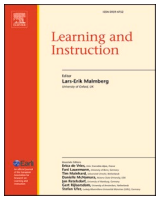
Learning and Instruction 92 (2024) 101907

Contents lists available at ScienceDirect

Learning and Instruction

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



Primary school students' awareness of their monitoring and regulation judgment accuracy

Sophie Oudman *, Janneke van de Pol, Eva Janssen, Tamara van Gog

Department of Education, Utrecht University, the Netherlands

ARTICLE INFO

Keywords:
Accuracy awareness
Monitoring judgment accuracy
Regulation judgment accuracy
Second-order judgments
Self-scoring
Primary school mathematics

ABSTRACT

Background: Improving students' monitoring and regulation judgment accuracy is necessary for improving the effectiveness of self-regulated learning, but might not be sufficient: Students presumably also need to feel confident about the accuracy of their judgments to act upon them. However, little is known about students' awareness of their monitoring judgment accuracy, and awareness of their regulation judgment accuracy has not yet been investigated.

Aims: We investigated (1) primary school students' awareness of their monitoring and regulation judgment accuracy in mathematics and (2) whether self-scoring, which is known to improve monitoring/regulation accuracy, would also improve awareness of their regulation judgment accuracy.

Sample(s): Primary school students (9-10 year-olds) from 34 classes ($N = 564$).

Methods: Students completed problem-solving tasks twice (parallel versions) on two different days and made monitoring/regulation judgments, rated their confidence in the accuracy of those judgments, self-scored their work, and again made (confidence) judgments, on both occasions. If an increase/decrease in judgment accuracy from day 1 to day 2 would be accompanied by an increase/decrease in their confidence in the accuracy of their judgements, students show accuracy awareness.

Results: Students' judgment accuracy did not predict their confidence in the accuracy of their judgments, indicating that students were not aware of their monitoring/regulation accuracy. Self-scoring improved students' awareness of their regulation judgment accuracy for students whose regulation judgment accuracy increased or stayed maximally accurate after self-scoring, but not for students whose regulation judgment accuracy decreased or stayed equally inaccurate after self-scoring.

Conclusions: Primary school students were not aware of their monitoring/regulation judgment accuracy. Self-scoring improved the awareness of their regulation judgment accuracy for some students.

1. Introduction

Primary school students are increasingly expected to become self-regulated learners (OECD, 2022). For self-regulated learning to be effective, it is critical that students' monitoring (evaluating one's own performance) and regulation (deciding on what subsequent learning actions should be taken to reach a learning goal) are accurate (Dunlosky & Rawson, 2012; Griffin et al., 2013). Accurate monitoring is a necessary (though not sufficient) condition for accurate regulation, which in turn determines how much students will learn (Dunlosky & Rawson, 2012): When students overestimate their own performance, they may quit studying or practicing too soon or not seek additional instructions or help they might need; when they underestimate their own performance, they spend valuable time working on learning tasks they can already perform. Prior studies have asked primary school students to make explicit monitoring judgments (i.e., scoring how well they think they performed on a task/expect to perform on a test) and regulation judgments (i.e., indicating what, if any, subsequent activity they would need to perform, such as restudying the material or attending additional instruction; Oudman et al., 2022; Baars et al., 2014; Boekaerts & Rozendaal, 2010; García et al., 2016; Van Loon & Roebers, 2017). These studies have shown, however, that primary school students' judgments are often inaccurate (i.e., do not correspond to their actual performance or their actual needs) when they are engaged in memorizing items or learning from texts (e.g., Van Loon and Roebers, 2017), or, as is the focus of the present study, in solving mathematics problems (e.g., Oudman

* Corresponding author. P.O. Box 80.140, 3508 TC, Utrecht, the Netherlands.
E-mail address: v.s.oudman@uu.nl (S. Oudman).

https://doi.org/10.1016/j.learninstruc.2024.101907
Received 2 December 2022; Received in revised form 26 January 2024; Accepted 9 March 2024
Available online 4 April 2024

et al., 2022; Baars et al., 2014; Boekaerts & Rozendaal, 2010; García et al., 2016).

Therefore, researchers have been looking for ways to help primary school students improve their monitoring and regulation judgment accuracy, with some success (e.g., Oudman et al., 2022; Baars et al., 2014; Van Loon and Roebers, 2017). However, only improving students' monitoring and regulation judgment accuracy might not be sufficient for improving the effectiveness of students' self-regulated learning. Students presumably also need to feel confident about the accuracy of their judgments to act upon them (suggested by Gabriele et al., 2016; Patterson et al., 2001), which is what we want them to do when their monitoring and regulation judgments are accurate. In contrast, when students make inaccurate judgments, it is helpful if they feel less confident about the accuracy of their judgments (as acting upon those would hamper their learning; see section 1.2). Students' ratings of their feeling of confidence in their monitoring and regulation judgment accuracy are also known as *second-order judgments* (SOJ[1]; Dunlosky et al., 2005; Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021). When students feel relatively more confident (i.e., providing a higher SOJ) about the accuracy of a more accurate judgment than of a less accurate judgment, they show *accuracy awareness* (e.g., Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021).

Previous studies on monitoring accuracy awareness seem to suggest that university students showed awareness of the (in)accuracy of their monitoring judgments (Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021), but that secondary school students were not (Nederhand et al., 2021). These findings might suggest that monitoring accuracy awareness is a metacognitive skill that only develops during adulthood or late adolescence, in which case one would not expect to find it in primary school students. However, given the fact that it has not yet been studied in this population, we set out to investigate primary school students' monitoring accuracy awareness. Moreover, students' awareness of their *regulation* judgment accuracy has not yet been investigated, and it is also an open question whether interventions that improve students' monitoring and regulation judgment accuracy would also improve their monitoring and regulation accuracy awareness. The present study aims to acquire more insight into these issues, which may ultimately help to design interventions that lead to more effective self-regulated learning.

1.1. Monitoring and regulation judgment accuracy

In the present study we defined students' monitoring judgment accuracy, or the degree to which students know how well they performed on a mathematical task, in terms of the absolute difference between their judgment of how many problems they answered correctly and the number of problems they actually answered correctly (cf. Baars et al., 2014; Dunlosky & Rawson, 2012). 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.

With regard to regulation judgment accuracy, most prior research was conducted in the field of word-pair learning, concept learning, and text comprehension, where regulation judgments involved students being asked to select word pairs, definitions, or texts for *restudy* (e.g., Dunlosky & Rawson, 2012; Metcalfe & Finn, 2008; Van de Pol et al., 2019; Van Loon & Roebers, 2017) or *allocate study time* to word pairs (e.g., Dufresne & Kobasigawa, 1989). The little research available on (improving) monitoring and regulation judgment accuracy in problem solving also mainly involved selecting worked examples for *restudy* (e.g., Baars et al., 2014, 2018, 2013), although there is some research going beyond that, training students to *select the complexity* of the subsequent problem-solving task (e.g., Kostons et al., 2012; Raaijmakers et al., 2018). In contrast to making relatively simple restudy decisions (where one can strongly rely on the monitoring judgment of whether an item is known or not), regulatory actions for learning problem-solving tasks are arguably more complex (and therefore interesting and important to study). For instance, in (Dutch) primary school when students have not yet mastered specific problem-solving skills two regulatory actions are most common. Either students receive or ask for additional instruction (by the teacher or another student) when they do not understand how to solve the problems, or they 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 regulation judgments in the present study as students' indications of what they would need: additional instruction, additional practice, both, or nothing. The concept 'regulation judgment accuracy' indicates the extent to which students' regulation judgments, meaning their evaluation of their need for additional instruction or practice, are in line with students' actual need for intervention, as indicated by experts (cf. Oudman et al., 2022).

As mentioned earlier, students' monitoring judgments influence their regulation judgments and accurate monitoring judgments seem to be a necessary (though not sufficient) precondition for accurate regulation judgments (Oudman et al., 2022; Dunlosky & Rawson, 2012). That is, if students overestimate their own performance, they are likely to terminate practicing and move on to another task while they do not yet master the skill and would need additional practice or instruction. If they underestimate their own performance (which seems rarer; De Bruin et al., 2017) they are likely to spend time on activities they already mastered rather than on those they need to learn. Interestingly, the unskilled-and-unaware effect (Garcia et al., 2016; Oudman et al., 2022) has shown that students who perform better also make more accurate monitoring judgments.

In sum, students who make inaccurate monitoring and, subsequently, regulation judgments may learn less than students who make more accurate judgments. Inaccurate monitoring and regulation judgments might be less problematic, however, when students are aware of the inaccuracy of their judgments.

1.2. Students' awareness of their judgment accuracy

Students show awareness of their monitoring and regulation judgment (in)accuracy when they indicate that they feel relatively more confident about the accuracy of a more accurate monitoring/regulation judgment than about the accuracy of a less accurate monitoring/regulation judgment (e.g., Fritzsche et al., 2018; Händel & Dresel, 2018). In other words, students who are aware of their judgment (in)accuracy are able to distinguish between their more and less accurate monitoring and regulation judgments in terms of their feeling of confidence, whereas students who are not aware of their (in) accuracy do not do so. Note that in some literature the term *awareness* is used for students' ability to estimate their performance, which is what we refer to as *monitoring judgment accuracy* here. Hence, some prior studies used the term *subjective awareness* for what we call *students' awareness of their monitoring accuracy* (e.g., Fritzsche et al., 2018). Similarly, the term *confidence judgment* is sometimes used for what we call a *monitoring judgment*. In the present study, confidence judgments concern students' rating of their feeling of confidence in their monitoring/regulation judgment accuracy (also known as *second-order judgments*).

[1] We use the abbreviation SOJ for second-order judgments. More specifically, SOJ-m: students' second-order judgment about their monitoring judgment accuracy; SOJ-r: students' second-order judgment about their regulation judgment accuracy.

Students' awareness of their monitoring and regulation judgment accuracy[2] might be an important predictor of what students will actually do during self-regulated learning, because students' feeling of confidence in the accuracy of their judgments might affect whether and how they act upon these judgments (suggested by Gabriele et al., 2016; Händel & Fritzsche, 2016; Patterson et al., 2001). When students are *aware of their accuracy*, they are likely to act on their accurate judgments, and this would be productive. When they are *unaware of their accuracy*, meaning that students make accurate monitoring and regulation judgments but do not feel confident that these are accurate, they might not act upon their accurate judgments (e.g., they might seek additional instruction just in case, whereas they correctly judged they would only additional practice), which would be unproductive. When students are *unaware of their inaccuracy*, students make inaccurate monitoring and regulation judgments, yet feel confident that they made accurate judgments, and consequently, are likely to act upon their inaccurate judgments, which would be unproductive. Finally, when students are *aware of their inaccuracy*, students make inaccurate monitoring and regulation judgments and (rightfully) feel unconfident about the accuracy of their judgments, which can be productive as they would be less likely to act upon their judgments and might ask their teachers for help. In sum, it is important to not only investigate students' monitoring and regulation judgment accuracy, but also their awareness of their monitoring and regulation judgment accuracy, as this might provide more insight into what is needed to improve students' actual self-regulated learning behavior.

1.2.1. Prior research into students' awareness of their judgment accuracy

Previous studies have shown that university students showed some awareness of their *monitoring* judgment accuracy—in that more accurate monitoring judgments were associated with a higher feeling of confidence in these judgments (e.g., Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021). These studies operationalized awareness differently: Nederhand et al. asked each student for one monitoring judgment and one confidence judgment after completing a whole exam, while Fritzsche et al. and Händel and Dresel asked students to make confidence judgments after completing each item of a performance test (consisting of 18 and 32 items respectively).

One study (Nederhand et al., 2021) also investigated secondary school students' awareness of their monitoring judgment (in)accuracy and showed that they were not aware of their monitoring judgment (in) accuracy with regard to exams for French, German, and Mathematics. This might suggest that monitoring accuracy awareness is a metacognitive skill that only develops during adulthood or late adolescence, in which case one would not expect to see it in primary school students. However, it could also be the case that monitoring accuracy awareness is related to task performance, and is found in students who show higher performance on the task but not in students with lower performance: Studies with university students (Fritzsche et al., 2018; Händel & Dresel, 2018) found that high-performing students (i.e., defined as students whose task performance fell in the fourth quartile) were aware of their monitoring judgment (in)accuracy and their low-performing peers (i.e., whose task performance fell in the first quartile) were not. It cannot be inferred from the data presented in the study by Nederhand et al. whether this could also explain the findings with secondary school students as the effect of performance on students' awareness of their monitoring accuracy was not investigated. Prior studies do suggest that better performing students make more accurate judgments of their problem-solving performance, an effect that was also found by a previous study based on the same data set as the present study (Oudman et al., 2022). However, it remains an open question whether performance also affects primary school students' *awareness* of their monitoring/regulation accuracy.

Moreover, another alternative explanation for why Nederhand et al. (2021) did not find secondary school students to be aware of their monitoring judgment accuracy is a methodological one. We presume, as also argued by Fritzsche et al. (2018), that students only show true awareness of their monitoring judgment accuracy when they are able to distinguish between their more and less accurate judgments in terms of their feeling of confidence, which asks for analyzing the effects at the *within*-student level (i.e., based on multiple measurements of accuracy and confidence per student). In contrast, Nederhand et al. (2021) analyzed the data at the *between*-student level (i.e., one measurement of accuracy and confidence per student), which answers a slightly different question, namely: do students, who make a more accurate monitoring judgment on a task, feel more confident about the accuracy of this judgment, than students who make a less accurate monitoring judgment on that task?[3] Thus, the present study aimed to explore whether primary school students' monitoring judgment accuracy predicts their feeling of confidence about their monitoring judgment accuracy at the within-student level: Students completed a task twice (we used parallel versions) on two different days, and made monitoring/regulation judgments, and rated their confidence in the accuracy of their monitoring/regulation judgments on both occasions. This allows us to investigate whether a student's increase in one variable (judgment accuracy) is accompanied by an increase in another variable (confidence in their judgment accuracy), and if a decrease in one variable (judgment accuracy) also results in a decrease in the other variable (confidence in their judgment accuracy). Moreover, as the unskilled-and-unaware effect (Garcia et al., 2016; Oudman et al., 2024) has shown that students who perform better also make more accurate monitoring judgments, an interesting open question is whether monitoring and regulation accuracy awareness would also depend on their problem-solving task performance.

To the best of our knowledge, students' awareness of their *regulation* judgment accuracy has not been investigated at all thus far. Yet, this might be at least as important for self-regulated learning as monitoring judgment accuracy awareness, because regulation judgments more directly influence whether and how students continue learning than monitoring judgments. Therefore, we also aimed to explore whether primary school students are aware of their regulation judgment accuracy, and whether this regulation accuracy awareness differs as a function of students' performance.

1.3. Effect of self-scoring on students' accuracy awareness

Self-scoring is frequently and increasingly used in primary education, not only to save teachers' time (in the sense that they have to spend less time on evaluating student work) but also as a simple and effective intervention to improve students' monitoring and regulation, skills that lay the foundation for students' lifelong learning (Bjork et al., 2013; OECD, 2022). Self-scoring is effective in improving students' monitoring accuracy, because when students compare their test responses to objectively correct information, they have access to information about the correctness of their answers (Rawson & Dunlosky, 2007).

Our previous study (Oudman et al., 2022) based on the same data set as the present study, showed that after self-scoring, primary school students' monitoring judgments of their performance on mathematics problem-solving tasks came close to being perfectly accurate. However, while the accuracy of students' regulation judgments also improved, these still deviated substantially from expert judgments of what (if any) activity the students would need to engage in (Oudman et al., 2022). To

[2] Although the concept 'students awareness of their monitoring/regulation judgment **(in)**accuracy' might be most correct, we use this concept interchangeable with 'students' awareness of their monitoring/regulation judgment accuracy', for the sake of readability.

[3] See appendix A for a more elaborate explanation of the differences between analyses at the within-student and between-student level.

effectively impact students' actual self-regulated learning behavior (i.e., whether or not they actually act upon their regulation judgments), it seems important to not only improve students' regulation judgment accuracy, but also the awareness of their regulation judgment accuracy. That is, when students' regulation judgment accuracy improves from self-scoring, it would be beneficial that they then also feel (more) confident about the accuracy of those judgments, and when they still make inaccurate regulation judgments after self-scoring, that they feel (more) unconfident. Hence, in the present study we explored the effect of self-scoring on students' awareness of their regulation judgment accuracy, a relation that has not been studied before.

1.4. Present study

The present study addressed three research questions (RQ) in the context of mathematics problem solving in primary school. Students participated on two different days with one week in between, working on parallel versions (i.e., with isomorphic problems that have the same solution procedure and difficulty, but different numbers) of a multiplication and a division task, each consisting of six problems. On both days, students made monitoring and regulation judgments, second-order judgments (SOJ; i.e., confidence in their monitoring and regulation judgments), self-scored their answers, and again made second-order regulation judgments for each of the tasks. The parallel measures on two days enabled us to investigate whether students showed accuracy awareness, that is, when students make more/less accurate judgments on the second day, do they also feel more/less confident about their accuracy?

In Fig. 1, we visualized this way of analyzing the data for monitoring (for regulation this is similar). Based on our data, for each student a line can be drawn between two measurement points (stemming from day 1 and day 2), which both contain a value for monitoring judgment accuracy and a value for students' feeling of confidence in their monitoring accuracy (i.e., SOJ-m). We wanted to know whether students assign higher confidence to their more accurate monitoring judgments than to their less accurate monitoring judgments. Simplifying our multilevel analyses, one could say that students show awareness of their monitoring judgment accuracy when their lines go down, because then, the more accurate monitoring judgment score (i.e., closer to zero) goes along with higher confidence (i.e., higher SOJ-m score) than the less accurate monitoring judgment score. Note that we are not interested in whether students' monitoring judgments become more accurate from day 1 to day 2 or whether they become more confident from day 1 to day 2. The monitoring judgment accuracy or SOJ-m of some students might increase from day 1 to day 2, while these variables might decrease for other students, and what we are interested in is: if there is an increase in one variable (monitoring judgment accuracy) is there also an increase in the other (SOJ-m), and if there is a decrease in one (monitoring judgment accuracy) is there a decrease in the other (SOJ-m)? As such, this approach of testing the relation between monitoring judgment accuracy and confidence (SOJ-m) within students based on multiple measurement points, is substantially different from the approach used by Nederhand et al. (2021) in which this same relation is tested between students using only one measurement point per student. With the latter approach, it is only possible to test whether students with higher monitoring judgment accuracy scores are, in general, more confident than students with lower accuracy scores; not whether they are able to distinguish between their more accurate and less accurate monitoring judgments, which is what indicates awareness.

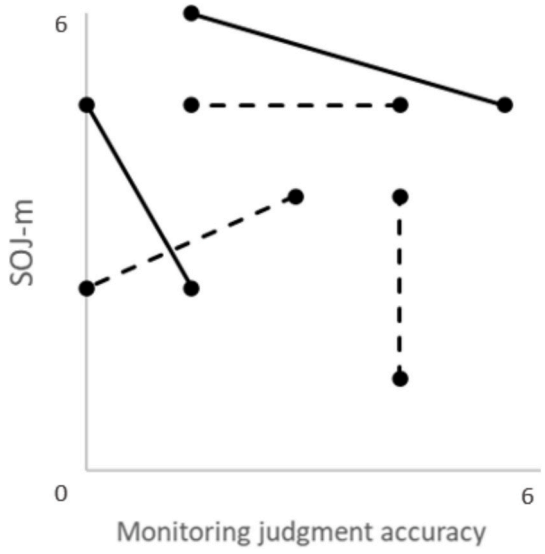


Fig. 1. Hypothetical Visualization of Modeling Students' Awareness of Their Monitoring Judgment Accuracy.
Note. Each line represents one fictional student and one task (for instance, the multiplication task). The dots at the end of the line represent the two measurement points (day 1 and day 2; the order is irrelevant). Students with solid lines show awareness of their monitoring judgment accuracy, as higher accuracy scores (scores closer to zero) are associated with higher SOJ-m (confidence) scores than lower accuracy scores. Students with dashed lines do not show awareness of their monitoring judgment accuracy.

First, we explored whether, before self-scoring, students showed monitoring accuracy awareness (meaning that they feel relatively more confident about the accuracy of a more accurate monitoring judgment than of a less accurate monitoring judgment; RQ1a) and whether and how this differs as a function of students' problem-solving performance (RQ1b). Second, we explored whether, before self-scoring, students showed regulation accuracy awareness (RQ2a) and whether and how this differs as a function of students' problem-solving performance (RQ2b). Third, we explored whether and how self-scoring affected students' regulation accuracy awareness (RQ3).

With regard to first research question, we had no specific hypothesis, because the only study about underaged students' awareness of their monitoring judgment accuracy (Nederhand et al., 2021) based their conclusions on analyses at the between-student level, whereas we are mainly interested in signs of students' accuracy awareness at the within-student level (see section 1.2.1), which might lead to different conclusions. Moreover, we had no specific hypotheses regarding the second and third research question, because there is no prior research about the students' *regulation* accuracy awareness.

2. Method

Data for the current study were collected in the context of a larger research project that also focusses on primary school students' monitoring and regulation judgment accuracy in mathematics (Oudman et al., 2022), teachers' judgments of their students' task performance (Oudman et al., 2023a) and teachers' judgments of their students' monitoring and regulation judgment accuracy (Oudman et al., 2023b). Ethical approval was obtained from the Ethics Committee of our institution in November 2018.

2.1. Design

The students participated on two different days with one week in between, working on parallel versions (i.e., with isomorphic problems that have the same solution procedure and difficulty, but different numbers) of a multiplication and a division task, each consisting of six problems. On both days, students made monitoring and regulation judgments, second-order judgments (i.e., confidence in their monitoring and regulation judgment accuracy), self-scored their answers, and again made (second-order) regulation judgments for each of the tasks. The parallel measures on two days enabled us to investigate whether students showed accuracy awareness (i.e., whether an increase/decrease in

accuracy would be accompanied by an increase/decrease in confidence).

2.2. Participants

The data set contains data of 34 Dutch sixth grade classes (Dutch sixth grade is similar to US fourth grade in terms of age, i.e., 9–10 years old). Of the 777 students who attended the 34 classes, data from 564 students were included in the analyses: 495 students were included in the analyses of the multiplication tasks and 359 in the analyses of the division tasks. The students in these two separate datasets partly overlapped: 290 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. Fig. 2 displays demographics and for which reasons (and how many) students had to be excluded. As Fig. 2 shows, a substantial number of tasks was excluded because students (1) did not answer any problem on one or both days, (2) did not correctly answer any of the problems on both days, or (3) correctly answered all problems on both days. The reason these tasks were excluded from the analyses is that making accurate judgments would be relatively easy for these students on these tasks, because the tasks were presumably far too complex (1 and 2) or far too easy (3) for them. Including these data could therefore have distorted the results. In the final sample, 5.9% of the students had a non-western background, meaning that the students themselves or both their parents were born in a non-western country, according to their teachers.

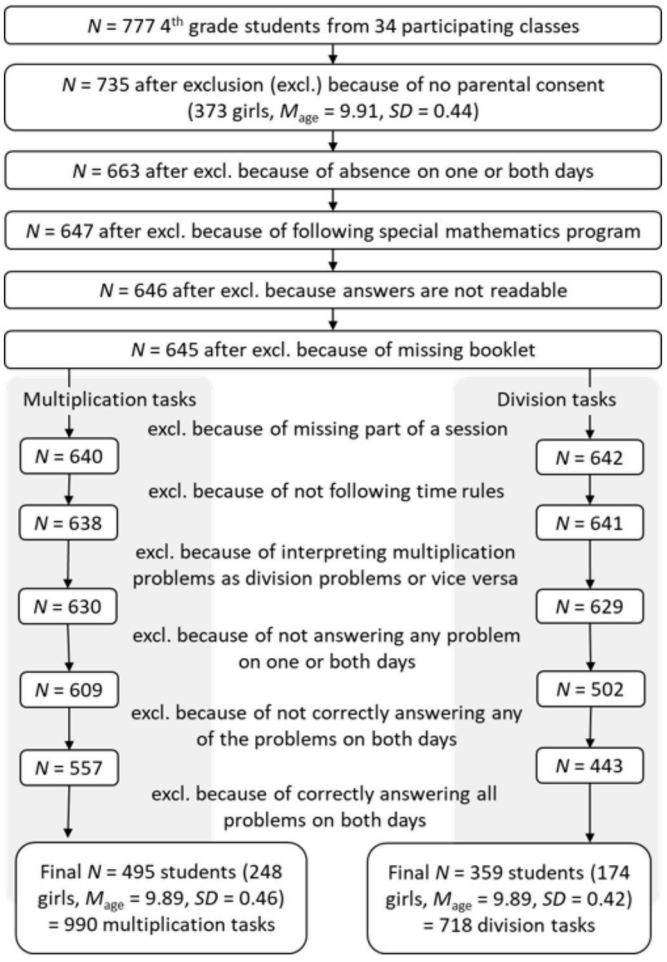


Fig. 2. Flowchart of why and how many students were excluded from all analyses.

2.3. Materials and measures

Materials were presented on paper, in booklets.

2.3.1. Problem-solving task performance

On both days, students worked on two problem-solving tasks: (1) a set of six multiplication problems (single-digit multiplicands multiplied by 3-digit multipliers, e.g., 6 × 472) and (2) 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 mathematical tasks were administered on the two days. Students received one point for each problem that was solved correctly, thus the task performance scores ranged between 0 and 6 per task. The internal consistencies of the task performance scores, in terms of Cronbach's alpha, were in the acceptable to good range (multiplication: 0.72 and 0.68; division: 0.78 and 0.81 on the first and second day, respectively).

2.3.2. Monitoring judgment accuracy

After students completed the multiplication or division task, they made a monitoring judgment by answering 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. Students also made a monitoring judgment after self-scoring, but this judgment was not used in the analyses (see section 2.3.4).

Absolute monitoring judgment accuracy was computed by taking the absolute difference between a student's monitoring judgment and their actual performance on a task (i.e., regardless of whether the difference was positive or negative), ranging from 0 to 6, with values closer to zero indicating more accurate monitoring judgments (Baars et al., 2014; Schraw, 2009).

2.3.3. Regulation judgment accuracy

Students also made a regulation judgment before and after self-scoring, by indicating which of the following choices was most applicable to them: additional instruction, additional practice, additional instruction and practice, or 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. As additional instruction is a more intensive intervention than additional practice, students' regulation judgments were treated as an ordinal variable and coded as follows: 0 = no intervention needed; 1 = additional practice needed, 2 = additional instruction needed (and practice afterwards). The needs "additional instruction" and "additional instruction and practice" were combined into one, as based on students' work, we were not able to determine which of the two was most suited (see Appendix B for an elaborate explanation).

To determine the accuracy of students' regulation judgments, we first coded students' actual need for intervention, based on a coding scheme we developed that is described in detail in Appendix B (as well as in Oudman et al., 2022). In short, we distinguished the same three categories as for students' regulation judgments (i.e., 0 = no intervention needed; 1 = additional practice needed, 2 = additional instruction needed [and practice afterwards]), based on the time students needed to complete the task and on whether they made computational or procedural errors.

Students' absolute regulation judgment accuracy was computed by taking the absolute difference between students' regulation judgment and their actual need for intervention according to the coding. It ranged from 0 to 2, with values closer to zero indicating more accurate regulation judgments (see Appendix B, Table B2, for the distribution of students' regulation judgments and needs).

2.3.4. Second-order judgments: confidence in monitoring and regulation judgments

Directly after students made the monitoring judgment before self-scoring, they made a second-order judgment (SOJ) about their confidence in their monitoring judgment accuracy (SOJ-m) by answering the question "How confident are you that you made a correct estimation during the previous question (question number …)?". Directly after the regulation judgments before and after self-scoring, students made a SOJ about their confidence in their regulation judgment accuracy (SOJ-r) by answering the question "How confident are you that you made a correct choice during the previous question (question number …)?". These SOJ-m and SOJ-r questions were answered on a 6-point Likert scale. In line with Fritzsche et al. (2018) this scale was labeled with smiley faces, see Fig. 3.

During a pilot study (in two classes) students also made a SOJ-m directly after the monitoring judgment after self-scoring. However, students experienced the SOJ-m question after self-scoring as "strange" because for them it felt evident that their monitoring judgments were perfectly accurate after seeing the answers. Therefore, we decided to remove the SOJ-m after self-scoring from the materials.

To check whether students understood the (second-order) judgment questions, we interviewed 12 of the pilot students (four low, four middle, and four high-performing students) one by one, after they completed the material. We asked them to describe the meaning of the (second-order) judgment questions on the multiplication task. All 12 students indicated that they understood the monitoring and regulation judgment questions. Eleven students correctly described the SOJ questions as their confidence in the 'correctness' of the previous judgment. One student described the SOJ questions as their confidence in the 'correctness' of the previous judgment and in the performance on the multiplication task. Therefore, we decided that each time students had to answer a SOJ question, it was emphasized in the written question (see above) and by the experimenter to what previous question the SOJ question referred.

2.3.5. Self-scoring

Students received a booklet in which each problem was stated on a separate line together with the correct answer (i.e. only the answer was given, no information on the solution procedure) and with two boxes: "correct" and "incorrect or not answered." They were instructed to tick the box that applied to their performance.

2.4. 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. Next, the students answered questions in their personal booklets (invested effort,[4] monitoring judgment, SOJ-m, regulation judgment, and SOJ-r). 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 experimenter did not read the correct answers aloud). The following monitoring judgment (not used in the present study), regulation judgment and SOJ-r 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 in a second session exactly one week later. The sessions lasted between 45 min and 1 h.

2.5. Analyses

As students' monitoring and regulation judgment accuracy can vary depending on the type of mathematical problem (Oudman et al., 2022; Boekaerts & Rozendaal, 2010), students' awareness of their judgment (in)accuracy might also differ across different mathematical tasks. Hence, all analyses were performed separately for the multiplication and the division task.

To answer the research questions we performed multilevel regression analyses in Mplus version 8.9 (Muthén & Muthén, 1998-2017), using maximum likelihood estimation with robust standard errors (MLR) which is robust to non-normality and suitable for ordinal outcome variables (such as our measure of regulation judgment accuracy, cf. Robitzsch, 2020). We defined three levels in our data: day (level 1), student (level 2), and class (level 3). The class level was modeled using 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. The Mplus input files are provided in Appendix C. To test RQ1a and 2a, we analyzed how students' monitoring or regulation judgment accuracy affected their SOJs. As explained in section 1.2.1 and Appendix A, we were mainly interested in the fixed effects at the day level: we analyzed whether an increase in individual students' monitoring/regulation judgment accuracy from day 1 to day 2 went along with an increase in students' SOJ-m/SOJ-r, and whether a decrease in students' monitoring/-regulation judgment accuracy from day 1 to day 2 went along with a decrease in students' SOJ-m/SOJ-r. This was reflected by regressing students' SOJ-m/SOJ-r on their monitoring/regulation accuracy at the day level (i.e., within-student level). If students would be aware of their monitoring/regulation judgment accuracy we would expect a significant and negative relationship between students' monitoring/regulation accuracy and their SOJ-m/SOJ-r (because judgment accuracy scores closer to zero indicate higher accuracy). As the differences within students were our primary interest, students' monitoring/regulation judgment accuracy were centered around the group mean (Enders & Tofighi, 2007). To enable comparison with the results of Nederhand et al. (2021), the fixed effects at the student level are additionally reported in Appendix A.

RQ1b and 2b concerned whether students' task performance affected students' monitoring and regulation accuracy awareness. This was analyzed by estimating the effect of the cross-level interaction term between students' monitoring/regulation judgment accuracy and their mean task performance on their SOJ-m/SOJ-r. This cross-level interaction effect reflects the effect of students' mean task performance on their monitoring/regulation accuracy awareness, as monitoring/regulation accuracy awareness is expressed as the linear relationship between students monitoring/regulation judgment accuracy and their SOJ-m/SOJ-r. Students' mean task performance concerned students' performance on the multiplication or division task, measured at the student level (thus averaged across two days) and centered around the grand mean, because we did not expect that students' relative task performance within their classes would be an important determinant of their confidence (Enders & Tofighi, 2007).

To test RQ3 (regarding students' change in regulation accuracy awareness from before to after self-scoring), we looked at students' change in SOJ-r from before to after self-scoring. We compared the group of students whose regulation judgment accuracy increased or

[4] The variable "invested effort" was not used in the present study, but collected for use in other studies.

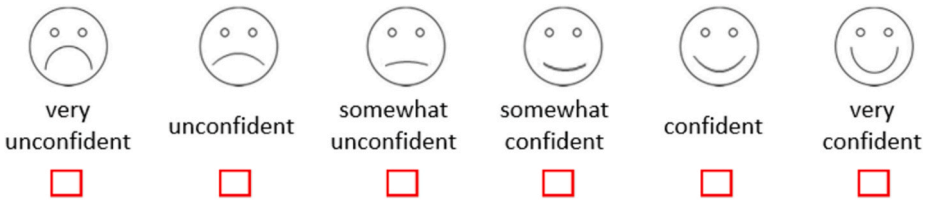


Fig. 3. Rating scale of the second-order judgments.

stayed maximally accurate (for them, an increase in SOJ-r would be desirable) to the group of students whose regulation judgment accuracy decreased or stayed equally inaccurate (for them, an increase in SOJ-r would be undesirable). Students' change in SOJ-r from before to after self-scoring (possible and observed range: −5 to 5) was regressed on students' change in regulation judgment accuracy, at the day level. When students' regulation judgment accuracy increased or stayed maximally accurate this was coded as 0; when students' regulation judgment accuracy decreased or stayed equally inaccurate this was coded as 1. This variable was centered around the grand mean as for this research question the differences within students were not of substantive interest (Enders & Tofighi, 2007).

Zero to 6.7% of the cases per variable were missing because students did not complete a question. After a missingness analysis and closer inspection of the data, we concluded it was unlikely that missingness would depend on the values of the unobserved variable. Missing values were deleted list-wise in the analyses. In each of the multilevel regression models, zero to four cases (a maximum of 0.62% 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, which not led to a difference in statistical significance.

The data of this study are openly available in an online depository at https://osf.io/36cak/?view_only=ba95fe7d0c6d4cd0a68a6bbed76edf90.

3. Results

Descriptive statistics are presented in Table 1. The students made more accurate regulation judgments after self-scoring than before, on both tasks (for statistical analyses, see Oudman et al., 2022). Before self-scoring, the students felt, on average, already quite confident about their monitoring judgments (means between 4 and 5 out of 6) and about their regulation judgments (means around 5). After self-scoring, students felt, on average, even more confident about their regulation judgments (means between 5 and 6; Table 1). Correlations between all measures are displayed in Appendix D.

3.1. Students' monitoring accuracy awareness before self-scoring (RQ1)

Table 2 shows the results of the analyses in which students' SOJ-m was regressed on their absolute monitoring judgment accuracy prior to self-scoring, as an indication of students' monitoring accuracy awareness. The main effect of students' monitoring judgment accuracy on their SOJ-m was not significant for either of the tasks (see M1 in Table 2), indicating that on average, students were not aware of their monitoring judgment (in)accuracy before self-scoring.

To test whether students' task performance affected students' monitoring accuracy awareness, we analyzed the effect of the cross-level interaction term between students' monitoring judgment accuracy and their mean task performance on their SOJ-m. For both the multiplication and division task, the interaction effect was not significant (see M2 in Table 2), indicating that students' mean task performance had no impact on their monitoring accuracy awareness before self-scoring.

3.2. Students' regulation accuracy awareness before self-scoring (RQ2)

Table 3 shows the results of the analyses in which students' SOJ-r was regressed on their absolute regulation judgment accuracy prior to self-scoring, as an indication of students' regulation accuracy awareness. Like for monitoring, the main effect of students' regulation judgment

Table 1
Descriptive statistics of students' task performance, monitoring/regulation judgment accuracy, and second-order judgments.

Variable	Range	Multiplication			Division		
		M (SD)	ICC [a]		*M (SD)*	ICC [a]	
			Student level	Class level		Student level	Class level
Task performance	0 to 6	3.67 (1.84)	0.278	0.094	3.20 (2.06)	0.357	0.101
Before self-scoring							
Absolute monitoring accuracy [b]	0 to 6	1.23 (1.18)	0.105	0.001	0.99 (1.01)	0.111	0.035
Absolute regulation accuracy [b]	0 to 2	0.61 (0.70)	0.258	0.001	0.43 (0.65)	0.176	0.005
SOJ-m	1 to 6	4.51 (0.94)	0.296	0.050	4.56 (1.07)	0.242	0.059
SOJ-r	1 to 6	4.91 (0.93)	0.294	0.040	5.06 (0.94)	0.347	0.022
After self-scoring							
Absolute regulation accuracy [b]	0 to 2	0.44 (0.62)	0.116	0.002	0.36 (0.60)	0.112	0.005
SOJ-r	1 to 6	5.29 (0.87)	0.295	0.014	5.30 (0.87)	0.296	0.012

Note. Means are across both days. All means significantly differ from 0, $p \leq .05$.

[a] Intraclass Correlation Coefficients (ICC) reflect the amount of between-student and between-class variability compared to the total amount of variability (within students, between students, and between classes).

[b] Values closer to zero indicate more accurate judgments.

Table 2
Effects of absolute monitoring judgment accuracy and students' mean task performance on SOJ-m before self-scoring.

	Multiplication		Division	
	M1: Effect of monitoring accuray	M2: Effect of monitoring accuracy and performance	M1: Effect of monitoring accuray	M2: Effect of monitoring accuracy and performance
Fixed effects	*B (SE)*	*B (SE)*	*B (SE)*	*B (SE)*
Intercept	4.51	4.51	4.57	4.57
	(0.05)***	(0.05)***	(0.07)***	(0.06)***
Monitoring accuracy	0.03	0.03	0.01	0.01
	(0.03)	(0.05)	(0.05)	(0.05)
Task performance		0.08		0.12
		(0.02)***		(0.03)***
Monitoring accuracy*performance		0.02		0.04
		(0.02)		(0.03)
Random effects	*SS (SE)*	*SS (SE)*	*SS (SE)*	*SS (SE)*
Intercept variance day	0.57	0.57	0.77	0.76
	(0.06)***	(0.14)***	(0.09)***	(0.11)***
Intercept variance student	0.30	0.29	0.36	0.32
	(0.05)***	(0.08)***	(0.08)***	(0.09)***
Slope variance accuracy		0.01		0.01
		(0.23)		(0.07)

Note. Monitoring accuracy is measured at the day level and group mean centered; Performance is measured at the student level, thus the mean across two days, and grand mean centered. The SS represents the amount of variability in students' SOJ-m that is not explained by the independent variables, at the respective level. ***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$.

Table 3
effects of absolute regulation judgment accuracy and students' mean performance on SOJ-r before self-scoring.

	Multiplication		Division	
	M1: Effect of regulation accuray	M2: Effect of regulation accuracy and performance	M1: Effect of regulation accuray	M2: Effect of regulation accuracy and performance
Fixed effects	*B (SE)*	*B (SE)*	*B (SE)*	*B (SE)*
Intercept	4.92	4.92	5.07	5.08
	(0.05)***	(0.05)***	(0.05)***	(0.04)***
Regulation accuracy	0.03	0.02	−0.09	−0.09
	(0.05)	(0.05)	(0.06)	(0.08)
Task performance		0.00		0.03
		(0.02)		(0.02)
Regulation accuracy*performance		−0.01		0.01
		(0.05)		(0.07)
Random effects	*SS (SE)*	*SS (SE)*	*SS (SE)*	*SS (SE)*
Intercept variance day	0.55	0.54	0.57	0.56
	(0.05)***	(0.07)***	(0.09)***	(0.25)*
Intercept variance student	0.29	0.29	0.26	0.26
	(0.04)***	(0.05)***	(0.06)***	(0.14)
Slope variance accuracy		0.03		0.04
		(0.20)		(1.19)

Note. Regulation accuracy is measured at the day level and group mean centered; Task performance is measured at the student level, thus the mean across two days, and grand mean centered. The SS represents the amount of variability in students' SOJ-r that is not explained by the independent variables, at the respective level. ***$p \leq 0.001$,**$p \leq 0.01$,*$p \leq 0.05$.

accuracy on their SOJ-r was not significant for either of the tasks (see M1 in Table 3), indicating that on average, students were not aware of their regulation judgment (in)accuracy before self-scoring. For both the multiplication and division task, the cross-level interaction term between students' regulation judgment accuracy and their mean performance had no significant effect on their SOJ-r (see M2 in Table 3), indicating that students' mean performance had no impact on their regulation accuracy awareness before self-scoring.

Table 4
Change in students' SOJ-r from before to after self-scoring.

	Regulation accuracy increased after self-scoring	Regulation accuracy = 0 before and after self-scoring	Regulation accuracy decreased after self-scoring	Regulation accuracy = 1 or 2 before and after self-scoring
	M (SD)			
Multiplication	$n = 195$	$n = 421$	$n = 74$	$n = 242$
	0.50 (1.23)	0.44 (0.88)	0.36 (1.11)	0.22 (0.95)
Division	$n = 83$	$n = 391$	$n = 44$	$n = 128$
	0.28 (1.13)	0.23 (0.85)	0.05 (1.08) [a]	0.18 (0.81)

Note. Sample sizes refer to number of tasks (per student two tasks were included). [a] This increase in SOJ-r, measured as the intercept of the intercept-only model, was not significant, $p = 0.778$. The other increases in SOJ-r were significant, $p < 0.05$.

3.3. Effect of self-scoring on students' regulation accuracy awareness (RQ3)

To explore the effect of self-scoring on students' regulation accuracy awareness, we analyzed the change in students' SOJ-r from before to after self-scoring across four different subsets of students, of whom the regulation judgment accuracy (1) increased after self-scoring, (2) decreased after self-scoring, (3) was maximally accurate both before and after self-scoring, and (4) was equally inaccurate before and after self-scoring. Table 4 presents the change in students' SOJ-r from before to after self-scoring, for these four subsets of students. As shown in this table, for students whose regulation judgment accuracy increased or stayed maximally accurate (subgroup 1 and 3), their confidence in their

Table 5
Effects of students' change in regulation judgment accuracy on students' change in SOJ-r from before to after self-scoring.

	Multiplication	Division
Fixed effects	*B (SE)*	*B (SE)*
Intercept	0.39 (0.04)***	0.21 (0.03)***
Change in regulation accuracy	−0.21 (0.09)*	−0.09 (0.09)
Random effects	*SS (SE)*	*SS (SE)*
Intercept variance day	0.99 (0.08)***	0.78 (0.07)***
Intercept variance student	0.01 (0.05)	0.02 (0.05)

Note. Change in regulation accuracy is measured at the day level and grand mean centered. When students' regulation judgment accuracy increased or stayed maximally accurate this was coded as 0; when students' regulation judgment accuracy decreased or stayed equally inaccurate this was coded as 1. Students' change in SOJ-r from before to after self-scoring ranged: −5 to 5 (possible and observed).
***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$.

judgment accuracy increased, on average. This is desirable as this might increase the likelihood that they act upon their regulation judgments. However, for students in the other two subgroups (2 and 4) whose accuracy decreased or stayed equally inaccurate, their confidence also increased, on average (with exception of the students whose regulation judgment accuracy decreased on the division task; their confidence did not increase significantly), which is not desirable.

Note though, that Table 4 also suggests that there are differences in the size of the increase in confidence: On both tasks, the SOJ-r increase was on average larger for students whose regulation judgment accuracy increased or stayed maximally accurate, than for students whose regulation judgment accuracy decreased or stayed equally inaccurate. Thus, even though an increase in confidence was observed in all subgroups, this increase was smaller when it was undesirable. Table 5 shows the significance of this difference in SOJ-r increase between the two left and the two right columns in Table 4, by regressing students' change in SOJ-r from before to after self-scoring on students' change in regulation judgment accuracy. The results in Table 5 show that this difference (in the increase in confidence between the groups of students for whom it was desirable and for whom it was not) was only significant for the multiplication task and not for the division task.

4. Discussion

Students' awareness of the (in)accuracy of their monitoring and regulation judgments could play a role in effective self-regulated learning behavior (e.g., seeking help when being rightfully unconfident about their monitoring or regulation judgment accuracy). Nevertheless, monitoring accuracy awareness has hardly been investigated, and the available research focused on adolescents and young adults. Regulation accuracy awareness has, to date, not yet been studied at all. In the present study, we explored to what extent primary school students (9–10 years old) showed awareness of their monitoring and regulation judgment accuracy on mathematical problem-solving tasks, whether this differed as a function of their problem-solving performance, and how students' regulation accuracy awareness was affected by self-scoring.

4.1. Are students aware of their monitoring and regulation judgment accuracy before self-scoring? (RQ1 and 2)

First, we explored whether primary school students were able to distinguish between their more and less accurate monitoring judgments in terms of their confidence in those judgments, as an indication of monitoring accuracy awareness. Previous studies concluded that university students are somewhat aware of their monitoring judgment accuracy (Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021), but that secondary school students were not (Nederhand et al., 2021). This might suggest that monitoring accuracy awareness is a metacognitive skill that only develops during adulthood or late adolescence, in which case one would not expect to see it in primary school students. However, another (not mutually exclusive) explanation could be that it could also have been the case that monitoring accuracy awareness would be related to task performance, and is found in students who show relatively higher task performance but not in students with lower performance: Studies with university students (Fritzsche et al., 2018; Händel & Dresel, 2018) found that high-performing students were aware of their monitoring (in)accuracy and their low-performing peers were not. Our findings, however, seem to provide more evidence for the metacognitive development explanation, as the primary school students in our sample showed no awareness of their monitoring judgment accuracy prior to self-scoring (RQ1a), regardless of their problem-solving performance (RQ1b).

Note that our findings are based on analyses at the *within*-student level, therewith answering the question: is a individual student's increase in judgment accuracy accompanied by an increase in confidence in their judgment accuracy, and is a decrease in judgment accuracy accompanied by a decrease in confidence in their judgment accuracy? Interestingly, had we performed the analyses of our data at the *between*-student level, as done by Nederhand et al. (2021), we could have concluded that students showed awareness of their monitoring and regulation accuracy on the division task (see appendix A). But again, analyses at the between-student level answer a slightly different question than the one we were interested in (i.e., do students, who make a more accurate monitoring judgment on a task, feel more confident about the accuracy of this judgment, than students who make a less accurate monitoring judgment on that task?). Future research should investigate whether the metacognitive development explanation holds, by testing effects at the within-student level across different age cohorts from the end of primary school until adolescence. Such developmental research should use tasks with a similar format (e.g., open vs. multiple choice questions, global vs. item-specific judgments) across all age groups, because we cannot entirely rule out the possibility that the difference in results between the different studies (i.e., the present study; Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021), can be explained by the different measures (see section 4.3 for further discussion about the difference in measures).

This was the first study to not only investigate students' monitoring accuracy awareness, but also their regulation accuracy awareness. Our primary school students showed no awareness of their regulation judgment accuracy prior to self-scoring (RQ2a) regardless of their problem-solving performance (RQ2b). It is possible that regulation accuracy awareness also develops only at a later age, or it could somehow be dependent on monitoring accuracy awareness, which could be tested in future research with (young) adults.

So why might age play a role in the development of monitoring (and possibly, regulation) accuracy awareness? A speculative explanation could lie in students' insights into their *cue use*, which might increase with age (cf. Roebers et al., 2019). When making monitoring judgments, students use cues such as, for example, their beliefs about their general mathematical ability or their fluency during learning (Ackerman, 2019; Thiede et al., 2010). These cues can be more or less diagnostic (i.e., predictive) of students' actual task performance and the use of more diagnostic cues will result in more accurate monitoring judgments (Koriat, 1997; Thiede et al., 2010). If students have (implicit) knowledge about what cues they use when making their monitoring judgments and about the diagnosticity of these cues, they can use this knowledge when rating their feeling of confidence in their monitoring judgment accuracy (also suggested by Fritzsche et al., 2018; Händel & Dresel, 2018). For instance, when students know or have the feeling that they used cues that are not diagnostic or that they missed highly diagnostic cues, they might not feel confident about their monitoring judgment accuracy. Possibly, primary school students lack insight into cue diagnosticity and their own cue use, whereas university students may have gained some more insight into this, for instance because they have been provided

with more (direct or indirect) instruction about metacognitive monitoring over the years. Primary school students' lack of insight in their own *monitoring* judgment accuracy might also be the reason for why the students were not aware of their *regulation* judgment accuracy: Students' feeling of confidence in their regulation judgment accuracy might be based on their feelings about their monitoring judgment accuracy, as the regulation judgments of 9–10 year old students on problem solving seem to be (at least partly) based on their monitoring judgments (Oudman et al., 2022).

4.2. What is the effect of self-scoring on students' regulation accuracy awareness? (RQ3)

It is important when students make (more) accurate regulation judgments as result of an intervention such as self-scoring, that they also feel confident about their accuracy, because otherwise, they might not act upon their accurate judgments. Vice versa, when students (still) make inaccurate judgments after an intervention like self-scoring, it might be helpful if they feel relatively less confident about the accuracy of their judgments, as they might then seek help.

Overall, students' confidence in their regulation judgment accuracy increased after self-scoring. This was not only the case for those students whose regulation judgment accuracy increased or stayed accurate—in which case it is desirable—, but also for students whose regulation judgment accuracy decreased or stayed inaccurate—in which case it is not desirable. This general increase in students' confidence in their regulation judgment accuracy might perhaps be a consequence of the fact that students presumably felt highly confident about their monitoring judgment accuracy after self-scoring (which we did not measure, but the pilot study strongly suggests that this is likely, see section 2.3.4). Self-scoring gives students information about the correctness of their answers, so this becomes a very salient cue to them. This cue is also highly diagnostic, as the students were quite accurate in self-scoring their work, resulting in very accurate monitoring judgments after self-scoring (see Oudman et al., 2022). Students' (implicit) knowledge that this cue (i.e., the self-rated correctness of their answers) was diagnostic could have resulted in feeling highly confident about their monitoring judgment accuracy and this, in turn, may have led to an increase in confidence in their regulation judgment accuracy after self--scoring—regardless of whether their regulation judgment accuracy had *actually* increased. Nevertheless, students' confidence in their regulation judgment accuracy after self-scoring increased more for students whose judgments indeed became more accurate or stayed accurate after self-scoring, than for students whose regulation judgment accuracy became more inaccurate or stayed inaccurate. However, this difference (in the increase in confidence between the groups of students for whom it was desirable and for whom it was not) was only significant for the multiplication task and not for the division task. So, to answer RQ3, based on these findings, one could cautiously conclude that on average, self-scoring seemed to have a positive effect on students' regulation accuracy awareness, especially for the multiplication task and for students whose regulation judgments became more accurate or stayed accurate after self-scoring. However, self-scoring might have a negative effect on students' regulation accuracy awareness for the students whose regulation accuracy decreases or stayed in accurate after self-scoring.

Prior studies showed that after self-scoring, improvements in primary school students' regulation accuracy lag behind improvements in their monitoring accuracy after self-scoring (Oudman et al., 2022; Van Loon & Roebers, 2017). The findings of the present study that these inaccurate regulation judgments after self-scoring are on average not associated with a *decrease* in students' feeling of confidence in their regulation accuracy but an *increase* (compared to their feeling of confidence before self-scoring) makes students' regulation inaccuracy after self-scoring even more worrisome. Hence, future research should further investigate other kinds of interventions for fostering students' monitoring and regulation accuracy awareness, beyond self-scoring. Intervention studies could try to increase students' regulation accuracy awareness by giving them feedback about their monitoring and regulation judgment (in)accuracy, increasing students' knowledge about cue diagnosticity, and giving them insight into their own cue use.

4.3. Limitations

The current study was the first to investigate primary school students' awareness of their monitoring and regulation judgment accuracy. This study has several limitations. First, a potential limitation is that our measures of monitoring and regulation accuracy awareness differed somewhat from two prior studies on accuracy awareness amongst adults, which measured this construct by asking for item-specific judgments and analyzing to what extent students' SOJ-m were higher for accurate than for inaccurate monitoring judgments (e.g., Fritzsche et al., 2018; Händel & Dresel, 2018). In contrast, we asked students for whole-task monitoring judgments (and so did Nederhand et al., 2021), measuring students' monitoring judgment accuracy on an interval scale, in our case ranging from zero to six (instead of accurate vs. inaccurate). 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 and can be useful when students reflect on which specific skills ask for an intervention (Hartwig & Dunlosky, 2017). Nevertheless, primary school (and older) students could also be asked to make item-specific (second-order) judgments and future studies could consider investigating potential effects of the grain size of (second-order) judgments on monitoring and regulation accuracy awareness.

Second, a large number of participants had to be excluded in the present study, because they answered all items correct or incorrect on both days (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; Borghouts et al., 2019). In future studies, researchers could consider to use adaptive tasks or otherwise showing non-adaptive tasks beforehand to the teachers, ask which of their students would normally not get a task of that difficulty, and only exclude these students.

Third, our findings only apply to procedural mathematics tasks, so future research could investigate to what extent these findings also apply to other subjects and tasks. Finally, although our results seem to imply that self-scoring could potentially have caused an increase in students' awareness of their regulation judgment accuracy, we cannot entirely rule out the possibility that students' awareness of their regulation judgment accuracy would also have changed had students not self-scored their performance. Hence, future studies could consider including a control group in which the students do not self-score their work, but perform a filler task in between the two moments of judgments.

4.4. Implications and conclusions

When aiming to improve the effectiveness of students' self-regulated learning, it seems important to not only focus on improving their monitoring and regulation judgment accuracy, but also on increasing their *awareness* of their judgment accuracy. The present study indicates that (without intervention) primary school students are not aware of their monitoring and regulation judgment accuracy. Our results lend some initial support to the idea that asking students to self-score their answers can potentially improve awareness of regulation judgment accuracy for students whose regulation judgment accuracy increased or stayed maximally accurate after self-scoring, but not for students whose regulation judgment accuracy decreased or stayed equally inaccurate after self-scoring. Future research on additional or other means to increase students' monitoring and regulation accuracy awareness (e.g.,

feedback or training) is needed, and ultimately, future research should address the question of whether the effectiveness of students' self-regulated learning indeed benefits more from interventions that not only focus on improving monitoring and regulation judgment accuracy but also on improving students' monitoring and regulation accuracy awareness.

CRediT authorship contribution statement

Sophie Oudman: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Janneke van de Pol:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. **Eva Janssen:** Formal analysis, Methodology, Writing – review & editing. **Tamara van Gog:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Acknowledgement

This work was supported by the Dutch Ministry of Education, Culture and Science (grant number OCW/PromoDoc/1065001).

Appendix A. Differences Between Analyses at the Within- and Between-student Level

We conducted within-subject analyses of monitoring and regulation accuracy awareness (based on multiple measurement points per student), which is important because in our view, students can only be said to show accuracy awareness when they are able to distinguish between their more and less accurate judgments in terms of their SOJs (as also argued by Fritzsche et al., 2018). Hence, we are interested in whether students, when they make more accurate monitoring/regulation judgments for the task on day 1 than for the task on day 2, feel more confident about their monitoring/regulation judgment accuracy for the task on day 1 than for the task on day 2. This approach differs from that of Nederhand et al. (2021) who measured students' monitoring accuracy awareness based on one judgment accuracy and one SOJ measure per student and thus, analyzed the data at the *between*-student level. However, analyzing data at the between-student level answers a slightly different question, namely: do students, who make a more accurate judgment on a task, feel more confident about the accuracy of this judgment, than students who make a less accurate judgment on that task?

Conclusions about whether or not students show monitoring/regulation accuracy awareness could be similar but can also differ depending on whether the analyses are conducted at the within or between-student level. Consider the theoretical example of student A and B displayed in Table A1. When analyzing this data at the between-student level, only including scores on the task on day 1, one would conclude that the students show *no* awareness of their judgment accuracy, as the student of whom the judgment is more accurate (i.e., Student A) does not feel more confident about the accuracy of their judgment (i.e., both students rate their confidence as five). In contrast, when analyzing this data at the within-student level, one would conclude that both students *do* show accuracy awareness as they both feel more confident of their more accurate judgments. Hence, in the Results section, we presented the results of the analyses at the within-student level (i.e., day level). To enable comparison with the study by Nederhand et al. (2021), the results at the between-student level are reported below in Table A2.

Table A.1
Numerical Example of Judgment Accuracy and SOJs for two Fictional Students

	Student A		Student B	
	Judgment accuracy	SOJ	Judgment accuracy	SOJ
Day 1	0	5	1	5
Day 2	2	3	3	3

Note. Accuracy scores closer to zero indicate that students' judgments are more accurate. A higher SOJ indicates that students feel more confident about the accuracy of the judgment.

Table A.2
Effects of Absolute Monitoring/Regulation Judgment Accuracy on SOJ-m/SOJ-r Before Self-Scoring, at the Student *Level*

	SOJ-m	SOJ-r
	B (SE)	
Effect of Judgment accuracy		
Multiplication	0.01 (0.04)	−0.04 (0.06)
Division	−0.14 (0.06)*	−0.17 (0.08)*

***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$.

Interpretation of the Results Displayed in Table A2

Table A2 shows the results of the analyses in which students' SOJ-m/SOJ-r were regressed on their absolute monitoring/regulation judgment accuracy at the between-student level—that is, only measures of students' judgment accuracy and SOJs of the first day were included. For the division task, both monitoring and regulation accuracy were significant and negative predictors of students' SOJ-m and SOJ-r respectively (whereas the effects at the within student level were nonsignificant, see Tables 2 and 3 in the Results section). This indicates that students who made a more accurate monitoring or regulation judgment on the division task of the first day, felt more confident about the accuracy of their judgment, than students who made a less accurate judgment on that task (which, in our view, does not actually indicate accuracy awareness). Thus, if we would have drawn conclusions about students' accuracy awareness based on the results at the between-student level (based on one measurement per student), we could have (erroneously) concluded that students showed awareness of their monitoring and regulation accuracy on the division task.

Appendix B. Coding Students' Actual Need for Intervention

Students were considered 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 Table B1. 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, which we coded as 0. Second, students who made computational errors or exceeded the time limit of 10 min were considered to need additional practice, which we coded as 1. 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), which we coded as 2. 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., 2019; 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 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 practice" and "additional instruction (and practice afterwards)" were both scored as accurate. The detailed coding scheme is depicted in Figure B1. 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 Figure B1). The interrater reliability was substantial for the multiplication tasks ($\kappa = 0.70$) and almost perfect for the division tasks ($\kappa = 0.85$; Landis & Koch, 1977). In case of disagreement, the coders reached consensus through discussion. The first author coded the other 90% of the tasks.

Table B.1
Examples of Procedural and Computational Errors

Type of Error	Example when problem is 6 × 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 to 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 B.2
Frequencies of Students' Regulation Needs vs. their Regulation Judgments

	Regulation need					
	Multiplication			Division		
	Nothing	Additional practice	Additional instruction	Nothing	Additional practice	Additional instruction
Regulation judgment before self-scoring						
Nothing	245	147	100	166	37	50
Additional practice	73	124	116	33	58	91
Additional instruction	22	13	133	15	9	233
Regulation judgment after self-scoring						
Nothing	291	123	56	188	41	35
Additional practice	35	132	117	17	50	91
Additional instruction	9	17	174	7	7	234

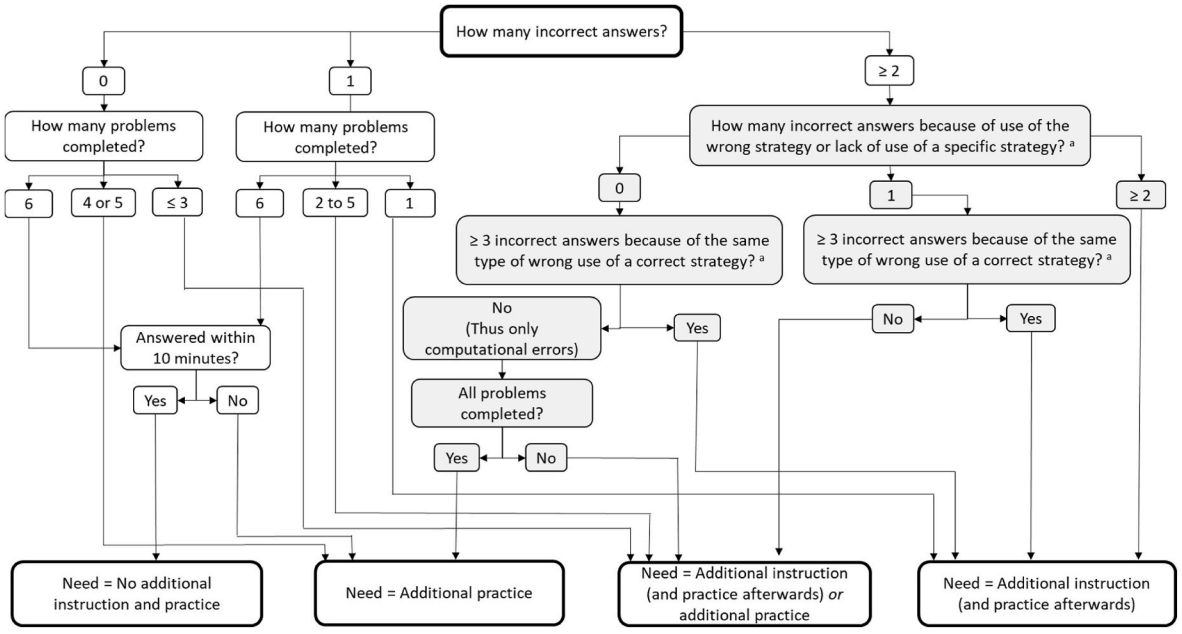


Fig. B.1. Coding Scheme for Actual Needs for Intervention
Note. Unshaded parts were coded automatically, shaded parts were coded manually, [a] See Table B1 for examples. A question mark/cross/line was coded as "lack of strategy."

Appendix C. Mplus Input

Note. Only the input files for analyzing the multiplication task data are provided, as the analyses of the division task data is similar, only the dataset differs. Moreover, the analyses for RQ2A and RQ2B are similar to the ones for RQ1A and RQ2B respectively, except for that the monitoring variables were substituted by regulation variables.

Explanation of variables

- Class: class identification number
- Student: student identification number
- Casenum: unique number based on student number and specific day (day 1 or day 2)
- Performance_mean: students' mean performance across two days, on the multiplication or division task
- Mon_acc_before: Students' absolute monitoring accuracy before self-scoring.
- Reg_acc_before: Students' absolute regulation accuracy before self-scoring.
- SOJ_m_before: Students' SOJ-m before self-scoring
- SOJ_r_reg_before: Students' SOJ-r before self-scoring
- SOJ_r_change: Change in students' SOJ-r from before to after self-scoring
- Reg_acc_change; Change in students' regulation accuracy from before to after self-scoring (0 = accuracy increase or stays accurate; 1 = accuracy decrease or stays inaccurate).

Mplus input for RQ1A

```
DATA: FILE = multiplication.dat;
VARIABLE: NAMES ARE Class Student Casenum performance_mean mon_acc_before
reg_acc_before SOJ_m_before SOJ_r_reg_before SOJ_r_change reg_acc_change;
USEVARIABLES ARE Class Student mon_acc_before SOJ_m_before;
Cluster = Class Student;
Within = mon_acc_before;
MISSING ARE ALL (99999);
DEFINE: CENTER mon_acc_before (GROUPMEAN);
ANALYSIS: TYPE = COMPLEX TWOLEVEL;
ESTIMATOR = MLR;
MODEL:
%WITHIN%
SOJ_m_before ON mon_acc_before;
%BETWEEN%
SOJ_m_before;
OUTPUT: TECH1 SAMPSTAT;
```

Mpluss input for RQ1B

```
DATA: FILE = multiplication.dat;
VARIABLE: NAMES ARE Class Student Casenum performance_mean mon_acc_before
reg_acc_before SOJ_m_before SOJ_r_reg_before SOJ_r_change reg_acc_change;
USEVARIABLES ARE Class Student mon_acc_before SOJ_m_before performance_mean;
Cluster = Class Student;
Within = mon_acc_before;
Between = performance_mean
MISSING ARE ALL (99999);
DEFINE: CENTER mon_acc_before (GROUPMEAN);
CENTER performance_mean (GRANDMEAN);
ANALYSIS: TYPE = COMPLEX TWOLEVEL RANDOM;
ESTIMATOR = MLR;
MODEL:
%WITHIN%
s | SOJ_m_before ON mon_acc_before;
%BETWEEN%
SOJ_m_before on performance_mean;
s with SOJ_m_before;
s on performance_mean;
OUTPUT: TECH1 SAMPSTAT;
```

Mplus input for RQ3

```
DATA: FILE = multiplication.dat;
VARIABLE: NAMES ARE Class Student Casenum performance_mean mon_acc_before
reg_acc_before SOJ_m_before SOJ_r_reg_before SOJ_r_change reg_acc_change;
USEVARIABLES ARE Class Student SOJ_r_change reg_acc_change;
Cluster = Class Student;
Within = reg_acc_change
MISSING ARE ALL (99999);
DEFINE: CENTER reg_acc_change (GRANDMEAN);
ANALYSIS: TYPE = COMPLEX TWOLEVEL;
ESTIMATOR = MLR;
MODEL:
%WITHIN%
SOJ_r_change ON reg_acc_change;
%BETWEEN%
SOJ_r_change;
OUTPUT: TECH1 SAMPSTAT;
```

Appendix D. Correlation Matrices

Table D.1
Zero-order Correlations Between Variables on the Multiplication Task

	1	2	3	4	5	6	7	8	9	10	11	12	13
Day 1													
1. Task performance	–												
2. Monitoring accuracy before self-scoring	−0.44***	–											
3. Regulation accuracy before self-scoring	−0.08	0.30***	–										
4. SOJ-m before self-scoring	0.10*	0.01	0.05	–									
5. SOJ-r before self-scoring	−0.02	0.01	−0.03	0.38***	–								

(*continued on next page*)

Table D.1 (*continued*)

	1	2	3	4	5	6	7	8	9	10	11	12	13
6. Regulation accuracy after self-scoring	−0.01	0.04	0.48***	0.02	−0.03	–							
7. SOJ-r after self-scoring	0.07	0.01	−0.09*	0.24***	0.32***	−0.07	–						
Day 2													
8. Task performance	0.41***	−0.14**	0.01	0.12*	0.04	−0.02	0.06	–					
9. Monitoring accuracy before self-scoring	−0.16***	0.11*	0.13**	−0.05	−0.02	0.06	−0.06	−0.24***	–				
10. Regulation accuracy before self-scoring	−0.15***	0.13**	0.26***	−0.04	−0.15***	0.16***	−0.14**	−0.22***	0.31***	–			
11. SOJ-m before self-scoring	0.11*	−0.05	−0.01	0.39***	0.20***	0.01	0.28***	0.10*	−0.01	−0.01	–		
12. SOJ-r before self-scoring	0.04	−0.05	−0.04	0.28***	0.36***	0.07	0.36***	−0.06	−0.03	−0.08	0.33***	–	
13. Regulation Accuracy after self-scoring	−0.15***	0.05	0.17***	0.01	−0.04	0.12*	−0.04	−0.17***	0.19***	0.61***	0.00	0.06	–
14. SOJ-r after self-scoring	0.06	−0.07	−0.08	0.16***	0.30***	0.00	0.32***	0.12**	−0.11*	−0.15***	0.22***	0.33***	−0.08

***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$.

Table D.2
Zero-order Correlations Between Variables on the Division Task

	1	2	3	4	5	6	7	8	9	10	11	12	13
Day 1													
1. Task performance	–												
2. Monitoring accuracy before self-scoring	−0.20***	–											
3. Regulation accuracy before self-scoring	0.00	0.23***	–										
4. SOJ-m before self-scoring	0.29***	−0.12*	0.06	–									
5. SOJ-r before self-scoring	0.08	−0.13*	−0.11*	0.29***	–								
6. Regulation accuracy after self-scoring	0.01	0.12*	0.63***	−0.01	−0.11*	–							
7. SOJ-r after self-scoring	0.11*	−0.14*	−0.06	0.28***	0.48***	−0.17**	–						
Day 2													
8. Task performance	0.50***	−0.05	0.04	0.16**	0.03	0.09	0.05	–					
9. Monitoring accuracy before self-scoring	−0.07	0.14**	0.08	−0.05	−0.03	−0.03	−0.04	−0.13*	–				
10. Regulation accuracy before self-scoring	−0.08	0.10	0.18***	0.00	−0.07	0.08	0.01	−0.03	0.29***	–			
11. SOJ-m before self-scoring	0.19***	−0.15**	−0.01	0.32***	0.42***	−0.02	0.36***	0.08	−0.06	0.07	–		
12. SOJ-r before self-scoring	0.08	−0.13*	−0.04	0.19***	0.38***	−0.05	0.29***	−0.01	0.02	−0.13*	0.35***	–	
13. Regulation Accuracy after self-scoring	−0.11*	0.08	0.12*	0.02	0.03	0.12*	0.01	−0.05	0.10	0.61***	−0.02	−0.15**	–
14. SOJ-r after self-scoring	0.06	−0.07	−0.04	0.19***	0.36***	−0.01	0.33***	0.06	−0.10	−0.17**	0.23***	0.48***	−0.17**

***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$.

References

Ackerman, R. (2019). Heuristic cues for meta-reasoning judgments: Review and methodology. *Psihologijske Teme, 28*(1), 1–20. https://doi.org/10.31820/pt.28.1.1

Oudman, S., van de Pol, J., Janssen, E., & Van Gog, T. *Primary school students' awareness of their monitoring and regulation judgment accuracy [Dataset]*. (). https://osf.io/36cak/?view_only=ba95fe7d0c6d4cd0a68a6bbed76edf90.

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., 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. (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

Baars, M., Visser, S., Van Gog, T., De Bruin, & 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

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. https://doi.org/10.1146/annurevpsych-13011-143823

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. (2019). *Pluspunt 4 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. 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

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., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second-order judgments about judgments of learning. *The Journal of General Psychology, 132*(4), 335–346. https://doi.org/10.3200/GENP.132.4.335-346

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121. https://doi.org/10.1037/1082-989X.12.2.121

Fritzsche, E. S., Händel, M., & Kröner, S. (2018). What do second-order judgments tell us about low-performing students' metacognitive awareness? *Metacognition and Learning, 13*(2), 159–177. https://doi.org/10.1007/s11409-018-9182-9

Gabriele, A. J., Joram, E., & Park, K. H. (2016). Elementary mathematics teachers' judgment accuracy and calibration accuracy: Do they predict students' mathematics achievement outcomes? *Learning and Instruction, 45*, 49–60. https://doi.org/10.1016/j.learninstruc.2016.06.008

García, T., Rodríguez, C., González-Castro, P., González-Pienda, J. A., & Torrance, M. (2016). Elementary students' metacognitive processes and post-performance calibration on mathematics 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.

Händel, M., & Dresel, M. (2018). Confidence in performance judgment accuracy: The unskilled and unaware effect revisited. *Metacognition and Learning, 13*(3), 265–285. https://doi.org/10.1007/s11409-018-9185-6

Händel, M., & Fritzsche, E. S. (2016). Unskilled but subjectively aware: Metacognitive monitoring ability and respective awareness in low-performing students. *Memory & Cognition, 44*(2), 229–241. https://doi.org/10.3758/s13421-015-0552-0

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

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

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

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

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

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

Nederhand, M., Tabbers, H., de Bruin, A., & Rikers, R. (2021). Metacognitive awareness as measured by second-order judgements among university and secondary school students. *Metacognition and Learning, 16*(1), 1–14. https://doi.org/10.1007/s11409-020-09228-6

OECD. (2022). *Trends shaping education 2022*. OECD Publishing. https://doi.org/10.1787/6ae8771a-en

Oudman, S., van de Pol, J., & Van Gog, T. (2022). Effects of self-scoring their math problem solutions on primary school students' monitoring and regulation. *Metacognition and Learning, 17*(1), 213–239. https://doi-org.proxy.library.uu.nl/10.1007/s11409-021-09281-9.

Oudman, S., van de Pol, J., & van Gog, T. (2023a). Effects of cue availability on primary school teachers' accuracy and confidence in their judgments of students' mathematics performance. *Teaching and Teacher Education, 122*, 103982. https://doi.org/10.1016/j.tate.2022.103982

Oudman, S., van de Pol, J., van Loon, M., & van Gog, T. (2023b). Primary school teachers' judgments of their students' monitoring and regulation skills. *Contemporary Educational Psychology, 75*, 102226. https://doi.org/10.1016/j.cedpsych.2023.102226

Patterson, M. L., Foster, J. L., & Bellmer, C. D. (2001). Another look at accuracy and confidence in social judgments. *Journal of Nonverbal Behavior, 25*(3), 207–219. https://doi.org/10.1007/s10919-009-0072-3

Raaijmakers, S. F., Baars, M., Schaap, L., Paas, F., Van Merriënboer, J., & Van Gog, T. (2018). Training self-regulated learning skills with video modeling examples: Do task-selection skills transfer? *Instructional Science, 46*(2), 273–290. https://doi.org/10.1007/s11251-017-9434-0

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.

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education, 5*, Article 589965. https://doi.org/10.3389/feduc.2020.589965

Roebers, C. M., Mayer, B., Steiner, M., Bayard, N. S., & van Loon, M. H. (2019). The role of children's metacognitive experiences for cue utilization and monitoring accuracy: A longitudinal study. *Developmental Psychology, 55*(10), 2077. https://doi.org/10.1037/dev0000776

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

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*(4), 331–362. https://doi.org/10.1080/01638530902959927

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