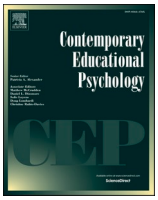
Contemporary Educational Psychology 75 (2023) 102226

Contents lists available at ScienceDirect

Contemporary Educational Psychology

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



Primary school teachers' judgments of their students' monitoring and regulation skills

Sophie Oudman [a,*], Janneke van de Pol [a], Mariëtte van Loon [b], Tamara van Gog [a]

[a] Department of Education, Utrecht University, the Netherlands
[b] Institute for Psychology, Department of Developmental Psychology, University of Bern, Switzerland

ARTICLE INFO

Keywords:
Teacher judgments
Student monitoring accuracy
Student regulation accuracy
Metacognitive judgments
Self-regulated learning
Primary education

ABSTRACT

To help students improve their self-monitoring and self-regulation skills, teachers should have an accurate idea of how well students can monitor and regulate their learning. We investigated how accurately primary school teachers can judge their students' monitoring and regulation accuracy and whether and how student characteristics are related to (the accuracy of) teacher judgments of student monitoring and regulation. Thirty-three teachers, teaching 9—10-year-old students, participated with their classes (N = 495 students). Students completed a multiplication and division task and made monitoring and regulation judgments before and after self-scoring their work. We measured (the accuracy of) teachers' judgments of their students' monitoring skills before self-scoring, and of their students' regulation skills before and after self-scoring. Additionally, we measured teachers' perceptions of student characteristics (e.g., conscientiousness, general mathematics ability, amount of teacher- student contact). Results showed that the teachers correctly estimated that, in general, their students made quite accurate monitoring and regulation judgments. However, they had difficulties with identifying those students who made substantially inaccurate monitoring and regulation judgments (for whom it is particularly important that the teachers can intervene). When taken together, teachers' perceptions of student characteristics explained substantial variance in (the accuracy of) teacher judgments of students' monitoring and regulation skills. Moreover, teacher judgments of students' monitoring accuracy were more accurate when students were perceived to have learning problems or to be relatively more skilled in mathematics. These findings and measures can ultimately contribute to the design of interventions to help teachers judge and develop their students' self-regulated learning skills.

1. Introduction

Preparing primary school students to become self-regulated learners is essential. Not only because self-regulated learning has beneficial effects on students' academic success (Dent & Koenka, 2016) but also because it is increasingly important that students can self-initiate and self-manage their learning outside school and throughout their entire lifetime (Bjork et al., 2013). There are many models describing the different phases and cognitive processes involved in self-regulated learning (e.g., Pintrich, 2000; Winne & Hadwin 1998; Zimmerman, 2000). These models share some general features. For instance, that three phases can be distinguished in self-regulated learning—a forethought (or planning), performance, and reflection (or evaluation) phase—between which learners switch whenever necessary (Panadero, 2017). Two central processes in most models of self-regulated learning, and in switching between the processes, are self-monitoring (evaluating one's own performance) and self-regulation (controlling one's own study activities; De Bruin & Van Gog, 2012; Panadero, 2017; Griffin et al., 2013).

The importance of monitoring and regulation processes is explained in Nelson and Narens' (1990) model of *meta*-memory, which has been extended to *meta*-comprehension (in learning from texts; e.g. Thiede et al., 2003), and *meta*-reasoning (including problem-solving, which the current study focuses on; Ackerman & Thompson, 2017). Across these three subdomains the specific monitoring and regulation decisions differ, but the main principles of how monitoring, regulation, and performance are interrelated are the same: These processes are interrelated by a flow of information between an object-level (the actual performance) and a *meta*-level (representation of the performance). The *meta*-level is informed by the object-level through monitoring. In turn, the

* Corresponding author at: Department of Education, Utrecht University, 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.cedpsych.2023.102226

Available online 23 September 2023

meta-level modifies the object-level through regulation. One assumption of this model is that monitoring influences the regulation decisions. Thus, accurate monitoring is a necessary, though not sufficient, precondition for accurate regulation, and therefore, for effective self-regulated learning (e.g., Dunlosky & Rawson, 2012; Thiede et al., 2003). That is, if students overestimate their performance, they may quit studying or practicing too early, and if they underestimate their performance (which seems more rare, De Bruin et al., 2017; Kruger & Dunning, 1999; Oudman et al., 2022) they will spend time on activities they already mastered rather than on those they need to learn.

Unfortunately, people's self-monitoring and self-regulation are often inaccurate, and primary school students are no exception (e.g., Baars et al., 2014; Oudman et al., 2022; Prinz et al., 2020; Van Loon & Roebers, 2017). Prior studies have found that interventions, such as asking students to self-score their work with the use of standards, improved primary school students' monitoring and regulation accuracy (in mathematics: Oudman et al., 2022; in text comprehension: Van Loon & Roebers, 2017). However, their regulation accuracy was still far from perfect after such interventions.

Students who cannot accurately monitor and regulate their learning process will need support from their teachers to develop these skills. To provide effective and efficient support for self-regulated learning, teachers first need to identify their students' need for support; that is, teachers need to be able to accurately judge their students' monitoring and regulation skills. This would allow teachers to instruct students on how to evaluate their performance, make appropriate subsequent decisions, and when and how to seek help (Azevedo et al., 2008; Dignath & Büttner, 2018). In other words, the model of Nelson and Narens (1990) could also be applied to the teacher's task: Teachers have a representation (*meta*-level) of their students' monitoring and regulation skills (the object-level). That is, teachers should accurately monitor their students' monitoring and regulation skills in order to be able to accurately regulate the activities to help students to improve these skills. However, it is unknown whether primary school teachers have accurate insights into how well their students can monitor and regulate their learning. Furthermore, to gain insight into how teacher judgments of students' monitoring and regulation skills can be improved, it is relevant to gain insight into how these teacher judgments are established, that is, into the kind of information teachers use when making judgments of their students' monitoring and regulation skills.

The present study addresses these questions in the context of mathematics problem-solving in primary school. Below, we explain how we define (teacher judgments of) student monitoring and regulation accuracy, the insights that previous research has acquired about these judgments, and how teachers' perceptions of student characteristics might be associated with these judgments.

1.1. Student monitoring and regulation judgments

Students are engaged in multiple monitoring and regulation processes before, during, or after completing a problem-solving task (Ackerman & Thompson, 2017). In the present study we focus on the monitoring and regulation judgments that primary school students commonly make after completing a mathematical problem-solving task, consisting of several items. In this context students' monitoring judgments can, for example, consist of their evaluations of how many items they answered correctly. When these evaluations lead to the conclusion that students do not yet master specific mathematical skills, common regulatory actions (in Dutch primary schools) are (1) getting additional instruction (from the teacher or another student) when students do not understand how to solve the problems, or (2) getting additional (similar) practice problems when they understand how to solve the problems, but not automated the procedure. When students master a certain type of problem, they can continue working on another/subsequent learning goal (Baak et al., 2018; Borghouts et al., 2019; Hollingsworth & Ybarra, 2018).

Different measures can be used to determine the accuracy of students' monitoring and regulation judgments (cf. Schraw, 2009, for a discussion of different measures). We are mainly interested in absolute accuracy, as this measure indicates the degree to which students know how they performed on a task and what their needs are. *Students' absolute monitoring accuracy* can be expressed by the absolute (unsigned) difference between a student's monitoring judgment of how many problems they answered correctly and the student's actual performance—that is, the number of problems they answered correctly (Baars et al., 2014; Dunlosky & Rawson, 2012; Oudman et al., 2022). We define *students' absolute regulation accuracy* as the extent to which a student's regulation judgment, meaning their evaluation of their need for additional instruction or practice, is in line with their actual need for intervention, as indicated by experts (cf. Oudman et al., 2022).

1.1.1. Effect of self-scoring on student monitoring and regulation accuracy

Before students make a regulation judgment, they will often have self-scored their task—that is, comparing their answers to the correct ones. Self-scoring seems a powerful tool to increase the accuracy of primary school students' monitoring and regulation judgments (Oudman et al., 2022; Van Loon & Roebers, 2017) as well as their learning outcomes (Hattie, 2009; Sadler, 1989), and is increasingly implemented in primary schools.

A prior study (Oudman et al., 2022) showed that students' absolute monitoring and regulation accuracy improved after they self-scored their solutions of procedural mathematics problems. However, whereas students' monitoring judgments came close to being perfectly accurate after self-scoring, students' regulation judgments (despite some improvement) frequently stayed inaccurate. The inaccurate regulation judgments after self-scoring were mostly too optimistic: Students indicated they needed no regulatory intervention (additional instruction or practice) whereas they actually did, or students indicated they needed a less intensive intervention than they actually did (i.e., indicating they only needed additional practice whereas they also needed additional instruction). As such inaccurate and overly optimistic regulation judgments can harm students' learning, these students need help from their teachers to improve the accuracy of their regulation judgments after self-scoring. To be able to provide such support and determine which students need support, teachers must be able to estimate how accurate their students' regulation judgments are. Hence, in the present study, we focus on teacher judgments of students' regulation accuracy before and after self-scoring.

1.2. Prior research into teacher judgments of student monitoring and regulation

There has been relatively little research on teachers' ability to judge their students' monitoring and regulation skills accurately. Two prior studies investigated primary school teachers' ability to judge their students' monitoring skills (Fleury-Roy & Bouffard, 2006; Jamain, 2019). The teachers were asked to classify their students into one of three categories: pessimists (i.e., students who underestimate their performance), optimists (i.e., students who overestimate their performance), or realists (i.e., students who accurately estimate their performance). Their judgments were then compared to whether students were actually realists, optimists, or pessimists. Both studies found that teachers were most accurate at classifying the realists and substantially less accurate at classifying optimists and pessimists. However, the methodological approach in these studies does not necessarily match educational practice, as their classification was based on *z*-scores, resulting in a fixed proportion of students in the class being classified as a realist, or in other words, as accurate. In contrast, in educational practice, the proportion of students in a class judging their performance accurately is likely to vary and may differ across tasks. Moreover, the approach of Jamain (2019) and Fleury-Roy and Bouffard (2006) does not enable us to establish the *degree* to which the actual student monitoring accuracy and teacher

judgments of student monitoring accuracy are different, which is important because a larger deviation is more problematic than a smaller one.

A study by Van de Pol and Oudman (in press) addressed absolute accuracy, but in a sample of secondary school teachers. It was investigated to what extent teachers were able to judge the accuracy of their students' monitoring judgments regarding their performance on a text comprehension test. On average, teachers' judgments deviated 3.44 points on a 24-point scale ($SD = 2.99$) from students' actual monitoring accuracy, which can be interpreted as fairly accurate.

None of these studies examined how accurately teachers could judge students' regulation decisions. Thus, it remains an open question how well primary school teachers can judge students' monitoring and regulation skills in terms of absolute accuracy.

1.3. (Potential) effects of student characteristics on teacher judgments of students' monitoring and regulation skills

When asked to judge their students' *performance*, teachers also seem to use their perceptions of general student characteristics (also called student cues) in making these judgments. For instance, some teachers seem to think (rightly so or not) that students will perform better on a task when they have higher general cognitive abilities (e.g., Kaiser et al., 2015), show more effort (e.g., Kaiser et al., 2013), are more interested in the task, have higher self-concept (e.g., Oudman et al., 2023b), have no disabilities (e.g., Hurwitz et al., 2007), or have no migration background (e.g., Furnari et al., 2017). Whether or not teachers also base their judgments of students' task performance on students' SES and sex or gender is less clear (for a review, see Urhahne & Wijnia, 2021).

There are some indications that teachers might also use their perceptions of such student characteristics when making judgments about students' monitoring and regulation skills. In interviews in the study by Dignath and Sprenger (2020), teachers reported using students' off-task behavior and (self-assessed) achievement level as indicators of students' self-regulated learning. Callan and Shim (2019) found that teachers reported seeing off-task behavior, disengagement, and poor academic performance as indicators of poor self-regulation. Correlational analyses by Carr and Kurtz-Costes (1994) suggest an association between teachers' perceptions of students' achievement level and self-concept and their judgments of students' metacognitive abilities. Correlational analyses by Friedrich et al. (2013) suggest an association between teachers perceptions of students' mathematics competence and their judgments of students' self-regulated learning strategies in the preactional/forethought phase (i.e., goal setting and planning behavior) when being engaged in mathematical tasks.

Because of a paucity of research, it is unclear whether teachers' perceptions of student characteristics would also be associated with their judgments of students' monitoring and regulation skills. We therefore explore what information about their students primary school teachers might use when making judgments of their students' monitoring and regulation skills.

Depending on the information teachers have available about each of their students, it might be easier for teachers to make accurate judgments for some students than others. For instance, it might be easier to make more accurate judgments when more relevant information is available (Funder, 2012, showed this for personality judgments). The degree to which teachers have information about the monitoring and regulating skills of their students, however, might very well differ across students, for instance, as a result of the amount of teacher-student contact or students' degree of extraversion. The halo effect could also play a role, that is, the tendency for positive impressions of a person in one area to influence one's opinion or feelings in other areas (Thorndike, 1920). Teachers could, for example, (erroneously) think that students who are better in mathematics, work more conscientiously, have less learning problems, or are more likeable, would have better monitoring and regulation skills.

By exploring how (perceived) student characteristics relate to (the accuracy of) teacher judgments about students' monitoring and regulation skills, we aim to gain more insight into how (in)accurate teacher judgments of student monitoring and regulation skills come about, which can ultimately contribute to interventions aimed at increasing teachers' ability to correctly judge their students' monitoring and regulation skills.

2. Present study

The present study aims to gain insight into how well primary school teachers can make judgments of their students' monitoring and regulation skills in the context of mathematics and what factors are related to these judgments. Because primary school students are often asked to self-score their work (using a standard of correct answers) before making their regulation decisions, it is relevant to know how well teachers can make judgments of their students' regulation skills before and after self-scoring.

This study has four aims. First, we aimed to investigate whether teachers had accurate insights into the monitoring accuracy of their students, by determining to what extent teacher judgments of their students' monitoring accuracy were in line with students' actual monitoring accuracy before self-scoring (Research Question [RQ] 1A). In the study of Van de Pol and Oudman (in press) secondary school teachers' judgments of students' monitoring accuracy of their text comprehension seemed to be on average fairly accurate (see section 1.2). Based on this finding, we expected that the teachers in our sample would also make fairly accurate judgments of students' monitoring accuracy in mathematics. Moreover, as it is particularly important for students with substantially inaccurate monitoring judgments that teachers can intervene, we explored to what extent teachers were able to identify the students of whom the monitoring judgments were substantially inaccurate (RQ1B).

Second, we aimed to investigate to what extent teachers had accurate insight into the regulation accuracy of their students. Therefore, we explored to what extent teacher judgments of their students' regulation accuracy were in line with students' actual regulation accuracy before and after self-scoring (RQ2A). Moreover, we explored to what extent teachers were able to identify the students who made inaccurate regulation judgments (RQ2B), as these students would be most in need of support with developing their regulation skills. We did not have specific hypotheses with regard to RQ2 because (the accuracy of) teacher judgments of students' regulation skills have not been studied before.

Third, we aimed to gain more knowledge about what information teachers might use to make judgments about their students' monitoring and regulation skills. We therefore investigated which student characteristics (as perceived by teachers) explained the magnitudes of the teacher judgments of their students' monitoring and regulation accuracy (RQ3). Based on studies of teacher judgments of students' self-regulated learning and metacognitive abilities (Callan & Shim, 2019; Carr & Kurtz-Costes, 1994; Dignath & Sprenger, 2020; Friedrich et al., 2013), we expected that teachers' perceptions of students' mathematics abilities, variables related to students' working behavior (such as effort and conscientiousness), and students' self-concept might be associated with their judgments of students' monitoring and regulation skills.

Fourth, we aimed to study whether it would be easier to make accurate judgments about some students' monitoring and regulation skills than others, depending on (perceived) students' characteristics. Therefore, we explored whether and to what extent student characteristics (as perceived by teachers) explained the degree to which teacher judgments of their students monitoring/regulation accuracy were in line with students' actual monitoring/regulation accuracy (RQ4). For instance, it might be that teachers have more information about students with whom they have more contact or about students who are more extravert, and that this results in more accurate judgments of students' monitoring and regulation skills (see section 1.3). However, because of a lack of prior research, we had no specific hypotheses.

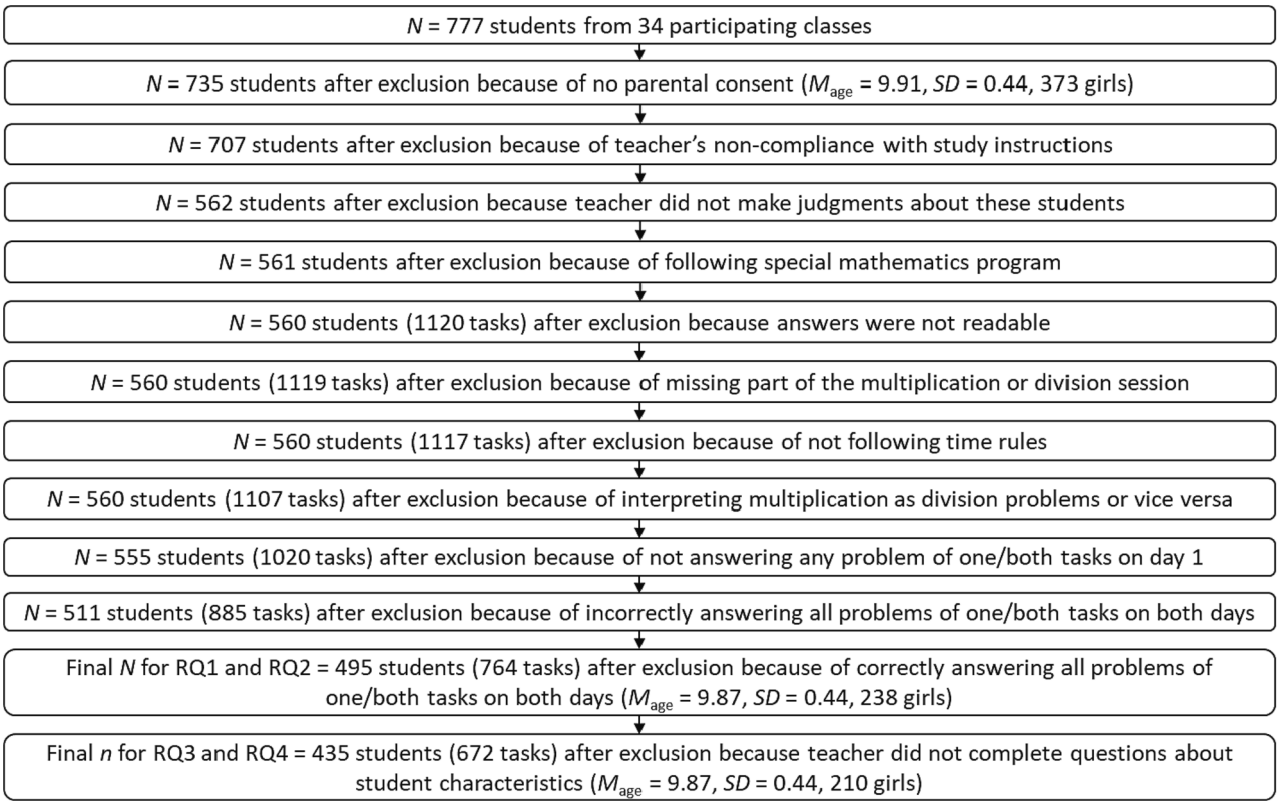


Fig. 1. *Flowchart of Reasons for and Number of Excluded Students and Tasks Note.* For some students, data from only one of the tasks was used. Multivariate outliers were defined after exclusion of students and for each analysis separately (and are thus still included in this flowchart).

3. Method

This study is based on a dataset of a larger project that also focuses on student monitoring and regulation judgments (Oudman et al., 2022; submitted) and teacher judgments of their students' performance (Oudman et al., 2023b) in the context of mathematics problem-solving in primary school. For completeness, in section 3.2. we shortly mention all measures that were part of the procedure, but we only elaborate on the measures that were used in the present study. Note that there may be some overlap in the description of the method section with other papers.

3.1. Participants

3.1.1. Teachers

Thirty-four teachers, teaching 9–10-year-old students (Dutch grade 6, comparable to US grade 4 in terms of age), volunteered to participate in this study. One teacher dropped out because of not feeling comfortable with completing the questionnaire about student characteristics. The other 33 teachers (25 female) taught at 21 different primary schools in the Netherlands. They were 23 to 59 years old (M = 37.71, SD = 12.10) and had one to 39 years of teaching experience (M = 12.33, SD = 10.18). They taught their classes between two and five days a week (M = 4.24, SD = 0.94). Data collection took place between January and May 2019. The teachers were teaching their students from the beginning of the school year, which, in the Netherlands, roughly spans from the end of August until half July, so they had known their students between 5 and 9 months. Eight of the teachers had also been teaching their class in a previous grade.[1] This study received approval from the ethics review board of the Faculty of Social and Behavioural Sciences, Utrecht University.

3.1.2. Students

Of the 777 students who participated, data from 495 students could be included in the analyses of RQ1 and 2, and data from 435 students in the analyses of RQ3 and 4 (as we only had data on teachers' perceptions of student characteristics for a part of the student sample, see section 3.2.2). Each student completed a multiplication and division task, but for some students, data from only one of the tasks could be used. Fig. 1 displays students' demographics and the number of students and tasks that were excluded, including the reason for exclusion. As Fig. 1 shows, data on a substantial number of tasks (i.e., 343 tasks: difference between 1107 and 764) was excluded because students did not answer any problem on day 1, or all problems were answered correctly or incorrectly on both days.[2] The reason these data were excluded from the analyses is that the tasks were presumably too complex or too easy for these students, and therefore, making accurate judgments would be relatively easy for these students and their teachers. Including these data from these students could have distorted the results (cf. Oudman et al., 2022).

[1] We found no significant differences in the (the accuracy) of teacher judgments of students' monitoring and regulation skills between the eight teachers who taught their class also in a previous grade and the other 25 teachers, $p > .05$.

[2] Similar tasks were administered on two days, but in the present study we only used student data of day 1, see Section 3.2.

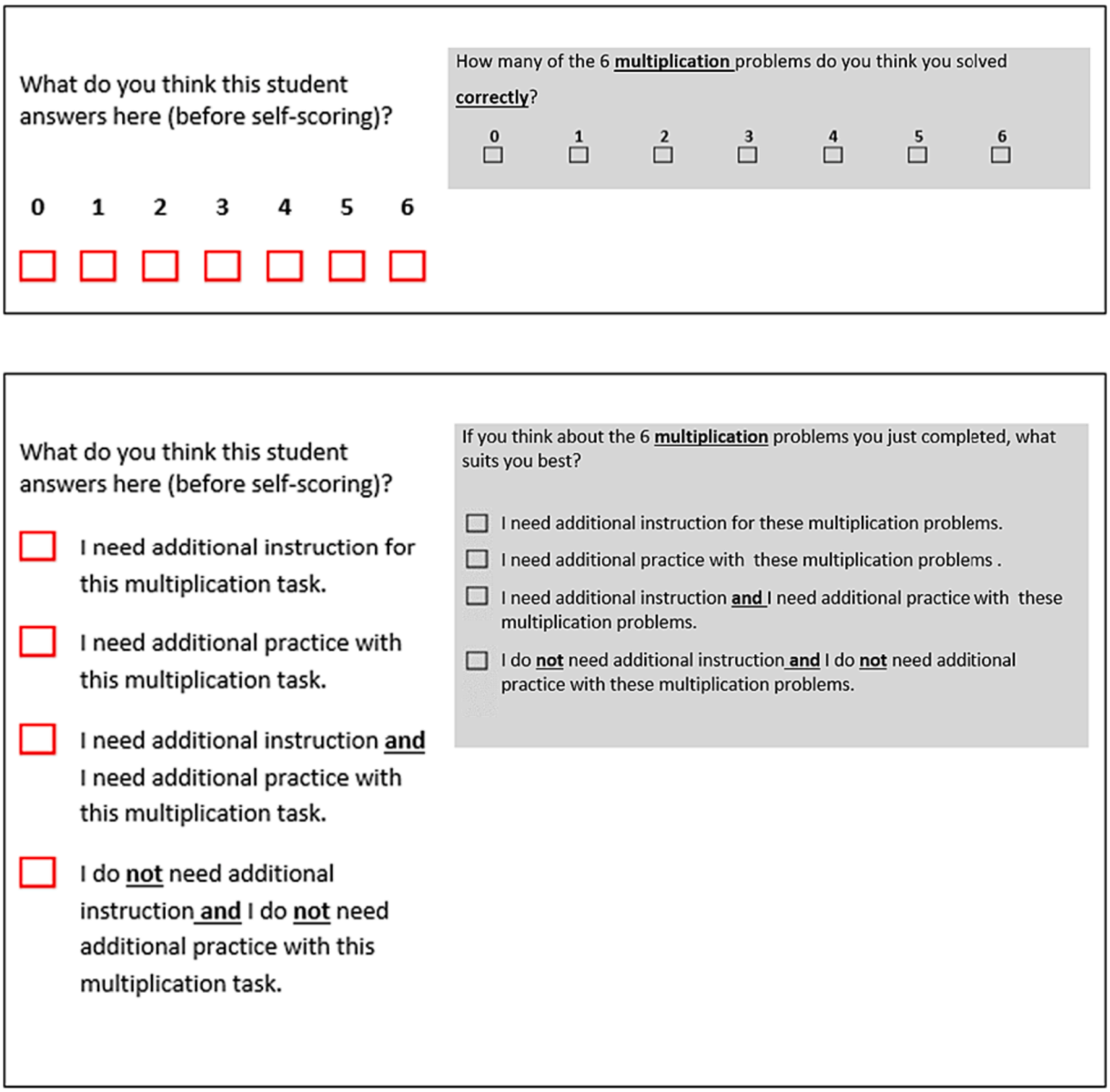


Fig. 2. Measures of Teacher Judgment of Student Monitoring/Regulation Judgment (TJSMJ & TJSRJ).

3.2. Materials and procedure

The data collection took place at participants' schools on two normal lesson days, with exactly one week in between. On both days, the student and teacher session took place simultaneously and lasted between 45 min and one hour.

3.2.1. Students

On day 1, after a short introduction by the experimenter, all students received the first booklet and a pen and completed the multiplication task, consisting of six multiplication problems (single-digit multiplicands multiplied by 3-digit multipliers, e.g., 6 × 472). They had 12 min to complete the task, but it was emphasized that there was no need to hurry. When students finished the task in less than 12 min, they were instructed to read the (fiction) books they kept in their drawers. After 12 min, the experimenter instructed that the students who had not yet finished all problems should stop working on the task. Next, the students made a monitoring judgment (Student Monitoring Judgment; SMJ) by answering the question "How many of the six multiplication problems do you think you solved correctly?" in their personal booklets. Then, the students made a regulation judgment (Student Regulation Judgment; SRJ) by answering the question "If you think about the six multiplication problems you just completed, what suits you best?", choosing one of the following options: additional instruction/ additional practice/ additional instruction and practice/ no additional instruction and no additional practice. These monitoring/regulation questions were read aloud and explained by the experimenter. In addition, students answered some other questions that were outside of the scope of the present study (which is, as mentioned earlier, based on a dataset from a larger project). This entire procedure was then repeated for the division task (consisting of six division problems: 3-digit dividends divided by single-digit divisors, e.g., 282: 6).

Next, all students received the second booklet and changed their blue pen to 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 one of the boxes (the experimenter did not read the correct answers aloud). After self-scoring, the students again made a monitoring judgment (SMJ; for

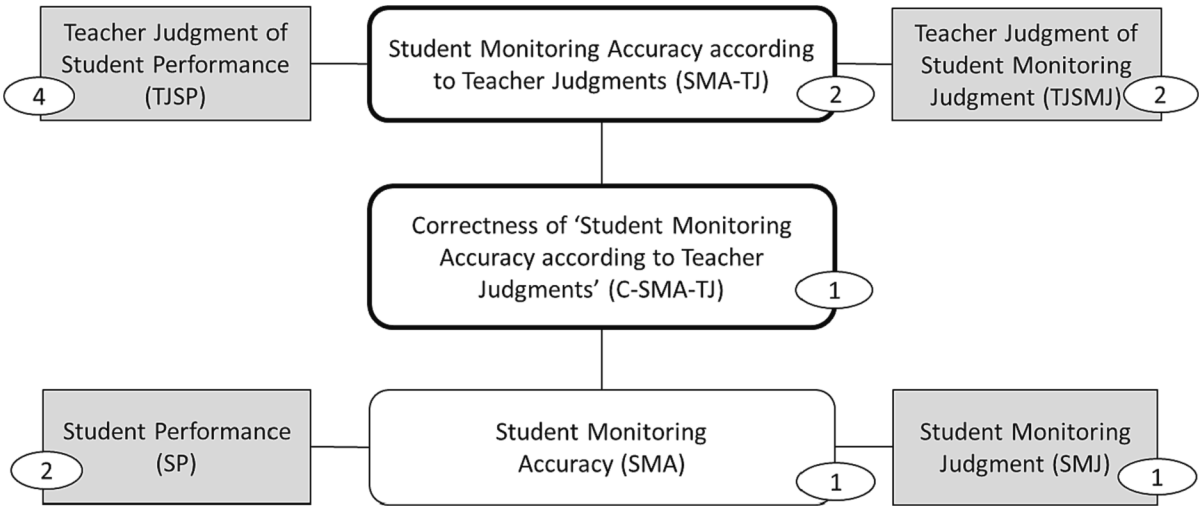


Fig. 3. *Measurement Framework of Teacher Judgments of Students' Monitoring Accuracy Note.* Shaded boxes are variables that we directly measured. Bold-lined boxes are the variables that this study mainly focuses on. All measures in this Figure range from zero to six in the present study; the (fictional) displayed values are those used for the calculation examples in the text.

another study, not used in the present study) and regulation judgment (SRJ) which were again read aloud by the experimenter. This self-scoring and judgment procedure was then repeated for the division task. This entire procedure (i.e., completing two booklets; but with isomorphic multiplication and division problems) was repeated exactly on the second day one week later (for another study; these data were not used in the present study).

3.2.2. Teachers

During the student session on day 1, teachers were provided with a laptop, a noise-canceling headphone, and a list with names of the students they had to make judgments about. For each teacher, 20 students were randomly selected. If a class consisted of 20 students or less, teachers made judgments about all their students. Teachers sat in or close to their classroom in such a way that they could not see their students' answers. On the laptop, teachers made five judgments for each selected student, all regarding the multiplication task. First, they made a judgment of student performance (Teacher Judgment of Student Performance; TJSP): Teachers were provided with the six multiplication items that students were asked to complete and answered the question "How many of these six multiplication problems do you think this student answers correctly within 12 min?". Second, teachers made a judgment of the student's need for intervention (Teacher Judgment of Student Need for intervention; TJSN). Hereto, teachers indicated which of the following needs was most applicable to the student with regard to the multiplication task: (1) additional instruction, (2) additional practice, (3) additional instruction and practice, or (4) no additional instruction and no additional practice. Third, teachers made a judgment of the student monitoring judgment, before self-scoring (Teacher Judgment of Student Monitoring Judgment; TJSMJ). Hereto, teachers were provided with the student monitoring judgment question and answered the question "What do you think this student answers here (before self-scoring)?", see Fig. 2. Fourth, teachers made a judgment of the student regulation judgment before self-scoring (Teacher Judgment of Student Regulation Judgment; TJSRJ): Teachers were provided with the student regulation judgment question and answered the question "What do you think this student answers here (before self-scoring)?", see Fig. 2. Fifth, teachers made a similar judgment of the student regulation judgment (TJSRJ), but after self-scoring. Then, teachers made the same five judgments, but with regard to the division task, after which they continued with making judgments for the next student.

During the student session on day 2, teachers completed a questionnaire about their perceptions of student characteristics for a part of the students for whom they made judgments on day 1. The student samples differed between day 1 and day 2, because on day 1, the students for whom teachers were asked to make judgments were selected randomly. On day 2, the student sample was optimized in terms of the variability in student performance (i.e., we avoided selecting students with similar scores as much as possible) to ensure variability in the teacher judgments of students' performance. After making the judgments of students' performance (for another study; not used in the present study), teachers' perceptions of the following student characteristics were measured: amount of contact (between teacher and student), conscientiousness (during mathematics lessons), effort (during mathematics lessons), extraversion (in general in class), sex, general interest in mathematics, general mathematics ability, likeability (how much the teacher likes the student), nationality, presence of learning problems, and self-concept (students' confidence in their mathematical skills).[3] Most perceptions of student characteristics were measured using one item per characteristic. An example item is: "This student works conscientiously during the regular mathematics lesson. *Examples: This student works orderly. This student works precisely*." The teachers answered this single question on a 4-point scale with the following options: strongly disagree, disagree, agree, strongly agree. Table 9 in Appendix A contains a list of the student characteristic measures, answer scales, and descriptive statistics per characteristic.

3.3. Judgment measures

To measure teacher judgments of students' monitoring and regulation accuracy, we used the same approach as used by Van de Pol and Oudman (in press) that builds on the literature about teacher judgments of their students' performance. Fig. 3 (monitoring) and 4 (regulation) display how the concepts related to teacher judgments of their students' monitoring and regulation accuracy can be operationalized, as well as numeric examples. The concepts in the boxes with bold lines are the main focus of this study. All measures are explained below (the tasks and questions are explained in the previous section 3.2). Additional

[3] Students' intelligence and parents' educational level were also included in the questionnaire with the intention of use in the present study. However, we removed students' intelligence from the analyses to prevent multicollinearity because the correlation with mathematics ability was very high (0.82). Parents' educational level was also removed from the analyses because teachers could not report this variable with certainty for most students.

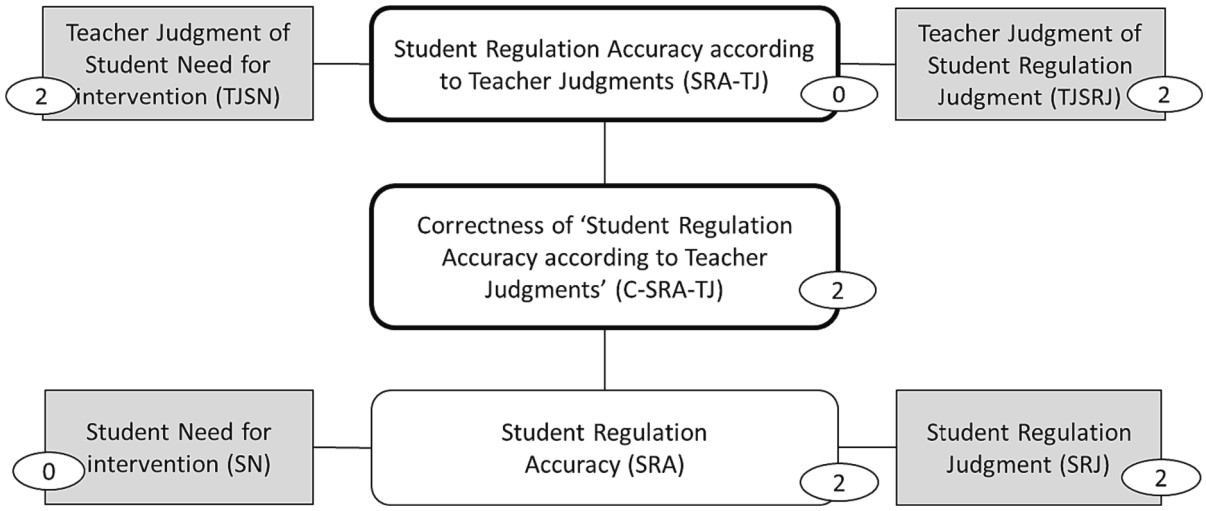


Fig. 4. *Measurement Framework of Teacher Judgments of Students' Regulation Accuracy Note.* Shaded boxes are variables that we directly measured. Bold-lined boxes are the variables that this study mainly focuses on. All measures in this Figure range from zero to two in the present study; the (fictional) displayed values are those used for the calculation examples in the text.

measures (the difference between TJSP and SP, TJSMJ and SMJ, TJSN and SN, and TJSRJ and SRJ) that can be derived from these measurement frameworks, but were not related to our specific research questions, are explained and reported in the online Supplement.

3.3.1. Student measures

Student Performance (SP). Students received one point for each problem that was solved correctly; thus the performance scores ranged between zero and six, separately for the multiplication and division tasks. In the numeric example in Fig. 3, the student scored two points.

Student Need for Intervention (SN). Student Need for intervention was coded based on a coding scheme we developed for a prior study (for a more detailed description, see Oudman et al., 2022). In short, we distinguished four categories, based on the time students needed to complete the task and whether they made computational or procedural errors. The types of errors could be inferred because students had been instructed to use space within the task booklets as scrap paper and write out their computations. First, students who correctly answered five or six out of six problems within 10 min were considered to *not need additional instruction or practice*. This category was coded as 0. Second, students who made computational errors or exceeded the time limit of 10 min (indicating that they had not sufficiently automated the procedures) were considered *to need additional practice*, which we coded as 1. Third, students who made procedural errors were considered to *need additional instruction (and practice afterwards)*, which we coded as 2. 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 student in the numeric example in Fig. 4 did not need additional instruction or practice, represented by the value '0'. The interrater reliability of coding students' needs was substantial for the multiplication tasks ($\kappa = 0.70$) and high for the division tasks ($\kappa = 0.85$; Landis & Koch, 1977).

Student Monitoring Judgment (SMJ). Students answered the monitoring judgment question on a scale ranging from 0 to 6. The student in the numeric example in Fig. 3 thought they scored one point.

Student Regulation Judgment (SRJ) Before and After Self-Scoring. Students'.

regulation judgments were coded as follows: (0) nothing needed, (1) additional practice needed, and (2) additional instruction needed (and practice afterwards). The needs 'additional instruction' and 'additional practice and additional instruction' were merged (for explanation see Oudman et al., 2022). The student in the numeric example in Fig. 4 thought they needed additional instruction (and practice afterwards), represented by the value '2'.

Student Monitoring Accuracy (SMA). Student monitoring accuracy was computed as the absolute difference between the judged and actual performance (i.e., regardless of whether it was positive or negative), ranging from zero to six, with scores closer to zero indicating that students know better how well they performed on a task. In the numerical example in Fig. 3, students' absolute monitoring accuracy is one on a scale ranging from zero to six, which can be interpreted as quite accurate.

Student Regulation Accuracy (SRA) Before and After Self-Scoring. Student Regulation Accuracy is the absolute difference between the Student Regulation Judgment (SRJ) of their need for intervention and the actual Student Need for intervention (SN), ranging from zero to two, with accuracy scores closer to zero indicating that students know better what their regulatory needs are. In the numeric example in Fig. 4, the student's absolute regulation accuracy has the value '2', indicating that the student regulation judgment maximally deviates from their actual need for intervention—and thus, is very inaccurate.

3.3.2. Teacher measures

Teacher Judgment of Student Performance (TJSP). Teachers judged their students' performance on a scale ranging from zero to six. The Teacher Judgment of Student Performance in the numeric example in Fig. 3 is four, so this teacher thought that the student scored four points on the task.

Teacher Judgment of Student Need for Intervention (TJSN). Teachers' judgments of their students' need for intervention were coded as follows: (0) nothing needed, (1) additional practice needed, and (2) additional instruction needed (and practice afterwards). As for the student regulation judgment, the needs 'additional instruction' and 'additional practice and additional instruction' were merged. The Teacher Judgment of Student Need for intervention in the numeric example in Fig. 4 is 'additional instruction (and practice afterwards)', represented by the value '2'.

Teacher Judgment of Student Monitoring Judgment (TJSMJ). Teachers judged their students' monitoring judgment on a scale ranging from zero to six. The Teacher Judgment of Student Monitoring Judgment in the numeric example in Fig. 3 is two, so this teacher thought that the student thought that they scored two points on the task. Teacher

judgments of student monitoring after self-scoring were not measured, as the teachers in the pilot study assumed students would make perfectly accurate monitoring judgments after self-scoring.

Teacher Judgment of Student Regulation Judgment (TJSRJ) Before and After Self-Scoring. The teachers' judgments of their students' regulation judgments were coded as follows: (0) nothing needed, (1) additional practice needed, and (2) additional instruction needed (and practice afterwards). In the numeric example in Fig. 4, the teacher thought that the student thought they needed 'additional instruction (and practice afterwards)', represented by the value '2′.

Student Monitoring Accuracy According to Teacher Judgments (SMA-TJ). Student Monitoring Accuracy according to Teacher Judgments is one of the four main variables in the present study. It is expressed by the student monitoring accuracy according to two teacher judgments: the Teacher Judgment of Student Performance (TJSP) and the Teacher Judgment of the Student Monitoring Judgment (TJMJ). In the numeric example in Fig. 3, the Teacher Judgment of Student Performance is four, and the Teacher Judgment of Student Monitoring Judgment is two. Thus, this student would inaccurately estimate (in this case: underestimate) their performance according to the teacher. This is what Student Monitoring Accuracy according to Teacher Judgments (SMA-TJ) expresses, as this measure is defined by the absolute (unsigned) difference between the teacher judgments of students' performance and monitoring (TSJP & TJSMJ) and indicates to what degree the teacher thinks that the student makes an accurate judgment of their performance. In the numeric example of Fig. 3, the Student Monitoring Accuracy according to Teacher Judgments is two (difference between four and two) on a scale ranging from zero to six, with scores closer to zero indicating the students are more accurate in the teachers' eyes, meaning that the teacher thought that the student made a somewhat inaccurate monitoring judgment of their performance.

Student Regulation Accuracy According to Teacher Judgments (SRA-TJ) Before and After Self-Scoring. The Student Regulation Accuracy according to Teacher Judgments is computed by subtracting the Teacher Judgment of Student Need for intervention (TJSN) from the Teacher Judgment of the Student Regulation Judgment (TJRJ). In the numeric example in Fig. 4, the Teacher Judgment of Student Need for intervention and the Teacher Judgment of Student Regulation are both 'additional instruction (and practice afterwards)', represented by the value '2′. Thus, in this case, the Student Regulation Accuracy according to Teacher Judgments (SRA-TJ), which indicates to what degree the teacher thinks that the student makes an accurate judgment of their own need for intervention, is zero on a scale ranging from zero to two, indicating that the teacher thought that the student made a perfectly accurate regulation judgment of their need for intervention.

Correctness of 'Student Monitoring Accuracy According to Teacher Judgments' (C-SMA-TJ). The Correctness of 'Student Monitoring Accuracy according to Teacher Judgments' is expressed by the absolute difference between the Student Monitoring Accuracy according to Teacher Judgments (SMA-TJ) and the actual Student Monitoring Accuracy (SMA). The Correctness of 'Student Monitoring Accuracy according to Teacher Judgments' indicates how well the teacher knows how accurately a student monitors their performance. In the numeric example in Fig. 3, the Student Monitoring Accuracy according to Teacher Judgments only deviates by one point from the actual Student Monitoring Accuracy (difference between two and one). Thus, the correctness (C-SMA-TJ) score is one on a scale ranging from zero to six (with zero meaning fully correct), indicating that the teacher knew quite well how accurately the student monitored their performance.

Correctness of 'Student Regulation Accuracy according to Teacher Judgments' (C-SRA-TJ) Before and After Self-Scoring. The Correctness of 'Student Regulation Accuracy according to Teacher Judgments' is expressed by the absolute difference between the Student Regulation Accuracy according to Teacher Judgments (SRA-TJ) and the actual Student Regulation Accuracy (SRA). The Correctness of 'Student Regulation Accuracy according to Teacher Judgments' indicates how well the teacher knows how accurately a student judges their need for intervention. In the numeric example in Fig. 4, the Student Regulation Accuracy according to Teacher Judgments deviates two points from the actual Student Regulation Accuracy. This results in a correctness (C-SRA-TJ) score of two on a scale ranging from zero to two, indicating that the teacher did not know how accurately the student judged their own need for intervention.

Table 1
Intraclass Correlation Coefficients (ICC) of Main Study Variables.

		Student Accuracy according to Teacher Judgments	Correctness of 'Student Accuracy according to Teacher Judgments'
Monitoring	ICC Student	0.411	0.103
	ICC Teacher	0.059	0.005
Regulation before self-scoring	ICC Student	0.308	0.085
	ICC Teacher	0.067	0.002
Regulation after self-scoring	ICC Student	0.292	0.172
	ICC Teacher	0.006	0.007

Note. The ICC reflects the amount of between-student and between-teacher variability compared to the total amount of variability (within students, between students, and between teachers).

3.4. Analyses

To answer RQ1 and 2 (about the degree of Correctness of 'Student Monitoring/ Regulation Accuracy according to Teacher Judgments' and whether teachers can identify which students made substantially inaccurate judgments), we provided descriptive statistics. RQ3 and 4 were analyzed by performing multilevel regression analyses in Mplus version 8 (Muthén & Muthén, 1998–2017), to account for the nested data structure. We treated the data as existing of three levels: tasks (level 1) clustered in students (level 2) and students clustered in teachers (level 3). The teacher level was modeled by use of the "Complex" function, because we were not interested in fixed effects on this level. We used full information maximum likelihood estimation (FIML) with robust standard errors, which is robust to non-normality. FIML handles missing values (0–9.7 % per variable) by using all available data when estimating parameters (Enders, 2001).

To answer RQ3, the Student Monitoring/Regulation Accuracy according to Teacher Judgments was regressed on the measured student characteristics (as perceived by teachers). To answer RQ4, Correctness of 'Student Monitoring/Regulation Accuracy according to Teacher Judgments' was regressed on the student characteristic variables. The fixed effects were modelled at the student level—meaning that conclusions are not specified for the multiplication or division task—because the student characteristics were measured at the student level. Moreover, we had no reason to expect differential findings across the two mathematics tasks regarding how (perceived) student characteristics would be associated with teachers' judgment process. Analyzing these effects at the student level was supported by the variance decomposition of the outcome variables—that is, the degree to which variability in (the Correctness of) Student Monitoring/Regulation Accuracy according to Teacher Judgments was due to differences within students, between students, and between teachers. There was substantial between-student variability, ranging from 8.5 to 41.1 % (Table 1).

Per multilevel multiple regression model, which we performed to answer RQ3 and 4, 0.15 to 0.45 % of the tasks were identified as multivariate outlier. 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

Table 2
Means (M), Standard Deviations (SD), and Calculations of the Measures in the Present Study.

Measure	Calculation	Range	*n*	*M (SD)*
Student Performance (SP)		0 to 6	764	3.20 (1.95)
Student Need for intervention (SN)		0 to 2	751 746	Before Self-scoring: 1.20 (0.84) [a] After Self-Scoring: 1.20 (0.84) [a]
Student Monitoring Judgment (SMJ)		0 to 6	760	3.79 (1.71)
Student Regulation Judgment (SRJ)		0 to 2	755 739	Before Self-scoring: 0.90 (0.80) After Self-scoring: 0.91 (0.83)
Student Monitoring Accuracy (SMA)	Absolute difference between Student Monitoring Judgment (SMJ) and Student Performance (SP)	0 to 6 [b]	760	1.15 (1.17)
Student Regulation Accuracy (SRA)	Absolute difference between Student Regulation Judgment (SRJ) and Student Need for intervention (SN)	0 to 2 [b]	744 728	Before Self-scoring: 0.54 (0.69) After Self-scoring: 0.39 (0.58)
Teacher Judgment of Student Performance (TJSP)		0 to 6	764	3.86 (1.70)
Teacher Judgment of Student Need for intervention (TJSN)		0 to 2	764	0.92 (0.84)
Teacher Judgment of Student Monitoring Judgment (TJSMJ)		0 to 6	764	4.02 (1.64)
Teacher Judgment of Student Regulation Judgment (TJSRJ)		0 to 2	764 762	Before Self-scoring: 0.94 (0.84) After Self-scoring: 0.95 (0.81)
Student Monitoring Accuracy according to Teacher Judgments (SMA-TJ)	Absolute difference between Teacher Judgment of Student Monitoring Judgment (TJSMJ) and Teacher Judgment of Student Performance (TJSP).	0 to 6 [b]	764	0.78 (0.83)
Student Regulation Accuracy according to Teacher Judgments (SRA-TJ)	Absolute difference between Teacher Judgment of Student Regulation Judgment (TJSRJ) and Teacher Judgment of Student Need for intervention (TJSN)	0 to 2 [b]	764 762	Before Self-scoring: 0.42 (0.60) After Self-scoring: 0.31 (0.53)
Correctness of 'Student Monitoring Accuracy according to Teacher Judgments' (C-SMA-TJ)	Absolute difference between Student Monitoring Accuracy according to Teacher Judgments (SMA-TJ) and Student Monitoring Accuracy (SMA)	0 to 6 [b]	760	1.01 (1.01)
Correctness of 'Student Regulation Accuracy according to Teacher Judgments' (C-SRA-TJ)	Absolute difference between Student Regulation Accuracy according to Teacher Judgments (SRA-TJ) and Student Regulation Accuracy (SRA)	0 to 2 [b]	744 726	Before Self-scoring: 0.65 (0.66) After Self-scoring: 0.48 (0.60)

[a] Means are calculated separately for before and after self-scoring because in some cases the codes of Student Need for intervention can differ from before to after self-scoring, see section 3.3.1.
[b] Values closer to zero indicate more accurate or correct judgments.

Table 3
Contingency table of student monitoring accuracy and student monitoring accuracy according to teacher judgments.

	Student Monitoring Accuracy according to Teacher Judgments							
	0	1	2	3	4	5	6	Total
Student Monitoring Accuracy								
0	105	114	22	4	2	0	0	247
1	126	128	34	2	2	2	0	294
2	45	66	12	4	1	0	0	128
3	23	25	9	0	2	0	0	59
4	9	5	2	0	0	1	0	17
5	3	4	1	0	1	0	0	9
6	1	4	1	0	0	0	0	6
Total	312	346	81	10	8	3	0	760[6]

[6] Sample sizes in Tables 3,4,5, and 8 are the number of tasks, of which the total number is slightly smaller than displayed in Fig. 1, because cases with missing values for one or more of the variables were excluded from this table.

additionally ran the analyses with outliers still included. When this led to differences in statistical significance of effects (this was the case for two of the fixed effects that were part of RQ4), we additionally reported the effects of the analyses with outliers in the Results section.

The data of this study are openly available on the Open Science Framework: https://osf.io/wh9r8/ (Oudman et al., 2023a).

4. Results

Descriptive statistics of all performance, need, and judgment variables are displayed in Table 2. Correlations between these measures are displayed in Table 10 in Appendix A.

4.1. Teacher judgments of Students' monitoring accuracy (RQ1)

4.1.1. Correctness of 'Student monitoring accuracy according to Teacher Judgments' (RQ1A)

On average, Student Monitoring Accuracy according to Teacher Judgments deviated 1.01 item or 16.83 % (1.01/6*100)[4] from Students' actual Monitoring Accuracy, on a scale ranging from zero to six (before self-scoring, see measure C-SMA-TJ in Table 2).

4.1.2. Identifying students with substantially inaccurate monitoring judgments (RQ1B)

As can be derived from the numbers presented in Table 3, of the 219 students who made monitoring judgments that deviated two or more items from their actual performance (which we considered as substantially inaccurate), only 34 students (15.53 %) were identified by their teachers as making monitoring judgments that deviated two or more items from their actual performance. Of the 541 students who made monitoring judgments that deviated less than two items from their performance, 473 students (87.43 %) were correctly identified by their teachers as making monitoring judgments that deviated less than two items from their actual performance. So, teachers were quite adept at recognizing which students could monitor their performance well, but not very good at identifying which students could not monitor their

[4] We report percentages to help readers to interpret the extent of the deviation because the scales for monitoring and regulation differ; however, this is also why these percentages should be interpreted with caution.

Table 4
Contingency table of student regulation accuracy and student regulation accuracy according to teacher judgments before self-scoring.

	Student regulation accuracy according to teacher judgments before self-scoring			
	0	1	2	Total
Student Regulation Accuracy before self-scoring				
0	265	131	26	422
1	160	67	12	239
2	53	23	7	83
Total	478	221	45	744 [4]

Table 5
Contingency table of student regulation accuracy and student regulation accuracy according to teacher judgments after self-scoring.

	Student Regulation Accuracy according to Teacher Judgments after self-scoring			
	0	1	2	Total
Student Regulation Accuracy after self-scoring				
0	363	101	17	481
1	145	57	5	207
2	25	10	3	38
Total	533	168	25	726 [4]

performance well (which are those who would be most in need of support).

4.2. Teacher judgments of students' regulation accuracy (RQ2)

4.2.1. Correctness of 'Student regulation accuracy according to Teacher Judgments' (RQ2A)

Before Self-scoring. On average, Student Regulation Accuracy according to Teacher Judgments deviated 0.65 or 32.50 %[4] from Students' actual Regulation Accuracy before self-scoring, on a scale ranging from zero to two (see measure C-SRA-TJ before self-scoring in Table 2).

After Self-scoring. On average, Student Regulation Accuracy according to Teacher Judgments deviated 0.48 or 24.00 %[4] from Students' actual Regulation Accuracy after self-scoring, on a scale ranging from zero to two (see measure C-SRA-TJ after self-scoring in Table 2).

4.2.2. Identifying students with inaccurate regulation judgments (RQ2B)

Before Self-scoring. As can be derived from Table 4, of the 322 students who made inaccurate regulation judgments before self-scoring, 109 students (33.85 %) were identified as such by their teachers. Of the 422 students who made accurate regulation judgments before self-scoring, 265 students (62.80 %) were identified as such by their teachers.

After Self-scoring. As can be derived from Table 5, of the 245 students who made inaccurate regulation judgments after self-scoring, 75 students (30.61 %) were identified as such by their teachers. Of the 481 students who made accurate regulation judgments after self-scoring, 363 students (75.47 %) were identified as such by their teachers.

Thus, teachers were quite adept at recognizing which students made accurate regulation judgments, both before and after self-scoring. Teachers did not seem to be very good at identifying which students made inaccurate regulation judgments before and after self-scoring (and who would, therefore, be most in need of support), especially when considering that the teacher judgments were made on a three-point scale (i.e., randomly made teacher judgments would result in values that have on average 33.33 % chance of being exactly in line with Students' actual Regulation Accuracy).

Table 6
The effects of teacher-perceived student characteristics on student monitoring and regulation accuracy according to teacher judgments.

	Monitoring $N = 671$	Regulation	
		Before self-scoring $N = 669$	After self-scoring $N = 670$
Fixed effects	*B (SE)*	*B (SE)*	*B (SE)*
Intercept	1.20 (0.33)***	0.20 (0.19)	0.44 (0.09)***
Fixed effects student level			
Amount of contact	0.10 (0.07)	0.07 (0.06)	−0.02 (0.05)
Conscientiousness	−0.17 (0.06)**	0.05 (0.04)	0.01 (0.03)
Effort	0.02 (0.07)	−0.01 (0.05)	0.01 (0.04)
Extraversion	−0.01 (0.04)	0.09 (0.04)*	0.04 (0.03)
Interest	0.09 (0.09)	−0.04 (0.06)	−0.04 (0.06)
Learning problems	0.09 (0.09)	0.05 (0.05)	0.08 (0.05)
Likeability	0.06 (0.06)	0.01 (0.05)	0.00 (0.04)
Mathematics ability	−0.13 (0.06)*	0.02 (0.04)	0.01 (0.03)
Nationality [a]	0.09 (0.03)***	0.03 (0.03)	0.04 (0.03)
Self-concept	−0.13 (0.07)*	−0.16 (0.04)***	−0.07 (0.04)
Sex	0.01 (0.06)	0.07 (0.05)	0.00 (0.05)
Random effects	*SS (SE)*	*SS (SE)*	*SS (SE)*
σ^2_e (task)	0.40 (0.05)***	0.22 (0.03)***	0.18 (0.02)***
σ^2_{u0} (student)	0.20 (0.06)**	0.12 (0.03)***	0.08 (0.02)***
R^2 (student level)	0.24	0.20	0.09

Note. Sample sizes are the number of tasks, which are slightly smaller than the ones displayed in Fig. 1, because multivariate outliers were removed from the analyses.

[a] See for coding footnote [5]. A higher value means a 'less Western background'.
*** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$.

4.3. Relation between perceived student characteristics and student Monitoring/Regulation accuracy according to Teacher judgments (RQ3)

4.3.1. Monitoring

Table 6 shows the results of the regression analysis of the teachers' perceptions of student characteristics on Student Monitoring/Regulation Accuracy according to Teacher Judgments. Teachers' perceptions of students' conscientiousness (B = -0.17, p =.007, β = -0.26), general mathematics ability (B = -0.13, p =.031, β = -0.25), nationality (B = 0.09, $p \leq 0.001$, β = 0.14), and self-concept (B = -0.13, p =.045, β = -0.20) significantly predicted Student Monitoring Accuracy according to Teacher Judgments. The direction of the coefficients (positive/negative) indicated that teachers' judgments of their students' monitoring accuracy were higher when their perception of students' conscientiousness, general mathematics abilities, and confidence in their mathematical skills was higher, and when students had 'more of a Western background' (i.e., students and their parents born in Western countries).[5] The effect size of all student characteristics together, in terms of f^2, was 0.32 (medium to large), indicating that teachers' perceptions of student characteristics explained substantial variance in teachers' judgments of how well students monitored their learning (0.02 is the criterion for a small effect, 0.15 for a medium effect, 0.35 for a large effect; Cohen, 1988).

4.3.2. Regulation before self-scoring

Students' extraversion (B = 0.09, p =.013, β = 0.22) and students' self-concept (B = -0.16, $p \leq 0.001$, β = 0.33) as perceived by the teachers significantly predicted Student Regulation Accuracy according to

[5] Teacher-perceived students' nationality was coded as follows (see also Appendix Table 9): (0) student, mother and father born in Western country (W); (1) student and mother or father born in W; (2) student born in W, mother and father not; (3) student not born in W, mother and father born in NL; (4) student, mother and father not born in W (it did not occur that student was not born in W, mother or father born in W).

Table 7
The Effect of Teacher-Perceived Student Characteristics on the Correctness of 'Student Monitoring/Regulation Accuracy According to Teacher Judgments'.

	Monitoring $N = 670$	Regulation	
		Before Self-scoring $N = 670$	After Self-scoring $N = 670$
Fixed effects	*B (SE)*	*B (SE)*	*B (SE)*
Intercept	0.92 (0.34)**	0.96 (0.27)**	0.68 (0.20)***
Fixed effects student level			
Amount of contact	0.13 (0.07)	0.02 (0.06)	0.04 (0.06)
Conscientiousness	0.00 (0.07)	−0.04 (0.05)	−0.07 (0.05)
Effort	−0.06 (0.10)	0.01 (0.07)	0.02 (0.07)
Extraversion	−0.01 (0.05)	0.02 (0.04)	−0.01 (0.04)
Interest	0.04 (0.06)	−0.06 (0.05)	−0.07 (0.06)
Learning problems	−0.21 (0.11)* [a]	−0.05 (0.08)	−0.03 (0.07)
Likeability	0.04 (0.06)	−0.04 (0.05)	−0.03 (0.04)
Mathematics ability	−0.08 (0.03)* [b]	0.05 (0.04)	0.02 (0.04)
Nationality [c]	−0.01 (0.06)	0.06 (0.03)	−0.02 (0.03)
Self-concept	0.02 (0.06)	−0.07 (0.05)	0.01 (0.05)
Sex	−0.06 (0.10)	0.01 (0.05)	0.06 (0.06)
Random effects	*SS (SE)*	*SS (SE)*	*SS (SE)*
σ^2_e (task)	0.86 (0.11)***	0.39 (0.03)***	0.30 (0.03)***
σ^2_{u0} (student)	0.08 (0.09)	0.03 (0.02)	0.04 (0.03)
R^2 (student level)	0.25	0.26	0.16

Note. Sample sizes are the number of tasks, which are slightly smaller than the ones displayed in Fig. 1, because multivariate outliers were removed from the analyses. The following effects were not significant when outliers were still included.

[a] B = -0.18 (0.12), $p = 0.126$.

[b] B = -0.06 (0.04), $p = 0.146$.

[c] See for coding footnote [5]. A higher value means a 'less Western background'.

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

Teacher Judgments before self-scoring (Table 6). The direction of the coefficients (positive/negative) indicated that teachers' judgments of their students' regulation accuracy were higher before self-scoring when their perception of students' extraversion was lower and their perception of students' confidence in their mathematical skills was higher. The effect size of all student characteristics together, in terms of f^2, was 0.25 (medium to large; Cohen, 1988), indicating that teachers' perceptions of student characteristics explained substantial variance in teachers' judgments of how well students regulated their learning prior to self-scoring.

4.3.3. Regulation after self-scoring

The (perceived) student characteristics did not significantly predict Student Regulation Accuracy according to Teacher Judgments after self-scoring. The effect size of all student characteristics together, in terms of f^2, was 0.10 (small to medium; Cohen, 1988), indicating that teachers' perceptions of student characteristics explained some variance in teachers' judgments of how well students regulated their learning after self-scoring.

4.4. Relation between perceived student characteristics and Correctness of Student Monitoring/Regulation accuracy according to Teacher Judgments' (RQ4)

Finally, we wanted to explore whether and to what extent student characteristics as perceived by the teachers would explain the degree of Correctness of 'Student Monitoring/Regulation Accuracy according to Teacher Judgments'. These results are displayed in Table 7.

4.4.1. Monitoring

The presence of students' learning problems (B = -0.21, p =.049, β = -0.27) and students' general mathematics ability (B = -0.08, p =.014, β = -0.25) as perceived by the teachers significantly predicted the Correctness of 'Student Monitoring Accuracy according to Teacher Judgments' (Table 7). Thus, when the Correctness of 'Student Monitoring Accuracy according to Teacher Judgments' was higher (i.e., teacher judgments were more accurate) teachers' perceptions of their students' general mathematical skills were higher and learning problems were (perceived to be) present. The effect size of all student characteristics together, in terms of f^2, was 0.34 (medium to large; Cohen, 1988), indicating that teachers' perceptions of student characteristics explained substantial variance in the correctness of their judgments of how well students monitored their learning.

Table 8
Descriptive Statistics of Teachers' Perceptions of Students' Mathematics Ability and Learning Problems per Combination of Student Monitoring Accuracy and Student Monitoring Accuracy According to Teacher Judgments.

	Student Monitoring according to Teacher Judgments			
	0	1	≥2	Total
Student Monitoring Accuracy				
0	$n = 97$	$n = 102$	$n = 26$	$n = 225$
Mathematic ability, *M (SD)*	3.58 (1.04)	3.40 (0.98)	3.19 (1.06)	3.45 (1.02)
Learning problems present, %	22.7	22.5	23.1	22.7
1	$n = 111$	$n = 113$	$n = 35$	$n = 259$
Mathematic ability, *M (SD)*	3.72 (0.97)	3.27 (0.98)	2.94 (0.94)	3.42 (1.01)
Learning problems present, %	16.2	25.7	37.1	23.2
≥2	$n = 63$	$n = 91$	$n = 26$	$n = 180$
Mathematic ability, *M (SD)*	3.56 (0.98)	3.20 (0.98)	2.77 (0.82)	3.26 (0.99)
Learning problems present, %	9.5	24.2	34.6	20.6
Total	$n = 271$	$n = 306$	$n = 87$	$n = 664$ [4]
Mathematic ability, *M (SD)*	3.63 (0.99)	3.29 (0.98)	2.97 (0.95)	3.39 (1.00)
Learning problems present, %	17.0	24.2	32.2	22.3

To gain more insight into the relationship between teachers' perceptions of students' learning problems or mathematics ability and teacher judgments of students' monitoring skills, we additionally explored the descriptive statistics of these student characteristics per combination of Students' Monitoring Accuracy and Student Monitoring Accuracy according to Teacher Judgments, see Table 8. This table shows that students who made inaccurate monitoring judgments were more often identified as such by their teachers when they were also perceived to have learning problems compared to students who were not perceived to have learning problems.

Table 8 also suggests that students of whom the teachers thought that they made *more* accurate monitoring judgments than the students actually made, were also perceived to be relatively *more* skilled in mathematics. Vice versa, students of whom the teachers thought that they made *less* accurate monitoring judgments than the students actually made, were perceived to be relatively *less* skilled in mathematics (note that these interpretations are based on the means reported in the Table, not on significance testing).

4.4.2. Regulation before self-scoring

The (perceived) student characteristics did not significantly predict the Correctness of 'Student Regulation Accuracy according to Teacher Judgments' before self-scoring (Table 7).The effect size of all student characteristics together, in terms of f^2, was 0.34 (medium to large; Cohen, 1988) indicating that teachers' perceptions of student characteristics explained substantial variance in the correctness of their judgments of how well students regulated their learning prior to self-scoring.

4.4.3. Regulation after self-scoring

The (perceived) student characteristics did not significantly predict the Correctness of 'Student Regulation Accuracy according to Teacher Judgments' after self-scoring (Table 7). The effect size of all student characteristics together, in terms of f^2, was 0.19 (medium; Cohen, 1988) indicating that teachers' perceptions of student characteristics explained substantial variance in the correctness of their judgments of how well students regulated their learning after self-scoring.

5. Discussion

To be able to help students improve their monitoring and regulation skills effectively and efficiently, teachers should have an accurate idea of how well students can monitor and regulate their learning. The present study aimed to investigate how well primary school teachers can make judgments of their students' monitoring and regulation skills when solving mathematics problems and how (the accuracy of) these judgments are related to teachers' perceptions of student characteristics.

We extended research on monitoring and regulation (Nelson & Narens, 1990) and in particular on the *meta*-reasoning framework (Ackerman & Thompson, 2017) in two ways. First, whereas the *meta*-reasoning framework focusses on students' monitoring and regulation judgments *during* task completion, we focused on monitoring and regulation judgments that primary school students commonly make *after* they completed a mathematical task (i.e., about what they should subsequently engage in to improve their learning), as well as on regulation judgments after self-scoring their answers. Second, whereas the *meta*-reasoning framework is developed to study *meta*-reasoning within students, we extended it to the teacher level, by investigating to what extent teachers can accurately monitor their students' monitoring and regulation skills, which is needed for them to help students improve these skills.

5.1. Teachers' ability to judge Students' monitoring and regulation skills (RQ1 & 2)

First, we investigated to what extent teacher judgments of their students' monitoring accuracy were in line with students' actual monitoring accuracy (RQ1A). The teachers in our study misjudged their students' monitoring accuracy by approximately 17 %, which is close to prior findings in text comprehension: Van de Pol and Oudman (in press) found that secondary school teachers misjudged their students' monitoring accuracy with regard to text comprehension with 14 %. The teachers in the present study correctly estimated that, on average, their students made accurate monitoring judgments. However, it did not seem easy for the teachers to identify which students would be most in need of help with making more accurate monitoring judgments: approximately 16 % of the students who made monitoring judgments that deviated two or more items (on a scale ranging from zero to six) from their actual performance were identified as such by their teachers (RQ1B).

This was the first study to not only investigate teacher judgments of their students' monitoring skills, but also of their students' regulation skills. We explored to what extent teacher judgments of their students' regulation accuracy were in line with students' actual regulation accuracy, before and after self-scoring (RQ2A). The teachers in our study misjudged their students' regulation accuracy before self-scoring with approximately 33 % and after self-scoring with 24 %. So, teachers correctly inferred that students' regulation judgments would become (on average) somewhat more accurate after self-scoring (see Tables 2, 4, and 5). With regard to identifying students who made inaccurate regulation judgments, only 34 % (before self-scoring) and 31 % (after self-scoring) of the students who made inaccurate regulation judgments were identified as such by their teachers, which is around chance level given the three-point scale. Hence, similar to the findings for monitoring, it did not seem easy for the teachers to identify which students would be most in need of help with making accurate regulation judgments. In other words, seen in terms of the Nelson and Narens (1990) model: If teachers do not accurately monitor (i.e., make accurate judgments about) which students are not able to accurately monitor and regulate their learning (i.e., the object-level for the teachers), then teachers cannot accurately regulate (i.e., effectively support) these students' development of monitoring and regulation skills. Or put more simply: Teachers' support directed at helping students to better monitor and regulate their learning will not be optimally adapted to their students' individual needs. Thus, it seems that teachers themselves need support in making more accurate judgments of students' monitoring and regulation skills, and in order to help teachers in doing so one first needs to look at the origin of inaccurate teacher judgments, which we discuss next.

5.2. Relations between perceived student characteristics and the (Accuracy of) Teacher judgments of Students' monitoring and regulation skills (RQ3 & 4)

To gain more insight into what information teachers might use to make judgments of their students' monitoring and regulation skills, we investigated which student characteristics (as perceived by teachers) explained the magnitude of the teacher judgments of their students' monitoring and regulation accuracy (RQ3). Based on prior studies (Callan & Shim, 2019; Carr and Kurtz-Costes, 1994; Dignath & Sprenger, 2020; Friedrich et al., 2013), we expected that teachers' perceptions of students' mathematics abilities, variables related to students' working behavior (such as effort and conscientiousness), and students' self-concept might be associated with their judgments of students' monitoring and regulation skills. Some of these expected relations indeed appeared. Teachers' judgments of students monitoring accuracy were higher when their perceptions of students' conscientiousness, general mathematics ability, and confidence in their mathematical skills were higher. Teachers' judgments of students' regulation accuracy prior to self-scoring were higher when their perception of students' confidence in their mathematical skills was higher. Unexpectedly, the teachers also seemed to associate a 'more Western background' with making more accurate monitoring judgments, and less extraversion with making more accurate regulation judgments prior to self-scoring. We should note, however, that considered individually, all effect sizes of teachers' perceptions of specific student characteristics on teachers' judgments of students' monitoring and regulation skills were small.

Interestingly, even if the effects of teachers' perceptions of student characteristics were not large when considering them individually, taken together, they explained a medium to large amount of variance in teacher judgments of students' monitoring and regulation accuracy before self-scoring. Although our data are correlational, this seems to suggest that teachers' overall picture of their students might influence their judgments of students' monitoring and regulation skills before self-scoring, and future research could attempt to confirm causality by more direct process measures of cue use. Another interesting finding was that teachers seemed to expect that self-scoring would improve the accuracy of students' regulation judgments, regardless of students' characteristics, as (1) teachers thought that most of their students would make perfectly accurate regulation judgments after self-scoring (Table 5), and (2) the student characteristics did not predict teacher judgments of students' regulation accuracy after self-scoring (effect size was small to medium, compared to a medium to large effect size before self-scoring). This seems to suggest that teachers might consider student characteristics to play less of a role in students' regulation (in)accuracy after self-scoring than before self-scoring.

Lastly, we wanted to know for which students it would be most easy for teachers to make accurate judgments about their monitoring/regulation skills, so we explored whether and to what extent teachers' perceptions of student characteristics explained the degree to which teacher judgments of their students' monitoring/regulation accuracy were in line with students' actual monitoring/regulation accuracy (RQ4). None

of the student characteristics as perceived by teachers predicted the accuracy of teachers' judgments of students' monitoring and regulation skills, except for the presence of learning problems and students' general mathematics ability. Teacher judgments of students' monitoring skills were more accurate when students were perceived to be relatively skilled in mathematics and to have learning problems, compared to no learning problems. Again, although our data are correlational, this may suggest that stimulating teachers to pay more attention to the monitoring process of students who are less skilled in mathematics but who do not have learning problems, might be an effective and efficient intervention to make teachers' judgments of students' monitoring skills more accurate, which is a hypothesis for future research to address.

We found no support for the hypothesis that teachers' judgement accuracy would be associated with having more relevant information available (which has been shown for personality judgments by Funder, 2012): Teachers did not make more accurate judgments for students with whom they had more contact or about students whom they perceived to be more extravert (about whom teachers could have more information).

Although the majority of the perceived student characteristics considered individually did not significantly predict (the accuracy of) teacher judgments (except for the effect of students' general mathematics ability and learning problems on the accuracy of teacher judgments of students' monitoring skills), our findings do indicate that the accuracy of teacher judgments of students' monitoring and regulation skills is associated with their perceptions of student characteristics: Considered together, (perceived) student characteristics explained medium to large amounts of variance in the accuracy of teacher judgments of students' monitoring and regulation skills before self-scoring, and a medium amount of variance in the accuracy of teacher judgments of students' regulation skills after self-scoring. An interesting question for future research is whether it would be easier for teachers to make accurate judgments for students with a specific profile of characteristics, compared to other students. Our findings also give reason to further investigate potential biases in teacher judgments of students' monitoring and regulation skills and to define the profiles of students for whom making accurate judgements is more difficult. Subsequently, it could be tested whether stimulating teachers to obtain more information before making judgments of the students with the 'difficult profiles', would be an effective an efficient way to enhance the accuracy of teacher judgments of students' monitoring and regulation skills.

5.3. Limitations and future research

One limitation of the present study is that we did not directly ask teachers how accurate they thought their students' monitoring and regulation judgments would be or what information teachers used to make their judgments. We calculated the variable Correctness of 'Student Monitoring/Regulation according to Teacher Judgments' by taking the difference between two other measures (Teacher Judgment of Student Performance/Need and the Teacher Judgment of Student Monitoring/Regulation Judgment). We decided for this difference score based on a small pilot study. Nevertheless, future research should establish whether these difference scores are similar to teachers' direct judgments of students' absolute monitoring/regulation accuracy. Another limitation is that in our study (and similar studies, e.g., Van de Pol and Oudman, in press) a single teacher is typically judging a student's performance or monitoring/regulation judgment (as would be the case in the classroom). Moreover, the different constructs are often measured with single items, of which statistical reliability cannot be computed. Furthermore, these single items are being subtracted from each other to compute the different measures, and with a very high/very low score on one of the items the room for overestimation or underestimation on the other item is impacted. Therefore, the absolute accuracy (i.e., deviation) scores should be interpreted with some caution.

As for the information about students that teachers used to make their judgments, we inferred this from correlations between our measures of teachers' perceptions of student characteristics and Student Monitoring/Regulation Accuracy according to Teacher Judgments. While this is a common approach in the emerging research on the information (i.e., cues) that students and teachers base their judgments on (cf. Furnari et al., 2017; Meissel et al. 2017; Palecnek et al., 2017) it does have some drawbacks. For instance, we cannot know whether teachers' perceptions of student characteristics influence their judgments of students' monitoring and regulation skills or whether this relationship is (partly) reversed or reciprocal. In addition, we cannot exclude the possibility that teachers did not base their judgments on the student characteristics we measured but instead used information that is highly related (both conceptually and in terms of correlations) to the characteristics we measured. They could also have used other student characteristics we did not measure, such as students' ability to reflect on their behavior. It could also be the case that specific student characteristics only influence (the accuracy of) teachers' judgments when they manifest to a specific degree, in a specific direction, or when combined with other student characteristics. For instance, in theory, teachers might think that especially students who are good at mathematics *and* show much effort would be skilled in making accurate regulation judgments.

So, while our findings provide an interesting starting point, especially given the fact that there were no prior studies that investigated what information teachers might use to make judgments of student' monitoring and regulation accuracy, future experimental research would be needed to further explore how (combinations of) student-related factors might influence (the accuracy of) teacher judgments of students' monitoring and regulation skills. To this aim, future studies could, for instance, use vignettes or more direct measures such as think-aloud procedures or questionnaires that directly ask teachers about the information they used. When using questionnaires, future research could also consider using multiple items per student characteristic, so that reliability can be assessed, instead of the single-item measures that we used here (although the latter is not uncommon, cf. Helwig et al., 2001; Zhu & Urhahne, 2020). However, using multiple items per variable will come at the expense of (1) the number of students that teachers can answer questions about, (2) the number of variables measured per student, or (3) the time that is taken from the teachers (who are experiencing a very high workload; Gemmink et al., 2020).

Future research should also establish whether our findings would replicate in a larger sample of teachers and schools, and generalize, for instance to other student ages and other types of tasks. Many students in the present study made quite accurate judgments and future research could investigate whether teachers would also recognize it when their students would, overall, make less accurate judgments. Another interesting question is whether the accuracy with which teachers can judge students' monitoring and regulation accuracy would differ between school subjects (cf. findings that accuracy of primary school teachers' judgments of students' performance seems to be domain-specific; Mack et al., 2023).

Finally, an important direction for future research would be to establish how we can help teachers to more accurately identify students who have difficulty with making accurate monitoring and regulation judgments, and to investigate whether more accurate teacher judgments of students' monitoring and regulation skills indeed help teachers to provide support and thereby have beneficial effects on students' (self-regulated) learning. For example, by means of (stimulated recall) interviews and classroom observations it could be investigated whether the amount and type of help teachers offer that is focused on developing students' monitoring/regulation skills, relates to teachers' judgments of students' monitoring/regulation skills.

6. Conclusions

Primary school students often make inaccurate monitoring and

Table 9
Explanation and Descriptive Statistics of Teachers' Perceptions of Student Characteristics.

Student characteristic	Question (translated from Dutch)	Source on which question is based	Range	Mean (SD)
Amount of Contact	I have a lot of contact with this student.	–	1 to 4 [a]	2.96 (0.66)
Conscientiousness	This student works conscientiousness during the normal mathematics lesson. *Examples: This student works orderly. This student works precisely.*	Big Five conscientiousness scale (Goldberg, 1992)	1 to 4 [a]	2.93 (0.76)
Effort	This student shows effort during normal mathematics lessons. *Examples: this student works hard; this student pays attention.*	Cf. Helwig et al. (2001)	1 to 4 [a]	3.19 (0.64)
Extraversion	This student is generally extravert in class. *Examples: this student is talkative; this student is **not** withdrawn.*	Big Five extraversion scale (Goldberg, 1992)	1 to 4 [a]	2.83 (0.89)
Interest	This student is generally interested in mathematics.	Cf. Karing (2009)	1 to 4 [a]	3.04 (0.66)
Mathematics ability	This student is in general strong in mathematics	Cf. Helwig et al. (2001)	1 to 5 [b]	3.39 (1.01)
Nationality	What is the country of Birth of this student/ the mother of this student/ the father of this student? Choose from: The Netherlands/ Another Country, namely:…	Cf. Driessen et al. (2015) and Van de Pol et al. (2021)	1 to 5 [c]	0.24 (0.78)
Learning problems	Does this student have learning problems (no diagnosis needed)? Choose from: No learning problems/ Dyslexia/ Dyscalculia/ ADHD/ ADD/ Autism/ Language delay/ Other, namely…,	Cf. Van de Pol et al. (2021)	no/ yes	77.5/ 22.5 %
Likeability	I like this student.	–	1 to 5 [b]	3.72 (0.72)
Self-concept	This student generally feels confident about their mathematical skills. *Examples: this student is convinced that he/she performs well on mathematics tasks and tests; this student knows that he/she can master the mathematics skills that he/she needs to learn.*	Perceived self-efficacy scale (Marsh et al., 2006)	1 to 4 [a]	2.88 (0.76)
Sex	Before the start of the experiment teachers were asked to provide the experimenter with a list of student names and their sex.	–	boy/ girl [d]	53.1/ 46.9 %

[a] Scale: Strongly disagree, disagree, agree, strongly agree.
[b] Strongly below average, below average, average, above average, strongly above average.
[c] Coded as follows: (0) student, mother and father born in Western country, (1) student and mother or father born in W, (2) student born in W, mother and father not, (3) student not born in W, mother and father born in NL, (4) student, mother and father not born in W (it did not occur that student was not born in W, mother or father born in W).
[d] This was an open question, but teachers only gave these two answers.

regulation judgments (Baars et al., 2014; Oudman et al., 2022; Prinz et al., 2020; Van Loon & Roebers, 2017). Although students' monitoring accuracy can improve from interventions—such as self-scoring—this often does not translate into more accurate regulation judgments (Oudman et al., 2022; Van Loon & Roebers, 2017). Hence, is essential that primary school teachers are able to identify whether their students need support with making accurate monitoring and (especially) regulation judgments. However, teachers' ability to do so had not yet been investigated.

Our findings show that the teachers in the present study correctly estimated that on average, their students made quite accurate monitoring and regulation judgments. However, they had difficulties with identifying those students who made substantially inaccurate judgments and who would be most in need of support. The findings also suggest that teachers' perceptions of their students' characteristics might play a role in this misidentification. The current study demonstrates that there is a need for research on ways to help teachers identify those students who need support for developing their monitoring and regulation skills, in order to ultimately help these students to become more effective self-regulated learners.

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. **Mariëtte van Loon:** Methodology, Writing – review & editing. **Tamara van Gog:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data of this study are openly available in an online depository at https://osf.io/wh9r8/?

Table 10
Zero-order Correlations Between Variables of the Measurement Framework of Teacher Judgments of Students' Monitoring/Regulation Accuracy.

	SP	SN [a]	SMJ	SRJ before	SRJ after	SMA	SRA before	SRA after	TJSP	TJSN	TJSMJ	TJSRJ before	TJSRJ after	SMA-TJ	SRA-TJ before	SRA-TJ after	C-SMA-TJ	C-SRA-TJ before
SN before	−0.81																	
SMJ	0.66***	-0.57***																
SRJ before	-0.51***	0.50***	-0.69***															
SRJ after	-0.75***	0.68***	-0.58***	0.64***														
SMA	-0.32***	0.25***	0.13***	0.01	0.21***													
SRA before	-0.00	0.19***	0.20***	-0.32***	-0.12***	0.32***												
SRA after	0.05	0.23**	0.10**	-0.22***	-0.29***	0.07*	0.55***											
TJSP	0.30***	-0.32***	0.30***	-0.29***	-0.27***	-0.05	-0.01	-0.03										
TJSN	-0.32***	0.33***	-0.31***	0.30***	0.27***	0.03	-0.01	0.03	-0.80***									
TJSMJ	0.25***	-0.28***	0.29***	-0.31***	-0.25***	-0.03	0.05	0.02	0.77***	-0.64***								
TJSRJ before	-0.26***	0.27***	-0.31***	0.30***	0.24***	0.02	-0.06	-0.04	-0.59***	0.62***	-0.75***							
TJSRJ after	-0.24***	0.27***	-0.28***	0.30***	0.24***	0.03	-0.00	-0.01	-0.73***	0.73***	-0.72***	0.71***						
SMA-TJ	-0.14***	0.11**	-0.11**	0.15***	0.12**	0.07*	0.02	-0.01	-0.29***	0.27***	-0.13***	0.18***	0.24***					
SRA-TJ before	-0.05	0.02	-0.05	0.12***	0.06	0.01	-0.01	0.00	-0.02	0.04	0.00	0.10**	0.05	0.36***				
SRA-TJ after	-0.06	0.05	-0.06	0.09*	0.05	0.05	0.05	0.06	-0.03	0.07	-0.04	0.10**	0.11**	0.20***	0.45***			
C-SMA-TJ	-0.20***	0.16***	0.12***	-0.05	0.09*	0.65***	0.27***	0.09*	0.05	-0.07*	0.08*	-0.07	-0.06	0.10**	0.01	0.09*		
C-SRA-TJ before	-0.08*	0.19***	0.07	-0.15***	−0.01	0.17***	0.49***	0.26***	0.02	-0.04	0.02	0.04	-0.01	0.11**	0.27***	0.18***	0.21***	
C-SRA-TJ after	0.05	0.20***	0.01	-0.06	-0.12***	0.07*	0.28***	0.53***	0.01	-0.01	0.00	0.04	0.03	0.12**	0.18***	0.37***	0.13***	0.36***

Note. For the full variable names see Table 2 or Figs. 1 and 2. Before/after means before/after self-scoring.

[a] As SN only differed in 2.24% of the cases from before to after self-scoring and the correlation between the two was 0.98***, we only included SN before self-scoring in this Table.

view_only=f2a1ba6a0b5748efa366735adb747450 (Oudman et al., 2023a).

Acknowledgements

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

Appendix A

(See Table 9 & 10).

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cedpsych.2023.102226.

References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences, 21*(8), 607–617. https://doi.org/10.1016/j.tics.2017.05.004

Azevedo, R., Moos, D. C., Greene, J. A., Winters, F. I., & Cromley, J. G. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research and Development, 56*(1), 45–72. https://doi.org/10.1007/s11423-007-9067-0

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

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. https://doi/abs/10.1146/annurev-psych-113011-143823.

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

Callan, G. L., & Shim, S. S. (2019). How teachers define and identify self-regulated learning. *The Teacher Educator, 54*(3), 295–312. https://doi.org/10.1080/08878730.2019.1609640

Carr, M., & Kurtz-Costes, B. E. (1994). Is being smart everything? The influence of student achievement on teachers' perceptions. *British Journal of Educational Psychology, 64*(2), 263–276. https://doi.org/10.1111/j.2044-8279.1994.tb01101.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* ((2nd ed.).). Lawrence Erlbaum Associates.

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

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

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review, 28*(3), 425–474. https://doi.org/10.1007/s10648-015-9320-8

Dignath, C., & Büttner, G. (2018). Teachers' direct and indirect promotion of self-regulated learning in primary and secondary school mathematics classes–insights from video-based classroom observations and teacher interviews. *Metacognition and Learning, 13*(2), 127–157. https://doi.org/10.1007/s11409-018-9181-x

Dignath, C., & Sprenger, L. (2020). Can you only diagnose what you know? The relation between teachers' self-regulation of learning concepts and their assessment of students' self-regulation. *Frontiers in Education, 5*, 1–17. https://doi.org/10.3389/feduc.2020.585683

Driessen, G., Elshof, D., Mulder, L., & Roeleveld, J. (2015). *Cohortonderzoek Cool 5–18: Technisch rapport basisonderwijs, derde meting 2013/14*. ITS.

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

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8*,128-141. https://doi.org/ 10.1207/s15328007sem0801_7.

Fleury-Roy, M. H., & Bouffard, T. (2006). Teachers' recognition of children with an illusion of incompetence. *European Journal of Psychology of Education, 21*(2), 149–161. https://doi.org/10.1007/BF03173574

Friedrich, A., Jonkmann, K., Nagengast, B., Schmitz, B., & Trautwein, U. (2013). Teachers' and students' perceptions of self-regulated learning and math competence: Differentiation and agreement. *Learning and Individual Differences, 27*, 26–34. https://doi.org/10.1016/j.lindif.2013.06.005

Furnari, E. C., Whittaker, J., Kinzie, M., & DeCoster, J. (2017). Factors associated with accuracy in prekindergarten teacher ratings of students' mathematics skills. *Journal of Psychoeducational Assessment, 35*, 410–423. https://doi.org/10.1177/0734282916639195

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21*(3), 177–182. https://doi.org/10.1177/0963721412445309

Gemmink, M. M., Fokkens-Bruinsma, M., Pauw, I., & van Veen, K. (2020). Under pressure? Primary school teachers' perceptions of their pedagogical practices. *European Journal of Teacher Education, 43*(5), 695–711. https://doi.org/10.1080/02619768.2020.1728741

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26–42. https://doi.org/10.1037/10403590.4.1.26

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.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Helwig, R., Anderson, L., & Tindal, G. (2001). Influence of elementary student sex on teachers' perceptions of mathematics achievement. *The Journal of Educational Research, 95*(2), 93–102. https://doi.org/10.1080/00220670109596577

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

Hurwitz, J. T., Elliott, S. N., & Braden, J. P. (2007). The influence of test familiarity and student disability status upon teachers' judgments of students' test performance. *School Psychology Quarterly*, *22*(2), 115-144. Doi: 10.1037/1045-3830.22.2.115.

Jamain, L. (2019). *Biais d'auto-évaluation de compétence en français et en mathématiques chez les élèves de primaire: évolution et implications pour l'adaptation et la réussite scolaire des élèves? [Self-assessment bias of proficiency in French and in mathematics among primary school pupils: Evolution and implications for psychosocial adaptation and pupils' academic success?]*. Université Grenoble Alpes). Doctoral dissertation,.

Kaiser, J., Möller, J., Helm, F., & Kunter, M. (2015). Das schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen. *Zeitschrift Fur Erziehungswissenschaft, 18*(2), 279–302. https://doi.org/10.1007/s11618-015- 0619-5

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction, 28*, 73–84. https://doi.org/10.1016/j.learninstruc.2013.06.001

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen. *Zeitschrift für Pädagogische Psychologie, 23*(34), 197–209. https://doi.org/10.1024/1010-0652.23.34.197

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

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

Mack, E., Gnas, J., Vock, M., & Preckel, F. (2023). The domain-specificity of elementary school teachers' judgment accuracy. *Contemporary Educational Psychology, 72*, Article 102142. https://doi.org/10.1016/j.cedpsych.2022.102142

Marsh, H. W., Hau, K. T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing, 6*(4), 311–360. https://doi.org/10.1207/s15327574ijt0604_1

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education, 65*, 48–60. https://doi.org/10.1016/j.tate.2017.02.021

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

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26*, 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5

Oudman, V. S., Van de Pol, J., Janssen, E. M., & Van Gog, T. (submitted). *Students' monitoring and regulation accuracy awareness*. Manuscript submitted for publication.

Oudman, V. 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*, 213–239. https://doi.org/10.1007/s11409-021-09281-9

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

Oudman, V. S., Van de Pol, J., Van Loon, M., & Van Gog, T.. *Primary School Teachers' Judgments of their Students' Monitoring and Regulation Skills*. Dataset published on the Open Science Framework. https://osf.io/wh9r8/.

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

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

Prinz, A., Golke, S., & Wittwer, J. (2020). To what extent do situation-model-approach interventions improve relative metacomprehension accuracy? Meta-analytic insights. *Educational Psychology Review, 32*, 917–949. https://doi.org/10.1007/s10648-020-09558-6

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. https://doi.org/10.1007/bf00117714

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., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25–29. https://doi.org/10.1037/h0071663

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review, 32*, Article 100374. https://doi.org/10.1016/j.edurev.2020.100374

Van de Pol, J., & Oudman, S. (in press). Teachers' judgment accuracy of students' monitoring skills: A conceptual and methodological framework and explorative study. *Metacognition and Learning.* Advance online publication Doi: 10.1007/s11409-023-09349-8.

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 de Pol, J., Van Gog, T., & Thiede, K. (2021). The relationship between teachers' cue-utilization and their monitoring accuracy of students' text comprehension. *Teaching and Teacher Education, 107*, Article 103482. https://doi.org/10.1016/j.tate.2021.103482

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

Zhu, C., & Urhahne, D. (2020). Temporal stability of teachers' judgment accuracy of students' motivation, emotion, and achievement. *European Journal of Psychology of Education, 36*, 319–337. https://doi.org/10.1007/s10212-020-00480-7

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