessing the correctness of one's concept definitions. Therefore, monitoring accuracy after self-scoring might become close to perfect when it comes to computational problem-solving tasks.

As for regulation, Van Loon and Roebers (2017) found that students still made substantially over-optimistic regulation judgments after self-scoring, which could possibly be a result of *hindsight bias*. That is, once students know the right answer, they assume that they knew it all along and would be able to reproduce it correctly in the future (Fischhoff, 1975). Therefore, students might think they do not need an additional intervention such as restudy, even though they made mistakes in their work and an additional intervention would actually be appropriate. As for concept learning, this hindsight effect might also play a role in students' regulation judgments of problem-solving tasks, thus improvements in monitoring might not always translate into improved regulation.

Another interesting question is whether potential differences in monitoring and regulation accuracy, and the relation between these two constructs, between low- and high-performing students would still exist after self-scoring. Self-scoring may close the gap between low- and high-performing students' accuracy as self-scoring provides both groups with information (to base their judgments on) that, in case students accurately self-score their answers, is highly predictive of their actual performances and equally predictive for all students.

The present study

The present study has four aims: First, we aimed to investigate whether the unskilled-and-unaware effect would apply to primary school students' monitoring and regulation judgments with regard to problem solving. Second, we investigated how self-scoring influences students' monitoring and regulation accuracy. A third aim was to explore whether there was a differential effect of self-scoring for high- and low-performing students. Fourth, we explored whether students base their regulation judgments on their monitoring judgments and whether potential improvements in monitoring due to self-scoring might also translate into improved regulation judgments. As monitoring accuracy and possibly also regulation accuracy can vary substantially depending on the type of math problem (Boekaerts & Rozendaal, 2010; Rutherford, 2017), we used two different math tasks here: a multiplication and a division task. The following four research questions (RQ) were addressed:

RQ1: Does the unskilled-and-unaware effect apply to primary school students who are involved in problem-solving tasks?

a. We expected low-performing students to make less accurate monitoring judgments than high-performing students (cf. García et al., 2016).
b. We explored whether low-performing students make less accurate regulation judgments than high-performing students.

RQ2: How does self-scoring affect students' monitoring and regulation accuracy?

a. We expected students' monitoring judgments to be more accurate (i.e., almost perfectly accurate) after self-scoring than before self-scoring (cf. Van Loon & Roebers, 2017).
b. We expected students' regulation judgments to be more accurate after self-scoring than before self-scoring (cf. Van Loon & Roebers, 2017). We did not necessarily expect these to become near-perfectly accurate, as hindsight bias may play a role here.

RQ3: Does the unskilled-and-unaware effect remain after self-scoring? We explored how low- and high-performing students differ in their:

a. monitoring accuracy after self-scoring.
b. regulation accuracy after self-scoring.

RQ4: Are students' regulation judgments related to their monitoring judgments? We explored whether students' regulation judgments are related to their:

a. monitoring judgments before self-scoring, and whether this differs between low- and high performing students.
b. monitoring judgments after self-scoring, and whether this differs between low- and high- performing students.

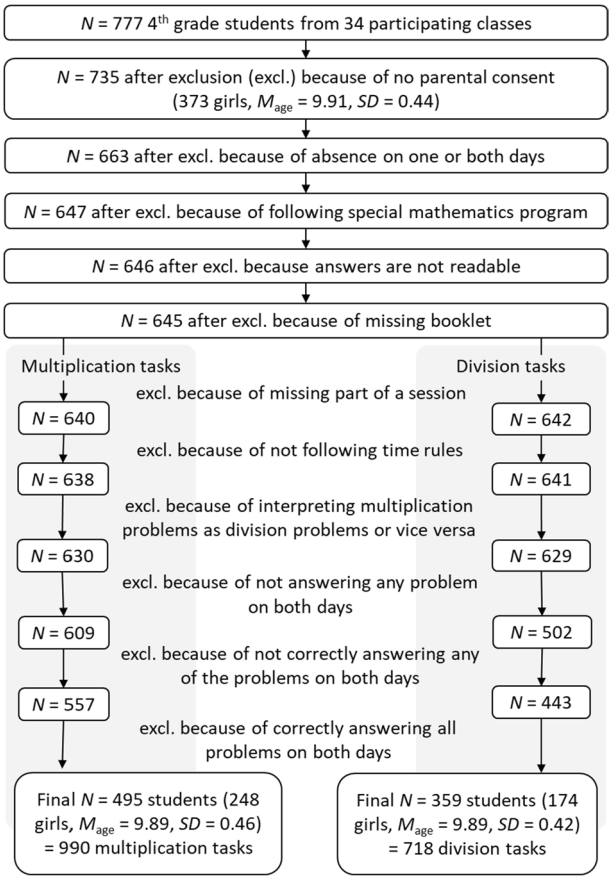

Fig. 1 Flowchart of why and how many students were excluded from all analyses (multivariate outliers were defined for each analysis separately and are still included in the numbers in this flowchart)

Method

Participants

Of the 777 fourth grade (Dutch sixth grade) students who attended the 34 participating classes,[1] data from 495 students were included in the analyses of the multiplication tasks and 359 in the analyses of the division tasks. Two hundred ninety students were included in the analyses for both the division and the multiplication task. The students participated in the current study on two different days with one week in between, working on parallel versions of the tasks.[2] Fig. 1 displays demographics and for which reasons (and how many) students had to be excluded. As Fig. 1 shows, a substantial number of students was excluded because they: 1) did not answer any problem on both days, 2) did not correctly answer any of the problems on both days, or 3) correctly answered all problems on both

[1] These data were collected in the context of a larger research project (that also focusses on teacher judgments).

[2] Data of two days was needed for other studies within this project.

days. The reason these students were excluded from the analyses is that making accurate judgments would be relatively easy for them, because the tasks are presumably far too complex (1 and 2) or far too easy (3) for these students. Including these students could have distorted the results. To draw meaningful conclusions about monitoring and regulation accuracy of low- and high-performing students, the tasks should be at a suitable level of complexity and not far beyond their reach or far too easy (Kostons et al., 2012).

Materials and measures

Student performances

On both days, students answered a set of six multiplication problems (single digit multiplicands multiplied by 3-digit multipliers, e.g., 6×472) and a set of six division problems (3-digit dividends divided by single-digit divisors, e.g., 282: 6). Parallel versions—with isomorphic problems that have the same solution procedure and difficulty, but different numbers—of the two math tasks were administered on the two days. Students received one point for each problem that was solved correctly, thus the performance scores ranged between 0 and 6 per task.

Monitoring judgment (accuracy)

After students completed the multiplication or division task they answered the question "How many of the 6 multiplication/division problems do you think you solved correctly?" on a 7-point scale ranging from 0 to 6 (i.e., monitoring judgment before self-scoring). After self-scoring a set of problems students answered the question "How many of the 6 multiplication/division problems did you solve correctly" on the same 7-point scale (i.e., monitoring judgment after self-scoring). Prior studies on primary school students' monitoring judgments in the context of problem solving used item-specific measures (Baars et al., 2014; Boekaerts & Rozendaal, 2010; García et al., 2016; Rutherford, 2017). The whole-task judgments that we used, on tasks measuring one specific skill, resemble the practice within (Dutch) upper primary school classes: In class students regularly first complete a whole task, then self-score their task, after which they can decide to practice more or ask for help, relying on the feeling they have about the whole task (Baak et al., 2018; Borghouts et al., 2019a, 2019b).

Monitoring bias was computed by subtracting students' actual performance from their monitoring judgment (Baars et al., 2013; Schraw, 2009). Monitoring bias ranged from −6 to 6, with values below zero indicating underestimation and values above zero indicating overestimation. The more the value deviates from zero, the larger students' overestimation or underestimation of their performance is. Because overestimation and underestimation cancel each other out when averaging scores, this measure does not gauge the extent to which judgments are actually accurate, when using it in the analyses. Therefore, monitoring bias is only reported in the descriptive results, but not used in the analyses. Hence, *absolute monitoring accuracy* was analyzed, which is the absolute difference between the judged and actual performance (regardless of whether it was positive or negative), ranging from 0 to 6, with values closer to zero indicating more accurate monitoring judgments (Baars et al., 2013; Schraw, 2009).

Regulation judgment (accuracy)

After the monitoring judgments students indicated which of the following choices was most applicable to them: 1) additional instruction, 2) additional practice, 3) additional instruction and practice, or 4) no additional instruction and no additional practice on the type of problems they just completed. The researchers made it clear to the students that they would not actually receive the additional intervention. Students made these regulation judgments before and after self-scoring. To check if students understood the regulation judgment questions we interviewed 12 fourth grade students (four low, four middle, and four high-performing students) individually after they completed the material, during a pilot study in two fourth-grade classes. All 12 students indicated that they understood the questions.

To determine the accuracy of students' regulation judgments, we first coded students' *actual need for intervention*, based on a coding scheme we developed. We considered students to be in need of an additional intervention when they made (1) procedural errors, which could consist of using a wrong strategy or making wrong use of a correct strategy (these errors are described by Van Zanten et al., 2007), (2) computational errors, indicating sloppiness or a lack of fluency with basic math facts (Calhoon et al., 2007), or (3) exceeding the time limit of 10 min (which, based on the opinion of two math experts and three experienced fourth grade teachers is the maximum amount of time students who have automated the procedures would need), indicating that students did not yet automatize the procedures or, again, lack fluency with basic math facts. Examples of procedural and computational errors are described in Appendix Table 8. We had insight into how students performed the computations, because they had been instructed to use space within the booklets as scrap paper and write out their computations. Students' tasks could not be coded item by item, because procedural errors could only be recognized as such when students made the same error multiple times. Therefore, students' needs were defined at the task level. We distinguished four categories. First, students who correctly answered five or six out of six problems within 10 min were considered to *not need additional instruction or practice*. Second, students who correctly answered five or six problems in more than 10 min or had more than one incorrect answer caused by computational errors were considered to *need additional practice*. Third, students who made procedural errors (specifically, students who gave more than one incorrect answer caused by the use of a wrong strategy or more than two incorrect answers caused by the wrong use of a correct strategy) were considered to *need additional instruction (and practice afterwards)*. We combined the needs "additional instruction" and "additional instruction and practice" into one, because we were not able to decide which of the two needs was more appropriate based on students' work (i.e., their answers and computations that were written out on the scrap paper). In (Dutch) classroom practice, teachers commonly decide during additional instruction to what extent a student needs additional practice afterwards, based on students' understanding during the additional instruction (cf. Baak et al., 2018; Borghouts et al., 2019a, 2019b; Van de Pol et al., 2010). Because actually giving additional instruction was not part of the procedure of our study, we did not know whether or not additional practice after instruction would be needed. However, it is arguably most important that students recognize their need for additional instruction, regardless of whether additional practice would then follow or not (because this can still be decided by the teacher during the additional instruction). Thus, when students' performance indicated they needed additional instruction (and

perhaps practice), the researchers scored both the student judgment "additional instruction" and the judgment "additional instruction and practice" as being accurate. Fourth, students who made one procedural error *and* gave one or more incorrect answers caused by computational errors, were considered to *need additional instruction (and practice afterwards)* or *additional practice* only (in other words, we did not know which intervention was most applicable to the student). When this double code was assigned by the researchers the student judgments "additional instruction", "additional instruction and practice" and "additional practice" were scored as accurate. The detailed coding scheme is depicted in Appendix Fig. 2. To check the interrater reliability of the coding scheme, two coders (the first author and a research assistant) independently coded 10% of the 409 multiplication and 201 division tasks that could not be coded by preprogrammed rules (see Appendix Fig. 2 for these rules). The interrater reliability was substantial for the multiplication tasks ($\kappa = .70$) and almost perfect for the division tasks ($\kappa = .85$; Landis & Koch, 1977). In case of disagreement, the coders reached consensus through discussion. The first author coded the other 90% of the tasks.

Students' *regulation bias* was measured by comparing their regulation judgment to their regulation need (determined by the researchers), which resulted in values ranging from −2 to +2, with values below zero indicating overestimation of their need for intervention and values above zero indicating underestimation of their need for intervention (see Appendix Table 9). Again, regulation bias is only reported in the descriptive results, but not analyzed. We used students' *absolute regulation accuracy* for the analyses, which is the absolute value of students' regulation bias. It ranged from 0 to 2, with values closer to zero indicating more accurate monitoring judgments.

Procedure

After a short introduction by the experimenter, all students received the first booklet and a blue pen, and then started to complete the multiplication task. They were instructed to write down at what time they finished (the time was projected on the digital board in front of the class), but it was emphasized that there was no need to hurry (if students had mastered the content, 10 min should be enough, even without hurrying). When students finished the task, they were instructed to read the (fiction) books they kept in their drawers. After 12 min, the experimenter gave the instruction that the students who had not yet finished all problems should quit the task.[3] Next, the students answered questions in their personal booklets (invested effort, monitoring judgment, second-order monitoring judgment, regulation judgment, and second-order regulation judgment[4]). Each question was separately read aloud and explained by the experimenter. This procedure was then repeated for the division task. Next, all students received a second booklet and changed their blue pen for a green one. In the second booklet, students first self-scored their multiplication answers. Each problem was stated on a separate line together with the correct answer and with two boxes: "correct" and "incorrect or not answered." The experimenter explained that students had to look at their answers in the first booklet and tick the right box (the

[3] Hence, waiting time differed across students. Waiting time did not significantly relate to students' monitoring or regulation accuracy before self-scoring ($p_{\text{monitoring\_multiplication}} = 0.051$, $p_{\text{regulation\_multiplication}} = 0.574$, $p_{\text{monitoring\_division}} = 0.635$, and $p_{\text{regulation\_division}} = 0.105$)

[4] The variables "students' invested effort" and the second-order judgments (i.e., "How confident are you that you made a correct estimation during the previous question?") were not used in in the present study, but collected for use in other studies.

experimenter did not read the correct answers aloud). The following monitoring judgment, regulation judgment and second-order regulation judgment were again read aloud by the experimenter. This procedure of completing the second booklet was then repeated for the division task. This entire procedure (but with isomorphic problems) was repeated exactly one week later.

Analyses

Low-performing and high-performing students were defined separately for both tasks based on their average performance on the problems across the two days. In line with previous studies (De Bruin et al., 2017; Kruger & Dunning, 1999), low-performing students were defined as those scoring in approximately the first (lowest) quartile; high-performing students as those scoring in approximately the fourth (highest) quartile. On the six multiplication problems, low-performing students ($n = 139$) correctly answered 0.5 to 2.5 problems on average. On the six division problems, low-performing students ($n = 101$) correctly answered 0.5 to 1.5 problems. Thirty-nine students were defined as low performing on both the multiplication and division task. High-performing students answered 5.0 or 5.5 problems correctly on the multiplication task ($n = 161$) and division task ($n = 105$). Forty-one students were defined as high performing on both the multiplication and division task.

All analyses were performed separately for the multiplication and the division task. We defined four levels in our data: self-scoring condition (before/after; level 1), day (level 2), student (level 3), and class (level 4). We performed multilevel regression analyses in Mplus version 8 (Muthén & Muthén, 1998 – 2017), using maximum likelihood estimation with robust standard errors (MLR) which is robust to non-normality.[5] The class level was modeled by use of the "Complex" function, because we were not interested in the (fixed or random) effects on this level, we only wanted to account for the non-independence of observations within classes. For the research questions 1A, 1B, 3A, and 3B, about the unskilled-and-unaware effect, the fixed effects were tested at the student level. This means that students' monitoring and regulation accuracy were averaged across the two days. For research questions 2A and 2B, on the effects of self-scoring on students' monitoring and regulation, the fixed effects were tested at the self-scoring condition (before/after) level. For research questions 4A and 4B, on the relation between monitoring and regulation, the fixed effects were tested at the day level. For each of the research questions 1A, 1B, 3A, and 3B, four models were analyzed: two for the multiplication task (intercept only model and model with predictors) and two for the division task (intercept only model and model with predictors; resulting in 16 models for RQ1 and RQ3). For both RQ2A and RQ2B, the four models as described for RQ1 and RQ3 had to be performed three times: for the whole sample and for the low- and high-performing students (resulting in 24 models for RQ2). For both RQ4A and RQ4B six models were estimated; three for the multiplication task (whole sample/ low-performing students/ high-performing students) and three for the division task (whole sample/ low-performing students/ high-performing students). For RQ4A and RQ4B, no intercept only models were analyzed, because nominal variables do not have a measure of variance, resulting in 12 models for RQ4. Thus, in total, 52 models were

[5] To answer research questions 4A and 4B, which asked for multilevel logistic regression models, we also performed the analyses with use of the Supermix software (Hedeker et al., 2008), which uses numerical quadrature instead of the MLR estimator that is used in Mplus. Although the coefficients differed, the statistical significance of the results did not differ between the two software programs.

analyzed, which are all presented as Online Resource. Only the results that are relevant for answering the research questions are presented in the Results section.

In each of the 52 multilevel models, zero to 13 cases (a maximum of 5.4% of the data) were identified as multivariate outliers. We were mainly interested in the results of the analyses without outliers to avoid drawing conclusions that are potentially affected by extreme cases in our data. For transparency we additionally ran the analyses also with outliers. When this led to differences in statistical significance of effects (this was the case for none of the fixed effects and for six variance components), we additionally reported the effects of the analyses with outliers in the Online Resource.

Results

The unskilled-and-unaware effect before self-scoring (RQ1)

Monitoring accuracy (RQ1A)

Descriptive statistics are presented in Table 1. Low-performing students on average substantially overestimated their performance, especially on the multiplication task ($M_{\text{multiplication}} = 1.34$; $M_{\text{division}} = 0.58$). High-performing students slightly underestimated their performances on the multiplication task ($M = -0.18$), but on the division task overestimation and underestimation cancelled each other out ($M = -0.10$, which was not significantly different from 0, see monitoring bias before self-scoring in Table 1). For both tasks, skill group was a statistically significant predictor of absolute monitoring accuracy before self-scoring, with high-performing students making more accurate monitoring judgments than low-performing students (Table 2).

Regulation accuracy (RQ1B)

Low-performing students underestimated their need for intervention, that is, they thought they needed less additional instruction and practice than they actually needed. High-performing students only slightly overestimated their need for intervention for the division task, but for the multiplication task overestimation and underestimation cancelled each other out (see regulation bias before self-scoring in Table 1). Skill group was a statistically significant predictor of absolute regulation accuracy on the multiplication task before self-scoring, with high-performing students making more accurate regulation judgments than low-performing students. On the division task, however, low- and high-performing students did not significantly differ in their absolute regulation accuracy (Table 2).

Effects of self-scoring on students' monitoring and regulation accuracy (RQ2)

Monitoring accuracy (RQ2A)

Table 3 shows that the whole sample of students and the subsets of low- and high-performing students made on average more accurate monitoring judgments after self-scoring, compared to before self-scoring, both on the multiplication and division tasks. The increase for the whole sample was on average about one problem on a set of six problems for the multiplication and division task. Monitoring became close to accurate

Table 1 Means (M) and Standard Deviations (SD) of the main variables of this study

		Multiplication						Division					
		Whole sample ($N=990$)		LP students ($n=278$)		HP students ($n=322$)		Whole sample ($N=718$)		LP students ($n=202$)		HP students ($n=210$)	
Variable	Range	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Performance	0 to 6	3.67	1.84	1.61	1.38	5.26	0.67	3.20	2.06	0.92	0.83	5.30	0.66
Before self-scoring													
Monitoring judgment	0 to 6	4.19	1.52	2.95	1.60	5.08	0.98	3.49	1.91	1.50	1.28	5.21	0.97
Monitoring bias	-6 to 6	0.52	1.62	1.34	1.74	−0.18	1.13	0.29	1.39	0.58	1.34	−0.10 [b]	1.12
Absolute monitoring accuracy [a]	0 to 6	1.23	1.18	1.67	1.42	0.83	0.80	0.99	1.01	1.06	1.00	0.77	0.82
Regulation bias	−2 to 2	0.34	0.86	0.61	0.73	−0.09 [b]	0.68	0.23	0.76	0.33	0.58	−0.09	0.56
Absolute regulation accuracy [a]	0 to 2	0.61	0.70	0.63	0.71	0.37	0.57	0.43	0.65	0.33	0.58	0.25	0.50
After self-scoring													
Monitoring judgment	0 to 6	3.78	1.84	1.75	1.46	5.33	0.67	3.32	2.09	1.05	1.10	5.38	0.67
Monitoring bias	−6 to 6	0.11	0.52	0.14	0.68	0.07	0.26	0.10	0.48	0.15	0.72	0.08	0.34
Absolute monitoring accuracy [a]	0 to 6	0.16	0.50	0.22	0.66	0.07	0.26	0.13	0.48	0.18	0.72	0.10	0.33
Regulation bias	−2 to 2	0.30	0.70	0.46	0.65	0.03 [b]	0.53	0.24	0.65	0.28	0.54	0.02 [b]	0.50
Absolute regulation accuracy [a]	0 to 2	0.44	0.62	0.47	0.62	0.24	0.47	0.36	0.60	0.28	0.54	0.20	0.45

HP high-performing, *LP* low-performing. Means are across both days

[a]Values closer to zero indicating more accurate monitoring judgments

[b]This value does not significantly differ from 0, $p > 0.05$

Table 2 Unstandardized and standardized coefficients for the comparison of low- versus high-performing students' absolute monitoring and regulation accuracy, before and after self-scoring

	Multiplication			Division		
	B	*SE*	β	*B*	*SE*	β
Absolute Monitoring Accuracy						
Before self-scoring	−0.84***	0.12	−0.91	−0.30**	0.11	−0.38
After self-scoring	−0.07*	0.03	−0.58	0.02	0.03	0.15
Absolute Regulation Accuracy						
Before self-scoring	−0.26***	0.05	−0.44	−0.03	0.05	−0.06
After self-scoring	−0.23***	0.04	−0.78	−0.02	0.04	−0.04

Note. Low-performing was coded as 0, high-performing as 1. The full output of the analyses, including intercepts and random effects, are displayed in Tables S1 and S5 (Online Resource)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 3 Effects of self-scoring on absolute monitoring/regulation accuracy

	Whole sample			Low-performing students			High-performing students		
	B	*SE*	β	*B*	*SE*	β	*B*	*SE*	β
Absolute Monitoring Accuracy									
Multiplication	−1.08***	0.04	−0.53	−1.37***	0.08	−0.59	−0.71***	0.04	−0.57
Division	−0.84***	0.05	−0.53	−0.93***	0.12	−0.55	−0.59***	0.05	−0.52
Absolute Regulation Accuracy									
Multiplication	−0.17***	0.02	−0.18	−0.16***	0.04	−0.16	−0.09***	0.02	−0.16
Division	−0.08***	0.02	−0.10	−0.05	0.03	−0.10	−0.02	0.03	−0.03

Note. Before self-scoring was coded as 0, after self-scoring as 1. The full output of the analyses, including intercepts and random effects, are displayed in Tables S2, S3, and S4 (Online Resource)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 4 Percentages of causes of inaccurate absolute monitoring after self-scoring

Cause	Multiplication	Division
Inaccurate self-scoring	79.6	76.1
Inaccurate judgment with unknown cause	5.3	7.5
Both above-mentioned causes applied	1.8	3.0
Students changed their original answers	13.3	13.4

after self-scoring (Table 1). However, 11.4% (for multiplication) and 9.3% (for division) of the students from the whole sample still inaccurately judged their performance even though they had been provided with the correct answers. Note that also high-performing students on average slightly overestimated their performance after self-scoring ($M = 0.07$ for multiplication; $M = 0.08$ for division; Table 1). We additionally explored the causes for the inaccurate monitoring judgments after self-scoring, which are presented in Table 4. The most frequent cause was that students did not accurately self-score their answers.

Table 5 Percentages of Students Whose Regulation Judgment = Nothing Needed, While Actual Need = Additional Practice or Instruction

		Before self-scoring	After self-scoring
Whole sample	Multiplication	25.4	18.8
	Division	12.6	11.3
Low-performing students	Multiplication	22.1	14.7
	Division	7.0	5.2
High-performing students	Multiplication	12.1	11.7
	Division	7.2	10.4

Regulation accuracy (RQ2B)

Across the whole sample, students made more accurate regulation judgments on both tasks after self-scoring, compared to before self-scoring. The regulation accuracy of the subset of low-performing students and high-performing students only increased significantly for multiplication, but not for the division task (Table 3). After self-scoring, the overestimation and underestimations of high-performing students cancelled each other out for both tasks. Low-performing students still underestimated their need for intervention (Table 1).

We additionally explored the frequency of regulation judgment errors. Students who are most "in danger" of not effectively self-regulating their learning process are those who think they do not need any further intervention whereas they actually need one (additional practice or instruction). Therefore it would be relevant to know how often this kind of error occurs. Table 5 shows that on the multiplication task 25% and on the division task 13% of all students made this specific judgment error before self-scoring. For the whole sample of students and for the subset of low-performing students the frequency of this judgment error decreased substantially after self-scoring, especially for the multiplication task. For the high-performing students the frequency hardly decreased after self-scoring for the multiplication task and they made this judgment error even more after self-scoring than before self-scoring, on the division task. At the same time, at least 75% of all students knew whether or not an intervention was needed (Table 5).

The unskilled-and-unaware effect after self-scoring (RQ3)

Monitoring accuracy (RQ3A)

Table 2 shows that on the multiplication task, high-performing students made significantly more accurate monitoring judgments after self-scoring than low-performing students, although this difference was only 0.07 on a seven-point scale. For the division task, high- and low-performing students' monitoring accuracy after self-scoring did not differ significantly.

Regulation accuracy (RQ3B)

High-performing students made more accurate regulation judgments after self-scoring than low-performing students on the multiplication task, but not on the division task (Table 2).

Table 6 Effect of monitoring judgments on regulation judgments, before and after self-scoring

	Whole sample			Low-performing students			High-performing students		
	B	*SE*	Odds ratio	*B*	*SE*	Odds ratio	*B*	*SE*	Odds ratio
Multiplication before self-scoring									
Practice vs. Nothing	−0.99***	0.08	0.37	−0.90***	0.15	0.41	−1.01***	0.17	0.36
Instruction vs. Nothing	−1.40***	0.10	0.25	−1.29***	0.17	0.28	−1.36***	0.19	0.26
Multiplication after self-scoring									
Practice vs. Nothing	−1.19***	0.11	0.30	−0.79***	0.16	0.45	−1.49***	0.31	0.23
Instruction vs. Nothing	−1.66***	0.11	0.19	−1.43***	0.20	0.24	−1.05***	0.11	0.35
Division before self-scoring									
Practice vs. Nothing	−1.18***	0.12	0.31	−0.76**	0.24	0.47	−1.99***	0.29	0.14
Instruction vs. Nothing	−1.67***	0.14	0.19	−1.07***	0.26	0.34	- [a]		–
Division after self-scoring									
Practice vs. Nothing	−1.20***	0.09	0.30	−1.97***	0.54	0.14	−2.18***	0.32	0.11
Instruction vs. Nothing	−1.70***	0.13	0.18	−2.23***	0.55	0.11	- [a]		–

Note. "Practice" refers to students' regulation judgment = additional practice needed. "Instruction" refers to regulation judgment = additional instruction (and practice) needed. "Nothing" refers to regulation judgment = no intervention needed. The full output of the analyses, including intercepts and confidence intervals of the odds ratios are displayed in Tables S6, S7, and S8 (Online Resource). [a] No students within this sample made the "Instruction" judgment, thus this model could not be analyzed

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Relation between monitoring and regulation judgments (RQ4)

Before self-scoring (RQ4A)

Table 6 presents the results of the logistic regression analysis, measuring the effect of students' monitoring judgments on their regulation judgments (i.e., whether students chose for additional practice versus no intervention or for additional instruction [and practice] versus no intervention). Before self-scoring, the magnitude of the monitoring judgments significantly predicted students' regulation judgments after self-scoring. This was the case for the whole sample of students as well as for the subsets of low- and high-performing students, for both the multiplication and division tasks. When students' monitoring judgments increased with one item, the odds of choosing for additional practice or instruction compared to no intervention became roughly two to seven times smaller.[6]

[6] Roughly two and seven times is calculated by dividing one by the odds ratio (1/0.47 = 2.13 and 1/0.14 = 7.14).

After self-scoring (RQ4B)

For the whole sample of students and the low- and high-performing students, the monitoring judgments after self-scoring significantly predicted students' regulation judgments after self-scoring, for both tasks. When students' monitoring judgments increased with one item, the odds of choosing for additional practice or instruction compared to no intervention became roughly two to nine times smaller (Table 6).

Discussion

The current study investigated two key components of primary school students' self-regulated learning, self-monitoring and self-regulation (De Bruin & Van Gog, 2012; Griffin et al., 2013; Panadero, 2017), as well as the interrelation between these two processes. Specifically, we investigated whether the unskilled-and-unaware effect, which states that low-performing students tend to overestimate their performance more than high-performing students, also applies to primary school students' monitoring and regulation judgments with regard to math problem solving (RQ1). In addition, we investigated whether self-scoring students' answers would improve their monitoring and regulation accuracy (RQ2), and whether this differed between low- and high-performing students (RQ3). Finally, we investigated whether students' regulation and monitoring judgments were related, before and after self-scoring (RQ4). Table 7 presents an overview of findings on each of those questions, which we discuss below.

Table 7 Overview of the findings of the present study

	Multiplication	Division
Is there an unskilled-and-unaware effect?		
Monitoring, before self-scoring (RQ1A)	Yes [a]	Yes [a]
Regulation, before self-scoring (RQ1B)	Yes [a]	No [a]
Monitoring, after self-scoring (RQ3A)	Yes [b]	No [a]
Regulation, after self-scoring (RQ3B)	Yes [a]	No [a]
Does self-scoring improve monitoring and regulation?		
Monitoring (RQ2A)	Yes [c]	Yes [c]
Regulation (RQ2B)	Yes [c]	Yes, for whole sample, not for low- and high-performing students
Are monitoring and regulation related?		
Before self-scoring (RQ4A)	Yes [c]	Yes [c]
After self-scoring (RQ4B)	Yes [c]	Yes [c]

[a]Findings were comparable when performance was added to the analyses as continuous variable (instead of low/high) and the whole sample was included (instead of only low- and high-performing students), see Tables S9 and S10 (Online Resource)

[b]This effect was not significant when performance was added to the analyses as continuous variable and the whole sample was included, see Table S10 (Online Resource)

[c]For whole sample and low- and high-performing students

Differences between high- and low-performing students before self-scoring (RQ1)

Very little research has investigated whether the unskilled-and-unaware effect also occurs in primary school students with regard to problem-solving tasks. Moreover, most research focuses on whether this effect is found in monitoring accuracy, not whether it also extends to regulation accuracy. Yet, regulation judgments are very important, as they directly influence whether and how students continue learning and indirectly influence whether students will master the learning goals or not. Therefore, in the present study we also investigated regulation accuracy, measured by asking students to indicate whether they needed additional instruction, practice, or not.

As for monitoring accuracy, in line with our expectations and the findings of García et al. (2016), high-performing students made more accurate monitoring judgments than low-performing students. Interestingly, we also found an unskilled-and-unaware effect in regulation accuracy, with high-performing students making more accurate regulation judgments than low-performing students, though only on the multiplication task and not on the division task. An explanation for why the unskilled-and-unaware effect was more pronounced on the multiplication task might lie in the fact that knowledge gaps were easier to identify for students on the division task than on the multiplication task. This explanation is supported by the fact that students made roughly twice as many omission errors (i.e., lack of answers) in the division task than in the multiplication task, where they made more commission errors (i.e., incorrect answers; Table S11). Thus it seems that, when working on multiplication tasks, students with low skill can easily come up with a "strategy," even if incorrect, by just multiplying "random" digits of the multiplier with the multiplicand (i.e., commission errors) whereas on the division task, students with low skill seem more likely to get stuck and realize they do not know how to solve the problem, leading to an omission error. This finding also underlines that besides primary school students' monitoring accuracy (Boekaerts & Rozendaal, 2010; Rutherford, 2017) also the unskilled-and-unaware effect seems to be influenced by the nature of the learning task, even if tasks are from the same domain (in our case, mathematics). Other explanations for differential findings across the two types of tasks could lie in (1) the fact that the division task was more difficult than the multiplication task (the difference in performance was 0.5 out of six items; Table 1), (2) the fact that students are more familiar with multiplication than with division, because in Dutch primary schools learning multiplications starts at the end of first grade, whereas learning divisions starts at the beginning of third grade (for the role of familiarity see Fitzsimmons et al., 2020), and (3) the possibility that order effects played a role, because students always first completed the multiplication task with corresponding judgments and then the division task with corresponding judgments.

Overall, we found substantial evidence for the unskilled-and-unaware effect. There are several possible (non-mutually exclusive) explanations for this difference in monitoring and regulation accuracy between high- and low-performing students (see also section Unskilled and Unaware). First, high-performing students may make more accurate monitoring judgments, because they have more knowledge of what good performance entails (Kruger & Dunning, 1999). More accurate monitoring judgments amongst the high-performing students compared to low-performing students might also translate into more accurate regulation judgments, as these two judgments are related (see our findings to the fourth research question). Second, because the tasks impose less cognitive load on high-performing than on low-performing students, high-performing students may have more cognitive capacity available for monitoring, which provides them with more information to base their monitoring judgments on, and subsequently their regulation judgments (Van Gog et al., 2011).

Third, low-performing students might suffer the most from wishful thinking as low-performing students' discrepancy between desired and actual performance is larger compared to high-performing students (Serra & DeMarree, 2016). The fourth possibility, which is not mentioned before, entails that, because of their high performance, there was less room for high-performing students to overestimate their performance and underestimate their need for intervention, compared to low-performing students, which may have enhanced the unskilled-and-unaware effect. Note though, that in the present study, this would have been mitigated at least partly by the fact that we excluded students who answered all problems correctly or incorrectly. Moreover, even if this "statistical" explanation would apply, our finding that high-performing students make more accurate judgments is still the reality in the classroom, as the math problems we used were part of the actual fourth grade curriculum.

Besides finding that high-performing students made more accurate monitoring and regulation judgments than low-performing students in general, the type of errors these two groups made, and thus the type of intervention they needed, also differed. While low-performing students mostly made procedural errors and therefore needed additional instruction (and practice afterwards) high-performing students mostly made computational errors and therefore needed additional practice (see Table S12). This finding indicates that low-performing students do not only need more support when regulating their learning than high performing students, but also a different kind of support.

The effects of self-scoring on monitoring and regulation accuracy (RQ2)

The second research question addressed the effectiveness of an intervention to improve monitoring and regulation accuracy: self-scoring of their solutions, based on a standard (i.e., the correct answers). In line with our expectation and prior research with other tasks (concept learning; Van Loon & Roebers, 2017), self-scoring improved the average monitoring accuracy of the whole sample of students and of the subsets of low- and high-performing students. Interestingly, approximately 10% of the students still incorrectly monitored their performance (they were almost always too optimistic), even though they had been provided with the correct answers. Inaccurate monitoring after self-scoring appeared to have three different causes (Table 4). First, most of these students did not accurately self-score their answers (in almost all cases students indicated that a specific answer was correct although it was not). This may be caused by students' limited ability or motivation to recognize differences between their answers and the objectively correct information (Dunlosky et al., 2005; Rawson & Dunlosky, 2007). Second, some of these students changed their original answers (we could see this because we changed the pen color in the self-scoring phase). Possibly, they tried to protect their (self-)image. Third, some of these students gave an incorrect monitoring judgment due to an unknown reason, possibly because they did not correctly add up the number of correct answers.

Across the whole sample, students made more accurate regulation judgments on both the multiplication and division task after self-scoring, compared to before self-scoring. Regulation accuracy of the subsets of low- and high-performing students only increased slightly for the multiplication task and not for the division task. The lack of improvement in regulation accuracy on the division task for the low- and high-performing students could be explained by the fact that students' regulation judgments before self-scoring were more accurate than on the multiplication task (because the knowledge gaps where easier to identify, see above). A possible explanation for the small improvement of the regulation judgments after

self-scoring in general could be that hindsight bias played a role here (i.e., the tendency of students to think that they master the computations, although they made mistakes; Fischhoff, 1975); students might have attributed their mistakes to computational errors (which could be an accurate judgment) and may therefore have concluded that no additional instruction and practice was needed in order to do better next time, although additional practice is also needed to prevent computational errors. Another explanation for why inaccurate regulation judgments did not improve, or improved only slightly after self-scoring, might be that students' standards of when they need an additional intervention differ from the standards of experts. For instance, students might think they need additional instruction or practice when they correctly answered three or less out of six problems, whereas we have set this standard at four or less correct answers (based on the opinion of experts).

Differences between high- and low-performing students after self-scoring (RQ3)

To find out whether low- and high-performing students also need a different focus in interventions after self-scoring, our third question addressed whether the unskilled-and-unaware effect would still be present after self-scoring. Fortunately, as for monitoring accuracy, the differences between low- and high-performing students almost disappeared after self-scoring, and both groups of students came close to perfect accuracy. As for regulation accuracy, on the multiplication task, high-performing students still were substantially more accurate than low-performing students after self-scoring. Nevertheless, the vast majority of low-performing students seemed to have realized after self-scoring that some intervention was needed, but not all of them chose the most suitable intervention; most low-performing students who did not make accurate regulation judgments after self-scoring, indicated they needed additional practice, while they actually needed additional instruction (followed by additional practice afterwards). This finding implies that differential interventions for improving students' regulation accuracy are needed. Whereas low-performing students seem to need help with choosing the most adaptive regulatory action after self-scoring, interventions for high-performing students should maybe focus on the hindsight effect, as their regulation accuracy seems relatively resistant to change. When the hindsight effect can be reduced, high-performing students might decide more often for additional practice when this is indeed an appropriate decision. In turn, this might lead to even higher performances.

Relation between monitoring and regulation judgments (RQ4)

Whereas theories on self-regulated learning generally assume that students' monitoring judgments are (partially) based on their regulation judgments (Pintrich, 2000; Winne & Hadwin, 1998; Zimmerman, 2000), findings of prior studies, which were only in the field of information recall, indicate that this relation starts to appear somewhere in the upper primary school years (Dufresne & Kobasigawa, 1989; Metcalfe & Finn, 2013; Roebers et al., 2014). Our findings showed that fourth grade students' monitoring and regulation judgments regarding math problem solving are, at least to some extent, interrelated. This finding indicates that these students might partially base their regulation judgments on their monitoring judgments, both before and after self-scoring. Interventions aimed at improving students' monitoring accuracy might therefore also, to some extent, translate into improved regulation judgments. However, importantly, our results also indicate that improved monitoring judgments after self-scoring do not always translate into improved

regulation judgments: We found that monitoring accuracy became much closer to perfect accuracy after self-soring than regulation accuracy (Table 1). Moreover, in many cases the proportions of students who inaccurately indicated that they did not need an intervention, hardly changed in regulation judgments from before to after self-scoring (Table 5). Students' regulation judgments thus seemed to be somewhat resistant to change (especially for high-performing students) or did change, but into another inaccurate decision (especially for low-performing students).

Limitations and future research

One limitation of the present study was that a large number of participants had to be excluded, due to several reasons (see Participants section). Note that in regular classroom practice (in the Netherlands), the excluded students would also be those who would get a different task because they are behind or ahead of the lesson aim for the majority of the students (cf. Baak et al., 2018; Borghouts et al., 2019a, 2019b). In future studies, researchers could consider showing the tasks beforehand to the teachers, ask which of their students would normally not get a task of that difficulty, and only exclude these students. Moreover, we still had a sizable sample overall and in the two subsamples of high- and low-performing students. Whereas there was substantial overlap in students included in the multiplication and division task analyses overall, there was only slight overlap within the low- and high-performing subsamples (i.e., students scoring low on division did not necessarily score low on multiplication and vice versa), which could have played a role in finding the unskilled-and-unaware effect for multiplication, but not for the division task.

The current study was the first to use those regulation judgment measures for problem-solving tasks that are highly relevant for teaching and learning in primary school. This measure gave us detailed insight into students' regulation decisions. Students were quite good at indicating whether they needed an intervention or not (both before and after self-scoring), but they often did not know whether additional practice sufficed, or additional instruction (and practice afterwards) was needed because they made procedural errors. There are several potential explanations for this finding, which also provide interesting avenues for future research. First, our way of coding students' needs required some interpretation and might have played a role in some of the discrepancies between students' judgments of their own needs and our judgments of their needs (e.g., our decision to use a time limit of 10 min for determining whether or not additional practice was needed, was based on the opinion of experts, yet for some judgments, a different cut-off could have led to a different classification). Second, our current data do not provide insight into students' motives for regulation decisions. The use of think aloud protocols or interviews might allow for investigating the motives of students with different profiles (e.g., students who noticed during self-scoring that they made many mistakes, but still indicate that they did not need an additional intervention vs. students who indicated they did). Third, (some) students might need additional interventions to be able to make more accurate regulation judgments and investigating their motives might provide valuable input for the design of such interventions. Future research should investigate what effective interventions would be to support students to choose for additional practice or instruction, adapted to their monitoring judgments after self-scoring.

Future studies might consider including item-by-item judgments in addition to whole task judgments when investigating students' monitoring and regulation judgments in the problem-solving context. The whole task judgments we used in this study are more specific than global judgments of one's own general mathematic skills, but somewhat less specific

than item-by-item judgments. Making judgments at this intermediate grain size, at which students judge the extent to which they master a specific skill, is regularly requested of students in primary education (see Method section) and can be useful when students reflect on which specific skills ask for an intervention (Hartwig & Dunlosky, 2017). However, primary school students also make item-specific judgments regularly when working on math problems and future studies could consider comparing the self-regulatory processes involved in solving a single problem and in a complete task (note that in a meta-analysis of Südkamp et al., 2012 no effect of judgment grain size on the accuracy of *teachers'* judgments of student performances was found).

Relatively little research on (improving) monitoring and regulation accuracy has focused on problem-solving tasks in primary education so far. Since the unskilled-and-unaware effect and the effect of self-scoring differed across the multiplication and division task, these effects should be more systematically investigated in different types of problem-solving tasks. For instance, when working on problems that are more ill-structured and more complex than the computational tasks in this study, making accurate judgments and accurately self-scoring one's answers might be more challenging. Moreover, nowadays, schools increasingly start using online learning environments with adaptive math learning programs, in which students receive immediate feedback on their performance. It would be valuable to investigate how different groups of primary school students differ in their help-seeking behavior and how this can be improved when working in these environments (cf. e.g., Roll et al., 2011, who investigated the latter for secondary school students).

Last but not least, our findings may be generalized to schools in which it is common practice (as it is in the Netherlands) that students self-score their answers and are encouraged to take self-regulatory actions such as asking for further instruction or terminating/continuing with practice tasks. For schools in which it is not common practice yet, that would consider implementing self-scoring and subsequent self-regulation, our finding that at least 75% of the students in this study accurately indicated whether or not they needed an additional intervention (Table 5) is very promising. However, future research should further investigate and confirm whether similar findings would be obtained in schools or countries where self-scoring and taking self-regulatory actions are not yet common.

Practical implications and conclusions

The current study, together with previous studies (Baars et al., 2014; Boekaerts & Rozendaal, 2010; García et al., 2016; Rutherford, 2017), showed that primary school students' self-monitoring and self-regulation when practicing with problem solving are not optimal and frequently too optimistic. Our study indicates that having fourth-grade students self-score their math problem solutions is an effective way to increase their monitoring accuracy, and that this partially translates into improved regulation judgments. Thus, the common practice in many Dutch primary schools to have students self-score their answers (Baak et al., 2018; Borghouts et al., 2019a, 2019b) seems to be good practice. While prior research investigated the unskilled-and-unaware effect with regard to monitoring judgments, our study indicated that this effect also applies to regulation judgments and after self-scoring, at least for one of the two tasks used here. Especially the finding that high-performing students still made more accurate monitoring and regulation judgments after self-scoring for one of the two tasks than low-performing students, suggests that low-performing students need more and different support with self-regulating their learning process than high-performing students, when practicing with problem solving.

Appendix

Table 8 Examples of procedural and computational errors

Type of Error	Example when problem is 6 x 472		Example when problem is 228 : 3	
Use of the wrong strategy or lack of use of a specific strategy (procedural error).	Not writing the numbers of the sum correctly under each other.	2400 420 12 + 7800	Split up in the wrong way.	228 : 3 = 200 28
Wrong use of a correct strategy (procedural error).	Forget to add the "small numbers that should be remembered" (the 1 from 12 and 4 from 42).	4 1 472 6 x 2422	Write down numbers double in a long division (in this case the 2 from the lowest 12 should be 8).	3/228\742 21 - 12 12 , 08 6 – 2
Computational error.	Make mistakes in the multiplication tables	6 x 2 = 10 6 x 70 = 480	Make mistakes in the division tables	210 : 3 = 80 18 : 3 = 7

Table 9 Cross tabulation of scoring students' regulation accuracy

	Actual need for intervention (as coded by the researchers)			
	No additional instruction (I) or practice (P)	Additional practice (P)	Additional instruction (and practice afterwards; IP)	P *or* IP
Student judgments				
No I or P	0	1	2	1
P	−1	0	1	0
IP	-2	-1	0	0

Note 0 = accurate; > 0 = underestimation of need for intervention; < 0 = overestimation of their need for intervention. Values closer to zero indicating more accurate regulation judgments

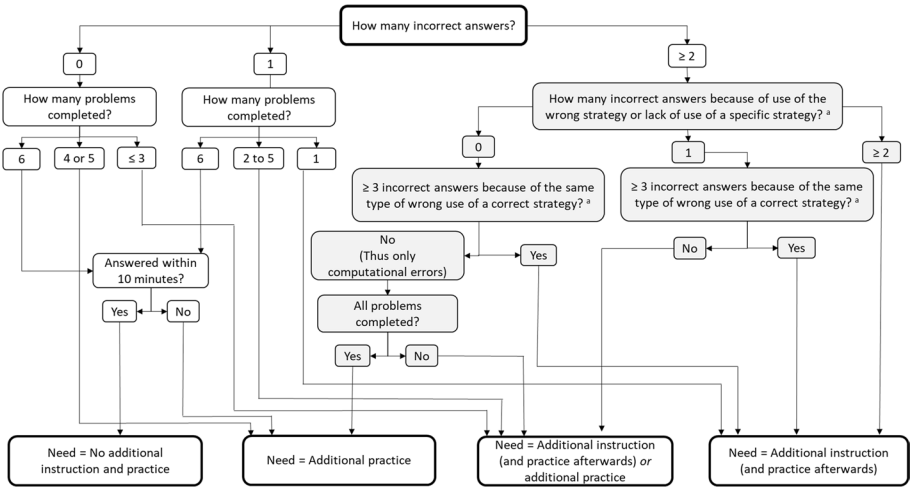

Fig. 2 Coding scheme for actual needs of intervention. Unshaded parts were code automatically, shaded parts were coded manually, [a] See Appendix Table 8 for examples. A question mark/cross/line was coded as "lack of strategy"

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11409-021-09281-9.

Acknowledgements This research was supported by funding from the Dutch Ministry of Education, Culture and Science, (grant number OCW/PromoDoc/1065001). The authors would like to thank the participating teachers and students for their time and effort.

Data availability Data have been stored anonymously on a secure server at our University and we uploaded our data on osf.io: https://www.osf.io/b4rkf/?view_only=70db7d9dffd84f3583a22d64e276a5f1.

Declaration

Conflict of interest The authors declare that they have no conflicts of interest.

Ethics approval Ethical approval was obtained from the Ethics Committee of the Faculty of our University in November 2018.

Consent to participate Before participation, all students and parents of all students were informed about the study procedure and aims, and given the opportunity to decline participation.

Consent for publication All authors agreed with the content and gave explicit consent to submit this study. We also obtained consent from the responsible authorities at our Department of our University to submit this work.

References

Baak, G., Boon, B., Bosma, G., Van der Brink, M., Cornelissen, F., Druif, D., … Wynia, F. (2018) *Getal & ruimte junior handleiding groep 6*. Noordhoff.

Baars, M., Visser, S., Van Gog, T., De Bruin, A. B. H., & Paas, F. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemporary Educational Psychology, 38*(4), 395–406. https://doi.org/10.1016/j.cedpsych.2013.09.001

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

Baars, M., Van Gog, T., De Bruin, A. B. H., & Paas, F. (2017). Effects of problem solving after worked example study on secondary school children's monitoring accuracy. *Educational Psychology, 37*(7), 810–834. https://doi.org/10.1080/01443410.2016.1150419

Baars, M., van Gog, T., de Bruin, A. B. H., & Paas, F. (2018). Accuracy of primary school children's immediate and delayed judgments of learning about problem-solving tasks. *Studies in Educational Evaluation, 58*, 51–59. https://doi.org/10.1016/j.stueduc.2018.05.010

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*(5), 372–382. https://doi.org/10.1016/j.learninstruc.2009.03.002

Borghouts, C., Buter, A., & Gool, A. (2019a). *Pluspunt 4 handleiding groep 6*. Malmberg.

Borghouts, C., Buter, A., & Gool, A. (2019b). *De wereld in getallen 5 handleiding groep 6*. Malmberg.

Calhoon, M. B., Emerson, R. W., Flores, M., & Houchins, D. E. (2007). Computational fluency performance profile of high school students with mathematics disabilities. *Remedial and Special Education, 28*(5), 292–303. https://doi.org/10.1177/07419325070280050401

De Bruin, A. B., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction, 22*(4), 245–252. https://doi.org/10.1016/j.learninstruc.2012.01.003

De Bruin, A. B. H., Kok, E. M., Lobbestael, J., & De Grip, A. (2017). The impact of an online tool for monitoring and regulating learning at university: Overconfidence, learning strategy, and personality. *Metacognition and Learning, 12*(1), 21–43. https://doi.org/10.1007/s11409-016-9159-5

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning, 10*(3). https://doi.org/10.1007/s11409-014-9133-z

Dufresne, A., & Kobasigawa, A. (1989). Children's spontaneous allocation of study time: Differential and sufficient aspects. *Journal of Experimental Child Psychology, 47*(2), 274–296. https://doi.org/10.1016/0022-0965(89)90033-7

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

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*(4), 551–565. https://doi.org/10.1016/j.jml.2005.01.011

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*(3), 288–299. https://doi.org/10.1037/0096-1523.1.3.288

Fitzsimmons, C. J., Thompson, C. A., & Sidney, P. G. (2020). Confident or familiar? The role of familiarity ratings in adults' confidence judgments when estimating fraction magnitudes. *Metacognition and Learning, 15*, 215–231. https://doi.org/10.1007/s11409-020-09225-9

García, T., Rodríguez, C., González-Castro, P., González-Pienda, J. A., & Torrance, M. (2016). Elementary students' metacognitive processes and post-performance calibration on mathematical problem-solving tasks. *Metacognition and Learning, 11*(2), 139–170. https://doi.org/10.1007/s11409-015-9139-1

Griffin, T. D., Wiley, J., & Salas, C. R. (2013). Supporting effective self-regulated learning: The critical role of monitoring. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 19–34). Springer.

Griffin, T. D., Mielicki, M. K., & Wiley, J. (2019). Improving students' metacomprehension accuracy. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 619–646). Cambridge University Press.

Hartwig, M. K., & Dunlosky, J. (2017). Category learning judgments in the classroom: Can students judge how well they know course topics? *Contemporary Educational Psychology, 49*, 80–90. https://doi.org/10.1016/j.cedpsych.2016.12.002

Hedeker, D., Gibbons, R., du Toit, M., & Cheng, Y. (2008). *Supermix: Mixed effects models*. Scientific Software International.

Hollingsworth, J. R., & Ybarra, S. E. (2018). *Explicit direct instruction (EDI): The power of the well-crafted, well-taught lesson*. SAGE Publications.

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*(2), 121–132. https://doi.org/10.1016/j.learninstruc.2011.08.004

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

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*(4), 307–318. https://doi.org/10.1037/a0017599

McClelland, M. M., & Cameron, C. E. (2011). Self-regulation and academic achievement in elementary school children. *New Directions for Child and Adolescent Development, 133*, 29–44. https://doi.org/10.1002/cd.302

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin and Review, 15*(1), 174–179. https://doi.org/10.3758/PBR.15.1.174

Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition and Learning, 8*(1), 19–46. https://doi.org/10.1007/s11409-013-9094-7

Muthén, L. K., & Muthén B. O. (1998-2017). *Mplus user's guide,* 8th edn. Muthén & Muthén.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology, 92*, 544–555. https://doi.org/10.1037/0022-0663.92.3.544

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*(4–5), 559–579. https://doi.org/10.1080/09541440701326022

Roebers, C. M., Krebs, S. S., & Roderer, T. (2014). Metacognitive monitoring and control in elementary school children: Their interrelations and their role for test performance. *Learning and Individual Differences, 29*, 141–149. https://doi.org/10.1016/j.lindif.2012.12.003

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*(2), 267–280. https://doi.org/10.1016/j.learninstruc.2010.07.004

Rutherford, T. (2017). Within and between person associations of calibration and achievement. *Contemporary Educational Psychology, 49*, 226–237. https://doi.org/10.1016/j.cedpsych.2017.03.001

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3

Serra, M. J., & DeMarree, K. G. (2016). Unskilled and unaware in the classroom: College students' desired grades predict their biased grade predictions. *Memory and Cognition, 44*(7), 1127–1137. https://doi.org/10.3758/s13421-016-0624-9

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6

Van de Pol, J., De Bruin, A. B. H., Van Loon, M. H., & Van Gog, T. (2019). Students' and teachers' monitoring and regulation of students' text comprehension: Effects of comprehension cue availability. *Contemporary Educational Psychology, 56*, 236–249. https://doi.org/10.1016/j.cedpsych.2019.02.001

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of concurrent monitoring on cognitive load and performance as a function of task complexity. *Applied Cognitive Psychology, 25*(4), 584–587. https://doi.org/10.1002/acp.1726

Van Gog, T., Hoogerheide, V., & Van Harsel, M. (2020). The role of mental effort in fostering self-regulated learning with problem-solving tasks. *Educational Psychology Review, 32*, 1055–1072. https://doi.org/10.1007/s10648-020-09544-y

Van Loon, M. H., & Roebers, C. M. (2017). Effects of feedback on self-evaluations and self-regulation in elementary school. *Applied Cognitive Psychology, 31*(5), 508–519. https://doi.org/10.1002/acp.3347

Van Loon, M. H., De Bruin, A. B. H., Van Gog, T., Van Merriënboer, J. J. G., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143–154. https://doi.org/10.1016/j.actpsy.2014.06.007

Van Zanten, M., Van den Brom-Snijders, P., Van den Bergh, J., Meier, R., & Vrolijk, A. (2007). *Reken-wiskundedidactiek: Hele getallen*. ThiemeMeulenhoff.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in education and practice. The educational psychology series* (pp. 277–304). Lawrence Erlbaum.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press.